ISSN 1847-3938

mipro 2014

organizer upro



37th international convention

May 26-30, 2014, Opatija – Adriatic Coast, Croatia

Lampadem tradere



mipro - path to knowledge and innovation

mipro proceedings

My profession.
My organization.
My IEEE.



Discover the benefits of IEEE membership.

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

Join today
www.ieee.org/join

IEEE

MIPRO 2014

37th International Convention

May 26 - 30, 2014
Opatija, Croatia

Proceedings

Conferences:

Microelectronics, Electronics and Electronic Technology /MEET

Distributed Computing, Visualization and Biomedical Engineering /DC VIS

Telecommunications & Information /CTI

Computers in Education /CE

Computers in Technical Systems /CTS

Intelligent Systems /CIS

Information Systems Security /ISS

Business Intelligence Systems /miproBIS

Digital Economy and Government, Local Government, Public Services / DE-GLGPS

MIPRO Junior - Student Papers /SP

Edited by:
Petar Biljanović

International Program Committee

Petar Biljanović, General Chair, Croatia
A. Abelló Gamazo, Spain
S. Amon, Slovenia
V. Anđelić, Croatia
M.E. Auer, Austria
M. Baranović, Croatia
L. Bellatreche, France
E. Brenner, Austria
A. Budin, Croatia
Ž. Butković, Croatia
Ž. Car, Croatia
M. Colnarič, Slovenia
A. Cuzzocrea, Italy
M. Čičin-Šain, Croatia
M. Delimar, Croatia
T. Eavis, Canada
M. Ferrari, Italy
B. Fetaji, Macedonia
T. Galinac Grbac, Croatia
L. Gavrilovska, Macedonia
M. Golfarelli, Italy
S. Golubić, Croatia
F. Gregoretti, Italy
S. Groš, Croatia
N. Guid, Slovenia
Y. Guo, United Kingdom
J. Henno, Estonia
L. Hluchy, Slovakia
V. Hudek, Croatia
Ž. Hutinski, Croatia
M. Ivanda, Croatia
H. Jaakkola, Finland
L. Jelenković, Croatia
D. Jevtić, Croatia
R. Jones, Switzerland
P. Kacsuk, Hungary
A. Karaivanova, Bulgaria
D. Knežević, Croatia
M. Mauher, Croatia
I. Mekjavić, Slovenia
B. Mikac, Croatia
V. Milutinović, Serbia
A.-I. Mincu, Slovenia
V. Mrvoš, Croatia
J.F. Novak, Croatia
J. Pardillo, Spain
N. Pavešić, Slovenia
V. Peršić, Croatia
G. Radić, Croatia
S. Ribarić, Croatia
J. Rozman, Slovenia
K. Skala, Croatia
I. Sluganović, Croatia
V. Sruk, Croatia
U. Stanič, Slovenia
N. Stojadinović, Serbia
J. Sunde, Australia
A. Szabo, IEEE Croatia Section
L. Szirmay-Kalos, Hungary
D. Šarić, Croatia
D. Šimunić, Croatia
Z. Šimunić, Croatia
D. Škvorc, Croatia
A. Teixeira, Portugal
E. Tijan, Croatia
A.M. Tjoa, Austria
R. Trobec, Slovenia
I. Turčić-Prstačić, Croatia
S. Uran, Croatia
T. Vámos, Hungary
M. Varga, Croatia
M. Vidas-Bubanja, Serbia
B. Vrdoljak, Croatia
R. Wrembel, Poland
B. Zajc, IEEE Region 8, Slovenia
D. Zazula, Slovenia

organized by

MIPRO Croatian Society

technical cosponsorship

IEEE Region 8

under the auspices of

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

patrons

University of Zagreb, Croatia
University of Rijeka, Croatia
IEEE Croatia Section
IEEE Croatia Section Computer Chapter
IEEE Croatia Section Electron Devices/Solid-State Circuits Societies Joint Chapter
IEEE Croatia Section Education Chapter
IEEE Croatia Section Communications Chapter
Ruđer Bošković Institute, Zagreb, Croatia
University of Zagreb, Faculty of Electrical Engineering and Computing, Croatia
University of Rijeka, Faculty of Maritime Studies, Croatia
University of Rijeka, Faculty of Engineering, Croatia
University of Rijeka, Faculty of Economics, Croatia
University of Zagreb, Faculty of Organization and Informatics, Varaždin, Croatia
T-Croatian Telecom, Zagreb, Croatia
Ericsson Nikola Tesla, Zagreb, Croatia
Končar - Electrical Industries, Zagreb, Croatia
HEP - Croatian Electricity Company, Zagreb, Croatia
VIPnet, Zagreb, Croatia
Kermas energija Zagreb, Croatia
Croatian Post and Electronic Communications Agency, Croatia
Croatian Post, Croatia
Selmet Zagreb, Croatia
CISEx Zagreb, Croatia
Rezultanta Zagreb, Croatia
River Publishers, Aalborg, Denmark

sponsors

Ericsson Nikola Tesla Zagreb, Croatia
T-Croatian Telecom Zagreb, Croatia
Končar-Electrical Industries Zagreb, Croatia
InfoDom Zagreb, Croatia
HEP - Croatian Electricity Company Zagreb, Croatia
VIPnet Zagreb, Croatia
Storm Computers Zagreb, Croatia
Transmitters and Communications Company Zagreb
King-ICT Zagreb, Croatia
IN2 Zagreb, Croatia
Altpro Zagreb, Croatia
Microsoft Croatia Zagreb, Croatia
Hewlett Packard Croatia Zagreb, Croatia
Micro-Link Zagreb, Croatia
Mjerne tehnologije Zagreb, Croatia
Selmet Zagreb, Croatia
ib-proCADD Ljubljana, Slovenia
Nomen Rijeka, Croatia

All papers are published in their original form

For Publisher:

Petar Biljanović

Publisher:

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

Printed by:

GRAFIK, Rijeka

ISBN 978-953-233-078-6

Copyright © 2014 by MIPRO

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

CONTENTS

LIST OF PAPER REVIEWERS

AUTHOR INDEX

FOREWORD

MICROELECTRONICS, ELECTRONICS AND ELECTRONIC TECHNOLOGY

INVITED PAPER 3

Nanotechnology, from the Recent History to (un) Predictable Future 5
J. Turković

PAPERS 11

Spectroscopic Evidence of Coexistence of Clusters Based on Low (α) and High Temperature (β) GeS_2 Crystalline Phases in Glassy Germanium Disulfide Matrix 13
V. Mitsa, R. Holomb, G. Lovas, M. Veres , M. Ivanda, T. Kovach

Synthesis and Characterization of Silver Colloidal Nanoparticles and Their Application in Surface Enhanced Raman Spectroscopy 17
L. Mikac, M. Ivanda, M. Gotić, T. Mihelj

Raman Spectroscopy of Tellurite Glasses 21
H. Gebavi, D. Ristić, V. Đerek, L. Mikec, M. Ivanda, D. Milanese

Coated Spherical Microresonators for Cutting-Edge Photonics Application 24
D. Ristić, M. Mazzola, A. Chiappini, C. Armellini, A. Rasoloniaina, P. Féron, R. Ramponi, G.N. Conti, S. Pelli, G.C. Righini, G. Cibiel, M. Ivanda, M. Ferrari

Examination of Silicon Material Properties Using THz Time-Domain Spectroscopy 28
B. Pejčinović

Impact of Microstrip Width and Annealing Time on the Characteristics of Micro-Scale Graphene FETs 33
M. Poljak , M. Wang, S. Žonja, V. Đerek, M. Ivanda, K.L. Wang, T. Suligoj

Graphene Base Transistors with Optimized Emitter and Dielectrics 39
S. Venica, F. Driussi, P. Palestri, L. Selmi

Phonon-Limited Hole Mobility in Sub-20 nm-Thick Double-Gate Germanium MOSFETs 45
V. Ivanić, M. Poljak, T. Suligoj

Comparison of RF Performance between 20 nm-Gate Bulk and SOI FinFET 51
S. Krivec, H. Prgić, M. Poljak, T. Suligoj

Simulation Study of Aging in CMOS Binary Adders 57
T. An, C. Hao, L. Alves de Barros Naviner

Analytical Method for Reliability Assessment of Concurrent Checking Circuits Under Multiple Faults 62
T. An, K. Liu, L. Alves de Barros Naviner

Synthesis of CMOS Multiplier Structures Using Multifunctional Circuits 66
C. Popa

Improved Accuracy Current-Mode Analog Function Synthesizer 70
C. Popa

Analysis and Optimization of a SAR ADC with Attenuation Capacitor 74
S. Brenna, A. Bonfanti, A. Abba, F. Caponio, A.L. Lacaita

A Simulation and Modeling Environment for the Analysis and Design of Charge Redistribution DACs Used in SAR ADCs 80
S. Brenna, A. Bonetti, A. Bonfanti, A.L. Lacaita

High-Speed Energy-Efficient 5:2 Compressor 86
A. Najafi, S. Timarchi, A. Najafi

An Electrochemical Sensor ASIC for Agriculture Applications 91
T.K. Yew, Y.Yusoff, L.K. Sieng, H.C. Lah, H. Majid, N. Shelida

Application of Charge Amplifiers with Polyvinylidene Fluoride Materials 97
A.S. Marinov, O.P. Stanchev, E.B. Bekov

Design of Transdermal Drug Delivery System with PZT Actuated Micropump 102
T. Dolžan, D. Vrtačnik, D. Resnik, U. Aljančič, M. Možek, B. Pečar, S. Amon

Modern Monitoring System for Water Pollution by Petrochemical Products Based on Optical Fibre Refractometer 106
J. Turán, Ľ. Ovseník

Broadband Characterization of SMA Connectors by Measurements 110
M. Magerl, T. Mandić, A. Barić

Application of TEM Cell in Measurement of Electromagnetic Fields of Electronic Modules 116
K. Ivanuš, A. Barić

Model of Low-Noise, Small-Current-Measurement System Using MATLAB/Simulink Tools 121
D. Nagradić, K. Pardon, D. Jurišić

Algorithm for Fast Determination of LiFePO4 Battery Nominal Current 127
I. Sirotić, V. Šunde, Ž. Ban

100 nA Power Management Circuit for Energy Harvesting Devices 131
D. Vinko, G. Horvat

Evaluation of the Operation of Depletion-Mode SiC Power JFET in DC-DC Converter Applications 136
J. Bačmaga, K. Bene, B. Pejčinović, A. Barić

Photocopier PID Controller with Iterative Error Compensation 142
K. Fukuda, V. Čeperić, A. Barić

Introduction of Some Design Aspects for Improved Performance of the DC/AC Induction Motor Converter 147
J. Podržaj, G. Babić, M. Podhraski, J. Trontelj

Microcontroller Implementation of Dynamically Adaptable Control of Stepper Motor with Continuous Second Derivative of Speed Curve 152
J. Divić, J. Đurić, K. Vrančić

Optimized Coupled Band-Pass Filters 156
N. Mijat, D. Jurišić

Development and Testing of Complex PLL Algorithm Using Hardware in the Loop System 162
T. Čihak, M. Puškarić, Ž. Jakopović

Software Simulation of LDPC Codes and Performance Analysis 168
C. Sandu, I. Florescu, C. Rotaru

Smart Home Automation System for Energy Efficient Housing 172
N. Skeledžija, J. Ćesić, E. Kočo, V. Bachler, H.N. Vučemilo, H. Džapo

Remote Laboratory for Testing Processor Cores in FPGA Device 178
K. Saksida, A. Trošt

Towards an Energy Efficient SoC Computing Cluster 184
Z. Krpić, G. Horvat, D. Žagar, G. Martinović

Electronic Interpretation of Selected Lessons in Fuzzy Mathematics 189
L. Perković, D. Kovačević, T. Karađole, M. Jović

Fuzzy Gates Interchangeability 195
D. Kovačević, T. Karađole, J. Vrsalović

Solving Mathematic Problems in Electronic Environment, Finding Global and Local Minimum and Maximum Function Values 201
I. Madunić, D. Kovačević, L. Perković, T. Karađole

DISTRIBUTED COMPUTING, VISUALIZATION AND BIOMEDICAL ENGINEERING

PAPERS 209

BIOMEDICAL ENGINEERING

Twenty-Nine-Year Stimulation of the Common Peroneal Nerve in a Hemiplegic Patient 213
P. Pečlin, J. Krajnik, R. Aćimović-Janežič, U. Stanič, J. Rozman

A Method for Time-Domain Analysis of Respiratory Sinus Arrhythmia 217
R. Trobec, V. Avbelj

Review of Regularization Techniques in Electrocardiographic Imaging 221
M. Milanič, V. Jazbinšek, R. Hren

Single-Dipole and Dual-Dipole Inverse Solutions in Electrocardiography 227
V. Jazbinšek, R. Hren

Automatic Knot Fastener Cor-Knot and United States Pharmacopeial Convention Class I Knot Pull-Apart Force Safety Regulations 233
B. Geršak, J. Sauer

Control and Optimization of Complex Biological Systems 235
D. Tomić, B. Pernarić

Classification of Benign and Malignant Massesin Breast Mammograms 241
A. Šerifović-Trbalić, A. Trbalić, D. Demirović, N. Prljača, P.C. Cattin

CHO Cell Viability in CO_2 Saturated Media 247
A. Štefanič, J. Jelenc, A. Osterman, M. Dular, D. Miklavčič, A. Maček Lebar

Automatic Cell Counter for Cell Viability Estimation 252
J. Lojk, L. Šajn, U. Čibej, M. Pavlin

Morphological Changes of Pressure Pulses in Oscillometric Non-Invasive Blood Pressure Measurements 258
V. Avbelj

Personalized Guiding Templates for Pedicle Screw Placement 262
U. Tominc, M. Vesel, S. Al Mawed, M. Dobravec, M. Jug, S. Herman, D. Kreuh

iHELP Emergency Care Network 265
A. Ogorevc, B. Lončarevič

Graph-Based Diagnosis and Treatment of Neuronal Communication Disorders 269
M. Veletić

Use of 3D Preoperative Planning Program in Understanding of Articular Fracture Treatment **275**
M. Tomaževič, A. Fischinger, A. Kristan, D. Kreuh, M. Cimerman

Education of Cosmetic Device Users in Slovenia **279**
Š.E. Soklič, J. Jelenc

Telehealth Using ECG Sensor and Accelerometer **283**
H. Gjoreski, A. Rashkovska, S. Kozina, M. Luštrek, M. Gams

Children Obesity Treatment Support with Telemedicine **288**
M. Kozlovszky, K. Karóczkai, B. Jókai, B. Ruda, D. Kacsuk, Z. Meixner

Analysis of sEMG Signals Associated with Isometric Contraction of Triceps Brachii Muscles Using Multifractals **292**
K. Marri, R. Swaminathan

DISTRIBUTED COMPUTING

Improving Data-Parallel Construction of δ_N-Nets with Maximum Dispersion **299**
P. Zinterhof

High Productivity Data Processing Analytics Methods with Applications **304**
M. Riedel, A.S. Memon, M.S. Memon

Optimal Block Size for Matrix Multiplication Using Blocking **310**
S. Ristov, M. Gusev, G. Velkoski

Extensible Framework for Execution of Distributed Genetic Algorithms on Grid Clusters **316**
D. Georgiev, E. Atanassov

The Performance Impact Analysis of Loop Unrolling **322**
G. Velkoski, M. Gusev, S. Ristov

Developing a Ticket Management SaaS Solution **328**
M. Gusev, S. Ristov, P. Gushev

Alert Notification as a Service **334**
M. Gusev, S. Ristov, G. Velkoski, P. Gushev

Crowd Computing to Support EGI Scientists **340**
J. Kovács, R. Lovas

Obtaining Massive Data Sets for Contextual Experiments in Quantum Information . **347**
M. Pavičić

Parallel Implicit Time Integration for Particle-Based Models on Graphics Clusters ... **351**
A. Sabou, D. Gorgan

Frequency Table Computation on Dataflow Architecture **357**
P. Škoda, V. Sruk, B. Medved Rogina

Virtue - a Different Approach to Human/Computer Interaction **362**
Z. Šojat, T. Ćosić, K. Skala

Dynamic Workflow Support in gUSE **369**
E. Kail, A. Bánáti, K. Karóczkai, P. Kacsuk, M. Kozlovszky

Formal Checking of WS-BPEL Orchestrations **375**
J. Matković, K. Fertalj

Hybrid MPI-OpenMP Parallel Optimization for Thermo-Fluid Problems **382**
M. Depolli

Delivering Bioinformatics MapReduce Applications in the Cloud **388**
L. Forer, T. Lipić, S. Schönherr, H. Weißensteiner, D. Davidović, F. Kronenberg, E. Afgan

Microspheres as Sensors to Measure Instant Refractive Index Variation During Phase Transition of Low Melting Substances **393**
H.-B. Fu, Y.-X. Huang

Multilingual Cloudlet-Based Dictionary **395**
A. Bahtovski, M. Gusev

Mjolnirr: Private PaaS as Distributed Computing Evolution **401**
D. Savchenko, G. Radchenko

VISUALIZATION

Visualization Support for Code Development in EUROfusion Integrated Modelling .. **409**
L. Kos

Fast Motion Estimation Algorithm Using Dual Bit-Plane Matching Criteria **415**
C. Choi, J. Jeong

Edge Detection in Segmented Non-Convex Space Polygon **419**
I. Kožar, D. Lozzi-Kožar, Ž. Jeričević

Visualization of Meteorological Data **423**
I. Hakštok, Ž. Mihajlović

3D Volume Localization Using Miniatures **428**
L. Šajn, M. Radojević , T. Dobravec

Evaluation of Image Processing Algorithms on ARM Powered Mobile Devices **434**
D. Demirović, A. Šerifović-Trbalić, N. Prljača, P.C. Cattin

Natural Eye Gaze Computer Interaction for Medical Oculography Diagnosis: Current Status and Future Prospects **438**
I. Grubišić, I. Grbeša, T. Lipić, K. Skala, O. Zrinšćak, R. Iveković, Z. Vatavuk

TELECOMMUNICATIONS AND INFORMATION

INVITED PAPER **445**

Green ICT for Sustainability: A Holistic Approach **447**
S. Aleksic

PAPERS **453**

TELECOM MARKET REGULATIONS

4G Mobile Internet, Services, Regulation and Mobile Operators in Bosnia and Herzegovina **457**
D. Šerval, Č. Marković, S. Kovačević

Creating Regulatory Environment for Cloud Services **461**
S. Hadžović, R. Šeremet

Emergency Communication in IMS Network: Challenges in Implementation **465**
I. Krevatin, I. Visković

Međupovezivanje i regulacija u telekomunikacijama s osvrtom na međupovezivanje i regulaciju u telekomunikacijama Bosne i Hercegovine **469**
J. Gabela, F. Gabela, A. Gabela

BUSINESS MODELLING

Connecting Clouds **477**
I. Vrbovčan, T. Pavić, M. Šoša Anić

Competence Transfer through Enterprise Mobile Application Development **481**
S. Vojvodić, M. Zović, V. Režić, H. Maračić, M. Kušek

An Empirical Study of Factors Affecting the Adoption of Self-Service Checkout Terminals in Slovenia **486**
B. Šumak, M. Pušnik, M. Heričko

Cash Register Lines Optimization System Using RFID Technology **492**
D. Budić, Ž. Martinović, D. Šimunić

Pozicija BH Telecoma d.d. Sarajevo prema Adižesovoj teoriji životnih ciklusa preduzeća 496
A. Gabela, L. Pamuk, J. Gabela

Unaprjeđenje upravljanja resursima uz pomoć sustava za podršku 500
D. Markulin, K. Musa

Predprodajna reklamna aplikacija telekomunikacijskih usluga..... 504
J. Mulabegović, H. Šabanović, A. Arslanagić

INTERNET & WEB APPLICATIONS

Multi-User Web Application for Content Management on Multi-Touch Devices..... 513
M. Cvijanović, P. Kirin, D. Pleša, J. Soko, S. Stamenov, A. Cikojević, S. Kelemen, A. Karin

Front End Optimization Methods and Their Effect 517
I. Iliev, G. Petrov Dimitrov

Comparative Analysis of Google and Bing SEO on leading Croatian News portals 526
I. Britvić, J. Đurić, D. Bužić

CHALLENGING ICTs

State of the Art in IPTV..... 531
B. Veselinovska, M. Gusev, T. Janevski

Kinect-based Presenter Tracking Prototype for Videoconferencing 537
D. Gadanac, M. Dujak, D. Tomić , D. Jerčić

Iterative Receiver for Qc-LDPC Coded Underwater Acoustic Communication Systems 543
L. Zhao

Software Platform for Wireless Indoor Localization Planning Based on the Strength of Received Radio Signal 547
I. Vitas, B. Lukovac, D. Šimunić

Management in M2M Networks 553
V. Čačković, Ž. Popović

Next-Generation Utilities Based on M2M Communications..... 559
M. Dujak, Z. Kljaić, M. Šerbec, H. Jelačić

Applications Methods of the Objects, and Suitability for Universal Applications in Computer Graphics Context 566
N. Zeqiri

On-line Rapid Diagnostics of the Human Immune Status 571
T.A. Hryhorova, L.V. Vorobiov

INFORMATION SYSTEMS

Information System Infrastructure Planning - Threshold Setting Based on Risk Analysis 579
I. Golub, B. Radojević

Interactive Application in Service of Popularizing Architectural Heritage 584
M. Jozić, T. Oštarčević, M. Šalković, D. Šimunić

Adresiranje i usmjeravanje e-dokumenata u sustavu informacijskog posrednika za uslugu e-Računa 587
D. Matić, A. Čičak, B. Vrdoljak

Informacijske pretpostavke za izgradnju infrastrukture za širokopojasni pristup 593
D. Mesarić, M. Štimac, F. Ambroš

Upotreba naprednih informacijskih sustava u kontekstu ubrzanja izrade studija izvodljivosti za potrebe izgradnje širokopojasne infrastrukture u RH 599
D. Medved, S. Vojvodić, S. Drezgić

QUALITY MANAGEMENT

Real Time Assuring QoE in the Lowest OSI/ISO Layers During Delivering of IPTV Services 607
N. Goran, M. Hadžialić, A. Begović

Measuring XML Schema Standard Quality Change through Versioning 611
M. Pušnik, M. Heričko, B. Šumak

N-BEB: New Backoff Algorithm for IEEE 802.11 MAC Protocol 615
M. Shurman, B. Al-Shua'b, M. Alsaedeen, M.F. Al-Mistarihi, K. Darabkh

Multi-Objective Parameter Optimization of S-MAC Protocol 620
E. Kaljić, S. Konjicija

Primjena FMEA metode u upravljanju kvalitetom telekomunikacijskih usluga 626
N. Karahmetović

SOFTWARE ENGINEERING & TOOLS

Experiences from Building a EUD Business Portal 635
N. Tanković, T. Galinac Grbac, M. Žagar

Modeling of Software Failure Cost in ERP Systems 641
F. Urem, K. Fertalj, I. Livaja

Automatizirano projektiranje telekomunikacijske infrastrukture softverskim alatom baziranim na geografskom informacijskom sustavu (GIS) 646
T. Pavičić, D. Medved, S. Vojvodić, N. Severinac, V. Roland

Upotreba alata za mjerenja na radijskom sučelju u nastavi mobilnih komunikacija... 652
L. Novosel

NETWORKS & ENVIRONMENT

Reselection of Base Station in TETRA Network 659
S. Punčikar

Inventory Management System for Water Supply Network 664
O. Jukić, I. Heđi

Procjena vrijednosti parametra SAR u blizini GSM/UMTS baznih stanica..... 668
D.S. Šuka, P.S. Međedović, M.I. Simić

Utjecaj LTE tehnologije na razine elektromagnetskih polja u mobilnim mrežama Hrvatskog Telekoma 674
R. Gajić, K. Janaček

NETWORKS PERFORMANCE

Incremental-Relaying Cooperative-Networks Using Dual Transmit Diversity and Decode and Forward Relaying Scheme with Best Relay Selection..... 681
A.M. Hayajneh, M. Khodeir, M.F. Al-Mistarihi

Closed-form Expression for Outage Probability in Relay-Based Cooperative Diversity Systems over Nakagami Fading Channels with Interference 687
R. Mohaisen, M.F. Al-Mistarihi, A. Sharaqa

On Practical Solutions of AQM Queues with Large Buffers 692
P. Mrozowski, A. Chydzinski

COMPUTERS IN EDUCATION

INVITED PAPERS 701

Children as Co-Creators of Video Stories: Mobile Videos for Learning..... 703
J. Multisilta, H. Niemi

Children's Safety on the Internet-Development of the School Curriculum 708
L. Kralj

PAPERS **713**

Introducing E-CRM into Academic Libraries: Exploration of Needs and Possibilities **715**
A. Papić, M. Primorac

Virtual Machines in Education **721**
V. Mateljan, V. Juričić, M. Moguljak

Open^n **726**
H. Jaakkola, T. Mäkinen, J. Henno, J. Mäkelä

From Learning to E-learning to M-learning to C-learning to ... ? **734**
J. Henno, H. Jaakkola, J. Mäkelä

The Evocation of Creativity and Critical Thinking: Analysis of Information Competencies and Behaviour Among Students of Electrical Engineering and Computing **741**
D. Vrkić, M. Pavlovski

Application of Modern Teaching Techniques in the Educational Process **746**
M. Filipović Tretinjak, A. Bednjanec, M. Tretinjak

Open Access Initiative and Its Impact on Use of Digital Information Resources for Research and Teaching in Higher Education Institutions in Croatia **751**
R. Vrana

The Organization of Classes and Assessment System Using Moodle **757**
T. Alajbeg, T. Horvat, T. Novak

Perceived Quality of Verificator in Teaching Programming **761**
T. Orehovački, D. Radošević, M. Konecki

E-learning Materials to Study Properties of Liquid Crystals at University Level and in Schools **767**
R. Repnik, A. Šabeder, M. Krašna

Distance Learning: Challenges, New Solution **771**
P. Panahi, F. Borna

Online vs. Paper-Based Testing: A Comparison of Test Results **775**
S. Čandrlić, M. Ašenbrener Katić, M. Holenko Dlab

Comparing Classification Models in the Final Exam Performance Prediction **781**
J. Gamulin, O. Gamulin, D. Kermek

The Role of Management in Career Development **787**
M. Lamza-Maronić, J. Glavaš, I. Uroda

E-learning Next Step – Learning Materials for Students **792**
M. Krašna, M. Duh, T. Bratina

Configurable Image-Processing System for Education in Automotive Applications 800
C. Madritsch, T. Klinger

Using Hybrid Mobile Applications for Adaptive Multimedia Content Delivery 804
A. Karadimce, D. Capeska Bogatinoska

A Model of Adaptive Learning with Interactive Images - ADELE 810
M. Gusev, S. Ristov, P. Gushev, G. Velkoski

User Satisfaction Evaluation of the Electronic Student Services System – iKnow 816
A. Guseva, M. Gusev, S. Ristov

Sign Language Tutor - Rebuilding and Optimizing 822
N. Ackovska, M. Kostoska

Interoperability Challenges Among University Electronic Services Systems 828
M. Kostoska, M. Gusev, S. Ristov

Cloud Based Laboratory for Distance Education 834
S. Ristov, M. Gusev, A. Donevski

Mediated Heritage Learning in Formal and Informal Contexts: Antiquity CyberEd.. 840
G. Zlodi, Ž. Miklošević

Development of an Online Multimedia Course 845
M. Trošelj, M. Matetić, M. Tkalčić, M. Brkić, P. Pahljina

Are they a Perfect Match?
Analysis of Usage of Author Suggested Keywords, IEEE Terms and Social Tags 850
D. Vrkić

Lab Assessments in Undergraduate Course in Compilers for Students with No Prior Knowledge in Assembly 856
V. Kirandziska, M. Jovanov, M. Mihova, M. Gusev

Analysis of Elective Courses Selection in Post - Bologna Programmes 862
I. Rožanc, B. Slivnik

Data Mining of Computer Game Assisted e/m-learning Systems in Higher Education..868
P. Jurić, M. Matetić, M. Brkić

E-education 2.0: Students' Digital Identity and Online Learning Activities..... 874
S. Babić, M. Krešić, K. Kućel

Usefulness of Prezi and PowerPoint Presentation..... 880
A. Špernjak

An On-line Solution Verification System for Programming Competitions..... 883
A.S. Dorofeev

Implementation of Information Competencies as Key Employment Skills at the Faculty of Humanities and Social Sciences **886**
S. Špiranec, T. Ivanjko, Đ. Pečarić

Learning Management Systems: Trends and Alternatives **891**
L. Abazi-Bexheti, M. Apostolova-Trpkovska, A. Kadriu

Information Competences - Students' Perspective **896**
K. Pavlina, S. Špiranec, A. Pongrac Pavlina

Development and Analyses of "ProSysEngine" M-Learning Course Management System **900**
J. Saveski, M. Fetaji, B. Fetaji, M. Ebibi, L. Abazi

Investigation of Natural User Interfaces and Their Application in Gesture-Driven Human-Computer Interaction **906**
A. Petrovski, B. Fetaji, M. Fetaji, M. Ebibi

ICT Supported Laboratory Testing of Installation Circuit Breaker **913**
D. Gadže, M. Puzak, D. Haldek

Web-based Control of Programmable Logic Controllers in Process Control Computers Education **917**
R. Buinac, D. Maršić, G. Malčić

Towards More Efficient Teaching **921**
J. Žufić, T. Žajgar

The Potential of Multimedia in Cross-Cultural Education **926**
Z. Buzaši

Computer Network Applications, Practical Implementations and Structural Control System Representations **931**
N. Zeqiri

The Significant Progress of the Microprocessors and Microcontrollers Course for Computer Science Students **936**
S. Ristov, N. Ackovska, V. Kirandziska, D. Martinovikj

State of e-Learning Projects in Croatian Companies **942**
T. Dragičević, M. Ćukušić, M. Jadrić

PaintArt 2.0 - Computer Program for Education of Visual Arts in Secondary Schools **948**
Z. Prohaska, Z. Prohaska, I. Uroda

Informatics as a Basis for a Knowledge-Based Society **952**
M. Nadrljanski, V. Domitrović

Information Education in a Knowledge-Based Society **955**
M. Nadrljanski, V. Domitrović

Visual Verification of the Properties of Pseudorandom Sequences 959
S. Predanić, G. Vujisić, T. Horvat

Integration of Conceptual Data Modeling Methods: Higher Education Experiences... 963
Lj. Kazi, B. Radulovic, I. Berkovic, Z. Kazi

Technology Use in EFL Learning 969
D. Pešut

Development of a Computer System to Support Knowledge Acquisition of Basic Logical Forms Using Fairy Tale "Alice in Wonderland" 974
A. Mihaljević Španjić, A. Jakupović, M. Tomić

Designing the Programming Assignment for a University Compiler Design Course.... 979
I. Budiselić, D. Škvorc, S. Srbljić

Micro Assessment SaaS Cloud Solution 985
A. Bundovski, M. Gusev, S. Ristov

The Role of IC Technology in Development and Application of Experimental Methods and Multivariate Analysis 991
M. Orlić, M. Marinović

The Views of Students and Teachers on Implementation of E-learning in Educational Process 997
D. Glušac, D. Radosav, D. Karuović, Ž. Juhas

The Model of Remote Video Surveillance in Hierarchical Autonomous E-testing *WbeTS* System for Knowledge Testing 1001
D. Purković, E. Ban

Web Application Development with Component Frameworks 1007
V. Okanović

Agile Management: A Teaching Model Based on SCRUM 1011
Ž. Požgaj, N. Vlahović, V. Bosilj-Vukšić

Ethical Aspects and Capabilities of Social Networks – Students, Teachers, and Facebook in Elementary Schools 1017
D. Vincek

Computer Classroom Operating System Security (Windows) 1023
M. Zovkić, T. Vrbanec

Elementary School Students Using Facebook: A Case Study Of Croatian Elementary School ban Josip Jelačić 1029
M. Draženović

Evaluation Study and Results of Intelligent Pedagogical Agent-led Learning Scenarios in a Virtual World 1032
M. Soliman, C. Guetl

LeCTo: a Rich Lecture Capture Solution 1037
P. Pale, J. Petrović, B. Jeren

Using Data Mining for Learning Path Recommendation and Visualization in an Intelligent Tutoring System 1042
I. Jugo, B. Kovačić, V. Slavuj

Security Analysis of Wireless Network Access Following 802.11 Standard in Educational Institutions of the Republic of Croatia 1047
A. Skendžić, B. Kovačić

Use of iPads in Foreign Language Classes 1055
K. Starkl Crnković

Tablet PC & Smartphone Uses in Education (TabletTours) 1058
K. Bedi

Informatika kao izborni predmet ili izvannastavna aktivnost u razrednoj nastavi 1064
J. Šurić, T. Pavičić, M. Dumančić

Primjer projekta u nastavi informatike – što je lijepo ljudskom oku 1070
V. Skočir

Primjena interaktivnog videa u obrazovnim nastavnim sredstvima 1075
D. Vrbanc

Primjena e-portfolija u nastavi 1080
S. Šalković

Bonton digitalnog doba 1083
M. Mirković

IKT, ljudski resursi te informacijska i računalna sigurnost u hrvatskom osnovnom školstvu 1088
M. Zovkić, T. Vrbanec

Učinkovitost elektroničkog učenja u nastavi 1094
M. Sertić

Analiza uspješnosti u usvajanju algoritamskog načina razmišljanja kod učenika u prvom razredu novog strukovnog kurikuluma za zanimanje Tehničar za računalstvo 1110
K. Blažeka

iPad u nastavi matematike i fizike 1104
M. Babić

#codeEU 1107
I. Ružić

Prijedlog entitetsko-relacijskog modela podataka registara Hrvatskog kvalifikacijskog okvira 1110
M. Balković, N. Bačun Lebarović, H. Balen

Istraživanje o odlukama o odabiru karijere maturanata i brucoša na platformi Facebooka 1116
T. Krištof, H. Balen, V. Šebek

Nastavni plan i program za predmet programiranje u prirodoslovno-matematičkim gimnazijama 1122
J. Novaković

Koliko programiranja i kada nam treba u kurikulu za osnovne i srednje škole? 1127
M. Čičin-Šain, S. Babić

***Webinar* u funkciji promicanja prava na pristup informacijama 1130**
T. Horvat, R. Pekorari

Izračun viših momenata distribucija pomoću VBA programskog jezika u Excelu 1134
M. Gardijan, B. Šego, T. Škrinjarić

Važnost uvođenja novih tehnologija u obrazovne institucije za razvoj lokalnog gospodarstva 1140
A. Vukašinović

Izrada programskih komponenti u TIA Portal programskom okruženju 1146
D. Maršić, G. Malčić, I. Vlašić

Primjena industrijskih računalnih mreža u nastavi sa energetskim pretvaračima 1151
D. Maršić, G. Malčić, V. Tomljenović

E-učenje: koncept i primjene u europskom vojnom školstvu 1156
V. Zekulić, D. Možnik

Model ubrzanog usvajanja znanja putem moodle Akademije 1162
N. Bačun Lebarović, V. Šebek, D. Kučak

Masovnost, uspješnost i prolaznost u masovnim otvorenim *online* kolegijima 1168
M. Melinščak

COMPUTERS IN TECHNICAL SYSTEMS

INVITED PAPERS 1175

Learning From Success Cases in Steel Industry Electrical and Automation Applications 1177
P. Paganini

The Multimedia Devices Portable Automated Test System 1180
V. Peković, I. Kaštelan, G. Velikić, N. Teslić

PAPERS 1185

Static Characteristic and Describing Function of Polynomial Compensated Nonlinearities 1187
O. Kuljača, K. Horvat, Z. Eškinja

A New Approach for Maximum Utilization of the Mother Steel Plates for Rectangular Steel Plates Orders 1193
L. Kasapinov, M. Jovanov, D. Trajanov

Simulation Environment for Creating Artificial Range Data in Underwater Object Reconstruction 1200
I. Aleksi , D. Kraus, Ž. Hocenski

Cooperative Ramp Metering Simulation 1204
M. Gregurić, E. Ivanjko, S. Mandžuka

Interpretation of Divers' Symbolic Language by Using Hidden Markov Models 1210
M. Menix, N. Mišković, Z. Vukić

Cooperative Control of Heterogeneous Robotic Systems 1216
N. Mišković, S. Bogdan, I. Petrović, Z. Vukić

Program Solutions for the Complex Cooling Process via LADDER Logic 1221
T. Špoljarić, M. Fruk, G. Vujisić

The Use of the Internet and Wireless Communications in the Monitoring and Control of Industrial Processes 1227
V. Tasić, M. Pavlov, D. Brodić, V. Despotović, D.R. Milivojević

Community Structure in Networks: Girvan-Newman Algorithm Improvement 1231
Lj. Despalatović, T. Vojković, D. Vukičević

Performance Evaluation of Websocket Protocol for Implementation of Full-Duplex Web Streams 1237
D. Škvorc, M. Horvat, S. Srbljić

Improving Thread Scheduling by Thread Grouping in Heavily Loaded Many-Core Processor Systems 1243
L. Milić, L. Jelenković

Raspberry Pi as a Wireless Sensor Node: Performances and Constraints 1247
V. Vujović, M. Maksimović

Remote Firmware Update for Constrained Embedded Systems 1253
G. Jurković, V. Sruk

Performance-Occupation Trade-off Examination in Custom Processor Design 1258
D. Ivošević, N. Frid

Critical Path Method Based Heuristics for Mapping Application Software onto Heterogeneous MPSoCs 1264
N. Frid, V. Sruk

Event Calendar for Internet Data Sources 1269
P. Rantanen, P. Sillberg

Simulation Library for Resource Constrained Project Scheduling with Uncertain Activity Durations 1275
M. Brčić, N. Hlupić

Methods and Software for Estimation of Information System Dependability 1281
N. Šamec, A. Jakupović

Internet of Things Cloud Mediator Platform 1286
K. Mišura, M. Žagar

Second Screen User Experience: A New Digital Television Frontier 1291
N. Šoškić, N. Kuzmanović, M. Vidaković, G. Miljković

Device Cloud Platform with Script-Based Agents for "Anywhere Access" Applications Development 1295
M. Bjelica, N. Ignjatov, I. Papp, N. Teslić

Sound Quality Assessment in VOIP Environment 1300
D. Lukša, S. Fajt, M. Krhen

Automatic Interpretation of Aircraft Piston Engine Monitor Log 1305
D. Miljković

Regime Dependent Aircraft Piston Engine Monitoring 1311
D. Miljković

Konfiguriranje Profinet I/O mreže u radu sa energetskim pretvaračem 1317
G. Malčić, D. Maršić, J. Čučuk

Primjena SoftPLC koncepta procesnog računala prilagođene izvedbe s IEC1131-3 normom 1322
I. Vlašić, D. Maršić, G. Malčić

INTELLIGENT SYSTEMS

PAPERS **1331**

Intelligent Support Used for Providing a Pleasant User Experience **1333**
J. Kaljun, G. Harih, B. Dolšak

Regular Equivalence in Informed Network Search **1338**
D. Helic

Algorithms for 2D Nesting Problem Based on the No-Fit Polygon **1344**
D. Domović, T. Rolich, D. Grundler, S. Bogović

An Optimal Control Module for Sustainable Natural Resources Consumption Control Synthesis for Decision Support Systems **1350**
E. Cherkashin, S. Badmatsyrenova, I. Vladimirov, A. Popova, A. Davydov

Evolving Genetic Algorithms for Fault Injection Attacks **1356**
S. Picek, L. Batina, D. Jakobović, R. Boix Carpi

An Overview of Free Software Tools for General Data Mining **1362**
A. Jović, K. Brkić, N. Bogunović

Intelligent Multi Agent Systems for Decision Support in Insurance Industry **1368**
I. Markić, M. Štula, J. Maras

Solving Tourist Trip Planning Problem Via Simulated Annealing Algorithm **1374**
K. Sylejmani, A. Muhaxhiri, A. Dika, L. Ahmedi

Generating Spatio-Temporal Streaming Trajectories **1380**
E. Mešković, D. Osmanović, Z. Galić, M. Baranović

Automatic Recognition of Handwritten Corrections for Multiple-Choice Exam Answer Sheets **1386**
M. Čupić, K. Brkić, T. Hrkać, Ž. Mihajlović, Z. Kalafatić

Discrimination Between Serbian and Slovenian Language by Texture Analysis **1392**
D. Brodić, Z. Milivojević, Č. Maluckov, M. Jevtić

Script Independent Feature Set for Handwritten Text Recognition **1397**
D. Khanduja, N. Nain

An Efficient Feature Extraction Method for Segmented Cursive Characters Recognition **1403**
S. Panwar, N. Nain

An Annotated Urdu Corpus of Handwritten Text Image and Benchmarking of Corpus **1409**
P. Choudhary, N. Nain

Non-Standard Words as Features for Text Categorization 1415
S. Beliga, S. Martinčić-Ipšić

Developing a Text Classifier with Constrained Development and Execution Time ... 1420
I. Budiselić, G. Delač, K. Vladimir

Application of the NOK Method in Sentence Modeling 1426
M. Rauker Koch, M. Pavlić, A. Jakupović

An Improved Image Least Significant Bit Replacement Method 1432
K. Darabkh, I. Jafar, R. Al-Zubi, M. Hawa

New Recognition Methods for Human Iris Patterns 1437
K. Darabkh, R. Al-Zubi, M. Jaludi

Feature Selection for Object Detection: The Best Group vs. the Group of Best 1442
L. Fürst, A. Leonardis

Movie Posters Classification into Genres Based on Low-level Features 1448
M. Ivašić-Kos, M. Pobar, L. Mikec

Detection of Roadside Vegetation Using Features from the Visible Spectrum 1454
I. Harbaš, M. Subašić

Predictions of Cloud Movements and the Sun Cover Duration 1460
A. Radovan, Ž. Ban

Intelligent Modelling for Predicting Students' Final Grades 1466
M. Simjanoska, M. Gusev, A. Madevska Bogdanova

Determining the Impact of Demographic Features in Predicting Student Success in Croatia 1472
B. Trstenjak, Dž. Đonko

Intelligent Question – Answering Systems: Review of Research 1478
J. Tomljanović, M. Pavlić, M. Ašenbrener Katić

Inteligentni sustav upravljanja tunelskom rasvjetom 1484
Z. Balaž, K. Meštrović, P. Jandrlić, M. Čaržavec

Inteligentni sustav održavanja aerodromskog sustava svjetlosne signalizacije 1488
Z. Balaž, K. Meštrović, G. Bijelić

Mobilni uređaji u funkciji racionalizacije upravljanja udaljenim računalima emuliranjem računalnog miša 1493
D. Vukelić, M. Franušić, I. Grakalić

Pametna komunikacija na Internetu preko REST protokola..... 1498
D. Lozić, A. Šimec

INFORMATION SYSTEMS SECURITY

PAPERS **1505**

Password Security – No Change in 35 Years? **1507**
V. Taneski, M. Heričko, B. Brumen

Brute Force Analysis of PsychoPass-generated Passwords **1513**
B. Brumen, A. Černezel

Security Assessment of Eucalyptus' Web Management Interface **1519**
A. Donevski, S. Ristov, M. Gusev

CROFlux –Passive DNS Method for Detecting Fast-Flux Domains **1523**
T. Gržinić, D. Perhoč, M. Marić, F. Vlašić, T. Kulcsar

A Survey of Static Code Analysis Methods for Security Vulnerabilities Detection **1528**
M. Kulenović, Dž. Đonko

Random Flip-Flop and Its Applications **1534**
M. Stipčević

Comparison of ID-Based Blind Signatures from Pairings for E-Voting Protocols **1541**
P. Ribarski, Lj. Antovski

Homomorphic Encryption in the Cloud
D. Hrestak, S. Picek **1547**

Secure and Privacy-Enhanced E-Mail System Based on the Concept of Proxies **1552**
I. Kounelis, S. Muftic, J. Löschner

Requirements of Practice in Relation to the Existing Information Technology and Security Management Competencies **1558**
K. Arbanas, D. Alagić

Development of Users' Information Security Awareness Questionnaire (UISAQ) - Ongoing Work **1564**
T. Velki, K. Šolić, H. Očevčić

Question Based User Authentication in Commercial Environments **1569**
K. Skračić, P. Pale, B. Jeren

Technique for Detection of Cooperative Black Hole Attack Using True-link in Mobile Ad-hoc Networks **1575**
G. Wahane, A.M. Kanthe, D. Šimunić

The Effects of Combined Application of SOM, ANFIS and Subtractive Clustering in Detecting Intrusions in Computer Networks **1582**
Z. Avdagić, A. Midžić

Customer and User Education as a Tool to Increase Security Level 1588
I. Sedinić, Z. Lovrić, T. Perušić

Izradba zaštitnog profila pomoću ISO 27k kontrola 1593
Z. Kummer, V. Lebinac

Djela protiv intelektualnog vlasništva, kod kojih se kao sredstvo izvršenja krivičnih djela javljaju računari, računarske mreže, kao i njihovi proizvodi 1598
F. Gabela, A. Gabela

BUSINESS INTELLIGENCE SYSTEMS

INVITED PAPER 1605

Log-Driven User-Centric OLAP 1607
P. Marcel

PAPERS 1613

bis_01
Combining Domain and Business Ontologies in a Modular Construction Method: EDF Study Case 1615
K. Royer, L. Bellatreche, S. Jean

Data Modeling Styles in Data Warehousing 1621
V. Jovanović, D. Subotić, S. Mrdalj

Big Data and Sentiment Analysis using KNIME: Online Reviews vs. Social Media ... 1627
A. Mihanović, H. Gabelica, Ž. Krstić

Implementing the Additional Knowledge in the Croatian Scientific Bibliography 1632
T. Jagušt, J. Stojanovski, M. Baranović

The Impact of Training Data Tailoring on Demand Forecasting Models in Retail 1636
M. Karan, D. Pintar, Z. Skočir, M. Vranić, A. Alajković, J. Milojević, M. Pleša

Data Mining Meets Network Analysis: Traffic Prediction Models 1642
T. Eterović, S. Mrdović, Dž. Đonko, Ž. Jurić

An Approach to Conceptual Modelling of ETL Processes 1648
S. Dupor, V. Jovanović

UNDERSTANDER Business Intelligence Seeker – User Agent 1654
V. Damjanovic, W. Behrendt

Spend Analysis Systems **1660**
S. Maržić, P. Krneta, M. Pavlić

Poslovna inteligencija u sustavu informacijskog posrednika za elektronički račun ... **1666**
D. Matić, D. Kalpić, A. Čičak

Analiza podataka u svrhu predlaganja studijskih programa u Nacionalnomu informacijskom sustavu prijava na visoka učilišta **1672**
I. Budišćak, V. Mornar, D. Mlinarić

Aplikacija za uvoz podataka iz vanjskih izvora u skladište podataka **1677**
K. Matejčić, Z. Mršo

DIGITAL ECONOMY AND GOVERNMENT, LOCAL GOVERNMENT, PUBLIC SERVICES

PAPERS **1685**

Information Security and the Right to Privacy in Digital Economy – The Case of Republic of Croatia **1687**
M. Boban

E-business Models for Web Based Social Networks in the Republic of Macedonia **1693**
D. Jovevski, E. Tijan

E-preduvjeti fiskalizacije računa u Hrvatskoj **1699**
I. Arbanas

Legal Regulation of Public Information Infrastructure in Republic of Croatia **1706**
G. Vojković, T. Katulić

Content Management Systems for Mobile, Context-Dependent Augmented Reality Applications **1712**
M. Lengheimer, G. Binder, T. Rößler

Elektronički račun u Republici Hrvatskoj i SEPA inicijativa **1718**
D. Matić, B. Vrdoljak, D. Kalpić, A. Čičak

Statistical Microdata – Production of Safe Datasets, Transparent Presentation of Contents and Advanced Services for Users Through Metadata Authorization System **1724**
M. Poljičak, H. Stančić

Choosing Broadband Access Solution for Modern Residential Buildings in Densely Populated Urban Center **1730**
J. Topić, Z. Begić, I. Kapetanović Serdarević

Odabir rješenja širokopojasne pristupne tehnologije u gusto naseljenom području sa dominantno individualnim stambenim objektima 1734
J. Topić, Z. Begić, I. Kapetanović Serdarević

E-poslovanje zasnovano na internetu u funkciji nabave 1738
Z. Buljubašić, N. Buljubašić, I. Kapetanović Serdarević

E-plaćanje u uslovima ne postojanja potpunog sistemskog rješenja e-poslovanja..... 1744
Z. Buljubašić, N. Buljubašić, I. Kapetanović Serdarević

ERP Systems in Public Sector 1749
R. Kelemen

Oslonjivost sustava e-poslovanja u oblaku 1756
M. Buković, M. Davidović, P. Špoljarić

Comparative Analysis of Tourist Satisfaction and Online Booking Services Usage for Incoming Tourists in Zadar County 1761
B. Krce Miočić, G. Vidić, T. Klarin

Economic and Financial Analysis of Investments in Information Security..... 1767
Z. Čapko, S. Aksentijević, E. Tijan

Seaport Cluster Information Systems - a Foundation for Port Community Systems' Architecture 1774
E. Tijan, S. Aksentijević, D. Čišić

Development of e-Clusters..... 1780
M. Davidović

E-Consultations in Croatian Central Government 1786
A.J. Klasinc

Evaluation of Shared Digital Economy Adoption: Case of Airbnb..... 1791
Lj. Zekanović-Korona, J. Grzunov

The Digital Economy in Developing Countries - Challenges and Opportunities..... 1797
V. Lazović, T. Đuričković

Secure Data Storage 1803
P. Ristov, A. Mrvica, T. Mišković

Implementation of Green ICT for Sustainable Economic Development 1809
M. Vidas-Bubanja

MIPRO Junior – STUDENT PAPERS

PAPERS 1817

Complex Networks Measures for Differentiation between Normal and Shuffled Croatian Texts..... 1819
D. Margan, A. Meštrović, S. Martinčić-Ipšić

Comparison of the Language Networks from Literature and Blogs..... 1824
S. Šišović, S. Martinčić-Ipšić, A. Meštrović

Performance Analysis of Parallel Master-Slave Evolutionary Strategies (μ,λ) Model Python Implementation for CPU and GPGPU..... 1830
D. Zubanović, A. Hidić, A. Hajdarević, N. Nosović, S. Konjicija

Simulating of Query Processing on Multiprocessor Database Systems with Modern Coprocessors 1835
K.Y. Besedin, P.S. Kostenetskiy

Simulacija dotoka autoregresivnim modelom prvog reda 1838
S. Bošnjak, E. Sutlović

Korelacijska analiza dotoka 1844
K. Zoričić, E. Sutlović

Statistička analiza funkcije vjerojatnosti prirodnih dotoka 1848
I. Perišić, E. Sutlović

Cloud Dilemma..... 1854
R. Petrunić, I. Orović, V. Bušelić

Microcontroller Managed Module for Automatic Ventilation of Vehicle Interior 1859
B. Raić, A. Radovan

Acquisition Card NI USB-6009 in Automatic Control Laboratory Exercises 1865
H. Benić, M. Fruk, G. Vujisić

Udaljeno praćenje ljudske aktivnosti prikupljanjem podataka s ugrađenih senzora pametnog mobitela 1870
J. Begovac, L. Celić

Visualization of Streaming Trajectories 1874
D. Osmanović, E. Mešković

SPAM Trends and Techniques Used on Exchange Online System Used for racunarstvo.hr Domain 1879
I. Pavleković, V. Gašparević

An Approach Using Simulation Techniques to Estimate Quality of Service Parameters in Communication Networks..... 1883
H. Kozačinski, P. Knežević

Mobile Data Wi-Fi Offloading 1889
M. Sraga, N. Momčilović, A. Teković

Testing Optical Splitters 1891
I. Žuti, A. Komparić

Comparison of Different DC Motor Positioning Control Algorithms 1895
N. Baćac, V. Slukić, M. Puškarić, B. Štih, E. Kamenar, S. Zelenika

LIST OF PAPER REVIEWERS

Abello Gamazo, A. (Spain)
Afgan, E. (United States)
Aksentijević, S. (Croatia)
Antolić, Ž. (Croatia)
Ašenbrener Katić, M. (Croatia)
Avbelj, V. (Slovenia)
Babić, D. (Croatia)
Babić, S. (Croatia)
Bača, M. (Croatia)
Bako, N. (Croatia)
Balarin, M. (Croatia)
Ban, Ž. (Croatia)
Banek, M. (Croatia)
Banek Zorica, M. (Croatia)
Baotić, M. (Croatia)
Baranović, M. (Croatia)
Bartolić, J. (Croatia)
Basch, D. (Croatia)
Bažant, A. (Croatia)
Blašković, B. (Croatia)
Bosiljevac, M. (Croatia)
Bosnić, I. (Croatia)
Botički, I. (Croatia)
Brkić, L. (Croatia)
Brkić, M. (Croatia)
Broz, I. (Croatia)
Budin, A. (Croatia)
Budin, L. (Croatia)
Butković, Ž. (Croatia)
Butorac, M. (Croatia)
Capan, I. (Croatia)
Car, Ž. (Croatia)
Cifrek, M. (Croatia)
Cihlar, D. (Croatia)
Correia, P. (Portugal)
Čandrlić, S. (Croatia)
Čapko, Z. (Croatia)
Čeperić, V. (Croatia)
Čičin-Šain, M. (Croatia)
Čupić, M. (Croatia)
Davidović, D. (Croatia)
Delač, G. (Croatia)
Dembitz, Š. (Croatia)
Depolli, M. (Slovenia)
Dešić, S. (Croatia)
Dobrijević, O. (Croatia)
Dragić, L. (Croatia)
Dumić, E. (Croatia)
Džapo, H. (Croatia)
Eškinja, Z. (Croatia)
Fajt, S. (Croatia)
Faundez-Zanuy, M. (Spain)
Feki, J. (Tunisia)
Filjar, R. (Croatia)
Fischer, D. (Croatia)
Frid, N. (Croatia)
Gajić, R. (Croatia)
Galinac Grbac, T. (Croatia)
Gamberger, D. (Croatia)
Gamulin, O. (Croatia)
Garcia Mateo, C. (Spain)
Garza, P. (Italy)
Genova, J. (Bulgaria)
Gerić, S. (Croatia)
Golfarelli, M. (Italy)
Golub, M. (Croatia)
Golubić, S. (Croatia)
Gotovac, S. (Croatia)
Granić, A. (Croatia)
Grbić, R. (Croatia)
Grgić, K. (Croatia)
Groš, S. (Croatia)
Grubišić, I. (Croatia)
Gulić, M. (Croatia)
Henno, J. (Estonia)
Hocenski, Ž. (Croatia)
Horvat, M. (Croatia)
Hrabar, S. (Croatia)
Hrkać, T. (Croatia)
Hudek, V. (Croatia)
Huljenić, D. (Croatia)
Humski, L. (Croatia)
Hutinski, Ž. (Croatia)
Ilić, Ž. (Croatia)
Ipšić, I. (Croatia)
Ivanda, M. (Croatia)
Ivašić-Kos, M. (Croatia)
Ivošević, D. (Croatia)
Jaakkola, H. (Finland)
Jakobović, D. (Croatia)
Jakopović, Ž. (Croatia)
Jelenc, J. (Slovenia)
Jelenković, L. (Croatia)
Jeličić, V. (Croatia)
Jevtić, D. (Croatia)
Ježić, G. (Croatia)
Jovanovic, V. (Netherlands)
Jović, A. (Croatia)
Jurčević, M. (Croatia)
Juričić, V. (Croatia)
Jurišić, D. (Croatia)
Kalafatić, Z. (Croatia)
Kalpić, D. (Croatia)
Karaivanova, A. (Bulgaria)
Katanić, N. (Croatia)
Katzy, B. (Germany)
Kaučič, B. (Slovenia)
Kišasondi, T. (Croatia)
Kittl, C. (Austria)
Kljajić Borstnar, M. (Slovenia)
Kolar, D. (Croatia)
Komljenović, T. (Croatia)
Koričić, M. (Croatia)
Korin-Lustig, A. (Croatia)
Kosec, G. (Slovenia)
Kovačić, A. (Croatia)
Kovačić, B. (Croatia)
Krašna, M. (Slovenia)

Križanović, K. (Croatia)
Krois, I. (Croatia)
Kuljača, O. (Croatia)
Kušek, M. (Croatia)
Lacković, I. (Croatia)
Lacković, M. (Croatia)
Lerga, J. (Croatia)
Lešić, V. (Croatia)
Lipić, T. (Croatia)
Lišnjić, A. (Croatia)
Lučić, D. (Croatia)
Lugarić, T. (Croatia)
Lukac, D. (Germany)
Magdalenić, I. (Croatia)
Mance, D. (Croatia)
Mandić, T. (Croatia)
Marčetić, D. (Croatia)
Marinović, I. (Croatia)
Marinović, M. (Croatia)
Marković, I. (Croatia)
Markulin, D. (Croatia)
Martinčić-Ipšić, S. (Croatia)
Martinović, G. (Croatia)
Mateljan, V. (Croatia)
Matetić, M. (Croatia)
Matić, T. (Croatia)
Matijašević, M. (Croatia)
Mauher, M. (Croatia)
Mekovec, R. (Croatia)
Mekterović, I. (Croatia)
Meng, L. (United Kingdom)
Mihajlović, Ž. (Croatia)
Mihelič, F. (Slovenia)
Mikac, B. (Croatia)
Miklić, D. (Croatia)
Mikšić Trontl, V. (Germany)
Mikuc, M. (Croatia)
Miličević, K. (Croatia)
Milišić, P. (Croatia)
Min Tjoa, A. (Austria)
Mišković, N. (Croatia)
Mlinarić, H. (Croatia)
Molnar, G. (Croatia)
Mošmondor, M. (Croatia)
Možek, M. (Slovenia)
Mrković, B. (Croatia)
Nađ, Đ. (Croatia)
Nain, N. (India)
Novaković, V. (Croatia)
Oletić, D. (Croatia)
Orlić, M. (Croatia)
Pavčević, M. (Croatia)
Pavešić, N. (Slovenia)
Pavković, D. (Croatia)
Pavlić, M. (Croatia)
Pavlović, T. (Croatia)
Pečar-Ilić, J. (Croatia)
Pejčinović, B. (United States)
Pelin, D. (Croatia)
Perić Hadžić, A. (Croatia)
Petošić, A. (Croatia)
Pintar, D. (Croatia)
Poljak, M. (Croatia)
Popović, M. (Serbia)
Popović, Ž. (Croatia)
Požgaj, Ž. (Croatia)
Pribanić, T. (Croatia)
Pripužić, K. (Croatia)
Prokopec, G. (Croatia)
Pucihar, A. (Slovenia)
Randić, M. (Croatia)
Rashkovska, A. (Slovenia)
Ribarić, S. (Croatia)
Seršić, D. (Croatia)
Siladi, V. (Croatia)
Skala, K. (Croatia)
Skočir, Z. (Croatia)
Skorin-Kapov, L. (Croatia)
Sluganović, I. (Croatia)
Sović, I. (Croatia)
Sruk, V. (Croatia)
Stojković, N. (Croatia)
Strle, D. (Slovenia)
Štruc, V. (Slovenia)
Subašić, M. (Croatia)
Suciu, G. (Romania)
Suligoj, T. (Croatia)
Sužnjević, M. (Croatia)
Svigelj, V. (Slovenia)
Sviličić, B. (Croatia)
Šarić, D. (Croatia)
Šarolić, A. (Croatia)
Šegvić, S. (Croatia)
Šilić, M. (Croatia)
Šimić, Z. (Croatia)
Šišul, G. (Croatia)
Škoda, P. (Croatia)
Škvorc, D. (Croatia)
Šnajder, J. (Croatia)
Šunde, V. (Croatia)
Tijan, E. (Croatia)
Tomašić, I. (Slovenia)
Tonković, S. (Croatia)
Trobec, R. (Slovenia)
Trontelj, J. (Slovenia)
Tržec, K. (Croatia)
Turan, J. (Slovakia)
Turković, J. (Croatia)
Varga, M. (Croatia)
Vasić, D. (Croatia)
Vasilijević, A. (Croatia)
Vidaček-Hainš, V. (Croatia)
Vladimir, K. (Croatia)
Vranić, M. (Croatia)
Vrček, N. (Croatia)
Vrdoljak, B. (Croatia)
Vučić, D. (Croatia)
Vukmirović, S. (Croatia)
Vuković, M. (Croatia)
Weber, M. (Croatia)
Wrembel, R. (Poland)
Zlatović, M. (Croatia)
Žagar, D. (Croatia)
Žagar, M. (Croatia)
Žemva, A. (Slovenia)
Žuvić-Butorac, M. (Croatia)

AUTHOR INDEX

Abazi L. 900
Abazi-Bexheti L. 891
Abba A. 74
Ackovska N. 822, 936
Aćimović-Janežič R. 213
Afgan E. 388
Ahmedi L. 1374
Aksentijević S. 1767, 1774
Al Mawed S. 262
Alagić D. 1558
Alajbeg T. 757
Alajković A. 1636
Aleksi I. 1200
Aleksic S. 447
Al-Mistarihi M.F. 615, 681, 687
Alsaedeen M. 615
Al-Shua'b B. 615
Alves de Barros Naviner L. 57, 62
Al-Zubi R. 1432, 1437
Aljančič U. 102
Ambroš F. 593
Amon S. 102
An T. 57, 62
Antovski Lj. 1541
Apostolova-Trpkovska M. 891
Arbanas I. 1699
Arbanas K. 1558
Armellini C. 24
Arslanagić A. 504
Ašenbrener Katić M. 775, 1478
Atanassov E. 316
Avbelj V. 217, 258
Avdagić Z. 1582
Babić G. 147
Babić M. 1104
Babić S. 874, 1127
Bachler V. 172
Bačmaga J. 136
Bačun Lebarović N. 1110, 1162
Baćac N. 1895
Badmatsyrenova S. 1350
Bahtovski A. 395
Balaž Z. 1484, 1488
Balen H. 1110, 1116
Balković M. 1110
Ban E. 1001
Ban Ž. 1460, 127
Bánáti A. 369
Baranović M. 1632, 1380
Barić A. 110, 116, 136, 142
Batina L. 1356
Bedi K. 1058
Bednjanec A. 746
Begić Z. 1730, 1734
Begovac J. 1870
Begović A. 607
Behrendt W. 1654
Bekov E.B. 97
Beliga S. 1415
Bellatreche L. 1615
Bene K. 136
Benić H. 1865
Berkovic I. 963
Besedin K.Y. 1835
Bijelić G. 1488
Binder G. 1712
Bjelica M. 1295
Blažeka K. 1100
Boban M. 1687
Bogdan S. 1216
Bogović S. 1344
Bogunović N. 1362
Boix Carpi R. 1356
Bonetti A. 80
Bonfanti A. 74, 80
Borna F. 771
Bosilj-Vukšić V. 1011
Bošnjak S. 1839
Bratina T. 792
Brčić M. 1275
Brenna S. 74, 80
Britvić I. 524
Brkić K. 1362, 1386
Brkić M. 845, 868
Brodić D. 1392, 1227
Brumen B. 1507, 1513
Budić D. 492
Budiselić I. 979, 1420
Budišćak I. 1672
Buinac R. 917
Buković M. 1756
Buljubašić N. 1738, 1744
Buljubašić Z. 1738, 1744
Bundovski A. 985
Bušelić V. 1854
Buzaši Z. 926
Bužić D. 524
Capeska Bogatinoska D. 804
Caponio F. 74
Cattin P.C. 241, 434
Celić L. 1870
Cherkashin E. 1350
Chiappini A. 24
Choi C. 415
Choudhary P. 1409
Chydzinski A. 692
Cibiel G. 24
Cikojević A. 513
Cimerman M. 275
Conti G.N. 24
Cvijanović M. 513
Čačković V. 553
Čandrlić S. 775
Čapko Z. 1767
Čaržavec M. 1484
Čeperić V. 142
Černezel A. 1513
Čibej U. 252

Čičak A. 1666, 587, 1718
Čičin-Šain M. 1127
Čihak T. 162
Čišić D. 1774
Čučuk J. 1317
Čupić M. 1386
Ćesić J. 172
Ćosić T. 362
Ćukušić M. 942
Damjanovic V. 1654
Darabkh K. 1432, 1437, 615
Davidović D. 388
Davidović M. 1756, 1780
Davydov A. 1350
Delač G. 1420
Demirović D. 241, 434
Depolli M. 382
Despalatović Lj. 1231
Despotović V. 1227
Dika A. 1374
Divić J. 152
Dobravec M. 262
Dobravec T. 428
Dolšak B. 1333
Dolžan T. 102
Domitrović V. 952, 955
Domović D. 1344
Donevski A. 834, 1519
Dorofeev A.S. 883
Dragičević T. 942
Draženović M. 1029
Drezgić S. 599
Driussi F. 39
Duh M. 792
Dujak M. 537, 55933
Dular M. 247
Dumančić M. 1064
Dupor S. 1648
Džapo H. 172
Đerek V. 21, 33
Đonko Dž. 1642, 1472, 1528
Đuričković T. 1797
Đurić J. 524, 152
Ebibi M. 900, 906
Eškinja Z. 1187
Eterović T. 1642
Fajt S. 1300
Féron P. 24
Ferrari M. 24
Fertalj K. 641, 375
Fetaji B. 900, 906
Fetaji M. 900, 906
Filipović Tretinjak M. 746
Fischinger A. 275
Florescu I. 168
Forer L. 388
Franušić M. 1493
Frid N. 1258, 1264
Fruk M. 1221, 1865
Fu H.-B. 393
Fukuda K. 142
Fürst L. 1442
Gabela A. 469, 496, 1598
Gabela F. 469, 1598
Gabela J. 469, 496
Gabelica H. 1627
Gadanac D. 537
Gadže D. 913
Gajić R. 674
Galić Z. 1380
Galinac Grbac T. 635
Gams M. 283
Gamulin J. 781
Gamulin O. 781
Gardijan M. 1134
Gašparević V. 1879
Gebavi H. 21
Georgiev D. 316
Geršak B. 233
Gjoreski H. 283
Glavaš J. 787
Glušac D. 997
Golub I. 579
Goran N. 607
Gorgan D. 351
Gotić M. 17
Grakalić I. 1493
Grbeša I. 438
Gregurić M. 1204
Grubišić I. 438
Grundler D. 1344
Grzunov J. 1791
Gržinić T. 1523
Guetl C. 1032
Gusev M. 810, 816, 828, 834, 856, 985, 1466, 531, 310, 322, 328, 334, 395, 1519
Guseva A. 816
Gushev P. 810, 328, 334
Hadžialić M. 607
Hadžović S. 461
Hajdarević A. 1830
Hakštok I. 423
Haldek D. 913
Hao C. 57
Harbaš I. 1454
Harih G. 1333
Hawa M. 1432
Hayajneh A.M. 681
Heđi I. 664
Helic D. 1338
Henno J. 726, 734
Heričko M. 486, 611, 1507
Herman S. 262
Hidić A. 1830
Hlupić N. 1275
Hocenski Ž. 1200
Holenko Dlab M. 775
Holomb R. 13
Horvat G. 131, 184
Horvat K. 1187
Horvat M. 1237
Horvat T. 757, 959
Horvat T. 1130
Hren R. 221, 227
Hrestak D. 1547
Hrkać T. 1386

Hryhorova T.A. 571
Huang Y.-X. 393
Ignjatov N. 1295
Iliev I. 517
Ivanda M. 13, 17, 21, 24, 33
Ivanić V. 45
Ivanuš K. 116
Ivanjko E. 1204
Ivanjko T. 886
Ivašić-Kos M. 1448
Iveković R. 438
Ivošević D. 1258
Jaakkola H. 726, 734
Jadrić M. 942
Jafar I. 1432
Jagušt T. 1632
Jakobović D. 1356
Jakopović Ž. 162
Jakupović A. 974, 1426, 1281
Jaludi M. 1437
Janaček K. 674
Jandrlić P. 1484
Janevski T. 531
Jazbinšek V. 221, 227
Jean S. 1615
Jelačić H. 559
Jelenc J. 247, 279
Jelenković L. 1243
Jeong J. 415
Jerčić D. 537
Jeren B. 1037, 1569
Jeričević Ž. 419
Jevtić M. 1392
Jókai B. 288
Jovanov M. 856, 1193
Jovanović V. 1621, 1648
Jovevski D. 1693
Jović A. 1362
Jović M. 189
Jozić M. 584
Jug M. 262
Jugo I. 1042
Juhas Ž. 997
Jukić O. 664
Juričić V. 721
Jurić P. 868
Jurić Ž. 1642
Jurišić D. 121, 156
Jurković G. 1253
Kacsuk D. 288
Kacsuk P. 369
Kadriu A. 891
Kail E. 369
Kalafatić Z. 1386
Kalpić D. 1666, 1718
Kaljić E. 620
Kaljun J. 1333
Kamenar E. 1895
Kanthe A.M. 1575
Kapetanović Serdarević I. 1730, 1734, 1738, 1744
Karadimce A. 804
Karađole T. 189, 195, 201
Karahmetović N. 626
Karan M. 1636
Karin A. 513
Karóczkai K. 288, 369
Karuović D. 997
Kasapinov L. 1193
Kaštelan I. 1180
Katulić T. 1706
Kazi Lj. 963
Kazi Z. 963
Kelemen R. 1749
Kelemen S. 513
Kermek D. 781
Khanduja D. 1397
Khodeir M. 681
Kirandziska V. 856, 936
Kirin P. 513
Klarin T. 1761
Klasinc A.J. 1786
Klinger T. 800
Kljaić Z. 559
Knežević P. 1883
Kočo E. 172
Komparić A. 1891
Konecki M. 761
Konjicija S. 620, 1830
Kos L. 409
Kostenetskiy P.S. 1835
Kostoska M. 822, 828
Kounelis I. 1552
Kovach T. 13
Kovács J. 340
Kovačević D. 189, 195, 201
Kovačević S. 457
Kovačić B. 1042, 1047
Kozačinski H. 1883
Kozina S. 283
Kozlovszky M. 288, 369
Kožar I. 419
Krajnik J. 213
Kralj L. 708
Krašna M. 767, 792
Kraus D. 1200
Krce Miočić B. 1761
Krešić M. 874
Kreuh D. 262, 275
Krevatin I. 465
Krhen M. 1300
Kristan A. 275
Krištof T. 1116
Krivec S. 51
Krneta P. 1660
Kronenberg F. 388
Krpić Z. 184
Krstić Ž. 1627
Kučak D. 1162
Kućel K. 874
Kulcsar T. 1523
Kulenović M. 1528
Kuljača O. 1187
Kummer Z. 1593
Kušek M. 481
Kuzmanović N. 1291
Lacaita A.L. 74, 80

Lah H.C. 91
Lamza-Maronić M. 787
Lazović V. 1797
Lebinac V. 1593
Lengheimer M. 1712
Leonardis A. 1442
Lipić T. 388, 438
Liu K. 62
Livaja I. 641
Lojk J. 252
Lončarevič B. 265
Löschner J. 1552
Lovas G. 13
Lovas R. 340
Lovrić Z. 1588
Lozić D. 1498
Lozzi-Kožar D. 419
Lukovac B. 547
Lukša D. 1300
Luštrek M. 283
Maček Lebar A. 247
Madevska Bogdanova A. 1466
Madritsch C. 800
Madunić I. 201
Magerl M. 110
Majid H. 91
Mäkelä J. 726, 734
Mäkinen T. 726
Maksimović M. 1247
Malčić G. 917, 1146, 1151, 1317, 1322
Maluckov Č. 1392
Mandić T. 110
Mandžuka S. 1204
Maračić H. 481
Maras J. 1368
Marcel P. 1607
Margan D. 1819
Marić M. 1523
Marinov A.S. 97
Marinović M. 991
Markić I. 1368
Marković Č. 457
Markulin D. 500
Marri K. 292
Maršić D. 917, 1146, 1151, 1317, 1322
Martinčić-Ipšić S. 1415, 1819, 1824
Martinović G. 184
Martinović Ž. 492
Martinovikj D. 936
Maržić S. 1660
Matejčić K. 1677
Mateljan V. 721
Matetić M. 845, 868
Matić D. 1666, 587, 1718
Matković J. 375
Mazzola M. 24
Medved D. 599, 646
Medved Rogina B. 357
Međedović P.S. 668
Meixner Z. 288
Melinščak M. 1168
Memon A.S. 304
Memon M.S. 304
Menix M. 1210
Mesarić D. 593
Mešković E. 1380, 1874
Meštrović A. 1819, 1824
Meštrović K. 1484, 1488
Midžić A. 1582
Mihajlović Ž. 1386, 423
Mihaljević Španjić A. 974
Mihanović A. 1627
Mihelj T. 17
Mihova M. 856
Mijat N. 156
Mikac L. 17
Mikec L. 1448
Mikec L. 21
Miklavčič D. 247
Miklošević Ž. 840
Milanese D. 21
Milanič M. 221
Milić L. 1243
Milivojević D.R. 1227
Milivojević Z. 1392
Milojević J. 1636
Miljković D. 1305, 1311
Miljković G. 1291
Mirković M. 1083
Mišković N. 1210, 1216
Mišković T. 1803
Mišura K. 1286
Mitsa V. 13
Mlinarić D. 1672
Moguljak M. 721
Mohaisen R. 687
Momčilović N. 1889
Mornar V. 1672
Možek M. 102
Možnik D. 1156
Mrdalj S. 1621
Mrdović S. 1642
Mrozowski P. 692
Mršo Z. 1677
Mrvica A. 1803
Muftic S. 1552
Muhaxhiri A. 1374
Mulabegović J. 504
Multisilta J. 703
Musa K. 500
Nadrljanski M. 952, 955
Nagradić D. 121
Nain N. 1397, 1403, 1409
Najafi A. 86
Najafi A. 86
Niemi H. 703
Nosović N. 1830
Novak T. 757
Novaković J. 1122
Novosel L. 652
Očevčić H. 1564
Ogorevc A. 265
Okanović V. 1007
Orehovački T. 761
Orlić M. 991
Orović I. 1854

Osmanović D. 1380, 1874
Osterman A. 247
Oštarčević T. 584
Ovseník Ľ. 106
Paganini P. 1177
Pahljina P. 845
Pale P. 1037, 1569
Palestri P. 39
Pamuk L. 496
Panahi P. 771
Panwar S. 1403
Papić A. 715
Papp I. 1295
Pardon K. 121
Pavičić M. 347
Pavičić T. 1064
Pavičić T. 646
Pavić T. 477
Pavleković I. 1879
Pavlić M. 1660, 1426, 1478
Pavlin M. 252
Pavlina K. 896
Pavlov M. 1227
Pavlovski M. 741
Pečar B. 102
Pečarić Đ. 886
Pečlin P. 213
Pejčinović B. 28, 136
Pekorari R. 1130
Peković V. 1180
Pelli S. 24
Perhoč D. 1523
Perišić I. 1848
Perković L. 189, 201
Pernarić B. 235
Perušić T. 1588
Pešut D. 969
Petrov Dimitrov G. 517
Petrović I. 1216
Petrović J. 1037
Petrovski A. 906
Petrunić R. 1854
Picek S. 1356, 1547
Pintar D. 1636
Pleša D. 513
Pleša M. 1636
Pobar M. 1448
Podhraski M. 147
Podržaj J. 147
Poljak M. 33, 45, 51
Poljičak M. 1724
Pongrac Pavlina A. 896
Popa C. 66, 70
Popova A. 1350
Popović Ž. 553
Požgaj Ž. 1011
Predanić S. 959
Prgić H. 51
Primorac M. 715
Prljača N. 241, 434
Prohaska Z. 948
Prohaska Z. 948
Punčikar S. 659
Purković D. 1001
Puškarić M. 162
Puškarić M. 1895
Pušnik M. 486, 611
Puzak M. 913
Radchenko G. 401
Radojević B. 579
Radojević M. 428
Radosav D. 997
Radošević D. 761
Radovan A. 1460, 1859
Radulovic B. 963
Raić B. 1859
Ramponi R. 24
Rantanen P. 1269
Rashkovska A. 283
Rasoloniaina A. 24
Rauker Koch M. 1426
Repnik R. 767
Resnik D. 102
Režić V. 481
Ribarski P. 1541
Riedel M. 304
Righini G.C. 24
Ristić D. 21, 24
Ristov P. 1803
Ristov S. 810, 816, 828, 834, 936, 985, 310, 322, 328, 334, 1519
Roland V. 646
Rolich T. 1344
Rößler T. 1712
Rotaru C. 168
Royer K. 1615
Rozman J. 213
Rožanc I. 862
Ruda B. 288
Ružić I. 1107
Sabou A. 351
Saksida K. 178
Sandu C. 168
Sauer J. 233
Savchenko D. 401
Saveski J. 900
Schönherr S. 388
Sedinić I. 1588
Selmi L. 39
Sertić M. 1094
Severinac N. 646
Sharaqa A. 687
Shelida N. 91
Shurman M. 615
Sieng L.K. 91
Sillberg P. 1269
Simić M.I. 668
Simjanoska M. 1466
Sirotić I. 127
Skala K. 438, 362
Skeledžija N. 172
Skendžić A. 1047
Skočir V. 1070
Skočir Z. 1636
Skračić K. 1569
Slavuj V. 1042

Slivnik B. 862
Slukić V. 1895
Soklič Š.E. 279
Soko J. 513
Soliman M. 1032
Sraga M. 1889
Srbljić S. 979, 1237
Sruk V. 1253, 1264, 357
Stamenov S. 513
Stanchev O.P. 97
Stančić H. 1724
Stanič U. 213
Starkl Crnković K. 1055
Stipčević M. 1534
Stojanovski J. 1632
Subašić M. 1454
Subotić D. 1621
Suligoj T. 33, 45, 51
Sutlović E. 1839, 1844, 1848
Swaminathan R. 292
Sylejmani K. 1374
Šabanović H. 504
Šabeder A. 767
Šajn L. 252, 428
Šalković M. 584
Šalković S. 1080
Šamec N. 1281
Šebek V. 1116, 1162
Šego B. 1134
Šerbec M. 559
Šeremet R. 461
Šerifović-Trbalić A. 241, 434
Šerval D. 457
Šimec A. 1498
Šimunić D. 492, 547, 584, 1575
Šišović S. 1824
Škoda P. 357
Škrinjarić T. 1134
Škvorc D. 979, 1237
Šojat Z. 362
Šolić K. 1564
Šoša Anić M. 477
Šoškić N. 1291
Špernjak A. 880
Špiranec S. 886, 896
Špoljarić P. 1756
Špoljarić T. 1221
Štefanič A. 247
Štih B. 1895
Štimac M. 593
Štula M. 1368
Šuka D.S. 668
Šumak B. 486, 611
Šunde V. 127
Šurić J. 1064
Taneski V. 1507
Tanković N. 635
Tasić V. 1227
Teković A. 1889
Teslić N. 1180, 1295
Tijan E. 1693, 1767, 1774
Timarchi S. 86
Tkalčić M. 845
Tomaževič M. 275
Tomić D. 537
Tomić D. 235
Tomić M. 974
Tominc U. 262
Tomljanović J. 1478
Tomljenović V. 1151
Topić J. 1730, 1734
Trajanov D. 1193
Trbalić A. 241
Tretinjak M. 746
Trobec R. 217
Trontelj J. 147
Trošelj M. 845
Trošt A. 178
Trstenjak B. 1472
Turán J. 106
Turković J. 5
Urem F. 641
Uroda I. 787, 948
Vatavuk Z. 438
Veletić M. 269
Velikić G. 1180
Velki T. 1564
Velkoski G. 810, 310, 322, 334
Venica S. 39
Veres M. 13
Vesel M. 262
Veselinovska B. 531
Vidaković M. 1291
Vidas-Bubanja M. 1809
Vidić G. 1761
Vincek D. 1017
Vinko D. 131
Visković I. 465
Vitas I. 547
Vladimir K. 1420
Vladimirov I. 1350
Vlahović N. 1011
Vlašić F. 1523
Vlašić I. 1146, 1322
Vojković G. 1706
Vojković T. 1231
Vojvodić S. 481, 599, 646
Vorobiov L.V. 571
Vrana R. 751
Vrančić K. 152
Vranić M. 1636
Vrbanc D. 1075
Vrbanec T. 1023, 1088
Vrbovčan I. 477
Vrdoljak B. 587, 1718
Vrkić D. 741, 850
Vrsalović J. 195
Vrtačnik D. 102
Vučemilo H.N. 172
Vujisić G. 959, 1221, 1865
Vujović V. 1247
Vukašinović A. 1140
Vukelić D. 1493
Vukičević D. 1231
Vukić Z. 1210, 1216
Wahane G. 1575

Wang K.L. 33
Wang M. 33
Weißensteiner H. 388
Yew T.K. 91
Yusoff Y. 91
Zekanović-Korona Lj. 1791
Zekulić V. 1156
Zelenika S. 1895
Zeqiri N. 931, 566
Zhao L. 543
Zinterhof P. 299
Zlodi G. 840
Zoričić K. 1844
Zović M. 481
Zovkić M. 1023, 1088
Zrinšćak O. 438
Zubanović D. 1830
Žagar D. 184
Žagar M. 635, 1286
Žajgar T. 921
Žonja S. 33
Žufić J. 921
Žuti I. 1891

FOREWORD

The *37th International ICT Convention MIPRO 2014* was held from 26th until 30th of May 2014 in Opatija, the Adriatic Coast, Croatia. The Convention consisted of nine conferences under the titles: *Microelectronics, Electronics and Electronic Technology (MEET), Distributed Computing, Visualization and Biomedical Engineering (DC VIS), Telecommunications & Information (CTI), Computers in Education (CE), Computers in Technical Systems (CTS), Intelligent Systems (CIS), Information Systems Security (ISS), Business Intelligence Systems (miproBIS), Digital Economy and Government, Local Government, Public Services (DE/GLGPS)*. A special conference was dedicated to works of the students: *MIPRO Junior-Student Papers (SP)*.

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

The convention was organized by the Croatian ICT Society MIPRO with the help of numerous patrons and sponsors to whom we owe our sincere thanks. We specially single out our gold sponsors Ericsson Nikola Tesla, T-Croatian Telecom and Končar-Electrical Industries and the silver sponsor InfoDom. Our bronze sponsors are HEP–Croatian Electricity Company, VIPNet, Transmitters and Communications Company and Storm Computers.

To all who helped organize the 37th International ICT Convention *MIPRO 2014* as well as editing of this Book of Proceedings we extend our heartfelt thanks.

Prof. Petar Biljanović, PhD
International Program Committee
General Chair

MEET

International Conference on

MICROELECTRONICS, ELECTRONICS AND ELECTRONIC TECHNOLOGY

Steering Committee

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

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

INVITED PAPER

Nanotechnology, from the recent history to (un)predictable future

Josip Turković B.Sc.Physics
CEI Mikroelektronika d.o.o. , Zagreb, Croatia
josip.turkovic@hi.t-com.hr

Abstract - The diversity of nano-themes (nano science, nanoelectronics, nano materials, and other nano-something), did not result in sharper focus on the substantial processes behind the all of it. The author suggests that nanotechnology originated from the methods of microelectronics from the late1980's: ever growing complexity of mixed CMOS and bipolar planar silicon technology, and concept of MEMS devices as a way of integrating functions (and thinking) from former mutually exclusive area of electronics, machine engineering, biology, chemistry etc. The fully developed diagnostic methods enabled deep and accurate insights into the structure of material, and the methods of producing thin layered structures produce new complex entities with exotic physical and/or chemical properties. What is the impact of nanotechnology on the world outside the usual scientific community, from something like small business to corporative imperial giants? The author presents two examples of nano structured products: substrates for TEM and new type of MOS cell based on graphene and silicon as very appropriate cases for the business environment. Nanotechnology without any doubt can give complex structures with enlarged computing and processing power in very near future. But, what is it all about? Do we really need a smart toilet bowl, wich processes Tbyte information to make our stay in such a place more meaningful? In other words, do we need such an intense development just to produce new gadgets? From commercial point of view, the answer is a big YES. From the point of view of global human society progress, the answer is no. The rest of the paper presents the concept of the anisotropic boundaries as a product of already reached state of the nanotechnology, what boundaries can change the global civilizational paradigm, turning the meaning of the nano-something from the gadgetry to a new chapter in the evolution of human society.

I. Introduction

Most of the people involved in the microelectronic litography process, either in laboratory or production facility, know what the descumming of photoresist is. In autumn 1979, a group of technologists from RIZ Semiconductor Factory published a paper at a Munich conference about the method of preparation of the photoresist to get increased quality of edge definition in planar silicon technology, naming it descumming. The members of this group were Branka Šmalcelj, Antonia Delač, Rudi Ročak and myself. It has been one month after I joined the RIZ Semiconductor Factory.

In 1983, one of my colleagues Radja Jakiša and me solved the problem of eutectic bonding of silicon chips on nickel plated headers for high reliability hermetic components. The technology passed on out to western companies through a well known Dutch corporation, and to eastern counterparts through a Polish company. Mr Radja Jakiša is today a leader of litography team in Texas Instrument, Huston.

It was really an honor to join the people who produced a second ever made chip in Europe. That chip was produced in 1968 using improvised laboratory equipment at the RIZ Institute, and as a kind of particularity, that chip was in a triangular form.

Moreover, I have to mention a chip design group originated from RIZ Semiconductor Factory and leaded by Srebrenka Ursić and Damir Vuk which are still designing very sophisticated IC's, at the real edge of the modern microelectronic technology.

Because of the limitation of time and number of pages of this paper, it is impossible to mention all of the people we had been in contact with, from all over the world, and who participated in the crucial development of microelectronic technology during the 1980, and I apologize to them. They will probably be a part of some other story-telling circumstances.

II. Big bang in nanotech, mystery or what?

In the close communication with the people deeply involved in the development of nano-science in the last 15 years, nobody between them has noticed anything spectacular about the fact that new knowledge is produced in an accelerated rate. The effect can be seen clearly when comparing today's processes with the situation after the conceptual definition of Quantum Mechanics (QM). The prediction possibility of QM did not result in the development of the specific proper

technology. The nuclear reactor and bomb, usually considered the most important achievement of QM, in fact, were not realized with essentially new technology, but by the conventional methods of chemistry, physics, machine engineering, etc. In the last 10 years the first pure QM effect of photon entanglement, gave hope of serious QM-based technology in the field of quantum computing. The time interval of virtual "latency" of QM was about 90 years.

On the other hand, the experimental data from nanoscience were not predicted by QM , but all of them were explained by the QM. A good example of the excellent knowledge, and at the same time the impotence in the sense of technology breakthroughs, and engineering power.

Where does differences between QM and nanoscience come from? The answer is a very simple one. The development of microelectronics in the 1980, left as a legacy a few things:

1. highly developed diagnostic tools, not improvised and specific to a given laboratory, but standardized in scientific and high tech community.
2. standardized methods to manipulate micro objects in reproducible way, so you can get what you want by design.
3. MEMS (Micro electromechanical systems) as a concept of the intermixing functions of the systems based on electronics, biology, physics, machine engineering, etc. on the same chip.

The first two items represent the real tools, and the last item represent the change in the way of thinking.

Till the expansion of MEMS components, one of the dogma accepted in microelectronics was that on the silicon chip you can have electrical components and nothing but electrical. After breaking up any dogma someone expects a kind of progress. And that is just what happened in nano area: an almost explosive growth of knowledge, technology and engineering.

Such a sudden growth is not something the responsible researchers usually dream of. The clear example are protective coatings filled with various types of nanoparticles. Due to the excellent properties of self cleaning and resistivity to weathering, such coatings are pushed in mass production and commercial use, without full understanding of the mechanisms behind it and their environmental impact. The pressure to push new concepts from science to final products was sometimes illicitly high.

III. Nanotech in small and corporate business

The scientific community is a little bit confused by t very fast development of nanotechnology. What to say about the politicians, who are definitely incompetent to understand the subject, and its impact on the social and financial state of the given environment. There are all together 5 essentially distinct business environments: USA, Europe, Russia, China and Japan. As the length of this paper is limited, we will discuss only two of them.

A. USA - small and medium enterprises (SMEs)

In the USA, SMEs are virtually free to make trials in high tech domain, and by some subtle regulations, they are even pushed into the direction of new advanced technologies. If you obey some simple regulations, there are no restrictions on purchasing raw materials. It is expected, and the reality confirms it, that after proof of concept, more or less successful start up and recognized market for the future product, the small company has to be sold to the (probably serious) investor, like some corporations or specialized state agencies. In the last 5 years when I met some young people who just started as operators in small business in the USA, all of them declared that their ultimate business goal is to sell company to somebody who has real money. None of them wanted to make extra effort to make their companies grow and mature in the direction of the increased volume of money, goods, employees and organization. The maturing of the small companies is rather complicated process in USA.

B. Europe - SMEs

Except in the case of some special circumstances, small business in Europe is in the domain of high tech practically impossible. To start up company you need special licensees, and the process of evaluation lasts from 3 to 5 years. And if you get all papers needed (and it includes the data like what your great-grandma ate for breakfast), you are faced with the problem of purchasing the equipment and raw materials.

You can start the company in a normal way if you have a supervisor like a well known science institute or some special state agencies. That increases the overall costs, but who cares? Free lancers in high tech small business are not allowed in Europe. In spite of that, there are a lot of subjects in small and medium high tech business in Europe, but they were all created by original European corporations which are part of the global European establishment. Therefore if you want to make some serious small business in high tech in Europe, think either about founding a secret (I mean really secret) society, or make a deal with some of the corporations, which is the simplest solution.

C. USA - corporate business

Corporations in the USA have all the attributes of imperial system. They have their own long range development plan, market share contracts and well defined area of influence. Their power is based on whichever high technologies they own today, and the one they will own in the near future, accompanied by all necessary human resources. Corporate task force

teams are in some cases at the level of police forces of some small countries, with the goal of assuring the domination of the corporation in the forseeable future. But what is the most important, these corporations are not a part of the establishment. Nevertheless, these corporations are under the tight control of the state in every sense.

D. Europe - corporate business

Corporations in Europe have attributes of feudal system with lords behind the stage. Their power is based primarily on political power, and high techs are of secondary character. All corporations are a part of the European establishment, and are rarely leaders in the field of global R&D. In the last 10 years German corporations started to copy US model, which was not welcomed in the rest of (old) Europe, and will probably become a seed of the future tensions between European feudal lords. The newcomers in Europe are without any importance regarding existing corporate structures, which has to be expected as obvious: otherwise, they will never become newcomers.

E. Example 1: very simple product from nanotech domain

Figure 1 shows the SEM photo of the silicon nitride membarane with etched holes in it. Thikness of the membrane and diameter of the holes are both 10^{-7}m. Such peices of membrane are produced on 0.2 m diameter silicon wafer. Each piece is of square shape $3 \cdot 10^{-3}$m by $3 \cdot 10^{-3}$m, mounted on ceramic frame 10^{-3}m thick. Such small membranes are carriers for the samples for transmission electron microscopy (TEM). But such carriers can be used for the few more purposes like:

- Cell biology: attached cells can be grown in their environment on the support film and subsequently analyzed
- Analysis of colloids, aerosols, nanoparticles
- Self-assembled mono-layers
- Polymer research
- Thin film research (directly deposited on the Silicon Nitride Support Film)
- Materials science
- Properties of nano-structures for semiconductor devices
- Semiconductor: characterization of thin films
- Catalyst development

If we focus only to the application of the membrane in TEM, let us estimate what is financial effect of producing (and selling) such membranes. The membranes are packaged as a set of 10 units, with purchase price of $400. Because of the simple structure, 0.2m diameter silicon wafer has yield of at least 80%. Purchase price

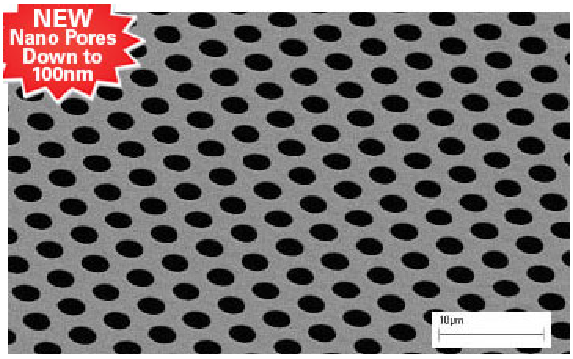


Figure 1: Thin silicon nitride membrane with 100 nm holes

of silicon wafer (which is processed in microelectronic laboratory or production facility, not in small business company) is app. $\$5,000.00$. Cutting the wafer into separate pieces is done in small company, and it requires clean room, diamond saw, and not more then 50 m^2 space and no more then 2 workers. With two silicon wafers per month, including all fixed expenses and taxes, it is a job of more then 2 million $ per year. Not so bad for such a simple product.

Last month I had a meeting with some young people from the "Faculty of Electrical Engineering and Computing" in Zagreb. They asked me if I can suggest a few types of products which can be produced in small nano laboratory. After about one week, I found about 120 examples of simple products like the just described silicon membrane for the TEM. Not a single case is from Europe, in spite of the fact that Europe has good enough laboratories for processing similar products, and that Europe is one of the biggest consumers of such products.

I would like now to say something about the nanotechnology in my homeland Croatia. We all know that politicians have a life full of stress.They are starting every day by checking wheather their pockets are deep enough, so that in every moment they can be ready to steal some more money from the people. The politicians in Croatia are even pushed to specialize themselves in those activities in an accelerated way, so they don't have time to analyze what could be the significance of the nanotechnology in such a small country like my homeland. So let us no disturb them in their everyday struggle for their better life. So much about nanotechnology in Croatia.

F. Example 2: graphen double layer photo detector

As a second example I choose a case which was announced as a note in the official magazine of AIOP (American Institute of Physics) for publishing the news from science and technology [1]. Before I explain why I made such a choice, let me quote the reporter Tushna C. of physicsworld.com : "*Graphene is a layer of carbon just one atom thick and since it was first isolated in 2004 its remarkable electronic and mechanical properties have been studied by physicists worldwide....*", and

"*....and colleagues from the University of Michigan have been developing photodetectors that can take advantage of graphene's unique material properties, including its ultra-broadband light-absorption capability. As graphene is a semimetal, it is capable of absorbing light across a wide spectrum – from ultraviolet to far infrared. This is unlike conventional photodetectors, which are made from semiconductors that can only absorb light at specific wavelengths.*".

The authors of the innovation are Chang-Hua Liu, You- Chia Chang, Theodore B. Norris & Zhaohui Zhong [2]. From the journalist's report and abstract of the paper [2], it looks like the team from the University of Michigen has done some everyday's routine job. But let us analyze few details from the Figure 2.

1. Two layers of the graphene are clearly sketched, and they represent the whole active part of circuitry. The first layer accumulate positive electric charge and electrostatically influences electric current in the layer beneath. The active layer is only two atoms thick.

2. Silicon dioxide and silicon thin layers are noth ing more than carriers for the graphen's active structure and electric contacts. From the whole legacy of microelectronics, a new type of active device use planar silicon technology "only" as a way to support the new structure. But without this support, there is no way to realize such a device. That is the true meaning of the valuable legacy, after all.

3. The graphen is recognized as an exotic material in 2004. In less than 10 years we are faced with a very sophisticated new concepts on the basis of the knowladge of about totally new and unusual phenomena.

4. By the authors of the innovation, the device is compared with MOSFET structure, but that is essentially wrong. The device is a new type of active device which can be called GCG (graphen-channel-graphen) device, or anything else, but not the MOS structure. Sensitivity, power consumption, fast switching and noise tolerance when designing circuits with such devices, must be reconsidered from the beginning.

5. Small power dissipation, and very thin active structure, give the possibility to stack the circuitry in the third dimension. By my modest opinion, that is the best promise of not so distant future.

6. Promoting the concept of the new device, from laboratory work to industrial product with reasonable application, that job cannot be done by the small and medium business subjects. That is typycal object case for corporative business, where financial power altogether with human resources can change the world as we know it.

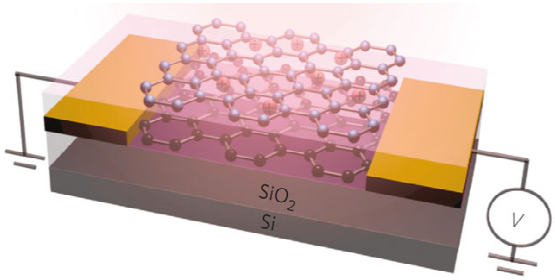


Figure 2: The graphen double-layer photo detector

IV. Ever growing circuit complexity, what is all about ?

The measure of the complexity of the electric circuits on microelectronic (or nanoelectronic) chips, is accepted as a measure of the state of the art of the technology achieved, or as the grade of excellency of some laboratory or facility. The question is, do I really need an intelligent toilet bowl whith controled temperature, humidity, light, and which connects me to the social networks on the internet and read me a status of my credit card, all presented in pretty way as a projection on my eye-glasses? Complexity of electronic circuits built into microcontrollers, with appropriate software behind, allows such solutions, and without any doubt such products have their market and satisfied customers.

In recent communication with some people from big corporations, I got an information that more then 80 % of their production based on microelectronic components, are in fact gadgetry. So, it seems that the main mission of microelectronic technology is to find new type of play-toys. Maybe it is good for business, but I find myself a little bit humiliated. The role of the science and the development of the technology are the means by which human species is getting better chance to survive, with additional responsibility for the rest of the planetary life-forms. It seems that people forgot that threats around are very real. One piece of the stone 10 km in diameter can end the history of civilization. Besides, we already have silent kiler(s) present in a form of new ugly diseases, and apparently slow climate changes. The role of the science has not to be the creation of the new computer games with multiple virtual realities.

What we really need is a kind of mission defined on a global level. The mission means a coherence in understanding what is the goal in front of us and how to move closer to them. Many years before, one american politician said in his famous speech, I quote:

"*No nation which expects to be the leader of other nations can expect to stay behind in this race for space. ...We choose to go to the Moon in this decade and do the other things, not because they are easy, but because they are hard.*"

By this words the whole nation (and great part of the rest of the world) recognized what the mission is.

The Europeans are very suspicious to any idea of coherency and mission. The reason is of historical nature. When somebody says "one homeland, one people, one state and one leader", and if you translate this words into German language, everything becoms much clearer.

I don't no either who is able to define the mission on the global level or what that mission would be, but I can try to show you how to open small window with a sight into not so distant future. Let us look!

V. Concept of the anysotropic boundaries

Suppose we have a plate (vertically positioned) in air as shown by the sketch in Figure 3. Having in mind the elementary collision processes from kinetic theory of gases, simple boundary on each side of plate is place where the molecules of air are reflected from the surface, either elastically or plastically. If one of these boundary does not reflect the molecules, but leaves them pass through the plate, we have a case of anysotropic boundary. In Figure 2, the boundary closer to viewer is standard boundary, and boundary on the other side of plate is anysotropic boundary.

What are the physical consequencies of the presence of the anysotropic boundary?

1. On the side where the boundary is anysotropic, there is no effect of the air pressure. The air pressure is created by the change in momentum of the molecules in the process of reflection from the body's surface. If the boundary does not reflect molecules, there is no pressure on the body.

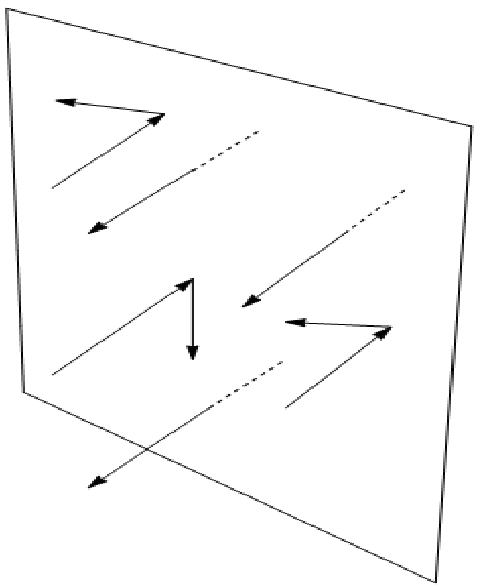


Figure 3: Anysotropic boundary is behind of the plate

2. On the side where the boundary is "normal" and reflects the molecules of the air, pressure is acting on the surface of the body.
3. Unbalanced pressure on oposite surfaces of the plate, results in the force in direction from "normal" boundary" to anysotropic boundary, and that is what we actually call the thrust.
4. If the body is moving in the direction so that anysotropic boundary is the attacked surface, there is no resistance to the flow of fluid, and there is no formation of the density discontinuities in the fluid. In other words, moving bodies with velocities greater than the speed of sound in the fluid do not produce so called bow shock and wake shock.

Modern airfighters usually have two jet engines, whith overall thrust of 75, 000 N per engine as an short time allowable maximum. Depending on the greatest inner diameter of the expansion chamber, I will make a rough estimation of specific thrust to about 60,000 (N/m^2). If we relize the anysotropic boundary with efficiency of **only 1 %**, our specific thrust on plate with anysotropic boundary would be 1,000 (N/m^2).

Now we must take into account the fact that jet engine is not shorter than 4 m, and that we can stack plates with anysotropic boundary in a sandwich system, satisfying the condition that distance between two plates must be greater than 100 times the mean free path of the air molecules. Let stack them at the distance of 5 mm between plates. Such a combination would give a specific thrust of 800,000 (N/m^2) what is almost 14 times greater of what modern airfighters can achieve.

And, yes...I forgot to emphasize: sandwich system of the plates with one anysotropic boundary does not need fuel! Knowing that *perpetum mobile* is not possible, where all this energy for such a trust is coming from? The answer is: from the kinetic energy of air molecules, or in other words from internal energy of the fluid. The air on the side of the anysotropic surface is in the state of the adiabatic expansion, which results in cooling, and the air on opposite side is in the state of adiabatic compression, which results in heating. The difference between heating and cooling is converted into usefull work delivered to the plate.

In the early days of the thermodynamics, analysis of the engines which use internal energy of the surrounding fluid, has been violently disregarded, almost prohibited. Only circular processes like Carnot and similar, were accepted and fully analyzed, probably because by the technology they had only such termodynymic engines could have been realized.

There are three different ways to realize anysotropic boundaries, and all three by the nanotechnology processes. We have some extras regarding the application of anysotropic boundaries outside the conventional fluids like gases and liquids. The vacuum state is by the QM theory a foam of boiling

short-living particles which oscillate between virtual states and real states in our space. This concept is approved by the experiment with so called Casimir's force in the vacuum.

There are two ways to realize anysotropic boundaries for the applications in vacuum environment. And of course, you guess : by the method of nanotechnology. What are the consequencies of such possibilities?

First, let us analyse how the conventional rocket technology has been used in space travel till today. After reaching the Earth low orbite, space module engages rocket engines and after a few minutes when the required velocity is obtained, the state of the module becomes passive. The module follows the so called ballistic trajectory. With ballistic trajectories you need several days to reach the Moon, 6 to 14 months to reach the Mars (depending on relative position of Mars and Earth), and more than 30,000 years to reach a nearby star.

If you have engines based on anysotropic boundaries, you can afford a kind of active trajectory. The reason is the enormous thrust possibility and no need for fuel. The fuel is right there in the space. If we decide to use active trajectory with acceleration of 1 g to the halfway, and deceleration of 1 g to the destination, the typical times to reach some space destinations are: 3,5 hours to the Moon, 49 hours ours to the Mars, 8 Earth years or 1,5 month of space module time to the nearby star. In this last case you have to account for the relativistic effect of time dilatation for space module. The calculation is very simple, and you can chack it by yourself.

Let us now close that window with our sight set on the near future. Maybe the drought from the future is not so harmless as we expected.

Because of the physical limitation of this paper, one question will be left unanswered: are you sure that you are ready to face the new space challenge?

We can reshape the question in a more familiar form: are you ready "*...boldly to go where nobody has gone before?*"

VI. Conclusions

A breif review of the fast rate of development in the domain of new technologies (nanotechnology), clearly shows that the achivement of the microelectronics during the 1980 is a legacy which boosts overall processes in science and technology progress. There can be no science progress without technology progress. Science and technology are the two faces of the same coin.

The lack of the coherent vision of the future on the global level, as a rule, results in pushing the huge potential of the nano science into the banality, which eventually results in the flurishing of gadgetry. It is necessary to remember what is the more profound mission of science and technology.

The lack of vision is probably the cause of the fear of the future in the human mind. It looks like humanity is punished to stay inside the borders of the Earth's habitat. Already for many years the humanity has been treating our planet as a jail. The environmental care is replaced by hate.

As a part of the possible vision of the future, the author announce an original concept of anysotropic boundary. A small number of possible repercussions are discussed, showing that humanity's habitat is much more spacious than we thought. It is certain that the hopes oriented in that direction will definitely change the existing civilizational paradigm. But, that is a story for some other paper.

Thanks

Before the daedline for submitting the paper, I did not received the permission to use a picture of double layer graphen photodetector, published in [1]. I hope that there is no reason to refuse such a permission, because papers submitted to MIPRO are not of the commercial character. Whatever, after all, I express my gratitude to the editors of IOP magazine for using the published drawing.

I would thank all the readers of this paper for having understood that it was imposible to cite all the authors whose papers are read to get the informative database about the fact of extraordinary fast progress in domain of the nanotechnology (more than 200 authors). But the abstracts of their papers can be found in IOP magazine collections for the year 2005-2014, avaliable online at physicsworld.com.

References

[1] Tushna Commisariat, "Graphene gains thermal vision", IOP Physics World, March 28 2014.

[2] Chang-Hua Liu, You-Chia Chang, Theodore B. Norris and Zhaohui Zhong ,"Graphene photodetectors with ultra-broad band and high responsivity at room temperature", Nature Nanotechnology Volume: 9, Pages: 273–278 , 16 March 2014.

[3] IOP Physics World magazine, Archive, collections 2005.-2014., avaliable online at url: physicsworld.com.

PAPERS

Spectroscopic evidence of coexistence of clusters based on low (α) and high temperature (β) GeS_2 crystalline phases in glassy germanium disulfide matrix

V. Mitsa[*], R. Holomb[*], G. Lovas[*], M. Veres[**], M. Ivanda[***], T. Kovach[*]

[*] Institute of Solid State Physics and Chemistry, Uzhhorod National University, 88000 Uzhhorod, Ukraine.
v.mitsa@gmail.com
[**] Wigner Research Centre for Physics, Hungarian Academy of Sciences, 1121 Budapest, Hungary
[***] Division of Materials Physics, Rudjer Boskovic Institute, Bijenicka 54, 10000 Zagreb, Croatia

Abstract – Technologically modified g-$GeS_2(T_iV_j)$ glasses prepared by melt quenching from different temperatures (T_i) and with different cooling rates (V_j) were studied using Raman spectroscopy and model calculations. Differential Raman spectra {$I_RGeS_2(T_iV_1)$-$I_RGeS_2(T_2V_1)$} showed the splitting of the main, most intensive wide-band, centered at 340 cm^{-1}. The position of the peaks found in the differential Raman spectra near 340 and 360 cm^{-1} and position of the main vibrational modes in the ordinary Raman spectra of low- and high-temperature crystalline phases (α-, β-GeS_2) were found to be in good agreement. Vibrational spectra of 4- and 6-member ring fragments selected from crystalline structure of β-GeS_2 were also determined using model calculations. Based on rings composed of GeS_4 tetrahedrons connected by corners and edges the structural interpretation of the vibrational bands of g-$GeS_2(T_iV_j)$ at 360, 370 and 433 cm^{-1} in their Raman spectra was performed. The existence of homopolar Ge-Ge and S-S "defect" bonds in the structure of germanium disulfide glasses was proved by formation of 5-member rings. Formation of fragments of both low- and high-temperature α-, β-GeS_2 crystalline phases in the structure of g-$GeS_2(T_iV_j)$ glasses indicates the existence of mixed medium-range ordering in g-GeS_2.

I. INTRODUCTION

Glassy germanium disulfide is among chalcogenide glasses attracting much attention and widely studied for many years. Thanks to the large infrared optical transmission window it found usage as material for night vision, acusto-optical modulators, chemical sensors, mid-infrared light delivery, thermal imaging, environmental monitoring, integrated optics and also in the area of fast ion conducting glasses for battery applications [1-6]. One of the promising new ways of its mid-IR application is to prepare glass-ceramics with controlled crystallization [7]. Due to their specific optical properties transparent glassy ceramics are considered as perspective materials for photonics [7]. A number of structural studies were performed on this material and it was concluded that the medium-range order existing in g-GeS_2 is connected with clustering of GeS_4 tetrahedral units by edges or corners and the two kinds of linkages produce both six- and four-member rings in a 2:1 ratio, like in layered β-GeS_2 crystalline structure [8-13]. In the low temperature α-GeS_2 crystalline phase the corner-sharing $GeS_{4/2}$ tetrahedra form a three-dimensional (3D) structure [14,15]. The density of α-GeS_2 is only about 0.055 g/cm^3 higher than that of β-GeS_2 [14]. The structure of α-GeS_2 crystal was used as starting configuration in a Reverse Monte Carlo (RMC) calculations [16]. In contrast to the common conclusions [8-13] about layered structure of g-GeS_2, the RMC simulation revealed that the structure of g-GeS_2 is somewhere between the structure of α-GeS_2 and β-GeS_2 crystals [16]. This conclusion was also confirmed by studies of g-GeS_2 performed at high pressures [17], where short-range structures of densified g-GeS_2 transformed from a mixed α-GeS_2 and β-GeS_2 type arrangement into an α-GeS_2-like structure [17]. Comparison of the Raman spectra of g-GeS_2 and α- and β-GeS_2 crystals in [18] showed that the middle-range structural order of g-GeS_2 is controlled by a corner-shared three-dimensional network which is topologically similar to that in the α-GeS_2 crystal. Similarity of medium-range ordering of g-GeS_2 and α-GeS_2 was also confirmed by the study of GeS_2 microcrystals grown in g-GeS_2 during annealing at 497 °C, below the glass transition temperature [18]. Linear crystal growth kinetics of both α-GeS_2 and β-GeS_2 polymorphs has also been observed in g-$Ge_{42}S_{58}$ over a relatively broad range of temperatures, *i.e.* 420 < T < 494 °C that correspond to viscosity of supercooled melt: $3 \times 10^9 > \eta > 8 \times 10^5$ Pa s [19].

In spite of the findings described above, the structural information about medium-range ordering in g-GeS_2 is controversial. It is still debated whether the structure of g-GeS_2-type glass is a continuous or a broken network of Ge–$(S_{1/2})_4$ tetrahedra [20]. The new potential applications, however, compel a need for thorough structural characterization of these glasses prepared in different technological regimes, including the utilization

of both experimental techniques and atomic scale modeling.

II. Materials and Methods

The g-GeS_2 samples were synthesized by melt quenching from different temperatures: 1173 K (T_1), 1273 K (T_2), 1373 K (T_3) and 1473 K (T_4) and cooling rates of 100 K/s (V_1) and 150 K/s (V_2). The four different g-GeS_2 samples will hereafter be denoted as (T_1V_1), (T_2V_2), (T_3V_2) and (T_4V_2). DFT calculations were performed on Ge_nS_m (n=2,3,5,6,12; m=6–9,14,16,30) units that represent the local structure of g-GeS_2 and some 'defect' Ge_nS_m clusters (Fig.1). The local short-range order structural units representing the beta form of GeS_2 crystal are corner- and edge-sharing tetrahedra (models I and II). Further extension of the length scale to the larger coordination spheres leads to an increase in the degree of freedom in the formation of amorphous structures. In this case, the structure formation may be studied in detail by *ab initio* method using a gas-phase cluster approach. The largest cluster which can be found in the crystal structure of GeS_2 is a six-member ring consisting of three $GeSe_4$ tetrahedra connected by their corners (III). It should be noted here that two edge-sharing tetrahedra can also be classified as a four-member ring. Models IV and V represent two six-member rings selected from two directions of 2D of β-GeS_2 crystal. In the first case the rings are connected by common Ge atom while in the other model rings are connected via the edge-sharing block. The largest cluster obtained from the crystal structure of GeS_2 monolayer (VI) consists of a big 16-member ring. Here, we represent few possible models with homopolar Ge–Ge (VII–VIII) and S–S (IX–XI) bonds. The simplest model of Ge–Ge bond being present in the hypothetical cluster is the so-called ethane-like geometry (VII). There is also possibility of S–S bond formation on the edge of the so-called 'outrigger raft' structure (model XI represents a part of this structure) [8]. In α-GeS_2 the $GeS_{4/2}$ tetrahedra pack to form a dense three-dimensional network with large hollows surrounded by twenty four-member rings together with six-member rings (Fig. 2).

Raman spectra of g-GeS_2 and β-GeS_2 were measured using Dilor–Labram and Renishaw system 1000 Raman spectrometers equipped with CCD detectors. Excitation was performed with lasers operating at 1064 and 785 nm.

III. Results and Discussion

Fig. 3 shows the Raman spectra of g-GeS_2 glasses measured using excitation photon energy of 1.17 eV (1064 nm), being less than the Tauc optical gap of GeS_2 glasses (E_g=3.2 eV) [21]. All these spectra are very similar to those measured with shorter excitation wavelengths (632.8 and 785 nm) and they can be treated as non-resonant Raman spectra. All spectra are normalized to their 340 cm^{-1} peak. A pronounced difference between the samples is that both the (T_1V_1) and the (T_3V_2) samples show similar small increases in the intensity of 370 and 433 cm^{-1} vibrational modes, as compared to (T_2V_2). For the (T_4V_2) sample these peaks broadened and merge with the main peak at 340 cm^{-1}. A very weak c-GeS-like Raman mode at ~200 cm^{-1} is observed for all samples while a weak S-S mode at 490 cm^{-1} is detected only for (T_2V_2) showing evidence of c-GeS type clusters (Fig.4) and S-rich clusters (IX–XI) respectively (Fig.1).

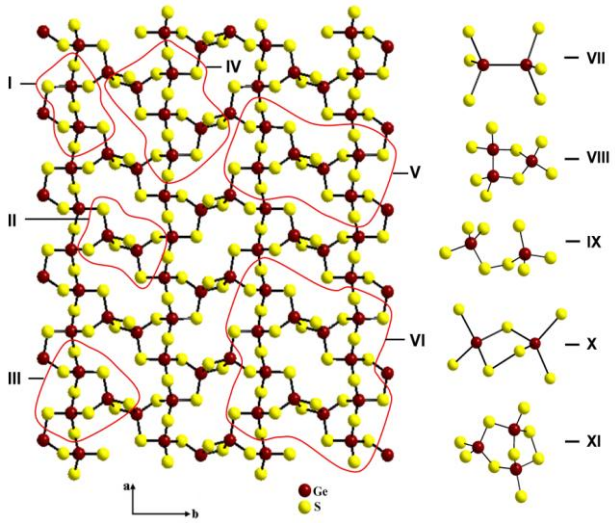


Figure 1. The crystal structure of high- temperature (β) GeS_2 monolayer [15] with the selected Ge_nS_m cluster models (I–VI) (left) and the 'defect' Ge_nS_m cluster geometries with homopolar Ge–Ge (VII–VIII) and S–S (IX–XI) bonds (right). Saturating hydrogen atoms and corresponding bonds are not shown for clarity.

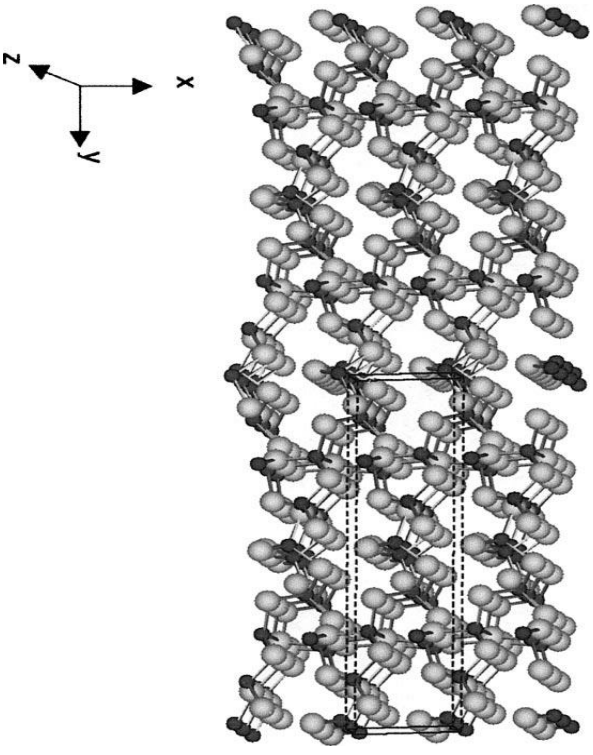


Figure 2. Crystal structure of α-GeS_2 (low-temperature modification) [14].

The Raman spectrum of g-$GeS_2(T_2V_2)$ didn't show band near 255 cm^{-1} being characteristic for Ge–Ge bonds in (VII–VIII) type of clusters (Fig.1). The Raman spectrum of this (T_2V_2) technological regime contains only minority of 'defect' clusters and was chosen as reference spectrum for the extraction of other differential Raman spectra of g-$GeS_2(T_iV_j)$ samples. The Raman modes at 360 and 370 cm^{-1} are related to the Ge-S stretching vibrations in so called corner- and edge-sharing tetrahedra, respectively [21]. However, the former modes can also be characteristic for an 'outrigger raft' cluster [8]. Our recent theoretical study on formation energy, stability and electronic properties of different Ge_nS_m clusters indicates that the formation of single ethane-like cluster, single corner sharing GeS_4 tetrahedra and cluster based on two GeS_4 tetrahedra connected by an S-S bridge are energetically not favorable structural motifs [22]. On the other hand, we have found that the ring-like structures were most favorable within our Ge_nS_m cluster models. One of our previous studies [21] showed that the vibrational mode at ~250 cm^{-1} observed in the Raman spectra of GeS_2 glass can be interpreted as the existence of GeS microphase rather than ethane-like Ge_nS_m cluster with Ge-Ge bond.

The differential Raman spectra {$I^RGeS_2(T_iV_2)$-$I^RGeS_2(T_2V_2)$}, (i=1, 3 and 4) clearly show two bands centered at 339 and 360 cm^{-1} (Fig.5) which are typical for Raman spectra of α-GeS_2 and β-GeS_2, respectively [23,24]. These two bands are more distinct in the differential Raman spectra {$I^RGeS_2(T_iV_1)$-I^Rβ-GeS_2} (Fig.6). Differential Raman spectrum {$I^RGeS_2(T_1V_1)$-I^Rβ-$GeS_2(T_2V_1)$} is mainly β-GeS_2 like (Fig.5). For this (T_1V_1) technological regime the fraction of edge-sharing tetrahedra in the glassy structure increases. By increasing temperature from T_2 to T_3 and T_4 the concentration of 5-member rings with Ge-Ge bonds and GeS-like clusters

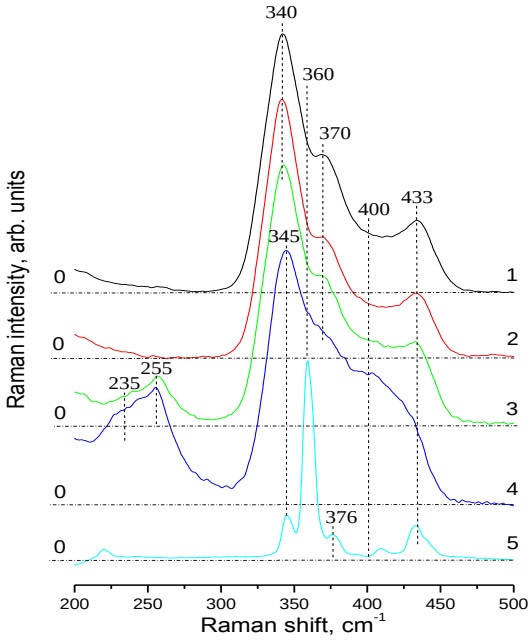


Figure 3. Raman spectra of GeS_2 glasses: 1- T_1V_1; 2- T_2V_2; 3- T_3V_2; 4- T_4V_2; 5- Raman spectrum of β-GeS_2.

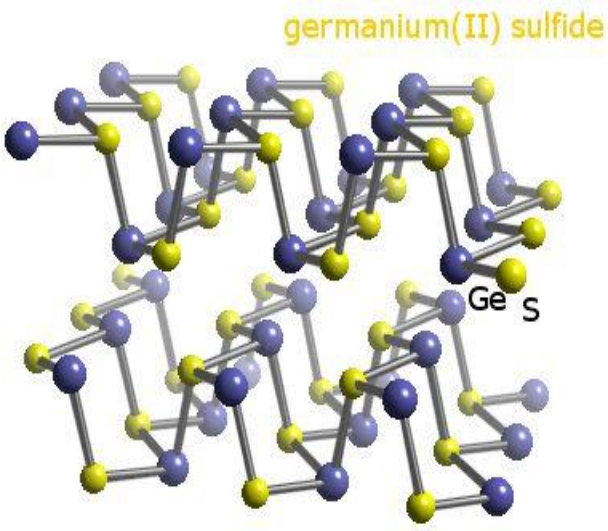


Figure 4. Crystal structure of c-GeS [14]

increases and concentration of 4-member rings with edge-sharing tetrahedra in the structure decreases. Therefore, in structural aspect, nucleates with medium-range ordering based on α-GeS_2 and β-GeS_2 crystal type clusters coexist in the structure of g-$GeS_2(T_iV_1)$ during nucleation of the melt by quenching from different temperatures.

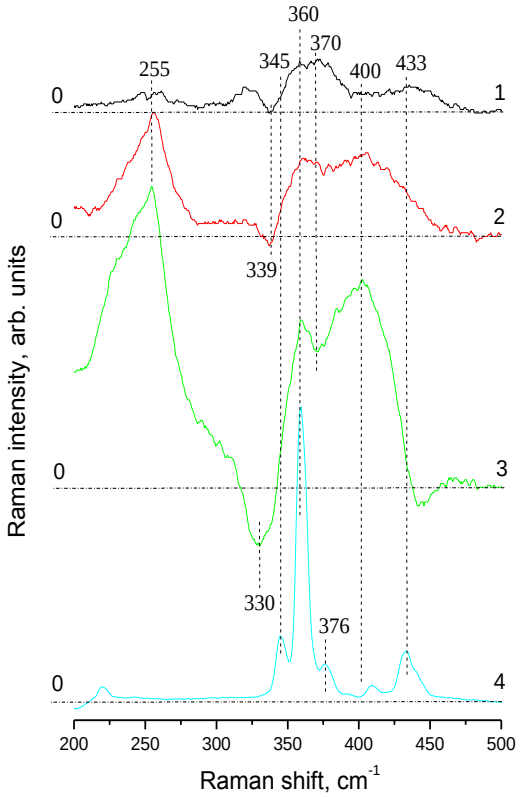


Figure 5. Differential Raman spectra {$I^RGeS_2(T_iV_1)$-$I^RGeS_2(T_2V_2)$}: 1- T_1V_1; 2- T_3V_2; 3- T_4V_2; 4 - Raman spectrum of β-GeS_2.

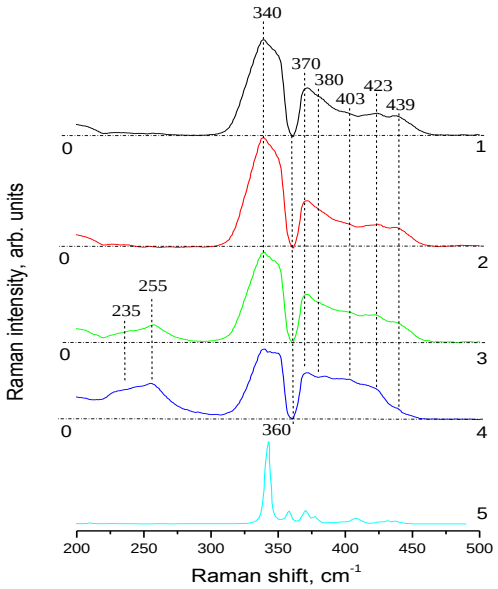


Figure 6. Differential Raman spectra {$I^RGeS_2(T_iV_1)$- $I^R\beta$-GeS_2}: 1 - T_1V_1; 2 - T_2V_2; 3 - T_3V_2, 4 - T_4V_2; 5 - Raman spectrum of α-GeS_2 [18].

IV. Conlusion

Two bands, centered at 339 and 360 cm^{-1} were observed in the differential Raman spectra {$I^RGeS_2(T_iV_1)$-$I^RGeS_2(T_2V_2)$, (i=1,3,4), which are typical for most intense bands in Raman spectrum of α-GeS_2 and β-GeS_2 crystals, respectively. These two peaks appear more clearly in the differential spectra {$I^RGeS_2(T_iV_1)$-$I^R\beta$-GeS_2), i=1-4. In structural aspect, during the melt clustering by quenching a middle-range structural ordering is created in g-GeS_2 as mixture of corner-shared three-dimensional clusters being topologically similar to those in α-GeS_2 crystal and middle range order which is topological similar to the two-dimensional β-GeS_2 crystal type.

Acknowledgements

R.H. gratefully acknowledges the support from the Hungarian Academy of Sciences within Domus Hungarica Scientiarum et Artium Programme.

References

[1] W.H. Zachariasen, "The crystal structure of germanium disulfide", J. Chem. Phys, vol.4, pp. 618-619, 1936.

[2] A.F. Ioffe, A.R. Regel, "Non–crystalline, amorphous and liquid electronic semiconductors", Progress in semiconductors. London. vol.4, №20, pp. 25-30, 1960.

[3] J. Savage, S.Nilsen, "Chalcogenide glasses transmitting in the infrared between 1 and 20μm- a state of the art review", J.Infrared Phys, vol. 5, pp. 195-204, 1965.

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

[5] Inseok Seo and Steve W. Martin. New Developments in Solid Electrolytes for Thin-Film Lithium Batteries, Lithium Ion Batteries - New Developments, Dr. Ilias Belharouak (Ed.), (2012). ISBN: 978-953-51-0077-5.

[6] V. Mitsa, M. Ivanda, O. Gamulin, R. Holomb, O. Kondrat, N. Popovych, N. Tsud, V. Matolín, K.C. Prince, G. Lovas, S. Petrecky,"Visible luminescence and synchrotron XPS spectra of amorphous Ge_2S_3 based films", 2013 MIPRO Proceedings of the 36th International Convention, pp. 34-39, 2013.

[7] Changgui Lin, Haizheng Tao, Xiaolin Zheng, Ruikun Pan, Haochun Zang, and Xiujian Zhao1, "Second-harmonic generation in IR-transparent-GeS_2 crystallized glasses", Optics Letters, vol. 34, No. 4, pp. 437-439, 2009.

[8] P.M. Bridenbaugh, G.P. Espinosa, J.E. Griffiths, J.C. Phillips, and J.P. Remeika ,"Microscopic origin of the companion A_1 Raman line in glassy $Ge(S,Se)_2$", Phys. Rev. B., vol.20,№10, pp. 4140-4144, 1979.

[9] R.J. Nemanich , F.L. Galeener, I.C. Mikkelsen G.A.N. Connel, "Configuration of chemically ordered continuous random network to describe the structure of $GeSe_2$ glass," Phisica B., vol.117–118, pp. 959–961, 1983.

[10] G. Lucovscky, C.K. Wong, W.B. Pollard, "Vibrational properties of glasses: intermediate range order", J. Non–Cryst. Sol, vol. 59–60, №2, pp. 839–846, 1983.

[11] P. Boolchand, J. Grothaus, M. Tenhaver, M. Halze, R.K. Grasseli. "Structure of GeS_2 glass: spectroscopic evidence for broken chemical order", Phys. Rev. B., vol. 33, pp. 5421-5434, 1986.

[12] P.S. Salmon, R.A. Martin, P.E. Mason. "Topological versus chemical ordering in network glasses at intermediate and extended length scales", Nature, vol.435, pp. 75–78,2005.

[13] P. Armand, A. Ibanez, H. Dexpert, D. Bittencourt, D. Raoux, E. Philippot, "Structural approach of Ge-X, GeX_2 -Ag_2X (X= S,Se) glassy systems", Journal de physique IVColloque C2, vol. 2, pp. c2-189-c2-194, 1992.

[14] C.T. Crewitt, H.S. Young, "Germanium and silicon disulfides: structure and synthesis", Science, vol. 149, pp. 535-537– 1965.

[15] G. Dittmar and H. Schäfer, "Die Kristallstruktur von Germaniumdiselenid", Acta Cryst. B., vol. B32, pp. 2726-2728, 1976.

[16] X. Zhao, H. Higuchi, Y. Kawamoto, "Structural analysis of GeS_2 glass by means of reverse Monte Carlo simulation**",** Physics and Chemistry of Glasses - European Journal of Glass Science and Technology Part B , vol. 39, pp. 98-100, 1998.

[17] K. Miyauchi, J. Qiu, M. Shojiya, Y. Kawamoto, N. Kitamura, "Structural study of GeS_2 glasses permanently densifed under high pressures up to 9 GPa", Journal of Non-Crystalline Solids, vol. 279, pp. 186-195, 2001.

[18] Z. Černošek, E. Černošková, L. Beneš, "Raman scattering in GeS_2 glass and its crystalline polymorphs compared ",Journal of Molecular Structure, vol. 435, pp. 193–198, 1997.

[19] J. Shanelova, J. Malek, M. Alcala, J.Criado, "Kinetics of crystal growth of germanium disulfide in $Ge_{42}S_{58}$ chalcogenide glass", Journal of Non-Crystalline Solids, vol. 351, pp. 557-567, 2005.

[20] V. Petkov, D. Le Messurier, "Atomic-scale structure of $GeSe_2$ glass **r**evisited: a continuous or broken network of Ge–$(Se_{1/2})_4$ tetrahedra?", J. Phys.: Condens. Matter., vol. 22, pp.115402-115408, 2010.

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

[22] R. Holomb, V. Mitsa, S. Akyuz , E. Akalin, "New ring-like models and *ab initio* DFT study of the medium-range structures, energy and electronic properties of $GeSe_2$ glass", Phil. Mag., vol. 93, pp. 2549–2562, 2013.

[23] K. Inoue, O. Matsuda, K. Murase, "Raman spectra of tetrahedral vibrations in crystalline germanium dichalcogenides, GeS_2 and $GeSe_2$, in high and low temperature forms", Solid State Commun., vol. 79, pp. 905–910, 1991.

[24] I.P. Kotsalas, C. Raptis, "Structural Raman studies of Ge_xS_{1-x} chalcogenide glasses", Journal of Optoelectronics and Advanced Materials, vol. 3, pp. 675–684, 2001.

Synthesis and Characterization of Silver Colloidal Nanoparticles and Their Application in Surface Enhanced Raman Spectroscopy

L. Mikac*, M. Ivanda*, M. Gotić** i T. Mihelj***

*Laboratory for Molecular Physics, Ruđer Bošković Institute, Bijenička c. 54, Zagreb, Croatia

**Laboratory for Synthesis of New Materials, Ruđer Bošković Institute

***Laboratory for synthesis and processes of self-assembling of organic molecules

ivanda@irb.hr

Abstract. – Silver colloids were produced by chemical reduction of silver salt (silver nitrate, $AgNO_3$) solution. As a reducing agents trisodium citrate, sodium borohydride, ascorbic acid, PVP (Polyvinylpyrrolidone) and glucose were used. The characters of silver nanoparticles were investigated using UV-Vis spectrophotometer, dynamic light scattering (DLS), zeta (ζ) potential measurements and scanning electron microscopy. The as prepared silver nanoparticles were used in SERS measurements of pyridine. The best enhancement of the signal was achieved using sample reduced and stabilized with citrate, with the particle size of approximately 40 nm and non-uniform distribution. As aggregating agents $NaNO_3$, KCl and KBr in different concentrations were tested. The Raman signal for pyridine was enhanced even more with the addition of sodium borohydride instead of other aggregating agents. Addition of borohydride had the largest impact to sample stabilized with DEAE-dextran which consists of large (0.5 μm) as well as very small silver nanoparticles. The mixture colloid-borohydride-pyridine was stable for hours.

I. INTRODUCTION

During the past decades metal nanostructures received great interest in many fields due to their interesting optical and electric properties. They have been used in electronics [(1)], catalysis [2], biomedicine [3] and surface-enhanced Raman scattering (SERS) [4],[5]. There is particular interest in silver nanostructures due to their outstanding plasmonic properties which lead to a great enhancement of the Raman signal. SERS is an advanced Raman technique that enhances the vibrational spectrum of molecules adsorbed on or in the vicinity of metal particles.

It is currently accepted that two mechanisms, electromagnetic and chemical, contribute to the SERS enhancement [6]. In the electromagnetic mechanism, the laser light excites collective oscillations of the surface electrons in a metal which results in the enhancement of the local electromagnetic field. The molecule which is near the metal surface or adsorbed to it, experiences this enhancement and as a result the intensity of the Raman scattered light is enhanced. In the chemical mechanism the polarizability of the molecule adsorbed on the metal surface increases. The total Raman signal enhancement is the product of both processes. In recent times a significant enhancement of the Raman signal has also been obtained from surfaces of oxide nanosensors (α-Fe_2O_3 thin films) [7] which indicates the importance of the chemical enhancement. Because of its readiness, sensitivity, and minimum sample preparation requirements, SERS is being considered as a powerful technique for detection of wide variety of analytes at very low concentrations, even down to the single molecule level [8].

The main issue today is the synthesis of metal (Ag, Au, Cu) nanoparticles with tunable size, shape and morphology [9]. Silver colloids are usually synthesized by in-situ reduction methods, such as chemical reduction, photo reduction [10], electrochemical reduction [11], or thermal decomposition [12]. Chemical reduction involves the reduction of a silver salt with a reducing agent in the presence of stabilizer. The most used methods for the preparation of silver nanoparticles in aqueous solution are reduction of silver nitrate, $AgNO_3$, using reducing agents such as citrate [13],[14], ascorbic acid [15], or borohydride [16]. The purity of water and reagents, cleanliness of the glassware and solution temperature are critical parameters in the synthesis process. In addition, the nanoparticles prepared by chemical reduction are usually mixtures of spherical and rod-like silver particles due to the poor balance of nucleation and growth processes [17]. Nevertheless, due to simplicity of preparation and the fact that they offer fairly good SERS enhancement (up to 10^6) under acceptably reproducible experimental conditions, silver colloids with stabilizing agents have been widely used. The stabilizing agents are chemicals used in the synthesis of nanoparticles that prevent their aggregation through electrostatic or steric repulsion (or both). In the case of silver colloids most commonly used stabilizing agents are citrate, sodium borohydride, polyvinylpyrrolidone (PVP) and CTAB (cetyltrimethylammonium bromide). According to the type of stabilizing agent particles of varying size, shape and stability are obtained.

II. MATERIALS AND METHODS

Sample AGC was prepared using chemical reduction method according to the slightly modified method of Lee and Meisel. Sample AGA was synthesized using ascorbic acid as reducing agent and citrate as stabilizer. Sample AGBH was synthesized using ice-cold sodium borohydride according to Creighton method. Sample AGD was synthesized using sodium borohydride as reducing agent in the presence of dextran as stabilizer. Sample AGP was prepared with PVP as reducing and stabilizing agent. Sample AGG was prepared by reducing the silver nitrate with glucose in PVP aqueous solution with sodium hydroxide which accelerates the reaction.

The absorption optical spectra of silver samples were recorded using UV–Vis–NIR spectrophotometer (UV-3600, Shimadzu). All the spectra were recorded at room temperature.

The morphology of silver nanoparticles in silver samples was investigated with Jeol JSM 7000F field emission scanning electron microscope (FE SEM) coupled with Energy-dispersive spectroscopy (EDS). 10 µL of colloid solution was pipetted on silicon wafer and allowed to dry in ambient conditions. Dynamic light scattering (DLS) measurements were performed by a Zetasizer Nano ZS (Malvern Instruments, UK) equipped with a 532 nm "green" laser.

The Raman spectra were recorded by Horiba Jobin Yvon T64000 triple Raman spectrometer in the subtractive mode. The 514.5 nm argon ion laser and 532 diode laser were used.

III. RESULTS AND DISCUSSION

A. Colloid characterization

Figure 1 shows UV-Vis spectra of the synthesized silver nanoparticles in the range from 200 to 800 nm. The absorption band in visible light region from 350 to 550 nm is typical for silver nanoparticles and the samples AGC, AGA, AGBH and AGD have the absorption maximum at 407, 402, 390 and 410 nm respectively. AGC sample has the broadest band indicating non-uniform distribution of silver nanoparticles. Samples AGA, AGBH and AGD have smaller values of absorption peak FWHM indicative of a narrow size distribution. FE-SEM measurements were carried out to observe the size and morphology of the silver nanoparticles prepared with different reducing and stabilizing agents. Figure 2 shows SEM image of synthesized nanoparticle. The sample AGC prepared with citrate consists of quasi-regular spherical nanoparticles with small amount of rod-like nanoparticles.

TABLE 1. CHARACTERIZATION OF THE PREPARED SILVER SAMPLES.

Sample	Reducing/ Stabilizing Agent	SEM (nm)	Zeta (mV)	Hydrodynamic diameter (nm)	UV-Vis-NIR (nm)	pH
AGC	citrate/citrate	35.8+/- 6.3	-33.4	14.6; 69.8	407	7.7
AGA	ascorbic acid /citrate	16.7 +/- 4.4	-46.8	5.4	402	7.9
AGBH	$NaBH_4/BH^{4-}$	10.9 +/- 2.9	-10.1	2.87	390	7.5
AGD	$NaBH_4$/DEAE-dextran	518 +/- 114	39.5	473	410	8.2
AGP	PVP/PVP	11.6 +/- 4.1	-0.867	4.62	420	5.6
AGG	glucose/PVP	12.6 +/- 4.2	-11.2	14.28	283	7.9

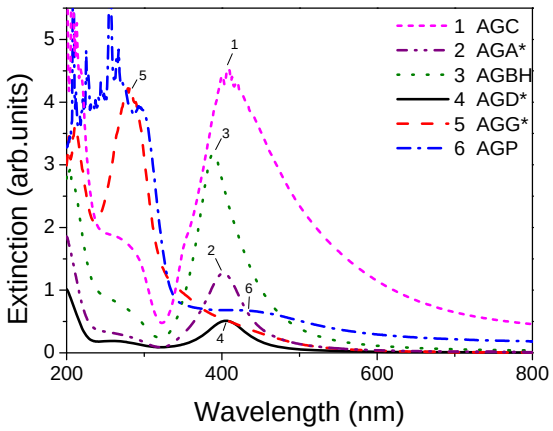


Figure 1. UV–Vis spectra of the silver nanoparticles synthesized. The samples marked with asterisk were diluted with water (sample AGA 10 times; samples AGD and AGG 300 times).

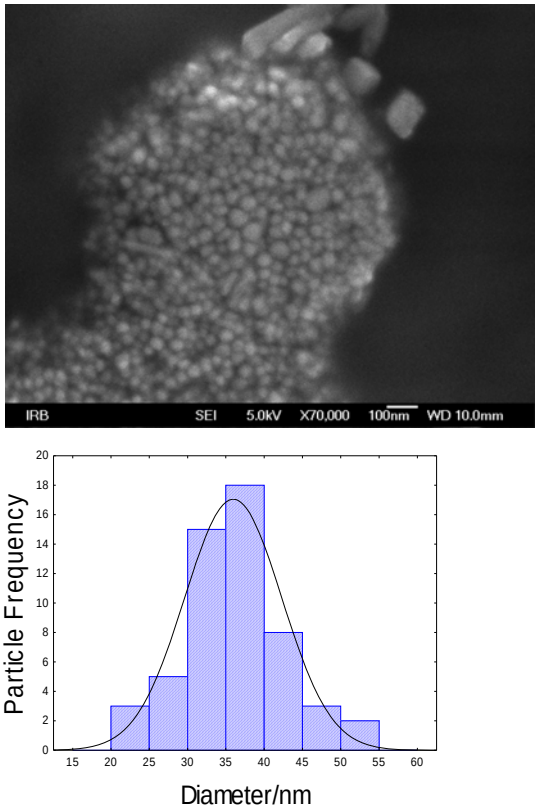


Figure 2. SEM image and particle diameter for AGC sample.

The particle size distribution was calculated from the corresponding SEM images using ImageJ computer program. The calculated mean particle sizes and their standard deviation are given in Table 1. Also, the hydrodynamic diameters and zeta potential results of DLS measurements are given in Table 1. Typically, nanoparticles with zeta potentials greater than 20 mV or less than -20 mV have sufficient electrostatic repulsion to remain stable in solution. The silver nanoparticles in AGC and AGA samples are stabilized with adsorbed negative citrate ions that cause repulsive force along particles and prevent aggregation while AGD, AGP and AGG samples are stabilized sterically.

The hydrodynamic diameter values in most cases are slightly lower than the diameter calculated from SEM images. Air dried colloids for SEM measurements are probably not a good representative of the same colloids in aqueous suspension.

B. SERS measurements.

The performances of the colloids were evaluated by their SERS activity using pyridine. To determine the suitable volume of colloid for further analysis, different volumes of AGC, AGA, AGD and AGBH samples were tested. The other samples show very low SERS enhancement. The SERS spectra of pyridine were also detected using KBr, KCl, $NaNO_3$ and $NaBH_4$ in different concentrations to find the optimal aggregating agent as well as optimum concentration for SERS enhancement. The differences of the SERS spectra of pyridine when adding different types and concentrations of ions to the colloids may be attributed to the both effects of degree of aggregation of the colloid and the change on the surface of silver structures.

Figure 3 shows enhanced Raman intensities as a function of concentration for two pyridine bands at 532 nm excitation wavelength.

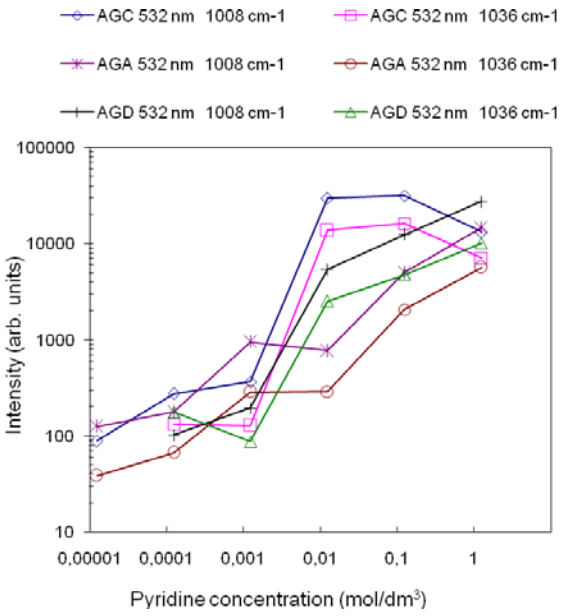


Figure 3. Pyridine SERS intensities at 1008 and 1036 cm^{-1} as a function of pyridine concentration in log-log scale in AGC, AGA and AGD samples.

The SERS enhancement factor values were calculated by comparing the intensity of the appropriate pyridine peak (~1008 cm^{-1}) measured in the SERS experiments to the corresponding peak measured from an aqueous solution of pyridine.

The SERS enhancement factor is defined as

$$EF = (I_{SERS}/I_{NR}) \times (c_{NR}/c_{SERS}) \quad (1)$$

where I_{SERS} and I_{NR} are intensities of the band of pyridine in the SERS spectrum and in the normal Raman spectrum, respectively (in this case 1008 cm^{-1}) and c is the pyridine concentration in the aqueous solution (c_{NR}) or in the colloid (c_{SERS}).

IV. CONCLUSION

A range of silver colloids, prepared by different methods, were investigated with respect to nanoparticle morphology and SERS activity. The results from UV-Vis, zeta potential, DLS and SEM confirm that stable silver nanoparticles of a certain size and charge are formed. The silver particles with 10–40 nm size and spherical shape were prepared by reducing silver nitrate with citrate, ascorbic acid, $NaBH_4$, PVP and glucose in the presence of stabilizing agents. Thus obtained silver nanoparticles were used as reproducible SERS substrate with pyridine as analyte. The volume of AGC, AGA, AGD and AGBH samples required for obtaining the best surface enhancement of the Raman signal was studied. Furthermore, the addition of various aggregating agents in terms of concentration was investigated. The surface-enhanced Raman spectroscopy of pyridine varied clearly when different kinds and concentration of aggregating agents were used. The best enhancement of the Raman signal was achieved for 1M salt solutions. The addition of sodium borohydride further amplified the Raman signal of pyridine. The best Raman enhancement for pyridine was achieved for AGC sample with the addition of 0.1M borohydride. The SERS enhancement factors for pyridine were calculated taking into account the number of molecules on the silver nanoparticles in laser activation volume and for the best results was estimated to be of the order 10^6.

REFERENCES

1. Li, Y., Y. Wu, and B.S. Ong, "Facile synthesis of silver nanoparticles useful for fabrication of high-conductivity elements for printed electronics," J Am Chem Soc, 2005, 127 (10): pp. 3266-3267.
2. Mallick, K., M. Witcomb, and M. Scurrell, "Silver nanoparticle catalysed redox reaction: An electron relay effect," Mater Chem Phys, 2006, 97 (2-3): pp. 283-287.
3. Thanh, N.T.K. and L.A.W. Green, "Functionalisation of nanoparticles for biomedical applications" Nano Today, 2010, 5(3): pp. 213-230.
4. Xu, H., et al., "Spectroscopy of Single Hemoglobin Molecules by Surface Enhanced Raman Scattering," Phys Rev Lett, 1999, 83 (21): pp. 4357-4360.
5. Ru, E.C.L. and P.G. Etchegoin, "Principles of Surface-Enhanced Raman Spectroscopy: And Related Plasmonic Effects," 2009, Elsevier Science & Technology Books.
6. Aroca, R., "Surface-Enhanced Raman Scattering, in Surface-Enhanced Vibrational Spectroscopy," 2007, John Wiley & Sons, Ltd. pp. 73-106.
7. Fu, X., et al., "Thin films of Î±-Fe2O3 nanoparticles using as nonmetallic SERS-active nanosensors for submicromolar detection," Front Chem China, 2011, 6 (3): pp. 206-212.
8. Kneipp, K., et al., "Single molecule detection using surface-enhanced Raman scattering (SERS)," Phys Rev Lett, 1997, 78 (9): pp. 1667-1670.
9. Wiley, B.J., et al., "Maneuvering the surface plasmon resonance of silver nanostructures through shape-controlled synthesis," J Phys Chem B, 2006, 110 (32): pp. 15666-15675.
10. Jia, H., et al., "Preparation of silver nanoparticles by photo-reduction for surface-enhanced Raman scattering," Thin Solid Films, 2006, 496(2): pp. 281-287.
11. Yin, B., et al., "Electrochemical Synthesis of Silver Nanoparticles under Protection of Poly(N-vinylpyrrolidone)," J Phys Chem B, 2003, 107 (34): pp. 8898-8904.
12. Jung, Y.K., J.I. Kim, and J.-K. Lee, "Thermal Decomposition Mechanism of Single-Molecule Precursors Forming Metal Sulfide Nanoparticles," J Am Chem Soc, 2009, 132 (1): pp. 178-184.
13. Munro, C.H., et al., "Characterization of the surface of a citrate-reduced colloid optimized for use as a substrate for surface-enhanced resonance raman scattering," Langmuir, 1995, 11(10): pp. 3712-3720.
14. Rivas, L., et al., "Growth of silver colloidal particles obtained by citrate reduction to increase the Raman enhancement factor," Langmuir, 2001, 17(3): pp. 574-577.
15. Sondi, I., D.V. Goia, and E. Matijević, "Preparation of highly concentrated stable dispersions of uniform silver nanoparticles," Colloids Surf A, 2003, 260(1): p. 75-81.
16. Dong, X., et al., "Synthesis of triangular silver nanoprisms by stepwise reduction of sodium borohydride and trisodium citrate," J Phys Chem C, 2010, 114(5): pp. 2070-2074.
17. Dong, X., et al., "Shape control of silver nanoparticles by stepwise citrate reduction," J Phys Chem C, 2009, 113(16): pp. 6573-6576.

Raman spectroscopy of tellurite glasses

H.Gebavi*, D. Ristić*, V. Đerek*, L. Mikec*, M. Ivanda* and D. Milanese**
* Institute Ruđer Bosković/Division of Materials Physics, Zagreb, Croatia
** Politecnico di Torino/Dipartimento di Scienza Applicata e Tecnologia, Torino, Italy
hgebavi@irb.hr

Abstract - Tellurite glasses are known as a promising material for optical fiber fabrication thanks to the specific optical and thermo-mechanical features. This paper shows tellurite glasses doped with different concentrations of various rare earth ions characterized by Raman spectroscopy. Especial attention was paid to the low frequency 'Boson peak' and the modes of vibration which correspond to TeO_4 and TeO_3 structural units. The results show interesting correlation between structural order and rare earth ions concentration.

I. INTRODUCTION

Tellurite glasses are considered as the promising material for active and passive optical fibers due to their specific characteristics. The most important and valuable features are: a) low phonon energy (among oxide glasses, tellurite glasses feature the lowest phonon energy of ~ 750 cm-1), b) high refractive index (at 633 nm ~2), c) wide transmission range (0.4 – 5 μm), d) thermo - mechanical stability and corrosion resistance (assuring sufficient stability towards crystallization), e) rheological properties (it is accepted that tellurite glasses exhibit Newtonian behavior which offers easy fibre drawing performance), f) high concentration solubility (tellurite glasses can accept up to 22 10^{20} cm^{-3} rare earth ions) [1, 2].

The infrared emission in active tellurite glasses can be obtained by utilizing particular energy level transitions of rare earths (RE) such as Yb^{3+} and Tm^{3+}. Up to now, the significant attention was dedicated to ~1 μm Yb^{3+} and ~ 2 μm Tm^{3+} emissions. Apart from the huge research efforts regarding RE fluorescence characteristics, the structural features should not be neglected since they are directly related with thermo-mechanical features linked with the fiber drawing process. The already published research papers consider mostly glasses doped with the low RE concentrations while the high dopant concentrations are still not well investigated. The incorporation of high RE concentrations in the glass host material can as the consequence cause glass crystallization or the phonon energy change.

Pure TeO_2 glasses are built of TeO_4 trigonal bi-pyramid (tbp) polyhedral [3]. In this paper we considered a ternary host glass (TeO2:ZnO:Na2O – 'TZN') codoped with Yb^{3+} and Tm^{3+}. The glass modifiers, ZnO and Na_2O as well as RE dopants, change glass structure which as the consequence influence thermal, mechanical and optical properties of the glass.

Identify applicable sponsor/s here. If no sponsors, delete this text box.

In TZN host several characteristic bands and corresponding structural units can be observed. Namely, at ~ 764 cm^{-1} (TeO_3 trigonal pyramid structural unit) and ~ 670 cm^{-1} (TeO_4 trigonal bi-pyramid structural unit) corresponds to stretching vibrations, while ~ 433 cm^{-1} ($Te_{eq} – O_{ax} – Te$) belongs to banding vibrations. The low frequency so called 'boson peak' at around 40 cm^{-1} [1] is assumed to be combination of acoustic phonons and rotational molecular modes [4]. Furthermore, the chemical glass composition can change the position or increase the intensity of the boson peak which is correlated with the medium range order decreases [5]. The structural order increase indicates the structural transformation trend from $TeO_4 \rightarrow TeO_3$ polyhedra [1].

II. EXPERIMENTAL

Raman spectroscopy features of tellurite glasses were monitored by Jobin Yvon T64000 instrument in micro – Raman configuration. The excitation source was Coherent Innova 100 argon ion laser operating at 514.5 nm and power of 20 mW.

The first glass group was doped with Yb^{3+} ions while the second group was codoped with Tm^{3+} ions. Dopant concentrations were in the range from 0.5 to 5 mol% and therefore the samples were labeled with letters 'Y' for Yb^{3+} and 'T' for Tm^{3+} followed with the concentration value. In order to determine dopant ion concentration in the unit volume it is necessary to determine glass density. Density measurements were obtained by precise mass measurements in air and water environment on selected specimens following the Archimedes's principle and using water as immersion liquid. Detailed dopant concentration are listed in Tables 1 and 2.

Table 1. Dopant concentration and density for Yb^{3+}-doped $75TeO_2$-$20ZnO$-$5Na_2O$ glasses.

Sample name	*Yb^{3+} (ppm)*	*$Yb^{3+}x10^{20}(cm^{-3})$*	*$\rho(g/cm^3)$*
Y0.5	6193	1.15	5.32
Y5	59485	11.24	5.43

* density measurement error ±0.05 (g/cm^3)

Table 2. Dopant concentration and density values for Tm^{3+}, Yb^{3+} - doped $75TeO_2$-20ZnO-$5Na_2O$ glasses.

Sample name	*Tm^{3+} $x10^{20}(cm^{-3})$*	*Yb^{3+} $x10^{20}(cm^{-3})$*	*$\rho(g/cm^3)$*
T4Y1	9	2.25	5.43
T4Y3	9	6.67	5.45
T4Y5	9	11.03	5.51
T1Y1	2.27	2.25	5.33
T2Y1	4.53	2.25	5.37
T4Y1	9	2.25	5.43
T5Y1	11.09	2.25	5.45

* density measurement error ±0.05 (g/cm^3)

III. RESULTS

A. Group I (TZN – Yb3+)

Raman spectroscopy spectra of TZN glasses doped with 0.5 and 5 %mol of Yb3+ are shown in Fig. 1.

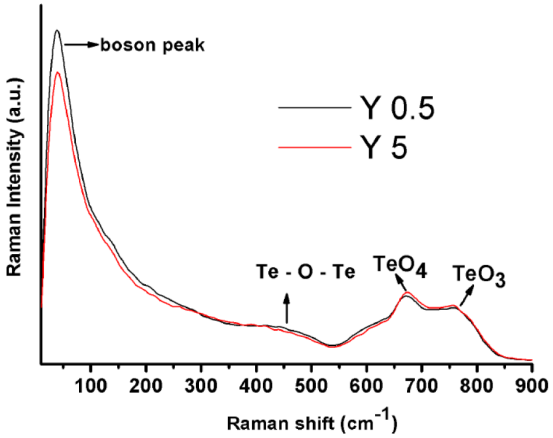


Figure 1. Raman spectra of TZN glass doped with Yb^{3+} with 514.5 nm laser excitation.

In the case of TZN doping with Yb^{3+}, boson peak position remained the same, at 40 cm^{-1} while the intensity decreased with dopant concentration increase. That showed that Yb^{3+} ion in TZN matrix has the same effect as Tm^{3+} ion [1, 2]. The mode of vibration which belongs to TeO_4 structural unit showed slightly wider width for larger dopant concentration.

B. Group II (TZN – Tm^{3+}, Yb^{3+})

In Fig. 2 Raman spectra of Tm - Yb doped TZN glass with 514.5 nm excitation line are displayed. As in the case of doping with only Yb^{3+}, characteristic boson peak and tellurite structural units can be observed.

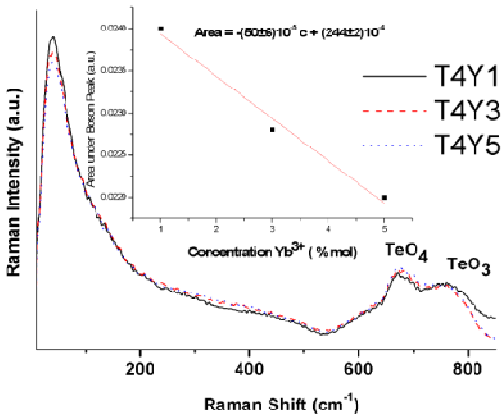


Figure 2. Raman spectra of TZN glass doped with Tm^{3+} - Yb^{3+} under excitation wavelength of 514.5 nm. The inset shows linear decrease of the area under boson peak with the increase of Yb^{3+} content.

In this case, the mode of vibration which corresponds to TeO_3 structural unit shows narrower shape. The highest frequency vibrational bend determinates phonon energy of the glass which influences numerous photonics features such as fluorescence quenching and cross relaxation in the case of thulium ions [1,2]. Therefore, the phonon energy is very important factor for the host material selection. This result shows indication of the phonon energy decrease although for the negligible value which would hardly influence the potential fiber laser. Another explanation for the TeO_3 band change could be clustering which probably occurs for higher dopant concentrations as in the case of the T4Y3 and T4Y5 samples. Boson peak linearly decreased with increasing dopant concentration as in [1]. Similar behavior is noticed for the TxY1 group (Fig. 3).

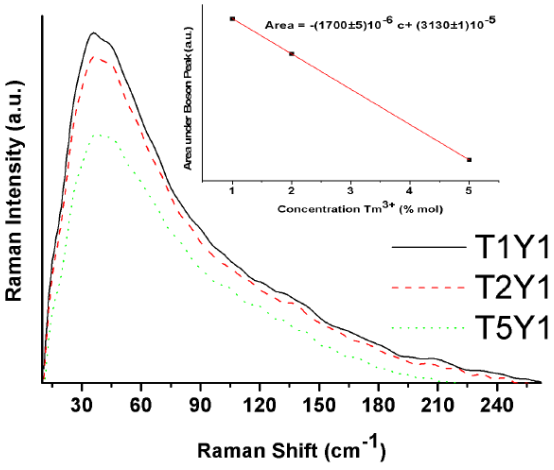


Figure 3. Raman spectra of group II (TZN - Tm^{3+}, Yb^{3+}) with 514.5 nm laser excitation..

From Fig. 2 it is obvious that TeO_4 structural unit changed its spectral line shape when Yb^{3+} concentration increased: the peak became narrower and its intensity slightly increased in comparison with TeO_3 band. Tm^{3+} and Yb^{3+} atomic and covalent radii are almost identical so the question is which property causes different band spectral shape?

One of the reasons can be that TeO_4 structural unit is more isolated from the amorphous network than in the case with Tm doping of TZN host which makes it narrower.

In Zhao paper [6] excitation wavelength for Tm/Yb doped tellurite glass was the same, 514.5 nm, but the change in TeO_4 band shape was not observed. Similar behavior of tellurite glass doped with Yb^{3+} at 514.5 nm laser excitation is not observed in literature to the best of author knowledge.

IV. CONLUSION

This paper reports structural changes of tellurite glasses observed by using Raman spectroscopy.

Raman spectroscopy showed that doping with Yb^{3+} influenced the TeO_3 structural unit and its role in glass matrix when the high concentration of thulium ions is present. A more precise definition and explanation would require additional measurements. As in the case of Tm^{3+}, position of TeO_4 and TeO_3 bands remained unchanged as well as the mutual ratio under 514.5 nm excitation.

All glass groups were characterized by the boson peak positioned at around 40 cm^{-1} and related to acoustic phonons. Boson peak linearly decreased with increasing dopant concentration indicating the increase of the middle range order.

ACKNOWLEDGMENT

The authors wish to thank the Regione Piemonte Converging Technologies ‘‘Hipernano’’ research project.

REFERENCES

[1] H. Gebavi, D. Milanese, G. Liao, Q. Chen, M. Ferraris, M. Ivanda, O. Gamulin, S. Taccheo, “Spectroscopic investigation and optical characterization of novel highly thulium doped tellurite glasses,” J J. Non-Cryst. Solids, vol. 355, 9, pp. 548 – 555, 2009.

[2] H. Gebavi, D. Milanese, R. Balda, M. Ivanda, F. Auzel, J. Lousteau, J. Fernandez, M. Ferraris, “Novel Tm^{3+}-doped fluorotellurite glasses with enhanced quantum efficiency,” Opt. Materials, vol. 33, pp. 428–437, 2011.

[3] J. Wu, Z. Yao, J. Zong, S. Jiang, “Highly efficient high-power thulium-doped germanate glass fiber laser,” Opt. Lett., vol. 32, pp. 638–640, 2007.

[4] S. Marjanovic, et al., “Character of new erbium – doped tellurite glasses and fibers,” J. Non-Cryst. Solids, vol.322, pp. 311-318, 2003.

[5] V.K. Malinovsky, A.P. Sokolov, “The nature of boson peak in Raman scattering in glasses,” Sol. State Comm., vol. 57(9), pp. 757-761, 1986.

[6] C. Zhao, Q. Zhang, G. Yang, Z. Jiang, “Laser-diode-excited blue upconversion in Tm3+/Yb3+-codoped TeO_2-Ga_2O_3-R2O (R=Li, Na, K) Glasses,” J. Fluorescence, vol. 18, pp. 87–91, 2008.

Coated spherical microresonators for cutting-edge photonics application

Davor Ristić[1,2], Maurizio Mazzola[2], Andrea Chiappini[2], Cristina Armellini[2], Alphonse Rasoloniaina[3], Patrice Féron[3], Roberta Ramponi[4], Gualtiero Nunzi Conti[5], Stefano Pelli[5], Giancarlo C. Righini[5,6], Gilles Cibiel[7], Mile Ivanda[1], Maurizio Ferrari[2]

[1]Institut Ruđer Bošković, Bijenička cesta 54, 10000 Zagreb, Croatia.

[2]IFN – CNR CSMFO Lab., Via alla Cascata 56/C Povo, 38123 Trento, Italy.

[3]FOTON-Systèmes Photoniques (CNRS-UMR 6082), ENSSAT, 6 rue de Kerampont, CS 80518, 22305 Lannion cedex, France.

[4]IFN-CNR and Department of Physics, Politecnico di Milano, p.zza Leonardo da Vinci 32, 20133 Milano, Italy.

5IFAC - CNR, MiPLab., 50019 Sesto Fiorentino, Italy.

[6]Museo Storico della Fisica e Centro di Studi e Ricerche "Enrico Fermi", Piazza del Viminale 1, 00184 Roma, Italy.

[7]Centre National d'Etudes Spatiales (CNES), 31401 Toulouse Cedex 9, France.

We present some results regarding the coating of spherical microresonators and their different applications in photonics. The spherical microresonators high Q-factor and low mode volume make them perfect for a number of photonics applications. A thin film coating on the surface of the resonator using an optically active or passive layer can be used to modify signifactly the optical properties of the resonator. We have succesfully coated the microresonators with a sol-gel derived coating and have studied the effect of the coating on different optical and spectroscopical properties of the microresonator.

IV. INTRODUCTION

Microresonators are small resonators of dimensions that can range from 30 μm to 1mm [1]. Generally they can be produced in different shape, of different materials and by different methods. The simplest microresonator is a silica microsphere which can be made by simply melting the tip of a standard telecom optical fibre by a oxy-gas torch or by an electrical spark. After the fibre melts, the surface tension will form it into a perfect sphere. This kind of silica microresonator can exhibit very high Q-factors (10^7 up to 10^9), Q-factor defined as the average number of round trips a photon can take inside a resonator before being absorbed or scattered inside the microresonators the light is confined by total internal reflection, so that the light is travelling along the circumference of the resonator. The modes of such resonators are called whispering gallery modes (WGM). In addition to the high quality factors the WGM also have a very low mode volume which means that we can confine light in a very small volume for a long period of time. This makes them very useful for different applications such as sensing [2], lasers and amplifiers[3], optoelectronic oscillators or non-linear optics[4].

V. THEORY

To calculate the eigenfrequencies of the whispering gallery modes of a coated sphere the Aden-Kerker generalization of the Mie scattering theory can be used [5,6]. The radial dependence of the electric field of the WGMs is given by the Bessel function in the core, a linear combination of the Bessel and Neumann functions in the coating and by the Neumann function outside of the sphere. The continuity of the tangential components of the electric and magnetic fields at the core-coating (r_1) and coating-air (r_2) interfaces leads to the following system of equations:

$$\begin{aligned}
&\psi_l(kn_s r_1) = \alpha\,\psi_l(kn_c r_1) + \beta\,\chi_l(kn_c r_1)\\
&P_s\,\psi_l'(kn_s r_1) = \alpha\,P_c\,\psi_l'(kn_c r_1) + \beta\,P_c\,\chi_l'(kn_c r_1)\\
&\alpha\,\psi_l(kn_c r_2) + \beta\,\chi_l(kn_c r_2) = \gamma\,\chi_l(kn_0 r_2)\\
&\alpha\,P_c\psi_l'(kn_c r_2) + \beta\,P_c\,\chi_l'(kn_c r_2) = \gamma\,P_0\,\chi_l'(kn_0 r_2)
\end{aligned} \qquad (1)$$

where $k=2\pi/\lambda$, n_s, n_c ,n_0 are the refractive indices of the sphere core, coating and surrounding medium respectively, $\psi_l(x)=xj_l(x)$ and $\chi_l(x)=xn_l(x)$ where j_l and n_l are the spherical Bessel and Neumann functions respectively, $P_s=n_s$, $P_c=n_c$, $P_0=n_0$ for the transversal electric (TE) and $P_s=1/n_s$, $P_c=1/n_c$, $P_0=1/n_0$ for the transversal magnetic (TM) modes. This system of equations can be solved numerically to obtain the eigenfrequencies of the WGMs. Each mode will be characterized by three integer number p, l and m, where p, $l-|m|$ and m correspond to the number of nodes of the wave function in the radial, azimuthal and polar directions respectively. Usually, the mode which can be coupled to the most efficiently is the $p=0$ fundamental mode because it is the mode which is confined closest to the surface and therefore will have the largest part of its electric filed extending outside of the sphere. This is

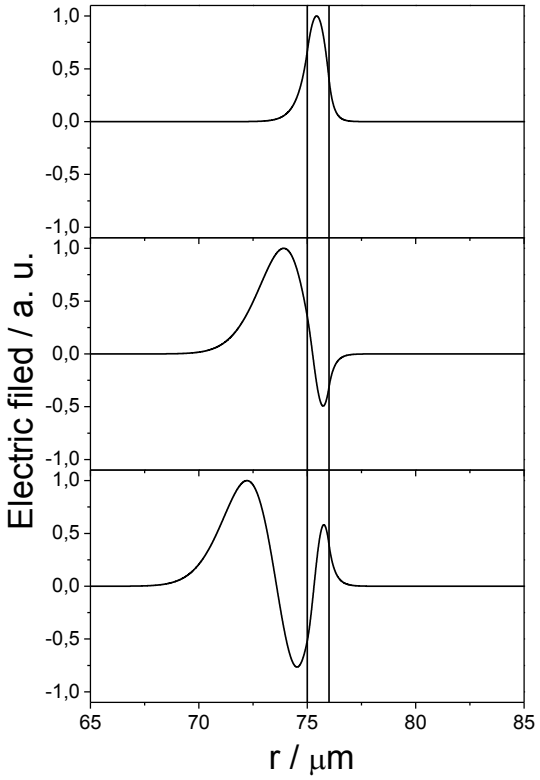


Figure 1. The electric field distribution of the fundamental (p=0) and first two internal (p=1,2) WGMs of a coated microsphere

demonstrated in Figure 1 which shown the WGM field profiles for a coated microsphere for the fundamental (p=0) and first two internal modes (p>1,2) for a 150 μm silica (n=1.44) microsphere coated with a $70SiO_2$-$30HfO_2$ coating(n=1.6) at the wavelength of 1550 nm.

VI. EXPERIMENT

The microspheres were produced by melting the tip of a standard optical telecommunications fiber. The diameter of the microspheres was found to be 130-160 μm using a standard optical microscope. The microsphere was coated with a film of composition $70SiO_2$-$30HfO_2$ activated with 0.3% mol Er^{3+}. The films were coated using the dip coating technique. The sol was prepared by mixing a solution of $Si(OC_2H_5)_4$ (TEOS), HCl, H_2O, CH_3CH_2OH (ethanol) and $HfOCl_2$. The molar ratio between TEOS, water and the catalyst (HCl) was $n(H_2O) : n(Si(OC_2H_5)_4) : n(HCl) = 200:100:1$. The molar ratio between the TEOS and $HfOCl_2$ was chosen to be $n(Si(OC_2H_5)_4) : n(HfOCl_2) = 70:30$ on the basis of previous experiments [18]. The overall amount of Si and Hf in the sol was controlled so that $c(Si(OC_2H_5)_4) + c(HfOCl_2) = 4.48\ 10^{-4}$ mol/mL in order to have a sol viscosity which makes the dipping rate 30 nm/dip. $Er(NO_3)_3$ x $5H_2O$ was added to activate the film so that the molar concentration of Er^{3+} in the sol is 0.3 % mol. The mixture of TEOS, HCl and H_2O was prehydrolised at 65 ^oC for 1 hour before adding the $HfOCl_2$ and the Er $(NO_3)_3$ x $5H_2O$. The entire mixture was then stirred for 16 h until the sol became transparent. Each microsphere was dipped 30 times into the sol. The dipping speed was 40 mm/min and the dipping time 20 s. After each dip the microspheres were densified at 900 ^oC for 50s and after every 10 dips an additional densification was made for 2 min. At the end of the deposition a final densification for 5 min at 900 ^oC was made.

The measurement of the whispering gallery modes was done using a microsphere-taper coupling setup. A 1480 nm fiber diode laser was used for the pumping. The coupling of light inside the sphere was achieved using a tapered fiber. The pump laser inside the microsphere is absorbed by Er^{3+} ions inside the coating which in turn emit in the C-band wavelength region around 1550 nm. The emission will be greatly enhanced at wavelengths which correspond to the WGMs of the microsphere. This emitted light is then coupled out of the sphere and into the taper and detected by an Anritsu MS9710B optical spectrum analyzer. The Q-factor of the spheres was measured by sweeping a tunable laser over a single resonance and monitoring the transmission, all in the C-band range. The Q factor is then the ration of the laser frequency to the width of the resonance dip.

VII. RESULTS

The spheres were produce by melting the tip of a standard SMF28 optical fiber. The Q-factor of the spheres was measured before the coating and a typical Q-factor measurement is presented in Figure 2. There is a dip in transmission clearly visible which corresponds to a single WGM. The 35 MHz width of the dip corresponds to a Q-factor of about $7x10^6$. This is a typical value of the Q-factor of our spheres before the coating. For most of our spheres the Q-factor before the coating falls in the range between $5x10^6$-$2x10^7$. After the coating the Q-factor is usually reduced by an order of magnitude ($5x10^5$-10^6) which is because of the erbium absorption. In fact the Q-factor of a coated sphere is lowest at about 1530 where there is the maximum of erbium absorption and then increases when going towards higher wavelengths (1550-1560 nm). A picture of the microsphere and the taper used

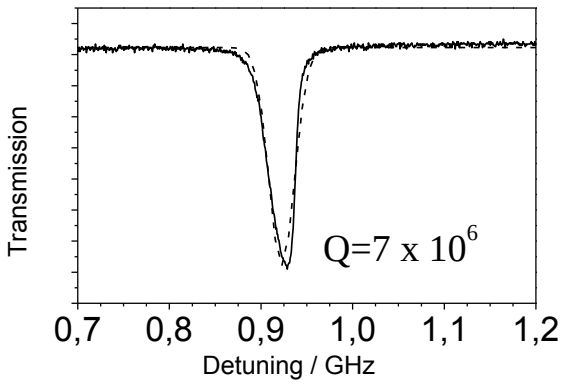


Figure 2. A typical Q-factor measurements of a silica microsphere. The dotted lines are the Lorenzian fit to the experimental data

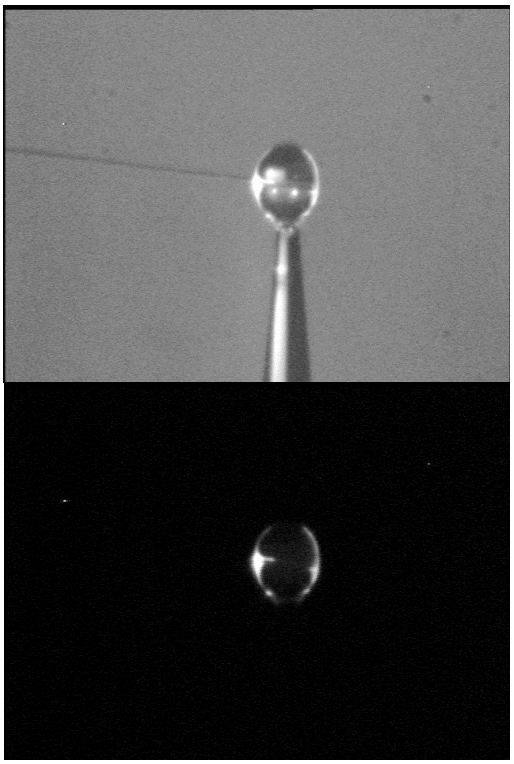


Figure 3. Optical microscopy images of the microsphere and the taper. The lower image shows the Er^{3+} green upconversion

for the pumping and signal recovery is shown in Figure 3. In the bottom part of Figure 3 can be seen the green Er^{3+} up-conversion from the WGMs of the microsphere. This up-conversion is the proof that we are efficiently coupling light inside the resonator.

A typical WGM luminescence spectrum is shown in Figure 4. The sharp peaks seen in the Figure are to be attributed to the WGMs with a different value of the *l* number. The modes with different values of the *m* number are frequency degenerate so that they cannot be observed unless the sphere is slightly elliptical. The modes presented in Figure 4 can be attributed to the fundamental (p=0) mode since the internal (p>0) modes are more difficult to couple light in and/or out of and therefore are less likely to be visible. In Figure 5 is shown a lasing spectrum of a coated microsphere. If the coating is thick enough so that the number of erbium ions interacting with

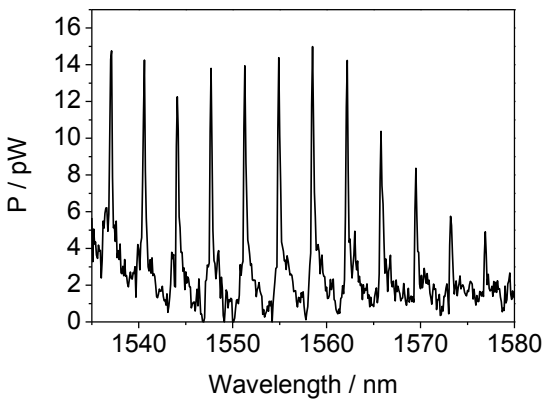


Figure 4. The WGM luminescence of a coated microsphere

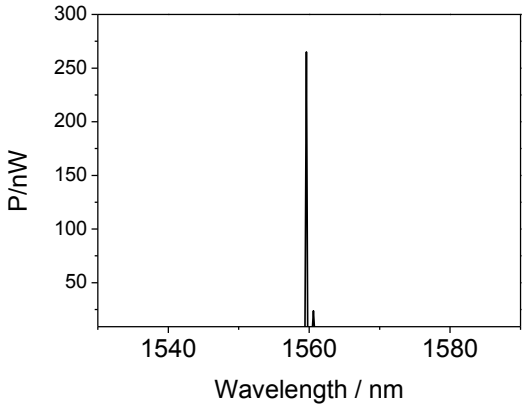


Figure 5. Lasing spectrum of a coated microsphere

the electric field of the WGMs is high enough, the stimulated emission gain can become greater than the cavity round trip loss and the sphere will start to lase. In Figure 5 we can see the lasing spectrum of a WGM with the peak power of 280 nW. To calculate the mode volume of the WGMs in Figures 4 and 5 we first need to consider the *m* values of the modes. For the studied system we can calculate the effective refractive index of the fundamental (l=m) mode around 1560 μm (l=430) which is around 1.53. Since the effective index of a body of a 3 μm taper is about 1.4 we can conclude that the taper is phase matched to the WGM with m=l*1.4/1.53=394. Thus we can conclude that the azimuthal mode distension is about $R\pi(l-m)/l$=18 μm. Since the radial distension for the p=0 mode is about 1μm and the polar distension is just the circumference of the sphere we conclude that the mode volume is around 8000 μm^3.

V. CONCLUSION

We have produced spherical microresonators and coated them with an optically active layer. We observed both the WGM luminescence and lasing. The maximum lasing peak power was found to be 280 nW

ACKNOWLEDGMENT

This research was performed in the framework of the CNES R&T project SHYRO (2011-2014). A. R acknowledges support from the Conseil Régional de Bretagne.

REFERENCES

[1] G.C. Righini, Y. Dumeige, P. Féron, M. Ferrari, D. Ristić, S. Soria "Whispering gallery mode microresonators:Fundamentals and applications", La Rivista del Nuovo Cimento vol. 34, pp. 435-488, 2011.

[2] S. Soria, F. Baldini, S. Berneschi, F. Cosi, A. Giannetti, G. Nunzi Conti, S. Pelli, and G. C. Righini, "High-Q polymer-coated microspheres for immunosensing applications", Opt. Express vol. 17, pp.14694-14699, 2009.

[3] Z.P. Cai, H.Y Xu, G. Stéphan, P. Féron & M. Mortier, "Red-shift in Er :ZBLALiP whispering gallery mode laser", Opt. Comm, vol. 229, pp. 311-315, 2004.

[4] Y. K. Chembo and N. Yu, "On the generation of octave-spanning optical frequency combs using monolithic whispering-gallery-mode microresonators", Opt. Lett., vol. 35, pp. 2696-2698, 2010.

[5] A. L. Aden and M. Kerker, J. Appl. Phys. "Scattering of Electromagnetic Waves from Two Concentric Spheres", vol. 22, pp. 1242-1246, 1951.

[6] D. Ristić, A. Rasoloniaina, A. Chiappini, P. Féron, S. Pelli, G. Nunzi Conti, M. Ivanda, G. C. Righini, G. Cibiel and M. Ferrari, "About the role of phase matching between a coated microsphere and a tapered fiber: experimental study", Opt. Express, vol. 21, pp. 20954-20963, 2013.

Examination of silicon material properties using THz Time-Domain Spectroscopy

B. Pejcinovic*
* Portland State University, Electrical and Computer Engineering dept, Portland, Oregon, USA
pejcinb@pdx.edu

Abstract - High-resistivity silicon is a very popular choice for substrate material when examining e.g. various thin films, biological samples etc. using THz radiation. This paper provides an overview of research on its THz properties, examines various practical issues appearing when determining its properties and reports results on complex index of refraction, permittivity and conductivity, obtained using modern instrumentation. Potential problems related to water absorption are highlighted. Various Drude models and parameters are also presented and discussed.

I. INTRODUCTION

THz Time-Domain Spectroscopy (TDS) is fast becoming a standard tool for examination of material properties of both electronic and optical materials, as well as detection of various biological and chemical compounds. Due to its relative transparency at THz frequencies and availability, many researchers use high-resistivity silicon wafers as the most convenient substrate to hold their samples. Therefore, it is imperative to understand hi-res silicon properties in detail and as accurately as possible.

A. History

Advantages of hi-res Si for THz use were recognized early, which prompted more detailed studies. THz TDS studies of silicon date back to early 1990-s and there have been a few follow-on studies [1-9]. Many other semiconductor materials were subject to TDS studies, which filled a gap between the electronic properties measured at DC to mm-wave frequencies, and optical characterization from far-IR to visible and beyond. Some exciting new materials, such as GaN, have been studied in some detail, see e.g. [16].

In the earlier days TDS systems were constructed from separate components, which allowed greater flexibility in, for example, choice of sources and detectors. Nowadays, several commercial systems exist, e.g. Advanced Photonix [11] and Teraview [12], each with is own specific advantages. One major advance has been in the speed of processing data. In the earlier literature it is rare to find use of averaging, which is either not mentioned or very few averages are reported. With T-Ray 4000 system used at Portland State University we can easily accumulate 5000 to 10000 responses in a matter of minutes. This brings the noise floor down which improves signal-to-noise ratio and increases the measurement bandwidth.

Theoretical foundation for most of the early experimental studies was provided by Drude theory, which showed reasonable accuracy and could fit most of the data across 0.2 – 2 THz or more. In cases where optical phonons contributed to THz response, such as ZnS, additional models had to be included [15]. More recently, detailed simulation and theoretical studies have been published which enable detailed and more first-principles-based comparison between theory and experiments [6] and modeling is moving beyond simple fitting of parameters to experimental data.

II. TDS MEASUREMENT SYSTEM

THz radiation can be generated and detected using differnt operating principles but current systems are predominantly based on photoconductive switching where ps-long pulses are generated by discharging a capacitor/antenna and a similar antenna/switch is used on output for detection [14]. Key advantage of TDS systems is coherent detection so that both amplitude and phase can be measured after FFT to frequency domain.

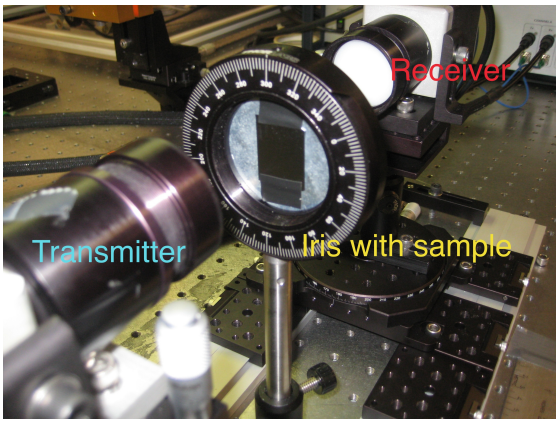


Figure 1 TDS measurement system in transmission mode with metal iris holding a sample. Iris prevents collimated THz beam from reaching the receiver without passing through the sample.

Experiments can be set up in reflection or transmission mode. The former is preferred for highly absorptive materials where very thin samples would be required for transmission measurements. Whenever possible, experimentalists prefer transmission measurement due to its simple setup and easyly obtained reference waveform.

When samples are big enough to cover the entier THz beam then all that is required is a simple holder with adjustments of hight and rotation to make sure that sample is normal to the beam. If samples are smaller than the beam, then some metal iris is used to restrict the beam and to hold the sample, see Figure 1. Of course, this leads to reduced signal strength. Alternatively, a focusing mirror or lens can be used to focus the beam on the

sample. Focusing should provide larger signal strength but its alignment is more involved.

In this report we use collimated beams with iris. Figure 1 shows a typical setup in transmission mode with a rotating iris. Focusing setup would look exactly the same but would use different lenses inside the transmitter and receiver heads.

Because of susceptability of THz radiation to absorption by water vapor, best results are obtained in enclosed environments filled with dry air or nitrogen. In this set of measurements our samples were in regular lab environment with relative humidity of around 30-35%.

A. Time-domain results

There are two modes in which TDS system can be used: transmission and reflection. Below, in Figure 2, are two examples of measurements on two silicon wafers: one is nominally high-resistivity (ρ > 10 kOhm-cm) and another has measured resistivity of 1.5 Ohm-cm. Interpretation of transmission measurements is fairly straightforward since we can easily observe two main effects: reduction of peak amplitude and delay. In addition, in hi-res sample we can observe multiple "echoes" of the signal as it is reflected within the sample multiple times. This is also called Fabry-Perot effect. The 1.5 Ohm-cm sample is more absorptive and it is difficult to distinguish secondary echoes.

Note that we could potentially utilize up to 320 ps of waveform, which is limited by the system, but internal reflections, e.g. reflections back-and-forth between the lenses, may appear and adversely affect parameter extraction. With around 125 ps included we can get 8 GHz resolution in frequency, which is more than sufficient for this study.

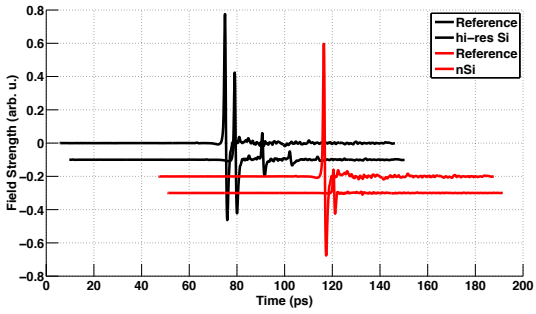


Figure 2 Time-domain waveforms from transmission TDS measurement of hi-res Si and 1.5 Ohm-cm n-Si wafers. They include multiple reflections within Si samples (Fabry-Perot) which have to be accounted for. Reference is obtained by removing the sample from the THz beam path.

Reflection mode results for the same samples are shown in Figure 2. Signal strength is reduced due to use of beam splitter but the advantage is that both incident and reflected beam are normal and there is no need to align transmit and receive heads, as is done in, e.g. elipsometry type of measurements. Reference waveform is provided by a "perfect" mirror, which in practice means metalic surface. In our case we used self-adhesive copper tape to cover the surface of another silicon wafer of similar thickness.

Interpretation is not as simple but we could get a first-order estimate of index of refraction by using the ratio between reference and sample waveform peaks in Fresnel coefficient equation. Reference waveform may have to be mathematically "shifted" in phase so that its front (reflection) surface lines up with sample one [17].

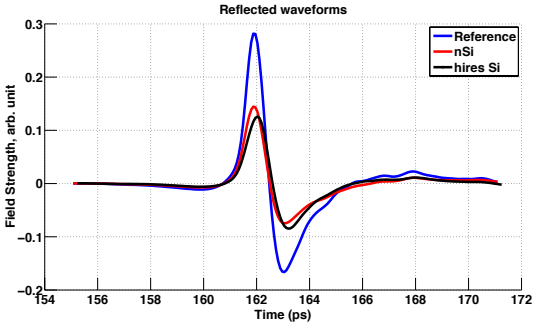


Figure 3 Time-domain waveforms from reflection TDS measurement of hi-res Si and 1.5 Ohm-cm n-Si wafers. Typically, only the first reflection is used and the rest of the waveform is windowed out. Reference waveform is from a perfect reflector (mirror).

Note that the waveforms have been windowed and only the first reflected pulse is retained. There is only around 15 ps of waveform which gives frequency resolution of around 65 GHz which would be sufficient for this study. The rest of this paper will deal exclusively with transmission mode measurements of silicon samples.

III. THEORY OF TEM WAVE PROPAGATION

Fundamental equation for transfer function of EM signal through sample is given in eq. (1) [13]. The summation part includes Fabry-Perot (FP) effects and is neglected when time-domain waveform is windowed to contain only the main (first) pulse. If more than 5 pulses (echoes) are included then a limit of infinite FP reflections can be applied resulting in eq. (2).

$$H(f) = t_{12}(f)t_{21}(f)exp\left(-\frac{i2\pi fd(\tilde{n}(f)-n_{air})}{c}\right) \cdot \sum_{l=0}^{m}\left[r^2(f)exp\left(-\frac{i4\pi fd\tilde{n}(f)}{c}\right)\right]^l \quad (1)$$

$$H_2(f) = \frac{t_{12}(f)t_{21}(f)exp\left(-\frac{i2\pi fd(\tilde{n}(f)-n_{air})}{c}\right)}{1 - r^2(f)exp\left(-\frac{i4\pi fd\tilde{n}(f)}{c}\right)} \quad (2)$$

$$r(f) = \frac{\tilde{n}(f)-n_{air}}{\tilde{n}(f)+n_{air}},$$

$$t_{12}(f) = \frac{2n_{air}}{\tilde{n}(f)+n_{air}}, \quad t_{21}(f) = \frac{2\tilde{n}(f)}{\tilde{n}(f)+n_{air}}$$

The transfer function given above is equated with ratio of measured response with sample and without it. The resulting set of nonlinear equations is solved for complex index of refraction $\tilde{n} = n_1 - i\,n_2$ using MATLAB function fsolve. Note that this derivation implies an EM

wave of the form $E(t) = E_0 \exp[i(\omega t - kx)]$ which has a $+\omega t$ term. Many authors use different definition which is complex conjugate of this one. As a result, our definitions of index of refraction, permittivity, and conductivity would have negative sign for their imaginary parts: $\tilde{\epsilon} = \epsilon_1 - i\,\epsilon_2$ and $\tilde{\sigma} = \sigma_1 - i\,\sigma_2$. If we take the complex conjugate of $\tilde{n}$ and use that to calculate permittivity and conductivity then we can keep the + sign in their definitions. In our definition n_2 is a positive number (also called extinction coefficient), as required to have attenuation in the sample.

Permittivity is calculated using $\tilde{\varepsilon} = \tilde{n}^2$ and to calculate conductivity we use this definition:

$$\tilde{\epsilon} = \epsilon_\infty + \frac{i\tilde{\sigma}}{\omega\varepsilon_0} \qquad (3)$$

where ϵ_∞ is permitivity in the limit of very high frequency where there is only dielectric response and no contribution from conduction. Unlike $\tilde{n}$ and $\tilde{\varepsilon}$, complex conductivity $\tilde{\sigma}$ has units and is typically reported as [1/(Ohm-cm)] or [S/cm].

Next, we need to model the frequency behavior of permittivity and conductivity, which is often done using Drude model and its variants. The simplest Drude model states

$$\epsilon = \epsilon_\infty + \frac{i\sigma}{\omega\varepsilon_0} = \epsilon_\infty - \frac{\omega_p^2}{\omega(\omega + i\Gamma)} \qquad (4)$$

$$\tilde{\sigma} = \frac{i\varepsilon_0\omega_p^2}{(\omega + i\Gamma)} \qquad (5)$$

where ω_p is plasma frequency and Γ is scattering or damping rate. These two parameters are optimized to fit the measured results and can be used to calculate carrier concentration and mobility [9]. In certain situations Cole-Davidson or Drude-Smith variants may provide a better fit [9], and a generalized form of conductivity was proposed in [6].

IV. Results

In this section we will present extracted values of index of refraction $\tilde{n}$, permittivity $\tilde{\varepsilon}$ and conductivity $\tilde{\sigma}$ vs. frequency for hi-res and 1.5 Ohm-cm silicon samples. These are just different representations of the same data so no new information is gained by switching from one to the other. However, modeling of the frequency response is almost never done for n; instead, Drude, Lorentz, Debye etc. models are applied to permittivity and/or conductivity. Representation in permittivity allows comparison with, or utilisation of, high frequency limit where we usually know optically measured value. Similarly, representation in conductivity allows comparison with measured DC values and calculation of carrier conecntration and mobility.

A. Index of refraction

Real part of index of refraction is shown in Figure 4 and imaginary part in Figure 5. When multiple reflections are included there is some residual oscillatory behavior despite optimization of sample thickness, as explained in [10]. Nonetheless, the results are the virtually the same independent of data processing method used. Extracted real values are very close to those reported for hi-res silicon and 1.5 Ohm-cm shows qualitative behavior observed by others [1-3]. Imaginary part of n is infrequently reported but, as will be shown later, other derived parameters are in good agreement with published results.

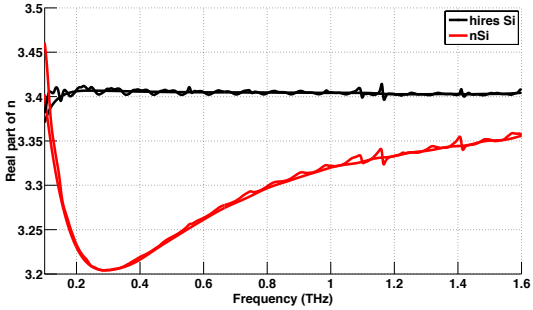


Figure 4 Real part of index of refraction for hi-res Si and 1.5 Ohm-cm n-Si wafers. Results for multiple-reflections and single pulse are included.

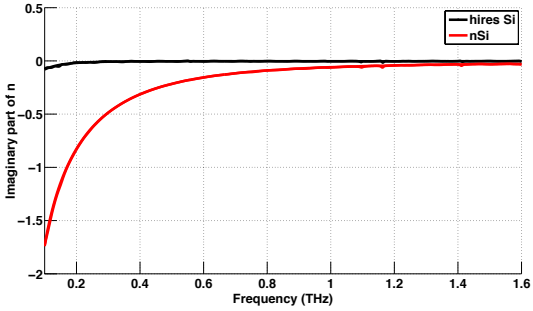


Figure 5 Imaginary part of index of refraction (= -n_2) for hi-res Si and 1.5 Ohm-cm n-Si wafers. Results for multiple-reflections and single pulse are included.

B. Permittivity

Given the simple square relationship between $\tilde{n}$ and $\tilde{\varepsilon}$, Figure 6 has no surprises. Real part for both samples converges to high frequency value of around 11.6. Imaginary part is related to conductive losses and it falls of with frequency but this is more easily interpreted by looking directly at conductivity.

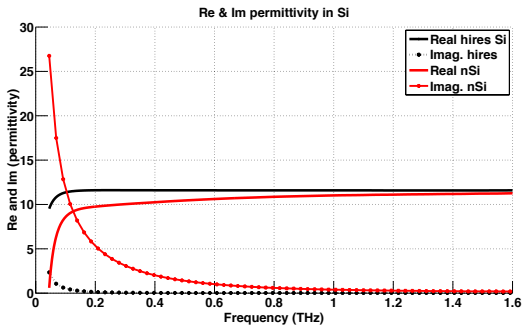


Figure 6 Real and imaginary permittivity (ϵ_1& ϵ_2) of hi-res Si and 1.5 Ohm-cm n-Si wafers. Only the first transmitted pulse is included.

C. Conductivity

Real and imaginary conductivity is shown in Figure 7. By extrapolating real part to DC we can extract resistivities of around 1.5 Ohm-cm which agrees well with four-point probe measurement results. The hi-res sample, however, shows an upswing at lower frequencies which would extrapolate to resistivity of around 20 Ohm-cm. We belive that this is due to a loss of precision at frequencies below 0.2 THz and that we should extrapolate from higher frequencies, in which case we obtain ≈100 Ohm-cm, which is closer to the nominal value. We will also do a four-point probe measurement to verify the nominal value.

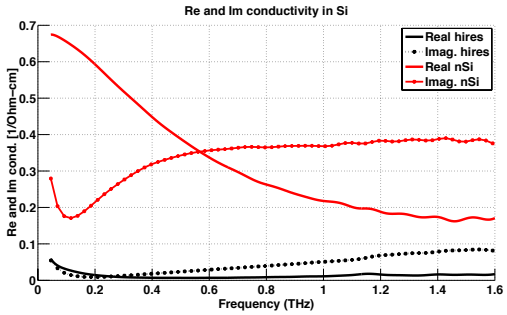


Figure 7 Real and imaginary parts of conductivity (σ_1 & σ_2) in hi-res Si and 1.5 Ohm-cm n-Si wafers. Only the first transmitted pulse is included.

Another interesting feature is frequency behavior of imaginary component. In both cases it plateaus at certain value. Others [1-3] have reported that it shows a maximum but it becomes less pronounced at lower resistivities. Reported data, however, is farily noisy which makes peak determination difficult. One possible reason for this behavior is the measurement environment and we will repeat these measurements in dry air and compare the results.

D. Extraction of Drude parameters

Despite some observed qualitative differences relative to earlier results for imaginary part of conductivity we attempted to extract parameters for simple Drude model. Fitting was done in MATLAB using lnsqnonlin and lsqcurvefit functions. Frequency range was limited to 0.31 to 1.1 THz . As observed by others, it can be difficult to fit Drude model if the peak in imaginary component is not well defined. As shown in Figure 8, fit for the real part is reasonably good but there is some divergence above 1 THz. Also shown are results from [3] which used samples with resistivity of 1.31 Ohm-cm.

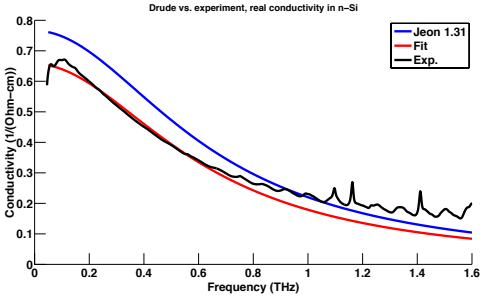


Figure 8 Real part of conductivity from [3] (blue), our measured results (black), and fitted Drude response (red). Multi-reflections.

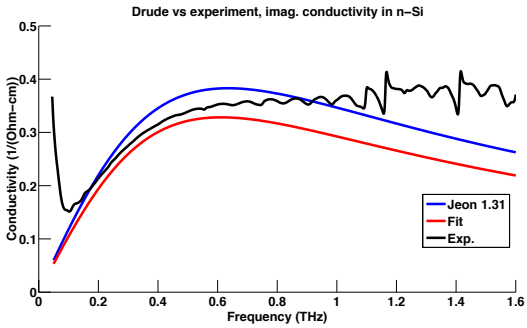


Figure 9 Imaginary part of conductivity from [3] (blue), our measured results (black), and fitted Drude response (red). Multiple reflections are included.

A summary of parameters from the literature and in this work is shown in Table 1. Imaginary component, shown in Figure 9, is not fitted as well, especially above 1 THz.

Table 1 Drude model parameters for silicon as reported in the literature. DC resistivity is from four-point probe measurements.

DC ρ (Ohm-cm)	ω_p (THz)	$\Gamma/(2\pi)$ (THz)	Ref.
n-type Si			
8.1	0.36	0.59	[1]
1.15	1.01	0.64	[1]
8.15	0.35	0.54	[2]
0.21	2.653	0.838	[2]
1.31	1.333	0.935	[3]
0.22	2.66	0.84	[4]
0.136	3.26	0.84	[5]
1.5	≅0.9	≅0.2	This work
p-type Si			
9	0.48	1.19	[1]
0.92	1.75	1.51	[1]
9.61	0.448	1.03	[2]
0.17	4.49	1.985	[2]

E. Thickness determination

Accurate material parameter extraction is critically dependent on determination of sample thickness. When possible, it can be measured directly using a micrometer but it has been shown that it can also be determined by optimization procedure that minimizes oscillations in transfer function. Additional geometry parameter that can also be used is angle of incidence which is usually assumed to be 0 degrees.

We have applied procedure given in [10] and obtained reasonable agreement with micrometer measurements. One problem is that in our case some “oscillations“ (or resonances) are real water vapor absorption lines and not due to Fabry-Perot effect. This makes thickness optimization less effective since it tends to weigh more heavily areas of rapid change, i.e. resonances. This problem was circuimvented by restricting the optimization

bandwidth but in the future we will perform the measurements in dry-air environment.

Figure 10 illustrates effects of 2 μm and 4 μm change in assumed thickness on real part of index of refraction. Relative changes are small, on the order of 1%, which is acceptable in many situations.

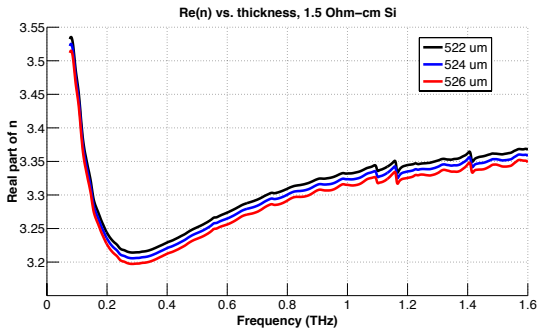


Figure 10 Variation of real part of index of refraction with assumed sample thickness. Silicon wafer has 1.5 Ohm-cm resistivity. Multiple reflections are included.

On the other hand, Figure 11 illustrates changes in real and imaginary part of conductivity. Real part is practically unaffected but imaginary part exhibits large changes and even shows a slightly different trend: 522 μm is starting to fall off with frequency while 526 μm increases.

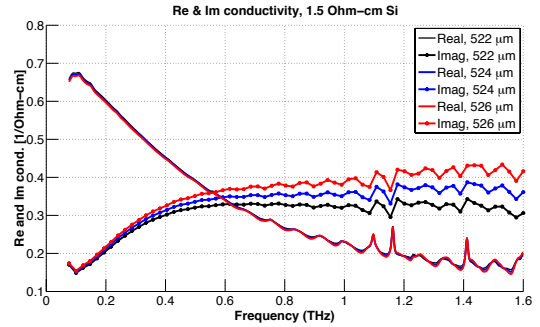


Figure 11 Variation of real and imaginary parts of conductivity with assumed sample thickness. Silicon wafer has 1.5 Ohm-cm resistivity. Multiple reflections are included.

Therefore, extracting material paramters with the best possible precision requires attention to details, from utilizing the best environemnt (dry air or nitrogen, if possible) to measuring thickness through independent means. Some parameters are more susceptible than others and we have to determine how much precision is needed for a specific problem. Some estimates, with 1% or 2% accuracy, can be obtained fairly quickly and that may be all that we need.

V. CONCLUSIONS

A brief review of various THz TDS techniques was given and pros and cons discussed, followed by a discussion of theory behind EM wave propagation in dielectric samples. Measurements and extracted parameters – index of refraction, permittivity and conductivity - were presented for two silicon samples. Real parts behaved as expected but imaginary part of conductivity did not show a distinct peak. Parameters of Drude model were extracted but only approximately due to lack of distinct peaks. Table of Drude models was compiled for comparison with presented results. Finally, influce of sample thickness and environement on extracted material parameters was discussed.

REFERENCES

[1] Martin van Exter and D. Grischkowsky, “Carrier dynamics of electrons and holes in moderately doped silicon,” Phys. Rev. B, vol. 41, no. 17, pp. 12140 – 12149, 15. June 1990.

[2] T.I. Jeon and D. Grischkowsky, “Nature of conduction in doped silicon,” Phys. Rev. Letters, vol. 78, no. 6, pp. 1106-1109, 10 February 1997.

[3] T.I. Jeon, “Characterization of Doped silicon from 0.1 to 2.5 THz using multiple reflections,” J. of the Optical Society of Korea, vol. 3, no. 1, pp. 10-14, March 1999.

[4] T.I. Jeon and D. Grischkowsky, “Characterization of optically dense, doped semiconductors by reflection THz time domain spectroscopy,” Appl. Phys. Lett., vol. 72, no 23, pp. 3032- 3034, 8 June 1998.

[5] S. Nashima, O. Morikawa, K. Takata, and M. Hangyo, “Measurement of optical properties of highly doped silicon by terahertz time domain reflection spectroscopy,” Appl. Phys. Lett., Vol. 79, no. 24, pp. 3923-3925, 10 December 2001.

[6] K. J. Willis, S. C. Hagness, and I. Knezevic “A generalized Drude model for doped silicon at terahertz frequencies derived from microscopic transport simulation,” Appl. Phys. Lett., vol. 102, no.12, id. 122113 (2013), 4 pages, DOI 10.1063/1.4798658 .

[7] T.I. Jeon and D. Grischkowsky, “Observation of a Cole-Davidson type complex conductivity in the limit of very low carrier densities in doped silicon,“ Appl. Phys. Lett., vol. 72, no. 18, pp. 2259 – 2261, 4. May 1998.

[8] Jianming Dai, Jiangquan Zhang, Weili Zhang, and D. Grischkowsky, “Terahertz time-domain spectroscopy characterization of the far-infrared absorption and index of refraction of high-resistivity, float-zone silicon,” J. Opt. Soc. Am. B, vol. 21, no. 7, pp. 1379–1386, July 2004.

[9] J. Lloyd-Hughes and T.I. Jeon, “A review of the terahertz conductivity of bulk and nano-materials,” J. of Infrared, Millimetre and Terahertz Waves, vol. 33, no. 9, pp. 871-925, 2012. DOI: 10.1007/s10762-012-9905-y .

[10] T.D. Dorney, R.G. Baraniuk, and D.M. Mittleman, “Material parameter estimation with terahertz time-domain spectroscopy,” J. Opt. Soc. Am A, vol. 18, no. 7, pp. 1562-1571, July 2001.

[11] Advanced Photonix, Inc, T-Ray 5000, http://advancedphotonix.com/thzsolutions/products/t-ray-5000/

[12] TeraView, http://www.teraview.com/

[13] O. Sushko, K. Shala, R. Dubrovka, and R. Donnan, “Revised metrology for enhanced accuracy in complex optical constant determination by THz-time-domain spectrometry,” J. Opt. Soc. Am. A, vol. 30, no. 5, pp. 979-986, May 2013

[14] Yun-Shik Lee, “Principles of Terahertz Science and Technology (Lecture notes in physics),” Springer, December 2008, ISBN-13: 978-0387095394

[15] L. Thamizhmani, A. K. Azad, Jianming Dai, and W. Zhang, “Far-infrared optical and dielectric response of ZnS measured by terahertz time-domain spectroscopy,” Appl. Phys. Lett., vol. 86, 131111, 2005. DOI 10.1063/1.1896451.

[16] J. Han, A.K. Azad, and W. Zhang, ”Far-Infrared Characteristics of Bulk and Nanostructured Wide-Bandgap Semiconductors,” J. Nanoelectronics and Optoelectronics, vol. 2, no. 3, pp. 222–233, 2007

[17] G.P. Kniffin and L.M. Zurk, ”Model-Based Material Parameter Estimation for Terahertz Reflection Spectroscopy,” IEEE Trans. Terahertz Science and Technology, vol. 2, no. 2, pp. 231-241, March 2012.

Impact of microstrip width and annealing time on the characteristics of micro-scale graphene FETs

M. Poljak[1,*], M. Wang[2], S. Žonja[1], V. Đerek[3], M. Ivanda[3], K. L. Wang[4], and T. Suligoj[1]

[1] Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia
[2] Intel Corporation, Boise, ID, USA
[3] Laboratory for Molecular Physics, Rudjer Boskovic Institute, Zagreb, Croatia
[4] Department of Electrical Engineering, University of California at Los Angeles, CA, USA
*Corresponding author: mpoljak@zemris.fer.hr

***Abstract* – Graphene field-effect transistors are fabricated using CVD-grown graphene and a wet transfer technique. Devices are characterized in terms of carrier mobility, contact resistance, output and transfer characteristics. We found that the mobility decreases and contact resistance increases with the decreasing graphene microstrip width. In addition, we found that graphene FET characteristics strongly depend on annealing time. The behavior of device characteristics is explained by discussing carrier scattering on defects and impurities, and the formation of *pn*-junctions in graphene.**

I. INTRODUCTION

Due to its high carrier mobility and two-dimensional structure, graphene-based field-effect transistors (FETs) have become promising alternatives to silicon MOSFETs for future devices [1]. Graphene exhibits a linear dispersion relation, which results in massless carriers and low density of states (DOS) [2]. The former is responsible for the high carrier velocity and mobility in graphene (in excess of 200.000 cm^2/Vs in suspended samples [3], ~10.000 cm^2/Vs in graphene on SiO_2 [4]), whereas the latter results in low carrier injection into graphene. Therefore, despite the high carrier mobility in the channel of the graphene FET, the ON-state performance is limited by the contact resistance that is governed by injection of carriers from the metal.

The feasibility of graphene FETs depends on the OFF-state current and ON-OFF current ratio, which is crucial for digital/logic applications, and on the ON-state performance, which is crucial for high-speed applications. Most of the papers reported so far have focused on reducing I_{OFF} through bandgap engineering, e.g. using graphene nanoribbons [5], [6], [7]. However, the number of reports on examining and improving the ON-state performance is limited [8], [9]. Hence, both the experimental study and theoretical analyses of carrier transport and contact resistance are necessary. Although the mechanically exfoliated graphene exhibits the highest mobilities at room temperature, the method is not viable for larger scale production. The growth of graphene by chemical vapor deposition (CVD) on certain metal foils has become a promising method [10], [11]. Nevertheless, a wet transfer technique is necessary to transfer the graphene layer onto substrate, which results in defects and may result in doping of graphene. In this paper, we study the feasibility of CVD-grown graphene FETs for micro-scale applications by examining the properties of carrier mobility, contact resistance and *I-V* characteristics. We focus on the analysis of size effects, i.e. the impact of graphene microstrip width, and annealing time (used with the initial aim of reducing the unwanted doping) on device figures of merit.

II. DEVICE FABRICATION

Chemical vapor deposition (CVD) of carbon on thin copper foils was used for the fabrication of mono-layer graphene. The process was performed in a tube furnace at 1000 °C and low pressure, using a mixture of a high purity methane (CH_4) and hydrogen (H_2). The details of the graphene growth process can be found elsewhere [10].

The substrate was a heavily doped *n*-type silicon with a 300 nm thick layer of SiO_2, where silicon is the gate electrode, while SiO_2 serves as the gate dielectric. The substrate was cut into several pieces (dies) each approx. 1.5 cm × 1.5 cm in size, after which the dies were cleaned (acetone rinse, isopropanol (IPA) cleaning, and O_2 plasma ashing). Graphene was transferred to the substrate using the wet transfer technique [12], which is briefly described here. The copper foil with graphene was fixed to an auxiliary wafer and a thick layer of polymethyl methacrylate (PMMA) was coated onto the foil with a spin coater (1000 rpm, 45 s), followed by bake-out at 180 °C for 2 – 3 min on a hot plate. The copper foil with graphene and PMMA on top of it was cut into smaller pieces, each with approx. dimensions of 1 cm × 1 cm. The copper-foil-graphene pieces were put onto the surface of the copper etchant ($FeCl_3$) for 10 min, keeping PMMA on the top side. A transparent layer of PMMA was floating on the surface with graphene attached to the bottom of the PMMA layer after the copper was etched out. After copper etching, the PMMA-graphene sample was cleaned in deionized (DI) water in two steps (10 min two times in two separate beakers). After DI cleaning, the floating PMMA-graphene sample was transferred to substrate using the previously prepared Si-SiO_2 dies. In this final step of the transfer process, we obtained the Si-SiO_2-graphene-PMMA layered structure. In order to allow the following fabrication steps, the PMMA was removed by acetone (30 min) and the dies were cleaned by IPA rinsing (5 min) and N_2 blow drying.

The next step was device fabrication using the obtained Si-SiO_2-graphene samples. The fabrication of graphene FETs consisted of three main steps: first lithography, metal deposition, and second lithography. During the first lithography, contacts and wires were defined. The die was coated with AZ5214 photoresist with

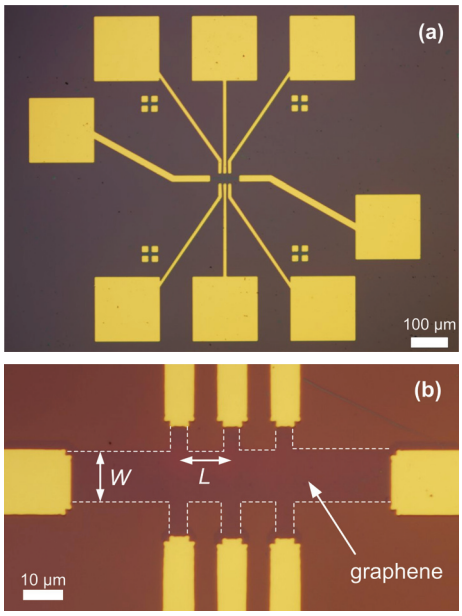


Fig. 1. Device layout showing (a) arrangement of contact pads and wires, and (b) magnified central part with the graphene layer.

a spin coater (4000 rpm, 40 s), followed by baking at 100 °C for 1 min. Next, the sample was exposed to UV light (365 nm) in the lithography system (hard exposure for 5 s, bake-out at 120 °C for 1 min, flood exposure 120 s). Photoresist was developed in 1 : 5 AZ400K : H_2O solution for 40 s, after which the dies were rinsed with DI water and N_2 blow-dried. Metal deposition (10 nm/100 nm Ti/Au) was performed using the e-beam metal deposition system. Metal lift-off and photoresist removal was done in acetone for 2 h, which was followed by additional acetone rinsing, cleaning by IPA and N_2 blow drying. The layout of the graphene layer was defined in the second lithography. In the lithography system, the exposure lasted 10 s in the hard-contact mode, and the photoresist was developed in 1 : 4 AZ400K : H_2O solution (1 min). After photoresist development, the die was cleaned with DI water and N_2 blow drying. The excess graphene was etched away by O_2 plasma (in plasma asher system for 10 min). Finally, the photoresist was removed by acetone (the samples were left in it for 2 hrs). The last step was cleaning with IPA and N_2 blow drying.

Figure 1(a) shows the final arrangement of contact pads (200 µm × 200 µm) and wires (10 or 20 µm in width), while the magnified central part in Fig. 1(b) shows the graphene area obtained after the second lithography. The width of the graphene microstrip in Fig. 2(b) is ~15 µm. The layout shown in Fig. 1 is standard for transport characterization using the four-probe setup [13]. We have fabricated devices with different graphene microstrip widths (5, 8, 10 and 15 µm) and different separations of inner electrodes (8 and 10 µm).

III. Characterization of Graphene FETs

A. *Raman Spectroscopy*

The number of layers and quality of graphene samples can be examined by Raman spectroscopy [14], [15].

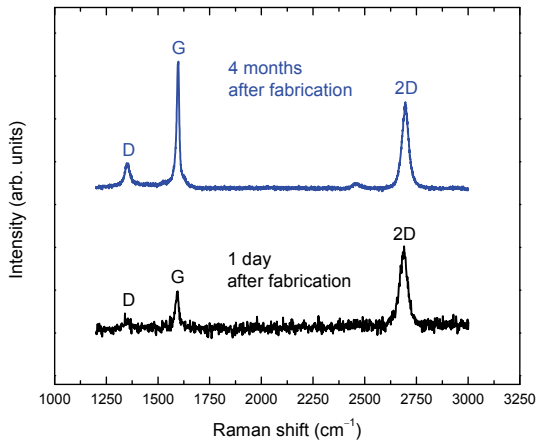


Fig. 2. Typical Raman spectra of graphene measured ~1 day and ~4 months after fabrication. The apparent noise in the spectrum of the fresh graphene sample is due to shorter integration time.

Typically, a fresh single-layer graphene (SLG) exhibits a Raman spectrum that contains two main peaks: G (graphite) peak centered at ~1580 cm^{-1} and 2D (double resonant defect) peak centered at ~2700 cm^{-1}. The G peak comes from phonon modes at the zone center (in the vicinity of Γ point). The fresh undoped SLG exhibits the intensity ratio I(G)/I(2D) of ~0.3. Furthermore, the 2D band can be approximated by a single Lorentzian curve in the case of SLG, which indicates a single Raman frequency. Yet another feature of SLG is that the FWHM of the 2D band equals ~30 cm^{-1}. The D (defect) peak centered at ~1350 cm^{-1} can be also observed, which depends on defect density in graphene. The D peak is a consequence of any kind of symmetry breaking; it can be induced by edge defects or defects at domain borders. In comparison to SLG, multi-layer graphene has a different 2D band shape and different values of I(G)/I(2D). For example, in bilayer graphene 2D band is approximated by four Lorentzian curves, with FWHM and I(G)/I(2D) ratio significantly higher than for SLG.

Raman spectroscopy was performed on all samples and a rather uniform response was obtained. Figure 2 presents a typical Raman spectrum of graphene devices, taken ~1 day after fabrication and ~4 months after fabrication. In the Raman spectrum obtained for fresh graphene devices, we observe a clear SLG signature; the G peak is centered at ~1590 cm^{-1} and 2D band is centered at ~2690 cm^{-1} with FWHM of ~35 cm^{-1}. In addition, the ratio I(G)/I(2D) equals ~0.4. Surprisingly, the spectrum taken ~4 months after fabrication exhibits I(G)/I(2D) ratio of ~1.46, while the G peak remains centered at ~1590 cm^{-1} and 2D band centered at ~2690 cm^{-1} with FWHM of ~35 cm^{-1}. The shape of the 2D band corresponds to that of SLG, however the values of I(G)/I(2D) is something we would expect from multi-layer graphene. Since the devices were kept at room temperature and in atmospheric conditions, we assume that a strong doping of graphene has occurred. Ni *et al.* [16] report that Raman spectrum of graphene can be modified by doping, i.e. that the G peak can increase compared to the 2D peak. Therefore, it seems that Raman spectroscopy can be used to analyze the doping and aging of SLG by monitoring the change in I(G)/I(2D).

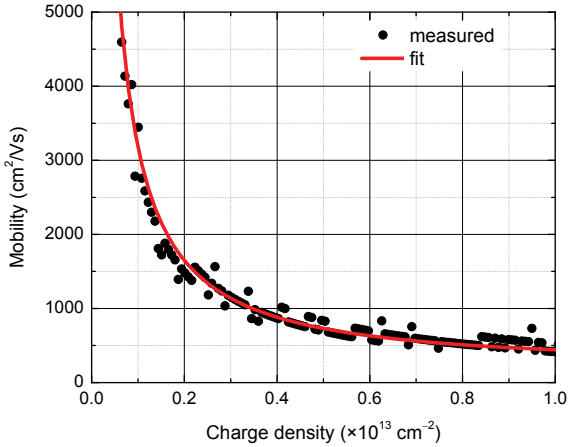


Fig. 3. Measured mobility of a typical graphene device with $W = 15$ µm and $L = 10$ µm at room temperature and in atmospheric conditions.

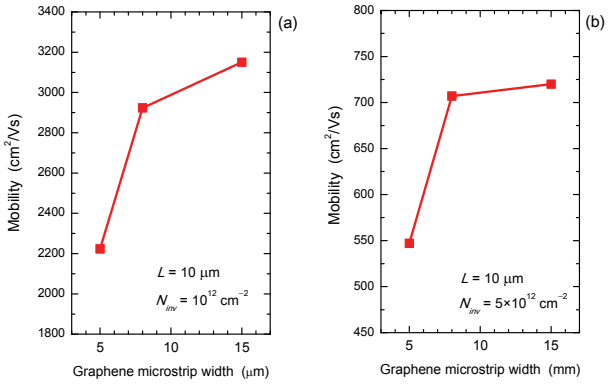


Fig. 4. Impact of downscaling of the graphene microstrip width on carrier mobility for N_{inv} of (a) 10^{12} cm^{-2} and (b) 5×10^{12} cm^{-2}.

B. *Mobility*

The mobility is measured by the four-probe method [13], [17] at room temperature and in atmospheric conditions. The measurement was performed on a probe station using HP 4145B Semiconductor Parameter Analyzer (HP-SPA). First, the low-field conductivity which is modulated by the back-gate voltage is extracted using the following equation:

$$\sigma = \frac{L}{W}\frac{I_{14}}{V_{23}}, \tag{1}$$

where W is the width of the graphene sample, L is the separation between inner electrodes 2 and 3, I_{14} is the current set between the outer electrodes 1 and 4, and V_{23} is the voltage drop measured between the inner electrodes. The mobility is calculated using:

$$\mu = \frac{\sigma}{qN_{inv}}, \tag{2}$$

where N_{inv} is the carrier density induced by the gate voltage calculated by the following expression:

$$N_{inv} = \frac{C_{ox}}{q}\left(V_G - V_D\right), \tag{3}$$

where C_{ox} is the oxide capacitance per unit area and $V_G - V_D$ is the gate voltage over-drive compared to the Dirac voltage. For $V_G - V_D > 0$, electrons are induced in graphene because the Fermi level is effectively shifted into the conduction band. In contrast, for $V_G - V_D < 0$ the Fermi level is in the valence band. For $V_G = V_D$ the Fermi level is precisely set at the crossing point of the valence and conduction bands, i.e. at the point of minimum DOS. This results in minimum conductivity at the Dirac point because carrier concentration is negligible for $V_G = V_D$.

The measured mobility of a typical graphene device with $W = 15$ µm and $L = 10$ µm at room temperature is shown in Fig. 3. Due to deviation of the measured mobility, which is a consequence of the low voltage resolution of the HP-SPA (1 mV), a fitted curve is also indicated. A rather high Dirac voltage is obtained from the measured conductivity, i.e. V_D is approx. 100 V, which means that graphene is heavily doped with *p*-type impurities. Consequently, the measured mobility shown in Fig. 3 is obtained for $V_G - V_D < 0$, i.e. it shows hole mobility. In order to measure the electron mobility $V_G > 100$ V is needed, which is beyond the range reachable by the HP-SPA. We report the mobility of 3180 cm^2/Vs at $N_{inv} = 10^{12}$ cm^{-2} (moderate inversion) and 450 cm^2/Vs at $N_{inv} = 10^{13}$ cm^{-2} (strong inversion). These values are comparable to other reported mobilities in CVD-grown and transferred graphene [17] and considerably lower than in exfoliated or suspended graphene samples [3], [4]. For the latter, the reasons are (i) the impact of remote phonon scattering that originates in the dielectric substrate, (ii) scattering on grain/domain boundaries that form during the CVD growth, (iii) scattering on lattice defects either on the edges of the graphene microstrip or on defects induced by the transfer process. Nevertheless, the fabricated graphene FETs exhibit greater mobility (3180 cm^2/Vs for holes) than bulk Si MOSFETs for either carrier type (~400 cm^2/Vs for holes, ~1500 cm^2/Vs for electrons). Hence, CVD-grown and transferred graphene is a feasible channel material as it exhibits a superior carrier mobility.

Devices with different microstrip widths were fabricated, which allows the investigation of the influence of edge-defect scattering on carrier mobility. Figure 4 presents the dependence of mobility on microstrip width for devices with L =10 µm, for two N_{inv} values. Both in weak and moderate inversion the mobility decreases with decreasing W. At $N_{inv} = 10^{12}$ cm^{-2}, the averaged mobility decreases from 3150 cm^2/Vs at $W = 15$ µm, over 2920 cm^2/Vs at $W = 8$ µm, to 2220 cm^2/Vs at $W = 8$ µm, i.e. mobility deteriorates by 29% when W is downscaled from 15 to 5 µm. In moderately strong inversion, i.e. for $N_{inv} = 5\times10^{12}$ cm^{-2} shown in Fig. 4(b), the decrease is approx. 24%. Mobility decrease reported in Fig. 4 indicates stronger carrier scattering in narrower microstrips. As shown in [7], [18], the contributions of all relevant scattering mechanisms increase when W decreases, and scattering induced by line-edge roughness exhibits an especially strong W-dependence. In the fabricated devices the relative ratio $\Delta W/W$ increases in narrower microstrips due to lithography resolution, which leads to stronger mobility degradation caused by the increased line-edge roughness scattering. In comparison to graphene nanoribbons ($W < 100$ nm) in which the strong W-dependence is naturally expected due to strong quantum confinement effects, mobility features reported

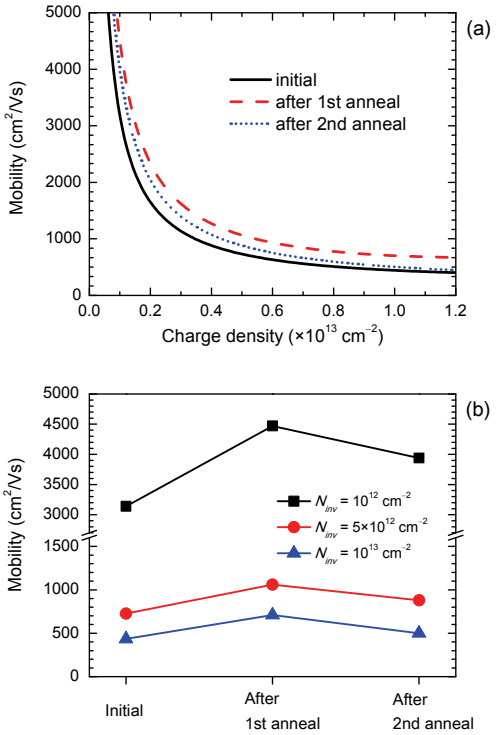


Fig. 5. Influence of annealing duration on carrier mobility in graphene with an emphasis on (a) N_{inv} and (b) annealing-step dependence.

in Fig. 4 exemplify the importance of device width engineering even at the micro scale.

The extracted high Dirac voltage indicates a heavy doping of graphene with *p*-type impurities. In addition to shifting the V_D, the doping also decreases the mobility due to increased carrier scattering on either neutral or charged impurities. In this work, we ask whether annealing can reduce doping, and hence we study the effects of hot-plate annealing at 300 °C performed in two steps, first lasting 30 min, the second 90 min. Figure 5 presents the initial mobility and two mobility curves obtained after each annealing step. The annealing and measurement were performed in atmospheric conditions and at room temperature for the same device as in Fig. 3 ($W = 15$ µm and $L = 10$ µm). After the first annealing (30 min at 300 °C), the mobility increases. We attribute this increase to evaporation of the adsorbed impurities from graphene. Surprisingly, the additional 90 min annealing decreases the mobility in comparison to the previous step, while keeping it above the initial mobility curve. The effects of annealing are quantified in Fig. 5(b) for three different N_{inv} values. The impact of annealing steps is qualitatively the same in weak, moderate and strong inversion, i.e. the first 30 min anneal increases the mobility, while the second 90 min step decreases the mobility. For example at $N_{inv} = 10^{12}$ cm^{-2}, the first anneal causes an increase of approx. 41% in comparison to the initial graphene sample, whereas the second annealing decreases the mobility by approx. 12%. Despite the decrease, after the second anneal the mobility is by 24% higher than in the initial sample. Therefore, annealing proved to be a viable method of reducing impurity density and increasing carrier mobility. However, the duration is a limiting factor due to defect generation, which is discussed in Subsect. III-C.

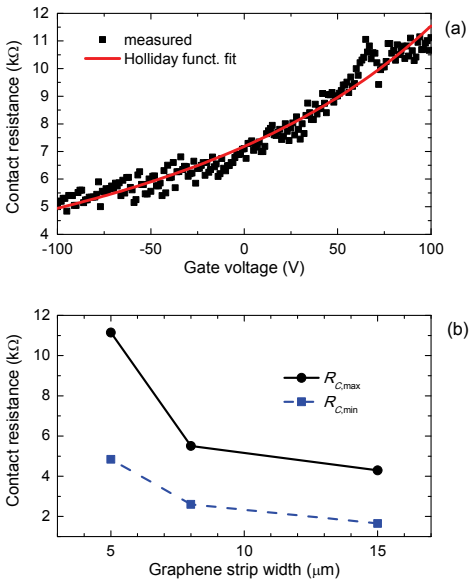


Fig. 6. (a) Contact resistance versus gate voltage for a device with $W = 5$ µm and $L = 10$ µm. (b) Dependence of $R_{C,\min}$ and $R_{C,\max}$ on the width of the graphene microstrip.

C. *Contact Resistance*

The contact resistance of the Ti/Au-graphene contacts is extracted by subtracting the resistance between the inner electrodes (R_{23}) from the resistance between the outer electrodes (R_{14}) and taking into account device dimensions. The following equation was used:

$$R_C = \frac{1}{2}\left(R_{14} - R_{23}\frac{L_{14}}{L_{23}} \right), \qquad (4)$$

where L_{14} and L_{23} are the separations of the respective electrodes [8], [13]. For all devices, the L_{14} equals 90 µm.

Dependence of the extracted contact resistance of the device with $W = 5$ µm and $L = 10$ µm on gate voltage is shown in Fig. 6(a). We observe that R_C depends on the gate voltage and ranges from ~5 kΩ (away from the Dirac point) to ~11 kΩ (close to the Dirac point). When the electrode width is taken into account, we find that $R_C W$ ranges from ~25 kΩ µm up to ~55 kΩ µm, which is comparable to the previously reported results for the contact resistance of Ti/Au-graphene samples [8], but is also much higher that those reported for Ni contacts (~1 kΩ µm [8]) and Pd/Au contacts (~200 Ω µm [9]).

The behavior of the contact resistance can be understood by examining the band structure of the metal-graphene interface, which is illustrated in Fig. 7. We model graphene as metal since it exhibits a zero bandgap, and we take into account the difference between work-functions for Ti and graphene (4.3 and 4.5 eV, respectively). Due to low DOS in graphene, the screening length is considerably larger in graphene than in metal [19], which leads to a finite spatial change of the vacuum level as illustrated in Fig. 7(a). Due to work-function difference, an electron charge transfer region is formed in graphene near the junction even without the applied gate voltage [Fig. 7(b)]. The situation in Fig. 7(b) corresponds to $V_G = V_D$ since E_F is in the neutrality point away from

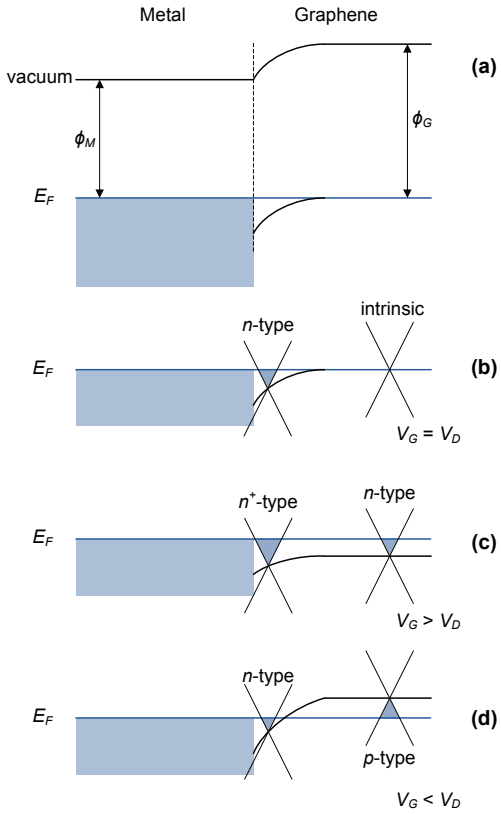


Fig. 7. Band diagram of the metal-graphene junction illustrating (a) band alignment, and specific cases when (b) $V_G = V_D$, (c) $V_G > V_D$ and (d) $V_G < V_D$.

the interfaces. In this case, an *n-i* junction is formed which results in the lowest possible conduction since the DOS in the intrinsic part equals zero. When V_G higher than V_D is applied, the neutrality level moves down and the Fermi level effectively shifts up, which results in high conduction since an n^+-*n* junction is formed [shown in Fig. 7(c)]. Finally, for $V_G < V_D$, the Fermi level effectively shifts down, which results in low conductance because a *p-n* junction is formed [see Fig. 7(d)]. Therefore, the experimental results in Fig. 6(a) can be explained using metal-graphene junction examples from Fig. 7(d) and (b). Namely, our graphene is heavily *p*-doped, which means that the cause of R_C increase with V_G in Fig. 6(a) is the evolution from *n-p* towards *n-i* junction type.

The influence of microstrip width downscaling on $R_{C,\min}$ and $R_{C,\max}$, extracted at V_G of −100 V and 100 V, respectively, is reported in Fig. 6(b). We observe that R_C increases with the decrease of the microstrip width. In turn, this means that the carriers are injected from the metal into graphene on the contact edges, which agrees with conclusions reported in [8].

Figure 8. shows the consequences of annealing on R_C in the device with $W = 15$ µm and $L = 10$ µm. After the first annealing (30 min at 300 °C), a decrease of both $R_{C,\min}$ and $R_{C,\max}$ is observed. The decrease equals ~7% and ~17% in the case of $R_{C,\min}$ and $R_{C,\max}$, respectively. The reduction of R_C indicates a change of the metal-graphene junction type from *n-p* to either *n-p*$^+$, *n-n* or *n-n*$^+$, since *n-p*$^+$, *n-n* and *n-n*$^+$ junctions are more conductive than the *n-p* junction, according to theoretical/numerical

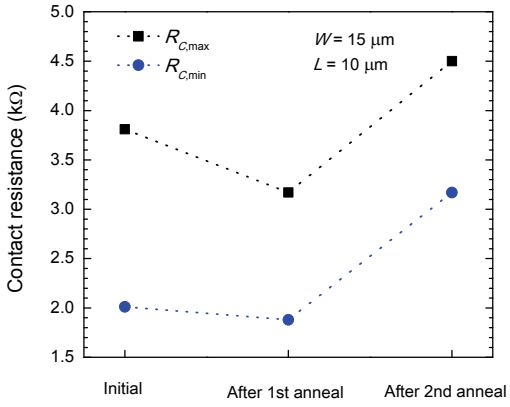


Fig. 8. Impact of annealing duration on contact resistances, $R_{C,\min}$ and $R_{C,\max}$, in a device with $W = 5$ µm and $L = 10$ µm. Annealing temperature is 300 °C (hot plate).

studies reported in [20], [21]. The most feasible assumption is the change from *n-p* to *n-n*$^-$ junction type because it seems unreasonable that a high doping, either of p^+ or n^+ type, would occur in graphene after the first annealing. The change of doping that is inferred from the R_C behavior agrees with the previously discussed mobility increase after the first annealing step, which was attributed to impurity evaporation. After the second annealing step (90 min at 300 °C), contact resistance increases considerably (~69% and ~42% for $R_{C,\min}$ and $R_{C,\max}$, respectively). We attribute this increase mainly to defect generation in graphene caused by evaporation of impurities and residual PMMA.

D. Output I-V Characteristics

Figure 9 shows the output characteristics of the graphene FET with $W = 15$ µm and $L = 10$ µm. The initial measurements are presented in Fig. 9(a), while (b) and (c) show the *I-V* curves obtained after the first and second annealing step, respectively. The curves are obtained for −100 V to 100 V range of V_G, and V_D from −5 V to 5 V, and the current is normalized by the microstrip width. Our devices do not exhibit current saturation, i.e. output conductance does not decrease as in semiconductor FETs. We observe similar behavior up to V_D of ±50 V, above which the high current densities destroy the devices. The absence of saturation is partly attributed to high V_D, so the pinch-off at the drain side cannot occur even for $V_D < 50$ V, and is partly caused by the absence of a high-κ dielectric that would be a source of remote (soft) phonon scattering [22]. The maximum current that is extracted at $V_D = 5$ V slightly increases from 55 µA/µm to 57 µA/µm after the first anneal (30 min), which is a consequence of increased carrier mobility and decreased R_C (see previous subsections). In contrast, the second annealing step (90 min) results in $I_{D,\max}$ of 42 µA/µm, which is lower than the current in the initial sample. We attribute this strong deterioration of $I_{D,\max}$ (~24%) to the increase of contact resistance as reported in Fig. 8.

IV. Conclusion

Graphene field-effect transistors are fabricated using CVD-grown graphene and a wet transfer technique. Devices are characterized in terms of carrier mobility,

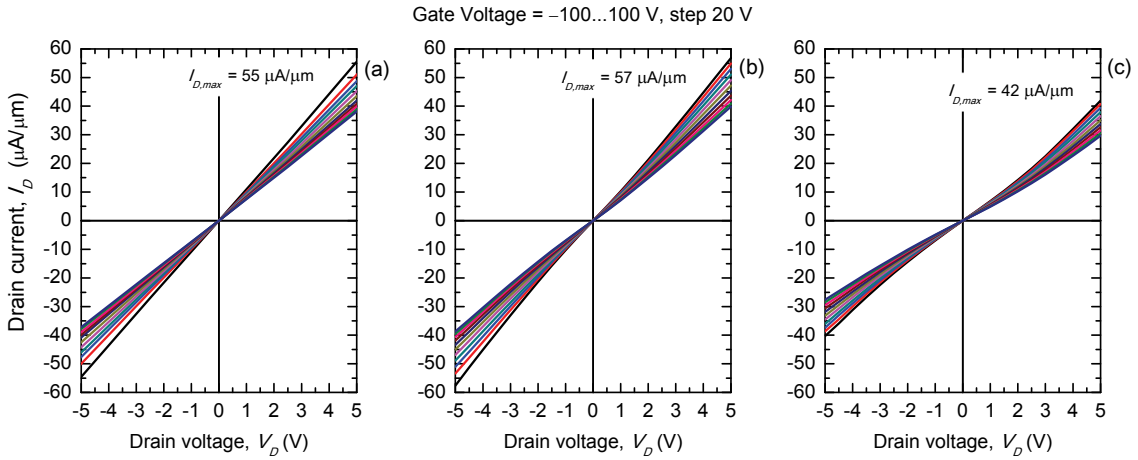


Fig. 9. Output characteristics of graphene FETs obtained for V_G range from −100 V to 100 V, and V_D range from −5 V to 5 V. The current is normalized by the microstrip width. Figures show (a) initial measurement, and measurement after the (b) first and (c) second annealing step.

contact resistance, output and transfer characteristics. Before annealing, typical device exhibits mobility of ~3200 cm^2/Vs at inversion charge density of 10^{12} cm^{-2}, with contact resistance of up to ~11 kΩ (~55 kΩ μm when microstrip width is taken into account). When graphene microstrip width is downscaled, the mobility decreases and contact resistance increases. Furthermore, device characteristics strongly depend on annealing time. For shorter anneal (30 min), mobility increases (+41%) and contact resistance decreases, while the additional longer anneal (90 min) has the opposite results. The impact of microstrip width and annealing on carrier mobility and contact resistance is explained by discussing the contributions from impurity and defect scattering and the consequences of *pn*-junction formation in graphene.

ACKNOWLEDGMENTS

M. P. would like to thank E. B. Song, M. Xue and C. Zeng for the help with device fabrication at UCLA Nanoelectronics Research Facility, and J. Žilak, T. Knežević and M. Koričić for the assistance with device characterization at FER-ZEMRIS. M. P. acknowledges support of the U. S. Department of State through the Fulbright Fellowship Program. M. P. and T. S. acknowledge partial financial support from the University of Zagreb and the Ministry of Science, Education and Sports of the Republic of Croatia. K. L. W. acknowledges financial support from the MARCO Focus Center on Functional Engineered Nano Architectonics (FENA).

REFERENCES

[1] F. Schwierz, "Graphene transistors," *Nat. Nanotechnol.*, vol. 5, no. 7, pp. 487–496, May 2010.

[2] A. H. Castro Neto *et al.*, "The electronic properties of graphene," *Rev. Mod. Phys.*, vol. 81, no. 1, pp. 109–162, Jan. 2009.

[3] K. I. Bolotin *et al.*, "Ultrahigh electron mobility in suspended graphene," *Solid State Commun.*, vol. 146, no. 9–10, pp. 351–355, Jun. 2008.

[4] K. S. Novoselov *et al.*, "Electric Field Effect in Atomically Thin Carbon Films," *Science*, vol. 306, no. 5696, pp. 666–669, Oct. 2004.

[5] M. Y. Han *et al.*, "Energy Band-Gap Engineering of Graphene Nanoribbons," *Phys. Rev. Lett.*, vol. 98, no. 20, p. 206805, May 2007.

[6] X. Li *et al.*, "Chemically Derived, Ultrasmooth Graphene Nanoribbon Semiconductors," *Science*, vol. 319, no. 5867, pp. 1229–1232, Feb. 2008.

[7] J. Baringhaus *et al.*, "Exceptional ballistic transport in epitaxial graphene nanoribbons," *Nature*, vol. 506, pp. 349-354, Feb. 2014.

[8] K. Nagashio *et al.*, "Metal/graphene contact as a performance Killer of ultra-high mobility graphene analysis of intrinsic mobility and contact resistance," in *Tech. Digest of International Electron Devices Meeting (IEDM)*, 2009, pp. 1–4.

[9] F. Xia *et al.*, "The origins and limits of metal-graphene junction resistance," *Nat. Nanotechnol.*, vol. 6, no. 3, pp. 179–184, 2011.

[10] H. I. Rasool *et al.*, "Continuity of Graphene on Polycrystalline Copper," *Nano Lett.*, vol. 11, no. 1, pp. 251–256, Jan. 2011.

[11] T. Kobayashi *et al.*, "Production of a 100-m-long high-quality graphene transparent conductive film by roll-to-roll chemical vapor deposition and transfer process," *Appl. Phys. Lett.*, vol. 102, no. 2, p. 023112, Jan. 2013.

[12] J. W. Suk *et al.*, "Transfer of CVD-Grown Monolayer Graphene onto Arbitrary Substrates," *ACS Nano*, vol. 5, no. 9, pp. 6916–6924, Sep. 2011.

[13] D. K. Schroder, *Semiconductor Material and Device Characterization*. John Wiley & Sons, 2006.

[14] A. C. Ferrari and D. M. Basko, "Raman spectroscopy as a versatile tool for studying the properties of graphene," *Nat. Nanotechnol.*, vol. 8, no. 4, pp. 235-246, Apr. 2013.

[15] A. C. Ferrari, "Raman spectroscopy of graphene and graphite: Disorder, electron–phonon coupling, doping and nonadiabatic effects," *Solid State Commun.*, vol. 143, no. 1–2, pp. 47–57, Jul. 2007.

[16] Z. H. Ni *et al.*, "The effect of vacuum annealing on graphene," *J. Raman Spectrosc.*, vol. 41, no. 5, pp. 479–483, 2010.

[17] V. E. Dorgan *et al.*, "Mobility and saturation velocity in graphene on SiO_2," *Appl. Phys. Lett.*, vol. 97, no. 8, pp. 082112–082112–3, Aug. 2010.

[18] M. Poljak *et al.*, "Influence of substrate type and quality on carrier mobility in graphene nanoribbons," *J. Appl. Phys.*, vol. 114, no. 5, p. 053701, Aug. 2013.

[19] J. M. Ziman, *Principles of the Theory of Solids*. Cambridge University Press, 1972.

[20] B. Huard *et al.*, "Transport Measurements Across a Tunable Potential Barrier in Graphene," *Phys. Rev. Lett.*, vol. 98, no. 23, p. 236803, Jun. 2007.

[21] T. Low *et al.*, "Conductance Asymmetry of Graphene p-n Junction," *IEEE Trans. Electron Devices*, vol. 56, no. 6, pp. 1292–1299, Jun. 2009.

[22] I. Meric *et al.*, "Current saturation in zero-bandgap, top-gated graphene field-effect transistors," *Nat. Nanotechnol.*, vol. 3, pp. 654–659, Sep. 2008.

Graphene Base Transistors with optimized emitter and dielectrics

Stefano Venica, Francesco Driussi, Pierpaolo Palestri and Luca Selmi
DIEGM, University of Udine, via delle Scienze 206, Udine 33100, Italy
Email: stevenica@alice.it

***Abstract*—The Graphene Base Transistor (GBT) is a very promising device concept for analog applications. The device operates similar to the hot electron transistor and exploits the high carrier mobility of graphene to reduce the base resistance that limits the unity power gain frequency (f_{max}) and the noise figure (NF) of RF devices. Although the DC functionality of the GBT has been experimentally demonstrated, at present RF performance can be investigated by simulations only. In this paper, we predict the DC current and the cutoff frequency of different GBT designs (including dimensions and various materials), with the aim to optimize the GBT structure and to achieve THz operation. In particular, optimized emitter/dielectrics combinations are proposed to maximize RF figures of merit.**

I. Introduction

The very large mobility of graphene makes this material very interesting, especially for analog and RF devices [1], [2]. Indeed, the graphene field effect transitor (GFET), obtained by integrating a graphene layer as channel material in a conventional FET structure, has been already largely investigated and it has shown intrinsic cutoff frequency of about 400GHz [3]. Potentially, the GFET can reach THz operation [4], however it suffers of a high output conductance and too large off currents [5]. This is a significant limitation mainly for digital applications and it is related to the zero band gap of graphene.

Recently, a new device concept, the Graphene Base Transistor (GBT), has been proposed and demonstrated [6], [7]. The first experimental results have been reported in [7], [8] and demostrate the basic DC functionality pf the GBT. The RF performance, instead, has been preliminary investigated by simulations [9], [10] and cutoff frequencies in the THz range for the intrinsic device have been predicted. The output resistance and intrinsic gain are expected to be larger than for GFETs [9], thus GBT may overcome one of the main limitations of GFETs.

In this work, we have extensively explored by simulations the design space of the GBT, in order to investigate the dependence of the GBT performance on the structure and material parameters. To this purpose, we made use of a refined version of the GBT model presented in [9], that has been extended by adding further physical mechanism relevant for the GBT operation. After this preliminary exploration, criteria for device optimization have been proposed and a few designs have been extensively analyzed in terms of cutoff frequency, intrinsic gain and output conductance.

II. GBT structure and operation

The Graphene Base Transistor has a structure very different from the GFET [6]. In the GBT, two insulators separate the three terminals of the device. The emitter, the graphene base and the collector are arranged vertically as in Fig. 1.

The device operates similarly to the hot electron transistor [11]. In particular, the electrons are injected by tunneling from the emitter towards the collector, thus crossing the emitter–base insulator (EBI), the graphene base and the base–collector insulator (BCI) [6]. Hence the charge travels perpendicular to the graphene sheet [12], [13]. Although the base is atomically thin, it is expected that the high mobility of the graphene reduces the base resistance.

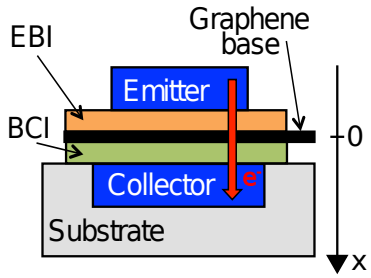


Fig. 1. Sketch of the GBT structure. Electrons leave the emitter by tunneling through the emitter–base insulator (EBI), cross the graphene base and then travel inside the conduction band of the base–collector insulator (BCI).

III. Simulation of the GBT

The model we used to simulate the GBT I–V characteristics is one–dimentional (in the x direction of the GBT structure in Fig. 1) and calculates the the potential profile self–consistently with the charge stored in the graphene base and that traveling inside the GBT.

Concerning the GBT current, we considered only electron's tunneling from the emitter and neglect any hole transport. Furthermore, since the physical origin of the base current is still unclear and debated [8], we assume a priori a negligible base current. The collector current density (J_C) is due to the electrons injected from the emitter.

The unity short circuit current gain frequency f_T is then estimated with the quasi–static expression [9]

$$f_T = \frac{1}{2\pi \frac{dQ_B}{dJ_C}}, \tag{1}$$

where Q_B is the charge density in the graphene base.

The basic equations describing the modeling approach can be found in [9]. However, the model has been extended in order

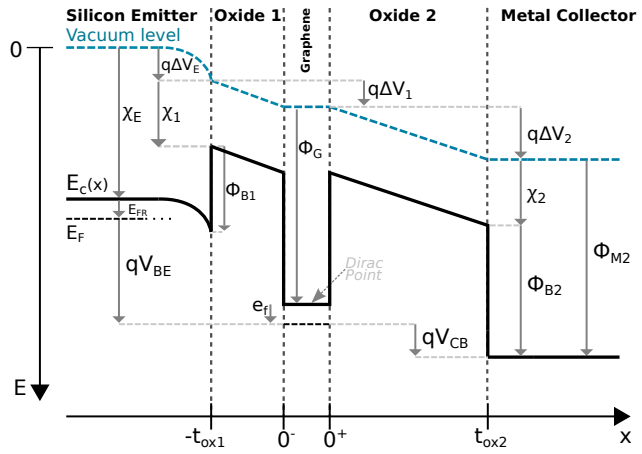


Fig. 2. Conduction band diagram of a GBT with Si–emitter. The band bending in the Si–emitter is calculated by the model self–consistently with the potential along the GBT structure. χ_E, χ_1 and χ_2 are the affinities of Si, the EBI and the BCI, respectively.

to consider GBTs with different emitter materials, including Si–emitters[14]. This allowed us also to verify the model against available experiments [7].

Furthermore, the impact of the charge traveling in the GBT structure is now accounted for when computing the electrostatics. As a result, space charge effects at high current (that usually reduce the maximum f_T value in bipolar transistors) are included in the model.

A. GBT with Si–emitter

Figure 2 reports the Conduction Band (CB) diagram of a GBT with an *n*–doped Si–emitter. As it can be seen, the applied base–emitter voltage (V_{BE}) is distributed among the voltage drops on the Si–emitter (ΔV_E), on the EBI (ΔV_1) and on the graphene. In particular, during normal operation of the GBT, due to the limited density of states in the graphene, the Fermi level (e_f) shifts below to the Dirac Point to induce holes in the base. The applied collector–base voltage (V_{CB}), instead, induces a voltage drop across the BCI (ΔV_2)[1] and, again, on the graphene.

Once ΔV_1, ΔV_2, ΔV_E and e_f are determined, the electrostatic of the GBT is obtained. The complete derivation of the the model is discussed in an upcoming paper [14]. Here we just underline that the graphene charge $Q_B(e_F)$ is computed according to a Fermi–Dirac distribution. Furthermore, by using an approach similar to what done in [15], we expressed the electric field (E_S) at the emitter side of the Si/EBI interface as a function of the potential drop ΔV_E. Now, with the $E_S(\Delta V_E)$ relationship and by resorting to the continuity of the electric displacement field at the Si/EBI interface we can write ΔV_1 as a function of ΔV_E

$$\frac{\Delta V_1}{t_{ox1}} k\varepsilon_0 = \varepsilon_E E_S(\Delta V_E), \qquad (2)$$

[1]Henceforth, to keep the notation as light as possible, we indicate the quantities related to the emitter section and collector section with the subscripts 1 and 2, respectively.

where ε_E is the electric permittivity of the emitter material (ε_E=ε_{Si} in our case), k and t_{ox1} are the relative permittivity and thickness of the EBI.

Equation 2 has been added to the model of [9] to link the electrostatic of the GBT to the Si band bending. The resulting system of non linear equations is solved self–consistently in MATLAB until the potentials inside the GBT structure are obtained. The GBT current is calculated following [9].

B. Impact of the electrons traveling inside the GBT

The high current effects (h.c.e.) impacting the maximum f_T of a HBT are due to the electrostatic feedback generated by the charge traveling inside the device itself [16]. We included this physical mechanism in our model, in order to investigate if GBTs are prone to similar effects and if THz operation is feasible.

In particular, the electron charge injected by the emitter and traveling inside the dielectrics induces holes inside the graphene base, thus reducing the electric field across the insulators and, consequently, the emitter injection current. To estimate this reduction, first we assume that the electrons in the CB of EBI and BCI are moving with constant drift velocity (v_d) and set v_d to a realistic value of 10^7 cm/s [17]. Second, we assume a uniform electron concentration n inside the insulators, hence n=$J_C/(qv_d)$. The impact of n on the GBT electrostatics becomes larger as J_C increases.

The calculated potential profile along the GBT now depends on n and, thus, on the current level. On the other hand, J_C can be calculated once the GBT electrostatic is solved and is known [9]. Therefore, the electrostatic calculation requires a coupled solution of the potentials and of the GBT current. We solved this problem by implementing an iterative procedure that obtains J_C from the potential profile (initial guess by considering n=0), which then becomes an input for the electrostatic solution. The model iterates until convergence is reached.

Figure 3(a) shows the calculated ΔV_1 versus the applied V_{BE}. The symbols are the reference curve calculated with the model of [9]. In the new model (lines), when the energy barrier height between the emitter and the EBI (Φ_{B1}) is relatively large, the tunneling current is low and, thus, the charge in the oxides is small. Since n is negligible, the new simulations agree with [9]. By reducing Φ_{B1}, instead, the tunneling of emitter electrons (thus J_C) increases. This raises the n value inside the oxides and, hence, changes the potential profile. In fact, Fig. 3(a) shows that by reducing Φ_{B1}, ΔV_1 degrades with respect to [9] (symbols).

Figure 3(b) shows how the k value affects the h.c.e. For low k values, the difference between the new model and that of [9] is relatively large, while for large k values, it is reduced. This means that GBTs with high–k dielectrics are less prone to h.c.e. and this is due to the fact that a larger permittivity of the dielectric reduces the impact of the space charge on the GBT electrostatics.

C. Impact of design and materials on GBT performance

We explored in detail the design space of the GBT, thus extending the work in [9]. In particular, we compared GBTs

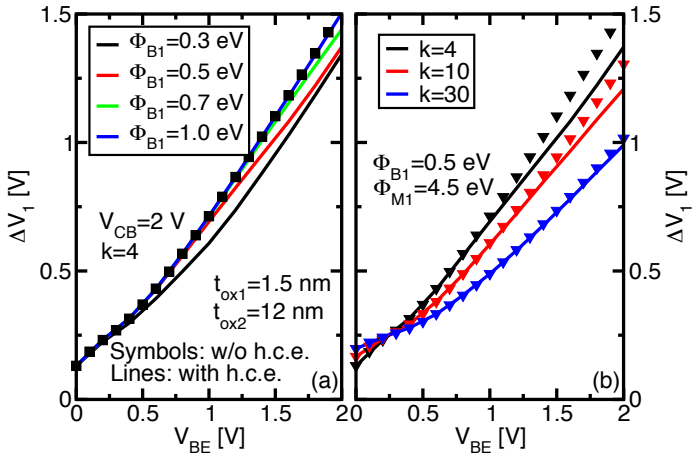


Fig. 3. Potential drop ΔV_1 on the EBI as a function of the applied V_{BE} for different EBI barrier heights Φ_{B1} (a) and for different dielectric constants (b) of metal emitter GBT designs. (a) The electrostatics visibly changes when relatively low Φ_{B1} are considered due to the J_C impact on the electrostatics. (b) The k value mediates the h.c.e.

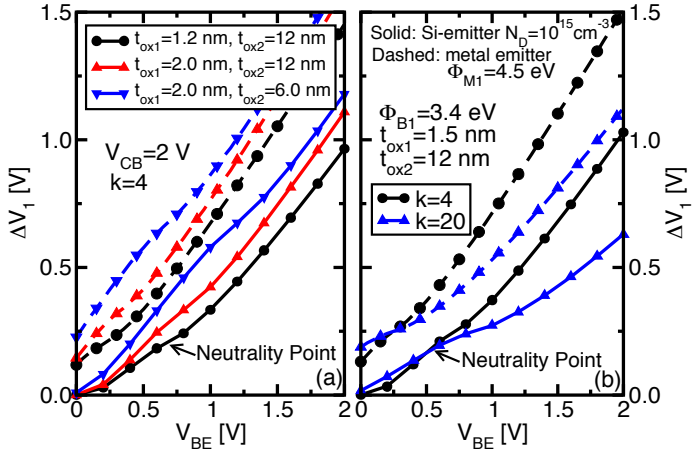


Fig. 4. Potential drop ΔV_1 as a function of V_{BE} for a few combinations of t_{ox1} and t_{ox2} (a) and for different k (b) for GBTs with metal (dashed lines) or Si–emitter (solid lines). The Si band bending reduces ΔV_1 and induces a shift of the neutrality point to higher V_{BE} with respect to metal emitter case.

with metal and Si emitter and investigated their performance as a function of the structure and material parameters. Fig. 4 shows ΔV_1 calculated for different EBI (t_{ox1}) and BCI (t_{ox2}) thicknesses and two k values. The inflection point in the ΔV_1 vs. V_{BE} curves indicates the Neutrality Point (NP), hence the zero base charge condition (Q_B=0) [9]. In normal GBT operation, the base is charged with holes, thus NP corresponds to the minimum V_{BE} that ensures the normal operation of the device. We see that for Si–emitter GBT, NP is moved towards higher V_{BE} and this is due to the additional voltage drop in the Si emitter (ΔV_E). This translates in larger threshold voltages to switch on the device. While t_{ox1} and t_{ox2} contribute to set the NP bias (Fig. 4(a)), this latter is not impacted by k (b).

Figure 5 reports the calculated J_C vs. V_{BE} for the metal (a) and the Si–emitter (b) GBT. It is worth noting the opposite trend of the current on the k value in the two cases. In the metal emitter device, for V_{BE}<1.5 V, the larger the k is, the lower the current is. For V_{BE}>1.5 V, instead, h.c.e. becomes

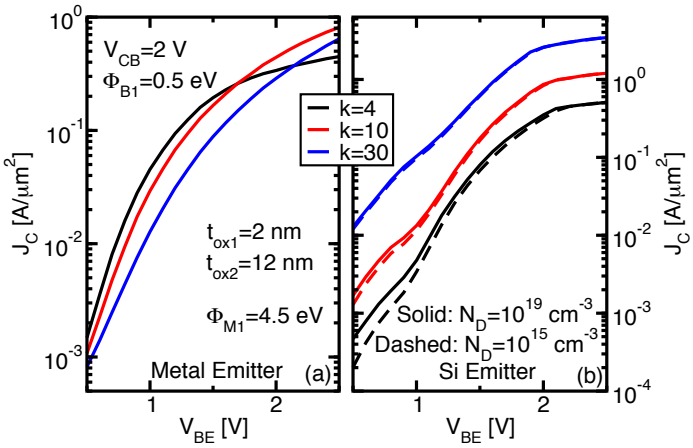


Fig. 5. Collector current density as a function of V_{BE} for a few values of k for GBTs with metal emitter (a) and Si–emitter (b). The Si doping is N_D=$10^{15}cm^{-3}$ (dashed lines) and N_D=$10^{19}cm^{-3}$ (solid lines).

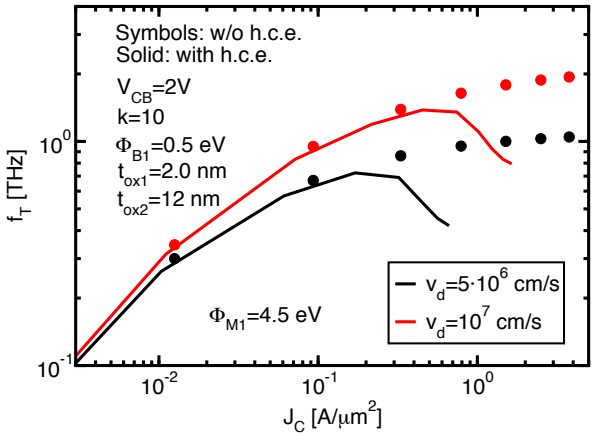


Fig. 6. Cutoff frequency f_T vs. J_C of a metal emitter GBT calculated with different models and for a few v_d values. H.c.e induce f_T drop at large J_C.

relevant and, since GBTs with low k oxides are more prone to h.c.e., a low k value results in a lower J_C. In fact, in Fig. 5(a), we see the black line crossing the other curves.

For the Si–emitter device (Fig. 5(b)), instead, the larger the k is, the higher the J_C is. This is due to Eq. 2, where larger k values result in larger E_S and, hence, larger ΔV_E. ΔV_E lowers the EBI energy barrier with respect to the Si Fermi level (E_F, see Fig. 2) and thus increases J_C. This trend is maintained also at high V_{BE}, where h.c.e. limit the current. Fig. 5(b) shows also that the J_C dependence on the Si doping (N_D) is limited, especially at large V_{BE}, where the Si emitter is driven in strong accumulation and comparable ΔV_E values are obtained, regardless of N_D.

Figure 6 shows f_T versus J_C estimated assuming different v_d values. The simulations that account for the h.c.e. (lines) show the drop off at large J_C typically seen also in HBTs. This effect is not visible in the simulations with the model of [9] (symbols), that are reported as a reference. When v_d is reduced, the maximum f_T of the device is lower. Note that with a realistic value of v_d=10^7 cm/s, the THz operation of the intrinsic device is maintained also when accounting for h.c.e.

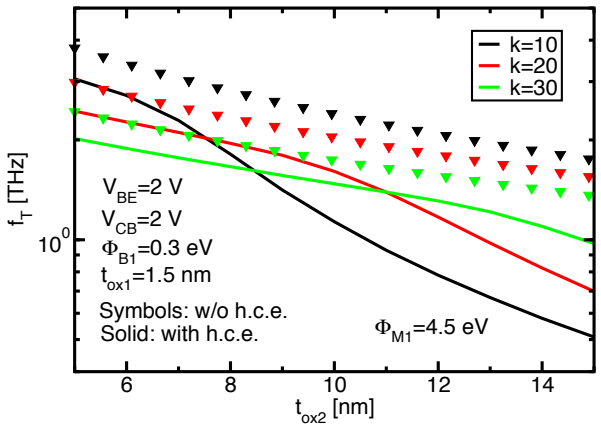


Fig. 7. Cutoff frequency f_T as a function of t_{ox2} for metal emitter GBTs, calculated using different models: thick BCI layers cause a serious decay in the GBT performance.

The cutoff frequency of the GBT is limited by the electrostatic impact of the charge inside the BCI layer, as demonstrated by Fig. 7, that shows f_T as a function of t_{ox2}. The BCI is typically much larger than the EBI, hence it contains a larger amount of charge that, via the electrostatics feedback, impacts the base hole density, thus reducing f_T (see Eq. 1). Indeed the model predicts that, by increasing t_{ox2}, f_T decreases below the values obtained with the model in [9]. Again, the k value is effective in mediating the impact of h.c.e. on f_T: the larger the k is, the smaller the f_T reduction is.

IV. Design and Optimization of the GBT

The above study of the GBT performance as a function of structure and material parameters allowed us to design and analyze optimized GBT structures. In particular, we simulated GBTs with Si or metal emitter, comparing them in terms of relevant RF figures of merit.

A. DEV.1: Silicon emitter GBT

Figure 5(b) shows that dielectrics with large k values ensure large J_C for Si–emitters devices. Therefore, to design the Si–emitter GBT structure, we selected Ta_2O_5 as EBI and BCI, because of its large k value (≈25) and the small CB offset with respect to silicon [21]. Furthermore we assume a Si emitter doping of N_D=10^{19} cm^{-3} and a BCI thickness of t_{ox2}=12 nm, that ensures that the NP is positioned at relatively low V_{BE} (Fig. 4).

Figure 8 shows the calculated J_C and transconductance (g_m) of the above GBT design (hereafter denoted as DEV.1). The model parameters used to simulate DEV.1 are listed in Tab. I. By assuming $t_{ox1} \leq$ 2 nm [9], J_C and g_m reach very large values (Fig. 8(a)). For $V_{BE} \geq$ 1.5 V, J_C is rather independent of t_{ox1} since it is limited by the h.c.e. induced by the charge in the BCI.

B. DEV.2: Metal emitter GBT

Figure 5(a) shows that GBTs with metal emitter are expected to have better performance if low–k dielectrics are selected for the EBI and BCI layers (at least until h.c.e. do not start to degrade the GBT operation). Hence, to design the metal

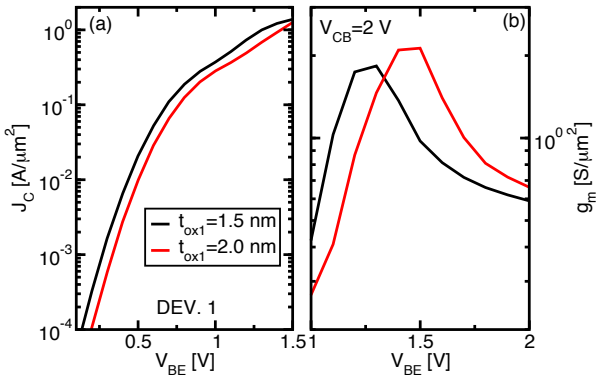


Fig. 8. J_C (a) and g_m (b) as functions of V_{BE} for a few t_{ox1} in DEV.1.

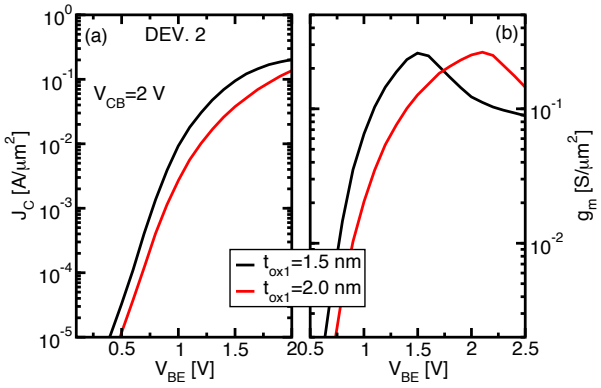


Fig. 9. J_C (a) and g_m (b) of DEV.2 as functions of V_{BE} for a few t_{ox1}.

emitter GBT (hereafter referred as DEV.2), we select a low–k material for the insulators (i. e. SiCOH, [22]). Furthermore, we assumed Ti emitter and collector terminals.

Figure 9 shows J_C and g_m for DEV.2, simulated by using the model parameters in Tab. I. We note that, if the two devices have similar EBI and BCI thickness, J_C and g_m of DEV.2 are lower than in DEV.1. This is due: firstly, to the smaller tunneling barrier between Si and Ta_2O_5 (0.36 eV) than between Ti and SiCOH (0.5 eV); secondly, to the band bending in the Si emitter, which lowers the energy barrier height for electrons tunneling at the emitter Fermi energy (see Fig. 2), thus increasing J_C.

TABLE I. Parameters used to simulate the optimized GBTs. m_I is the tunneling mass inside EBI and BCI to calculate the tunneling current [9]. Collector metal is Ti.

	DEV.1	**DEV.2**	**DEV. 3**
Emitter	Si N_D [cm^{-3}] 10^{19}	Ti Φ_{M1} [eV] 4.33	Si N_D [cm^{-3}] 10^{19}
EBI	Ta_2O_5	SiCOH	Ta_2O_5
BCI	Ta_2O_5	SiCOH	SiCOH
m_I [m_0]	0.5 Ref. [18], [19]	0.5	0.5
k [ε]	25 Ref. [20]	2.5 Ref. [22]	25 / 2.5
Φ_{B1} [eV]	0.36 Ref. [21]	0.53 Ref. [22]	0.36
Φ_{M2} [eV]	4.33	4.33	4.33
t_{ox2} [nm]	12	12	12

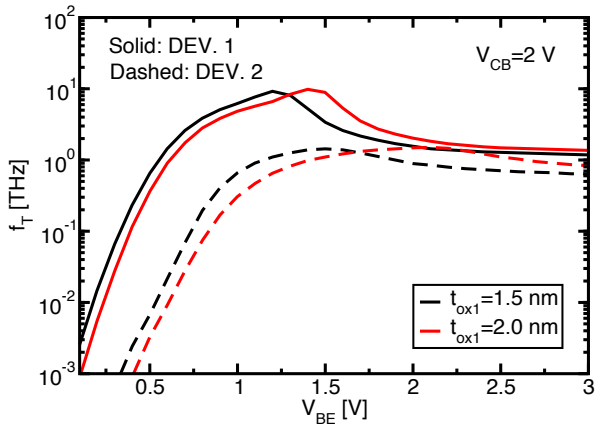


Fig. 10. Comparison of f_T values of DEV.1 (solid lines) and of DEV.2 (dashed lines) as a function of V_{BE} for different EBI thickness.

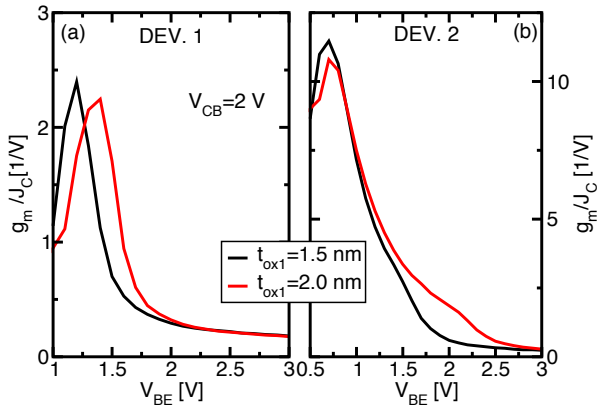


Fig. 11. g_m/J_C ratio as a function of V_{BE} for a few values of EBI thickness, for DEV.1 (a) and DEV.2 (b) GBTs. Mind the different y–axis scale in the two plots.

C. RF performance of GBTs with Si and metal emitters

Figure 10 reports the simulated f_T for DEV.1 and DEV.2. As it can be seen, f_T for DEV.1 (solid lines) is larger than for DEV.2 (dashed) at similar EBI and BCI thicknesses, as expected based on the larger current of DEV.1. This latter transistor shows maximum f_T values up to ≈10 THz, but also for DEV.2 THz operation is reachable by the intrinsic device.

Concerning other figures of merit for analog applications, Fig. 11 reports the transconductance efficiency (g_m/J_C) for the two devices. We note that DEV.2 has a large g_m/J_C ratio, that is comparable to those of scaled CMOS technologies [23]. Furthermore, Fig. 12 compares the instrinsic voltage gains (g_m/g_d) and, again, DEV.2 shows a better behavior. In particular, the metal emitter GBT can reach a g_m/g_d of 20, while for the Si emitter case, the gain is rather low (≈4).

The low intrinsic gain for DEV.1 is due to the high output conductance g_d, in spite of a large g_m compared to that of DEV.2, as demonstrated by Fig. 13. Indeed, the use of a high–k BCI induces a large capacitance coupling between the collector and the base. The base potential can be appreciably modulated by the collector bias, thus increasing the output conductance of the device. This problem can be alleviated by integrating in the Si–emitter GBT structure a BCI with a lower k value, as

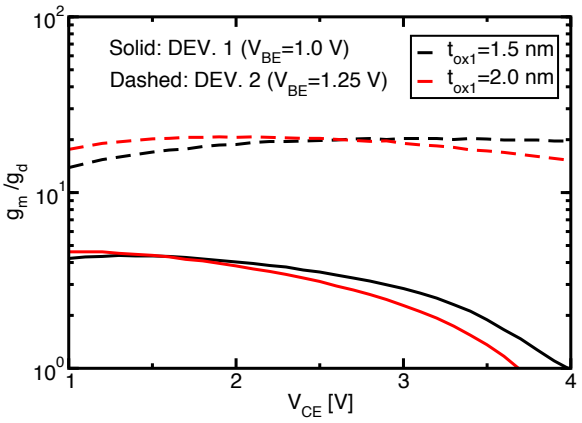


Fig. 12. Intrinsic gain g_m/g_d as a function of V_{CE} for a few values of the EBI thickness, for DEV.1 (a) and DEV.2 (b) GBTs.

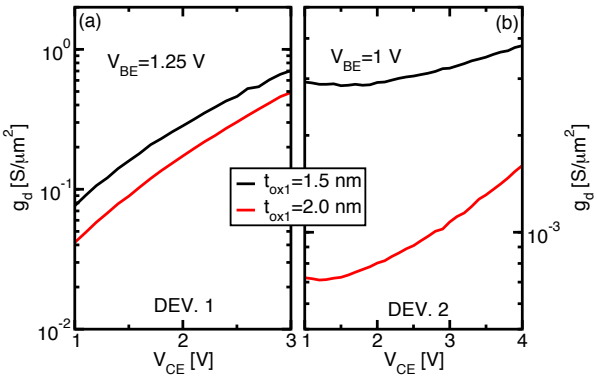


Fig. 13. Output conductance g_d versus V_{CE} for DEV.1 (a) and DEV.2 (b).

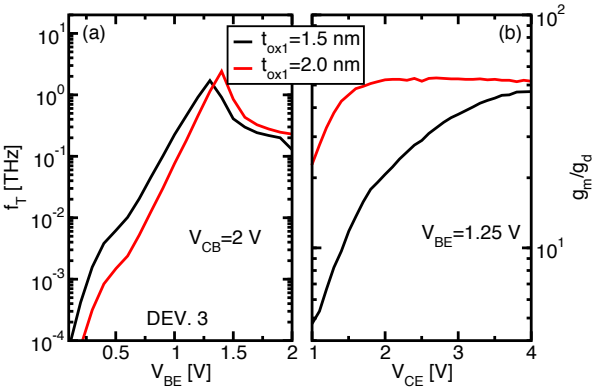


Fig. 14. Cutoff frequency (a) and intrinsic gain (b) for DEV.3.

demonstrated by Fig. 14(b), where we simulated a Si emitter GBT with Ta_2O_5 EBI and SiCOH BCI (named DEV.3). The use of a low–k BCI reduces the base–collector capacitance and increases g_d and, consequently, the gain. In fact g_m/g_d for DEV.3 is around 50 in the best case (Fig. 14, red curve), that is one order of magnitude larger than for DEV.1 (Fig. 12, solid lines). DEV.3 also achieves a maximum f_T above 1 THz (Fig. 14(a)).

V. Conclusion

We explored the design space of the GBT with an improved model for metal or Si emitter devices. The model accounts also for the electrostatic feedback of the charge traveling in the EBI and BCI regions during conduction. We showed that this charge limits the GBT performance at high currents, but the THz operation is still feasible for the GBT.

The model suggests that GBTs with Si/Ta_2O_5 emitters can ensure the largest J_C and f_T values, while GBTs with Ti/SiCOH barriers have better transconductance efficiency and intrinsic gain. Indeed, these figures of merit are strongly related to the permittivity of the EBI and BCI. In particular, a high–k EBI can reduce the tunneling barrier for the electrons in case of Si emitters, while a low–k BCI is the best option to achieve high intrinsic gain and low output conductance.

Acknowledgment

The authors would like to thank Prof. Esseni for helpful discussion. This work was supported by the EU through the FP7 STREP Project "GRADE" under the grant 317839 and by the Italian MIUR through the FIRB project RBFR10XQZ8.

References

[1] S. J. Han, A. V. Garcia, S. Oida, K. A. Jenkins and W. Haensch, "Graphene radio frequency receiver integrated circuit", *Nature Communications*, vol. 5, no. 3086, doi:10.1038/ncomms4086, 2014.

[2] Y. Wu, K. A. Jenkins, A. V. Garcia, D. B. Farmer, Y. Zhu, A. A. Bol, C. Dimitrakopoulos, W. Zhu, F. Xia, P. Avouris and Y. M. Ling, "State of the Art Graphene High Frequency Electronics", *Nano Letters*, vol. 12, no. 6, pp. 30623067, 2012.

[3] R. Cheng et al., *Proceedings of the National Academy of Sciences*, doi:10.1073/pnas.1205696109, 2012.

[4] A. Paussa, M. Geromel, P. Palestri, M. Bresciani, D. Esseni and L. Selmi, "Simulation of graphene nanoscale RF transistors including scattering and generation/recombination mechanisms", *IEDM Technical Digest*, pp. 11.7.1–11.7.4, 2011.

[5] S. O. Koswatta, A. Valdes-Garcia, M. B. Steiner, Yu-M. Lin, and P. Avouris, "Ultimate RF Performance Potential of Carbon Electronics", *IEEE Transaction on Microwave Theory and Techniques*, vol. 59, pp. 2739–2749, 2011.

[6] W. Mehr, J. Dabrowski, J. C. Scheytt, G. Lippert, Y.-H. Xie, M. C. Lemme, M. Ostling, G. Lupina, "Vertical Graphene Base Transistor", *IEEE Electron Device Letters*, vol. 33, no. 5, pp. 691–693, 2012.

[7] S. Vaziri, G. Lupina, C. Henkel, A. D. Smith, M. Ostling, J. Dabrowski, G. Lippert, W. Mehr, and M. C. Lemme, "A Graphene–Based Hot Electron Transistor", *Nano Letters*, vol. 13, no. 6, pp. 2370–2375, 2013.

[8] C. Zeng, E. B. Song, M. Wang, S. Lee, C. M. Torres, J. Tang, B. H. Weiller, and K. L. Wang, "Vertical Graphene–Base Hot–Electron Transistor", *Nano Letters*, vol. 13, no. 4, pp. 1435–1439, 2013.

[9] F. Driussi, P. Palestri, L. Selmi, "Modeling, simulation and design of the vertical Graphene Base Transistor", *Microelectronic Engineering*, vol. 109, pp. 338–341, 2013.

[10] V. Di Lecce, R. Grassi, A. Gnudi, E. Gnani, S. Reggiani and G. Baccarani, "DC and Small–Signal Numerical Simulation of Graphene Base Transistor for Terahertz Operation", *Proceedings of ESSDERC*, pp. 314–317, 2013.

[11] M. Heiblum, "Tunneling hot electron transfer amplifiers (theta): Amplifiers operating up to the infrared", *Solid State Electronics*, vol. 14, pp. 343–366, 1981.

[12] L. Britnell, R. V. Gorbachev, R. Jalil, B. D. Belle, F. Schedin, A. Mishchenko, T. Georgiou, M. I. Katsnelosn, L. Eaves, S. V. Morozov, N. M. R. Peres, J. Leist, A. K. Gerim, K. S. Novoselov and L. A. Ponomarenko, "Field–Effect Tunneling Transistor Based on Vertical Graphene Heterostructures", *Science*, 335, 6071 (2012).

[13] S. B. Kumar, G. Seol and J. Guo, "Modeling of a vertical tunneling graphene heterojunction field–effect transistor", *Applied Physics Letters*, 101, 033503 (2012).

[14] S. Venica, F. Driussi, P. Palestri, D. Esseni and L. Selmi, "Simulation of DC and RF performance of the Graphene Base Transistor", *submitted to IEEE Transaction on Electron Devices*.

[15] E. H. Nicollian and J. R. Brews, "MOS (Metal Oxide Semiconductor) Physics and Technology", *John Wiley and sons (1982)*.

[16] S. M. Sze, "Physics of Semiconductor Devices", Wiley, New York, 1981.

[17] P. Palestri, L. Selmi, E. Sangiorgi and M. Pavesi, "Cathode Hot Electrons and Anode Hot Holes in Tunneling MOS Capacitors", *Proceedings of ESSDERC*, pp. 296–299, 2000.

[18] A. Shanware, H. Z. Massoud, E. Vogel, K. Henson, J. R. Hauser and J. J. Wortman, "Modeling the Trends in Valence–Band Electron Tunneling in NMOSFETs with Ultrathin SiO_2 and SiO_2/Ta_2O_5 Dielectrics with Oxide Scaling", *Microelectronic Engineering*, vol. 48, pp. 295-298, 1999.

[19] B. C. Lai, N. Kung, and J. Y. Lee, "A study on the capacitance–voltage characteristics of metal–Ta_2O_5–silicon capacitors for very large scale integration metal–oxide–semiconductor gate oxide applications", *Journal of Applied Physics*, vol. 85, pp. 4087–4090, 1999.

[20] T. Devoivre, F. Martin, C. Papadas and M. Setton, "Silicon Nitridation by Nitric Oxide (NO) for Ta_2O_5 Gate Dielectric Application in MOS Devices", *Solid-State Device Research Conference, 1999. Proceeding of the 29th European*, vol. 1, pp. 420–423, 1999.

[21] J. Robertson, "Band offsets of wide–band–gap oxides and implications for future electronic devices", *Journal of Vacuum Science & Technology B*, vol. 18, pp. 1785–1791, 2000.

[22] C. Guedj, E. Martinez, C. Licitra, G. Imbert, J. P. Barnes, D. Lafond, A. Toffoli, V. Arnal and L. Anaud, "Bulk and interface band diagrams of advanced intermetal dielectrics", *Electron Devices Meeting, IEDM 2007*, pp.977-980, 2007.

[23] B. Parvais, A. Mercha, N. Collaert, R. Rooyackers, I. Ferain, M. Jurczak, V. Subramanian, A. De Keersgieter, T. Chiarella, C. Kerner, L. Witters, S. Biesemans and T. Hoffman, "The Device Architecture Dilemma for CMOS Technologies: Opportunities & Challenges of FINFET Over Planar MOSFET", *Proceedings of VLSI-TSA Symposium*, pp. 80–81, 2009.

Phonon-limited hole mobility in sub-20 nm-thick double-gate germanium MOSFETs

Vedran Ivanić[1], Mirko Poljak[2,*], Tomislav Suligoj[2]
[1]Department of Physics, Faculty of Science, University of Zagreb, Croatia
[2]Department of Electronics, Microelectronics, Computer and Intelligent Systems,
Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia
*Corresponding author: mpoljak@zemris.fer.hr

***Abstract*—Physics-based modeling of hole mobilities in ultra-thin germanium layers is calculated with self-consistent Schödinger-Poisson solver and Kubo-Greenwood formula. Quantum confinement is taken into account for heavy, light and split-off hole band in a double-gate germanium MOSFET structure. Acoustic and optical phonon scattering is taken into consideration in the calculation of hole momentum relaxation time. The observed reduction of mobility in thinner layers is explained by examining the influence of field-induced and geometry-induced confinement. Contributions from different hole bands are investigated by calculating band population and respective band mobilities.**

I. Introduction

Metal-oxide-semiconductor field-effect transistor or MOSFET is the basic element in modern semiconductor electronics. In order to increase transistor performance new materials and device architectures are explored [1]. For example, ultrathin-body (UTB) semiconductor on insulator (SOI) MOSFET in combination with strain in crystal lattice in order to suppress short channel effects (SCE) and improve drive current [2], [3]. Among advanced multi gate structures, FinFET is the most promising device due to high scalability, improved immunity to SCEs and compatibility with CMOS [4], [5]. In order to meet ITRS requirements for future CMOS technology nodes [6], germanium is considered to be a promising material for next generation pMOS devices due to its high hole mobility [7], [8]. The UTB Ge MOSFETs are often coupled with III-V MOSFETs, e.g. InGaAs, in a dual-channel CMOS [9].

In this work, we study hole mobility behavior with the downscaling of body thickness in UTB double gate (DG) germanium MOSFETs using semiclassical numerical modeling. Physics-based modeling of the hole mobility is indispensable for the performance prediction of UTB Ge transistors. Since ultra-thin bodies are needed to suppress the SCEs in the extremely scaled MOSFETs [1], we investigate the effects of germanium layer thickness (T_{Ge}) on the mobility. Schrödinger's and Poisson's equation is solved self-consistently, which takes quantum confinement of holes into account. Scattering rates (on acoustic and optical phonons), form factor and hole distribution in the channel are calculated and used to find the mobility for different hole bands (heavy, light and split-off). Some of the mobility features are explained by examining the field-induced and geometry-induced confinement.

II. Theory and Numerical Modeling

A. Schrödinger's and Poisson's Equation

The DG germanium MOSFET is a multi-layered structure, with a potential energy profile that has a shape of a quantum well with the assumed barrier heights of 3 eV. As a consequence, holes in the inversion channel experience quantum confinement. Inside the quantum well holes have discrete energy levels with accompanying wavefunctions [10], [11]. The z-axis is perpendicular to the oxide-semiconductor interface, at which end the voltage bias is applied to create the inversion layer. Density of states is represented with 2D Fermi gas because holes are confined in xy-plane. An effective mass model is used with a simplification of spherical equipotential energy surface in the **K** space. Each hole band is described by three different masses: effective quantization or confinement mass (m_q) for calculating discrete energy levels, effective mass used for calculating density of states (DOS) (m_d) and effective conductivity mass (m_c) [12]. We assume $m_q = m_d = m_c$ for each hole band due to assumption of spherical bands, which, in turn, is made because of lack of effective mass data in the literature.

Schrödinger's equation in z-direction for DG MOS structure is

$$\left(-\frac{\hbar^2}{2m_{q,\nu}}\frac{\partial^2}{\partial z^2}+U(z)\right)\phi_{\nu,n}(z)=\varepsilon_{\nu,n}\phi_{\nu,n}(z). \quad (1)$$

First term describes kinetic energy of a hole, $U(z)$ potential energy of a hole. $\phi_{\nu,n}$ and $\varepsilon_{\nu,n}$ are resulting eigen energy levels and eigen wavefunctions. ν determines hole band and n determines state inside quantum well. Equation is solved numerically using finite differences approach on a uniform grid [13], i.e.

$$\frac{\partial^2\phi(z_k)}{\partial z^2}\approx\frac{\phi_{k+1}+\phi_{k-1}-2\phi_k}{a^2}. \quad (2)$$

The result from (1) and (2) are $\varepsilon_{\nu,n}$ and $\phi_{\nu,n}(z_k)$. $|\phi_{\nu,n}(z_k)|^2$ is probability of finding a hole in band ν and subband n at node z_k. Energy levels and wavefunctions are used for calculation of hole density per unit volume and inversion charge density per unit area which is perpendicular to inversion channel. Hole density per unit volume

is

$$p_{\nu,n}(z_k) = |\phi_{\nu,n}(z_k)|^2 \int_{-\infty}^{+\infty} g_{\nu,n}(E) f_{FD}(E) dE, \quad (3)$$

where $g_{\nu,n}$ is density of states for 2D Fermi gas, and $f_{FD}(E)$ is Fermi-Dirac distribution function. Total hole density per unit volume, $p(z_k)$, is calculated by summation over all available indices ν, n. Inversion density per unit area for is calculated from

$$P_{\nu,n}^{inv} = \int_0^{T_{Ge}} p_{\nu,n}(z) dz, \quad (4)$$

where T_{Ge} is thickness of Ge layer. For studying band occupancy for separate hole type

$$P_{\nu}^{\%} = \sum_n \frac{P_{\nu,n}^{inv}}{P_{inv}} \quad (5)$$

is used.

If we apply external voltage bias the $U(z)$ term in Schrödinger's equation becomes

$$U(z) = -eV(z) - \chi(z), \quad (6)$$

$V(z)$ describes external voltage and $\chi(z)$ is position-dependent carrier affinity which is determined by height of barriers in DG MOS structure. Spatial distribution of holes changes electrostatic potential $V(z)$ via Poisson's equation:

$$-\frac{d}{dz}\left[\varepsilon_r(z)\frac{V(z)}{dz}\right] = \frac{e}{\varepsilon_0}\left[p(z) + N_I(z)\right], \quad (7)$$

where $p(z)$ is volume charge density and $\varepsilon_r(z)$ is the spatially variable dielectric constant. $N_I(z)$ is volume density of ionized dopants which is here assumed to be zero as the Ge body is undoped. The Poisson's equation also contains a second derivation and it is solved similarly to (1), the result being the electrostatic potential $V(z)$. The iterative solving of the Schrödinger's and Poisson's equation is done numerically until convergence is achieved, i.e. until a desired error is obtained which in this work equals $\varepsilon_{error} = 10^{-3}$ [12], [10].

B. Scattering Rates in Inversion Layer

When passing through Ge, or any other semiconductor, holes are scattered by crystal lattice vibrations (phonons), impurities or defects in lattice [14]. Using time dependent perturbation theory, transition from initial to final state is calculated with Fermi's golden rule [11]. Scattering rates for various scattering mechanisms are calculated [15] by linearizing the scattering term in the Boltzmann's transport equation. In this paper, only intra-band scattering for holes is considered, nevertheless, both intra- and inter-subband transitions are taken into account. The band designates hole type (HH, LH, SO), while the subband designates the eigenstate inside the quantum well. Important quantity for scattering is wavefunction form factor (FF)

$$F_{\nu;i,j} = \int_0^{T_{Ge}} |\phi_{\nu,i}(z)|^2 |\phi_{\nu,j}(z)|^2 dz \quad (8)$$

TABLE I
PARAMETERS USED FOR MOBILITY CALCULATIONS

m_{HH}	$0.33m_0$
m_{LH}	$0.043m_0$
m_{SO}	$0.084m_0$
D_{AP}	11 [eV]
D_{OP}	$6 \cdot 10^8$ [eV/cm]
E_{OP}	38 [meV]
v_s	$5 \cdot 10^5$ [cm/s]
ρ	5.32 [g/cm^3]

which describes overlap of probability functions of initial state i and final state j for band ν. In this work, the mobility simulation is based on calculating momentum relaxation time (MRT). The MRT for acoustic phonons (AP) is calculated from:

$$\frac{1}{\tau_{\nu,i}^{AP}(E)} = \sum_j \frac{n_\nu m_{d,\nu} k_B T D_{AP,\nu}^2}{\hbar^3 \rho v_s^2} \times F_{\nu;i,j} H(E - \varepsilon_{\nu,j}), \quad (9)$$

[15], [16]. Same index ν on both sides shows that this is intra-band scattering, i and j designates initial and final state inside quantum well, n_ν is band multiplicity for band ν, $m_{d,\nu}$ is effective density of states mass for band ν, k_B is Boltzmann constant, T is temperature, $D_{AP,\nu}$ is the acoustic deformation potential for band ν, $\hbar$ is reduced Planck constant, ρ is density of Ge, v_s is speed of sound in Ge and H is Heaviside step function.

Similar expression is given for MRT for optical phonons (OP):

$$\frac{1}{\tau_{\nu,i}^{OP}(E)} = \sum_j \frac{n_\nu m_{d,\nu} D_{OP}^2}{2\hbar\rho E_{OP}} \times \left(f_{BE}(E_{OP}) + \frac{1}{2} \pm \frac{1}{2}\right) \frac{1 - f_{FD}(E \pm E_{OP})}{1 - f_{FD}(E)} \times F_{\nu;i,j} H(E - (\varepsilon_{\nu,j} \mp E_{OP})), \quad (10)$$

but in this case scattering is inelastic [15], [16]. D_{OP} is optical phonon deformation strength, $E_{OP}(E)$ is energy of optical phonon. Numbers of states of phonons, which are bosons, is given with Bose–Einstein distribution f_{BE}. Holes, which are fermions, are represented with Fermi-Dirac distribution f_{FD}. Sign $\pm$ represent two possibilities of absorption or emission of optical phonon. Spin multiplicity factor, n_{sp}=1, is omitted because holes are scattered on phonons. Holes which are fermions and have half-integer spin do not change their spin while interacting with phonons which are bosons and have integer spin. Parameters need for calculation are taken from [17], [3] and are listed in Table I.

There can be many different sources of scattering present at the same time: AP, OP, polar OP, surface roughness, Coulomb scattering, alloy disorder. In this preliminary work, only AP and OP scattering is taken into account in order to assess the properties of the intrinsic mobility in ultra-thin DG UTB Ge MOSFETs. For comparison with experiments and other work more scattering mechanisms should be taken in consideration and this will be done in future work.

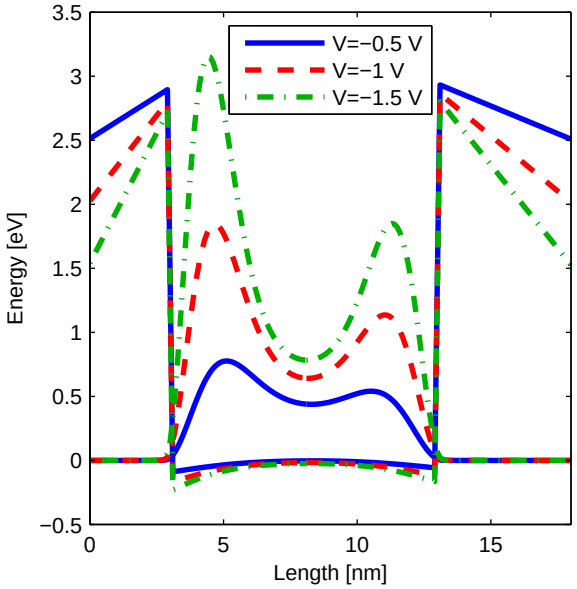


Fig. 1. Energy diagram and hole density per unit volume for UTB GOI MOSFET with gate voltage as parameter. Straight lines shows modified potential well. Curved lines shows hole density per unit volume. Greater localization of holes can be seen near thinner oxide for greater voltages.

Generally, combining MRT for more scattering mechanisms when they are mutually independent is done by Matthiessen's rule:

$$\frac{1}{\tau_{\nu,n}(E)} = \sum_{mech} \frac{1}{\tau^{mech}_{\nu,n}(E)}. \tag{11}$$

From total MRT mobility for n-th subband in band ν is calculated with Kubo-Greenwood formula [15]

$$\mu_{\nu,n} = \frac{e}{m_{c,\nu}} \frac{\int_{\varepsilon_{\nu,n}}^{+\infty} (E-\varepsilon_{\nu,n})\tau_{\nu,n}(E) \left|\frac{\partial f_{FD}(E)}{\partial E}\right| dE}{\int_{\varepsilon_{\nu,n}}^{+\infty} (E-\varepsilon_{\nu,n}) \left|\frac{\partial f_{FD}(E)}{\partial E}\right| dE}. \tag{12}$$

Total mobility is calculated from combining subband and band mobilities weighted by subband and band hole inversion densities in total hole inversion densities:

$$\mu = \sum_{\nu,n} \frac{P^{inv}_{\nu,n}}{P_{inv}} \mu_{\nu,n}. \tag{13}$$

III. Results and Discussions

A. Self-Consistent Schrödinger-Poisson Solver and Momentum Relaxation Time

Results of self-consistent Schrödinger-Poisson solver are shown on Fig. 1. Main results are eigen energies and eigen wavefunctions. Thickness of oxide layer is deliberately chosen to be asymmetrical, 3 and 5 nm, to illustrate the consequences of quantum confinement. From Fig. 1 it is seen that there is greater concentration of holes in inversion channel at side with thinner oxide layer due to stronger field-induced inversion. Greater concentration of holes is achieved for greater voltages. For three voltages deformation of rectangle quantum well is shown with greater deformation for greater voltages. Small penetration of wavefunction inside oxide layer is due to finite barrier height.

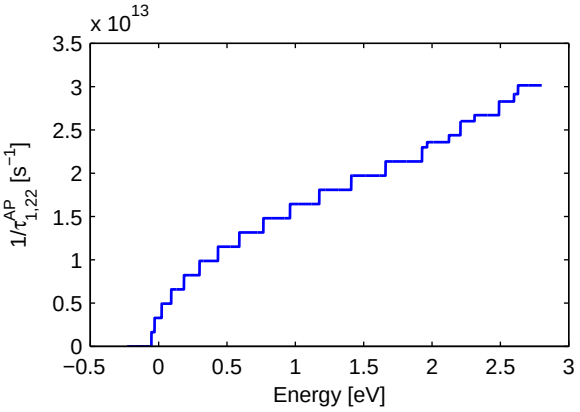


Fig. 2. Inverse MRT for AP. Indices (1,22) labels heavy hole, 22 state in subband of quantum well. Dimensions of oxide-semiconductor-oxide are 3, 10, 5 nm respectably and voltage bias on DG is -1V.

Eigen energies and eigen wavefunction from self-consistent Schrödinger-Poisson solver are used for calculation of MRT according to (9) and (10). The MRT for the twenty second state in the HH band is shown on Fig. 2, each step is determined by step function from (9), which stems from DOS of 2D Fermi gas. Heavier holes will have shorter MRT and increase of FF will also lead to shorter MRT. Greater voltages accelerates holes to greater speeds and consequently to greater energies which leads to shorter MRT. Therefore, holes under greater voltages will scatter more often than holes under smaller bias.

B. Effects of Body-Thickness Downscaling

For the following results, the thickness of oxide layers is constant (3 nm), while the thickness of the germanium layer is varied. Instead of plotting various data vs. voltage bias, data can be plotted vs. inversion charge, P_{inv}, which is more suitable for comparison with different materials and structures. In Fig. 3 dependence of hole mobility on inversion charge is shown for several body thicknesses that range from 2 to 20 nm, and it is clear that the mobility decreases with the downscaling of T_{Ge}. Thicker layers, $T_{Ge} > 10$ nm, exhibit inversion modulation of the mobility that is induced by external field. For $T_{Ge} = 20$ nm mobility drops by 38% from $P_{inv} = 3 \times 10^{12}$ cm^{-2} to $P_{inv} = 10^{13}$ cm^{-2}. Mobility in thinner layers, $T_{Ge} < 10$ nm, is constant and determined only by T_{Ge}, i.e. geometry of Ge layer. Two important effects have to be considered, field-induced confinement and geometry-induced confinement.

If we examine Fig. 1, greater spatial confinement dominates for higher voltages, and consequently higher P_{inv}. For explanation, more detailed discussion about the wavefunction form factor is needed. Higher voltage on gates, which also induces greater P_{inv}, confides holes in smaller space near the oxide layer and their wavefunctions will overlap more leading to higher FF. Equation (8) shows how much probabilities of different hole states overlap.

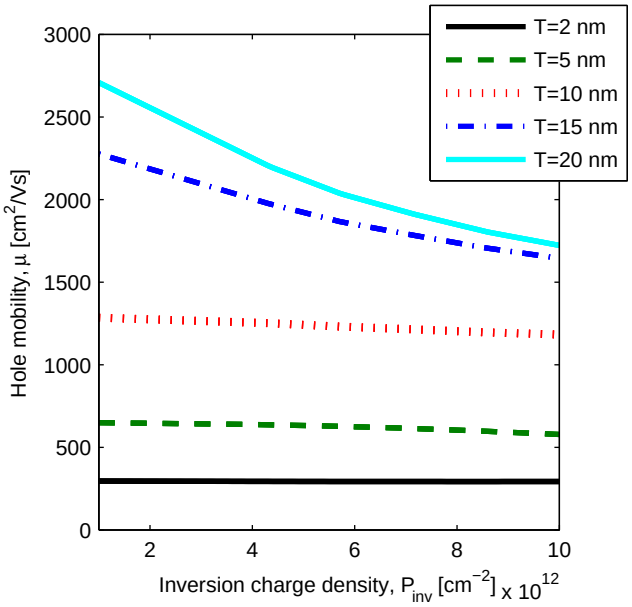


Fig. 3. Inversion charge density, P_{inv}, vs. hole mobilities, μ, with body-thickness, T_{Ge}, ranging from 2 to 20 nm as parameter.

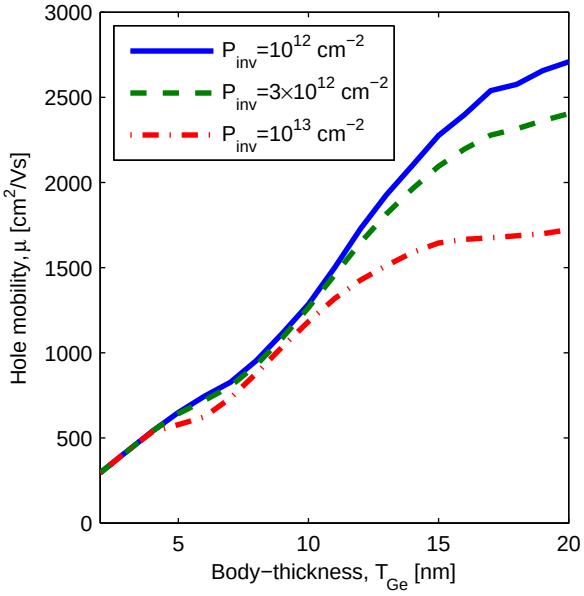


Fig. 5. Body-thickness vs. hole mobility with inversion charge as parameter.

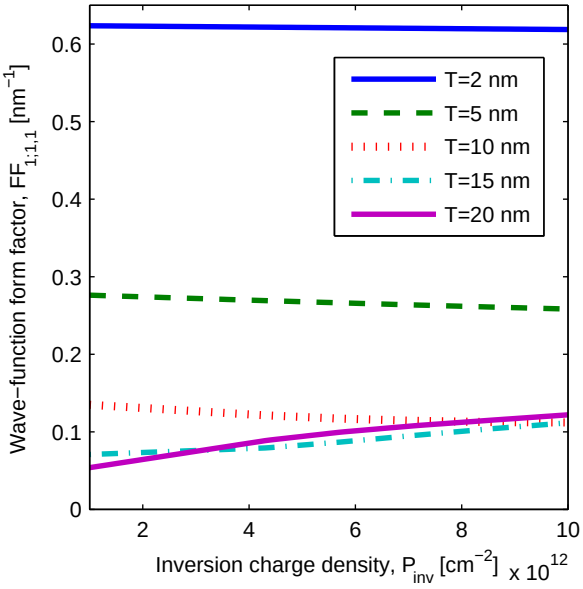


Fig. 4. Inversion charge density, P_{inv}, vs. wavefunction form factor, $FF_{1;1,1}$, with body-thickness, T_{Ge}, ranging from 2 to 20 nm as parameter. Indices (1;1,1) labels heavy hole band scattering from first energy subband to first energy subband.

Correlation between P_{inv} and $FF_{1;1,1}$ is shown on the Fig. 4. Scaling down T_{Ge} rises FF. External field induces inversion charge modulation of FF in thicker layers. For $T_{Ge} = 20$ nm FF drops by 39% from $P_{inv} = 10^{13}$ cm^{-2} to $P_{inv} = 3 \times 10^{12}$ cm^{-2}. Also for large inversion density around $P_{inv} = 10^{13}$ cm^{-2} FF is almost the same for T_{Ge} ranging from 10 to 20 nm. This is evidence of field-induced confinement in thicker layers, during which localization of holes near oxide-semiconductor interface increases with voltage bias. For thinner layers FF is constant for different P_{inv} and it is evidence of geometry-induced confinement. Higher FF in (9) and (10) lowers MRT in (12) and μ. In the same way lower FF raises μ.

Effect of body-thickness on hole mobility is shown on Fig. 5. It shows that scaling down T_{Ge} from 20 to 2 nm lowers mobilities by 88% for $P_{inv} = 3 \times 10^{12}$ cm^{-2} and 83% for $P_{inv} = 10^{13}$ cm^{-2}. Geometry-induced confinement is again seen for thinner layers because there is no modulation with P_{inv}. Field-induced confinement becomes dominant for layers above 12 nm. At $P_{inv} = 10^{13}$ cm^{-2} mobility starts falling more rapidly for $T_{Ge} < 12$ nm. From 12 nm to 4 nm mobility drops from 1400 cm^2/Vs to 500 cm^2/Vs. It is interesting to compare this results with other works for hole mobilities in silicon. Donetti L., et al. [18], also taking into account only AP and OP reports hole mobilities at $P_{inv} = 3 \times 10^{12}$ cm^{-2}: $\mu_\nu = 200$ cm^2/Vs for $T_{Si} = 10$ nm, $\mu_\nu = 150$ cm^2/Vs for $T_{Si} = 5$ nm, $\mu_\nu = 120$ cm^2/Vs for $T_{Si} = 3$ nm. If we compare it with our results from Fig. 3 and Fig. 5 reported hole mobilities for silicon are 70 to 80 % smaller than for our results for germanium. This has to be taken with caution because our model is uncalibrated and results in [18] are from calibrated simulation.

More details for smaller mobility in thinner layers are found inspecting FF, but this time its dependence on Ge layer thickness shown on Fig. 6. FF increases when scaling down T_{Ge} from 20 to 2 nm. Drop of 88% for $P_{inv} = 3 \times 10^{12}$ cm^{-2} and 81% for $P_{inv} = 10^{13}$ cm^{-2} is obtained. Similarity with figures for mobilities when downscaling T_{Ge} shows that rise in FF is one of dominating effect of lowering mobility for thinner layers. Geometry induced confinement is dominant for thinner layers because wavefunction will be forced to overlap more. From Fig. 1 we can picture what happens

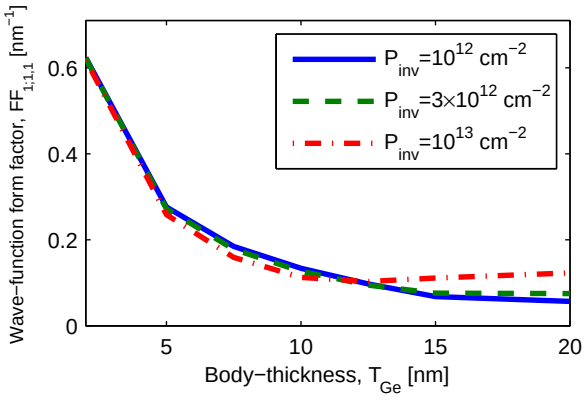


Fig. 6. Body-thickness vs. wavefunction form factor, $FF_{1;1,1}$ with inversion charge as parameter. Indices (1;1,1) labels HH band scattering from first energy subband to first energy subband.

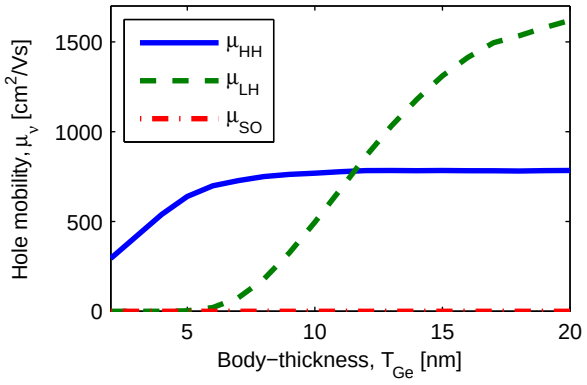


Fig. 7. Effect of body thickness on different hole mobility for $P_{inv} = 3 \times 10^{12}$ cm^{-2}

to wavefunctions with downscaling of T_{Ge}. Maximum of wavefunction will be shifted closer to center and they will overlap more with wavefunction of holes confined at other oxide layer. For thicker layers, wavefunctions are shifted apart and overlap is smaller. At $T_{Ge} = 20$ nm FF is the largest for $P_{inv} = 10^{13}$ cm^{-2} because higher voltages forces wavefunction to smaller space near the oxide-semiconductor interface and will overlap more. As said before, larger FF leads to smaller μ so this explains smaller mobility in thicker layers for high inversion in Fig. 3 and Fig. 5.

Contributions of different hole bands to the total mobility is reported in Fig. 7 for T_{Ge} ranging from 2 to 20 nm at $P_{inv} = 3 \times 10^{12}$ cm^{-2}. Picture is qualitative the same for different inversion densities which means that this is geometry related effect. For thicker layers mobility of light holes is greater than mobility of heavy holes. At $T_{Ge} = 20$ nm μ_{HH} is 52% lower than $\mu_{LH} = 1600$ cm^2/Vs. Around $T_{Ge} = 12$ nm mobilities are the same for both bands. Light holes exhibit a steep descend in mobility from $\mu_{LH} = 1400$ cm^2/Vs for 16 nm to $\mu_{LH} = 20$ cm^2/Vs for 6 nm. We can conclude that μ_{LH} has strong geometry dependence. Mobility of heavy holes is almost constant for thickness above 6 nm, while the geometry dependence

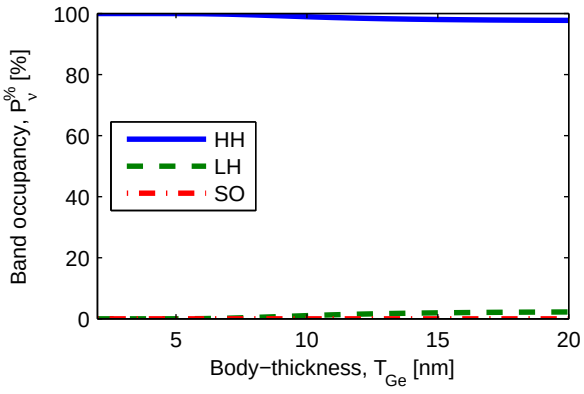


Fig. 8. Band occupancy of heavy, light and split-off holes for various body thickness at $P_{inv} = 3 \times 10^{12}$ cm^{-2}.

becomes noticeable below $T_{Ge} = 6$ nm. Mobility of heavy holes drops by 58% for $T_{Ge} = 6$ nm from 700 cm^2/Vs to 300 cm^2/Vs for $T_{Ge} = 2$ nm. Mobility of split-off holes is $\mu_{SO} = 0.002$ cm^2/Vs at best and can be assumed to be zero for all practical purpose. To explain greatly varying μ_{LH}, separate P_{inv} for every hole has to be studied, as different hole mobilities are weighted by their P_{inv} in total mobility as in (5).

Band occupancy for separate hole bands is shown in Fig. 8. For every $T_{Ge} = 2 - 20$ nm population of heavy holes dominates, with $P^{\%}_{HH}$ over 99.99% for $T_{Ge} = 2$ nm and slowly declining to $P^{\%}_{HH} = 97.75\%$ for $T_{Ge} = 20$ nm. Population of LH is $P^{\%}_{LH} \approx 0\%$ for $T_{Ge} = 2$ nm. If we look at Fig. 7 again, interesting thickness are $T_{Ge} = 6$ nm with $P^{\%}_{LH} = 0.07\%$ where μ_{LH} begins to contribute to total μ, $T_{Ge} = 12$ nm with $P^{\%}_{LH} = 1.54\%$ where $\mu_{HH} = \mu_{LH}$ and maximal thickness $T_{Ge} = 20$ nm with $P^{\%}_{LH} = 2.25\%$. Question arises how can LH with such small populations contribute so much to total mobility. We see in (12) and (9) or (10) that

$$\mu_\nu = \frac{1}{m_{c,\nu}} < \tau > \sim \frac{1}{m_{c,\nu}} \frac{1}{m_{d,\nu}}. \qquad (14)$$

Great difference in masses between LH and HH, $m^2_{HH}/m^2_{LH} \approx 60$ is responsible for μ_{LH} of the same order as μ_{HH} even for small population as $P^{\%}_{LH} \approx 1-2\%$.

IV. Conclusion

In this work mobility of holes in UTB DG MOSFET was assessed numerically. Self-consistent Schrödinger-Poisson solver was used for physics-based modeling of energy levels and distribution of holes inside UTB GOI MOSFET. Heavy, light and split-off hole band was taken in consideration with effect of quantum confinement in Ge layer between two oxides. Simple effective mass model is used were effective confinement, density of state and conductivity masses are the same. Mobility of holes is calculated from scattering rates of holes on AP and OP for Ge layers from 2 to 20 nm. It is clearly shown that mobility of holes decreases for thinner layers of Ge. When scaling down germanium layer from 20 to 2 nm mobility drops by 88% for lower inversion ($P_{inv} = 3 \times 10^2$ cm^{-2}) and 83%

for higher inversion ($P_{inv} = 10^3$ cm^{-2}). For layers below 12 nm decrease in mobility is independent of voltage on gate. Main limiting factor for hole mobility is geometry-induced confinement. Thickness of Ge changes overlap of hole wavefunctions shown through form-factor. In layers above 12 nm field-induced confinement is dominant and voltage bias is main modulating quantity. Greater mobility of light holes than of heavy holes for low light hole band occupancy, $P_{LH}^{\%} \approx 1-2\%$, is explained by great difference of hole masses $m_{HH}^2/m_{LH}^2 \approx 60$. When compared with silicon, germanium is more promising when downscaling because of greater hole mobilities. Only acoustic and optical phonon limited results show 70-80% lower hole mobilities for silicon.

REFERENCES

[1] Skotnicki T., Fenouillet-Beranger C., Gallon C., Boeuf F., Monfray S., Payet F., et al., "Innovative materials, devices, and CMOS technologies for low-power mobile multimedia," IEEE Trans. Electron Devices, vol. 55, no. 1, pp. 96-130, Jan .2008.

[2] Park J.-T., Colinge J.-P., "Multiple-gate SOI MOSFETs: device design guidelines," IEEE Trans. Electron Devices, vol 49, no. 12, pp. 2222-2229, Dec. 2002.

[3] A. Pham, C. Jungemann, B. Meinerzhagen, "Physics-based modeling of hole inversion-layer mobility in strained-SiGe-on-insulator," IEEE Trans. Electron Devices, vol. 54, no. 9, pp. 2174–2182, Sep. 2007.

[4] Hisamoto D., Lee W.-C., Kedzierski J., Takeuchi H., Asano K., Kuo C., et al., "FinFET - a self-aligned double-gate MOSFET scalable to 20 nm," IEEE Trans. Electron Devices, vol. 47, no. 12, pp. 2320-2325, Dec. 2000.

[5] Poljak M., Jovanović V., Suligoj T., "Improving bulk FinFET DC performance in comparison to SOI FinFET," Microelectron. Eng., vol. 86, no. 10, pp. 2078–2085, Oct. 2009.

[6] *International Technology Roadmap for Semiconductors, 2012. Update Overview*, (2012.), "www.itrs.net", [accessed 31.01.2014.]

[7] Takagi S., Iisawa T., Tezuka T., Numata T., Nakaharai S., Hirashita N., et al. "Carrier-transport-enhanced channel CMOS for improved power consumption and performance," IEEE Trans. Electron Devices, vol. 55, no. 1, pp. 21-39, Jan. 2008.

[8] C. Claeys, E. Simoen, *Germanium-based technologies: From Materials to Devices*. Elsevier, UK, 1. edition, 2007.

[9] Takagi S. and Takenaka M., "III-V/Ge CMOS technologies on Si platform," in VLSI Symp., 2010, pp. 147-148.

[10] S. Datta, *Quantum Transport: Atom to Transistor*. Cambridge University Press, UK, 1. edition, 2005.

[11] M. Lundstrom, *Fundamentals of carrier transport*. Cambridge University Press, UK, 2. edition, 2000.

[12] Poljak M., "Carrier transport in low-dimensional nanoelectronic devices," Ph.D. dissertation, Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia, 2013.

[13] Aleksandar Maksimović, *Numeričke metode i praktikum*, (2006.), Institut Ruđera Boškovića, Zagreb, "http://www.irb.hr/users/maks/numericke.html", [accessed 31.01.2014.]

[14] C. Kittel, *Introduction to Solid State Physics*. John Wiley & Sons, USA 8. edition, 2005.

[15] D. Esseni, P. Palestri, L. Selmi, *Nanoscale MOS Transistors*. Cambridge University Press, UK, 1. edition, 2011.

[16] M. Poljak, V. Jovanović, D. Grgec, T. Suligoj, "Assessment of electron mobility in ultrathin-body InGaAs-on-insulator MOSFETs using physics-based modeling," IEEE Trans. Electron Devices, vol. 59, no. 16, pp. 1636–1643, Jun. 2012.

[17] *New Semiconductor Materials. Characteristics and Properties*, (2008.), Ioffe Physical-Technical Institute, St. Petersburg, RF, "http://www.ioffe.rssi.ru/SVA/NSM/", [accessed 31.01.2014.]

[18] Donetti L., Gamiz F., Rodriguez N., Godoy A., "Hole mobility in ultrathin double-gate SOI devices: the effect of acoustic phonon confinement," IEEE Electron Device Letters, Vol. 30, No. 12, Dec. 2009.

Comparison of RF performance between 20 nm-gate bulk and SOI FinFET

S. Krivec, H. Prgić, M. Poljak and T. Suligoj*

Department of Electronics, Microelectronics, Computer and Intelligent Systems
Faculty of Electrical Engineering and Computing, University of Zagreb, HR-10000, Zagreb, Croatia
*Corresponding author: tomislav.suligoj@fer.hr

Abstract – **Due to their 3D architecture, FinFETs exhibit a strong dependence of RF figures-of-merit on device geometry due to large extrinsic resistances and capacitances. We analyze the RF performance of 20-nm gate length FinFETs implemented on either bulk or SOI substrate by using 3D numerical device simulation. SOI and bulk FinFET are compared in terms of cut-off and maximum oscillation frequency, gate capacitance and trans-conductance. We have proposed and investigated the influence of different doping methods that aim at improving the RF figures-of-merit of 20 nm bulk FinFETs while maintaining acceptable DC performance.**

I. Introduction

The successful scaling of the CMOS technology for digital/logic applications has also demonstrated its potential for high frequency applications [1]. Namely, smaller channel length intrinsically results in higher drive currents and higher level of integration. Regardless of these advantages, the continuous scaling results in a group of problems known as short-channel-effects (SCEs) which are caused by the decreased control of gate over the channel. The SCEs involve subthreshold swing (SS) degradation, increase of the drain-induced barrier lowering (DIBL), device variability and increased leakage currents. In order to suppress the SCEs, new device architectures have been introduced. Multiple-gate MOSFETs have shown excellent down-scaling characteristics and high-performance when compared to the conventional single-gate device structures, due to better gate controllability of the channel [2].

The most feasible implementation of the multiple-gate architecture is the FinFET structure [3], [4], [5] which has been implemented in microprocessors of the latest generation. FinFETs are classified into two types, bulk FinFET and SOI FinFET [6]. Structurally, bulk FinFETs and SOI FinFETs are similar except that in SOI FinFETs the gate electrode and fin are isolated from the bulk silicon substrate. Although SOI devices have shown excellent short channel characteristics, they exhibit certain disadvantages over bulk devices such as higher wafer cost, heat transfer problems and may suffer from the floating body problem. Bulk FinFET structure draws attention mainly due its ability to be integrated with the conventional bulk planar CMOS technology. Although FinFETs alleviate electrostatic limitations in further scaling, their 3D structure plays a significant role in forming a large parasitics thus degrading radio frequency (RF) performance [7], [8], [9]. Therefore, it is necessary to optimize FinFET device geometry, materials and doping profiles for each technology generation in order to meet DC and RF performance metrics as given by the ITRS [10].

In this paper, we analyze the RF performance of 20 nm gate length FinFETs implemented on either bulk or SOI substrate. Analysis is performed on structures with identical geometrical dimensions, in order to assess the suitability and compare bulk and SOI FinFET technologies. Devices are compared both in terms of their DC performance and RF figures of merit.

II. Numerical Simulations

Bulk FinFET and SOI FinFET structures are analyzed and compared by using 3D numerical device simulator [11]. Figure 1(a) and (b) show 3D illustrations of both structures while the cross-sectional views are shown in Fig. 1(c). We assumed *p*-type body doping of 10^{15} cm^{-3}, source and drain doping of 10^{20} cm^{-3}, gate oxide thickness of 2 nm and top-gate nitride thickness of 10 nm. The top gate nitride layer cancels the influence of the gate electrode on the top of the fin which makes the devices effectively double-gate. Source and drain junctions are extended 2 nm below the gate. Fin height is set to 30 nm and gate length is set to 20 nm, while the fin width is set to 8 nm. In order to set the threshold voltage to positive values of around 0.2 V we used metal gate with a workfunction of 4.5 eV. In case of SOI FinFET, the thickness of the buried oxide (BOX) layer is set to 50 nm.

Electrostatic analysis is performed with Sentaurus SDevice [11]. The transport model used in simulations is based on the drift-diffusion formalism with band to band tunneling included in the Shockley-Read-Hall recombination. We used the concentration-dependent mobility model which includes the effect of impurity scattering and the mobility model which includes degradation of mobility due high electric field. Since the structures analyzed in this work have sub-20 nm dimensions in the active area, one can expect a strong influence of quantum effects such as carrier energy quantization or mobility dependence on fin width [12]. These effects are ignored in this study since we focus on qualitative comparison of the performance of SOI and bulk FinFET structures. Nevertheless, we acknowledge

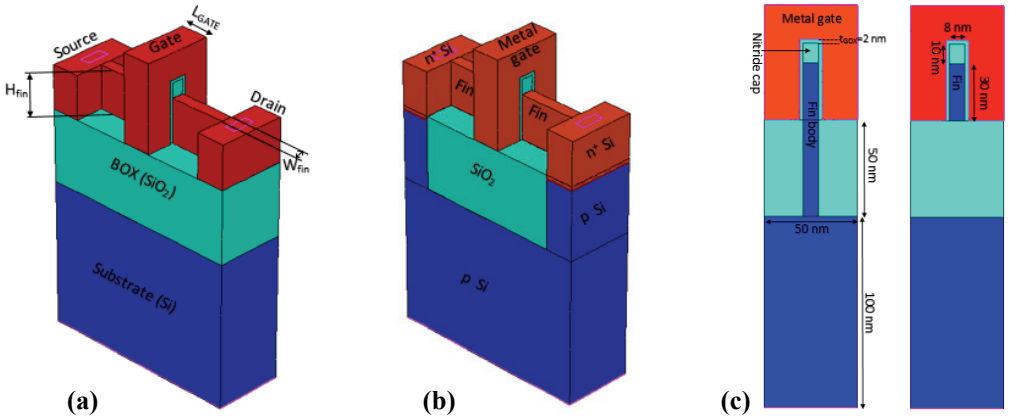


Fig. 1. 3D illustration of (a) SOI FinFET and (b) bulk FinFET . (c) 2D cross-section along transistor gate of bulk and SOI FinFET

the importance of the aforementioned effects and they will be accounted for in future work.

III. Results And Discussion

A. DC Analysis

Transfer functions for SOI and undoped bulk FinFET, obtained for V_{DS} of 0.05 V and 0.9 V are shown in Fig. 2. We use the attribute "undoped" to emphasize the initial bulk FinFET structure that has uniform doping in the whole fin (active part of the fin and the body-tied region). In this case, SOI FinFET has DIBL of 139 mV/V and SS of 91 mV/dec, whereas the undoped bulk FinFET has DIBL of 211 mV/V and SS of 186 mV/dec. The higher SS and DIBL of undoped bulk FinFET can be attributed to higher dielectric constant of depleted silicon compared to silicon oxide, which guides the drain electric field towards the source under gate. Figure 2 shows that the leakage current in OFF state (at V_{DS}= 0.9 V, V_{GS}= 0 V) is greater in the undoped bulk FinFET in comparison with SOI FinFET. This can be explained by examining the 2D cross-sectional current density plot reported in Figure 3, where we clearly see that current flow exists in the body-tied region, i.e. in the lower part of the fin that is out of reach of the gate electrode. Due to decreased control of the fin, both the DIBL and SS increase for the undoped bulk FinFET.

In order to make the undoped bulk FinFET comparable with SOI FinFET in terms of DC performance, the leakage current in bulk FinFET must be suppressed, which will, in turn, improve its sub-threshold performance. One approach to suppress leakage current in the body-tied (inactive part of the fin) region is to increase the body doping in that region. This reduces the punch-through effect because drain and source depletion regions become smaller, which prevents the generation of the parasitic current path [13]. This type of doping is also called Punch Trough Stopper (PTS) [14]. The doping profile in the body-tied region is a Gaussian function, with the maximum concentration of 5.2×10^{18} cm^{-3} set at 8 nm depth below the bottom of the gate where maximum

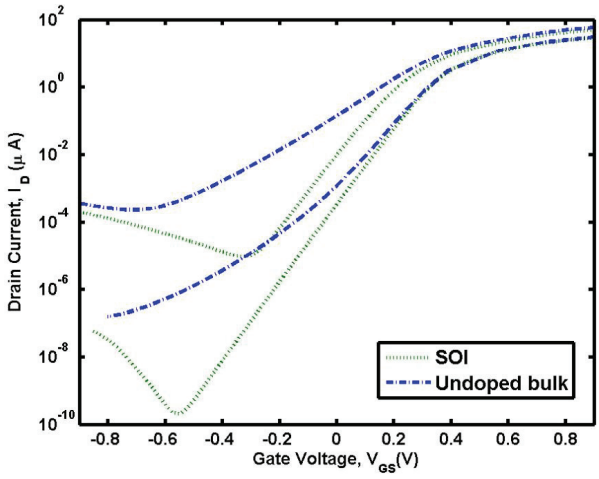


Fig. 2. Transfer characteristics for SOI and undoped bulk FinFET (at V_{DS}=0.9 V and V_{DS}=0.05 V). Increased leakage deteriorates subthreshold performance of bulk FinFET.

current leakage was determined. From that point, the PTS region spreads 20 nm toward substrate and also 20 nm toward upper fin. Two PTS-doping approaches are examined in this work; in plane-doped bulk FinFETs the PTS extends under the active fin and source/drain regions, which can be implanted before the gate fabrication whereas in the fin-doped bulk FinFET the PTS is implanted only under the active part of the fin, which can be implemented in gate-last process flow.

The results of DC analysis for all the examined structures are listed in Table I, and Fig. 4 shows the transfer characteristics for all the examined structures. With the PTS region, the bulk FinFET exhibits a subthreshold slope of 94 mV/dec, DIBL of 132 mV/V and OFF state current is 8 nA which is approximately 20 times smaller than the OFF state current of the undoped bulk FinFET. This improvement is partially attributed to increase in threshold voltage for doped bulk structures. Also, the DIBL has improved by 38 % and SS by 50 % in comparison with undoped bulk FinFET.

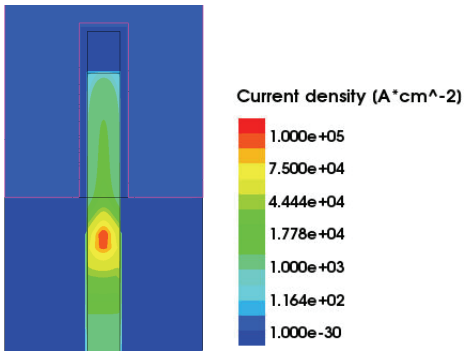


Fig. 3. 2D cross-sectional current density plots in OFF state for undoped bulk FinFET (at V_{DS}= 0.9 V, V_{GS}= 0 V). Current flow exists in the body-tied region, i.e. in the lower part of the fin that is out of reach of the gate electrode.

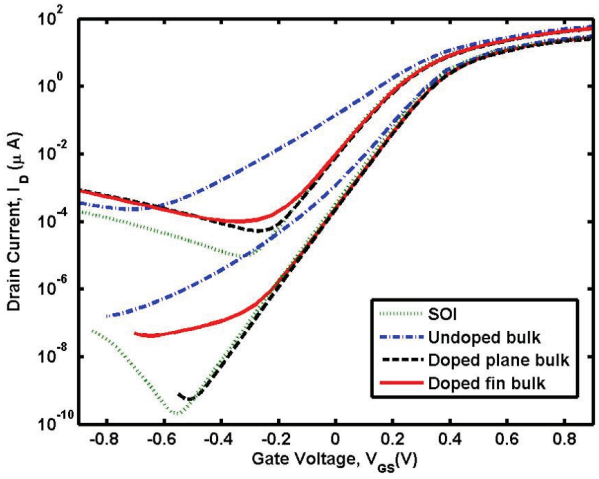


Fig. 4. Transfer characteristics for all examined structures (at V_{DS} = 0.9 V and V_{DS} = 0.05 V). Better subthreshold performance for doped bulk structures is observed.

It is shown that doped bulk structures have better subthreshold performance and lower leakage current than the initial undoped bulk FinFET. Doped-plane and doped-fin bulk FinFET, have approximately 4 % and 10 % higher SS than SOI FinFET, respectively. As expected, DIBL is the highest for undoped bulk, while in both cases of bulk FinFET with PTS, the achieved DIBL is lowered by about 40 % in regard to initial bulk FinFET. In the end that is 6 % lower than SOI for doped plane bulk and 3 % higher than SOI for doped fin bulk FinFET. For all examined structures, the ON state currents (at V_{DS} = 0.9 V, V_{GS} = 0.9 V) are around 50 µA for the fin height of 30 nm. The ON state current per channel width is approximately between 0.8 mA/µm and 1 mA/µm, where channel width is $W = 2 \times H_{fin}$, which is still significantly above the requirements for High Performance Logic Technology stated in ITRS [10]. Therefore, both approaches for suppressing leakage in bulk FinFETs give satisfying results. Slightly better results, especially in terms of current in OFF state and DIBL are achieved in the case of doped plane bulk FinFET. It is shown that doped plane bulk and doped fin bulk structures have about 3 and 2 times smaller OFF current than SOI FinFET, respectively. Since PTS in both bulk doped structures slightly penetrates in upper fin making the fin more doped compared to undoped bulk or SOI FinFET, both bulk doped FinFETs have higher linear subthreshold voltage ($V_{TH,LIN}$), about 65 % higher than undoped bulk and 15 % higher than SOI FinFET.

Figure 5 shows comparison of output characteristics for all structures obtained for V_{GS} = 0.9 V. Output characteristics of undoped bulk FinFET show higher drain currents than in the case of SOI FinFET. For

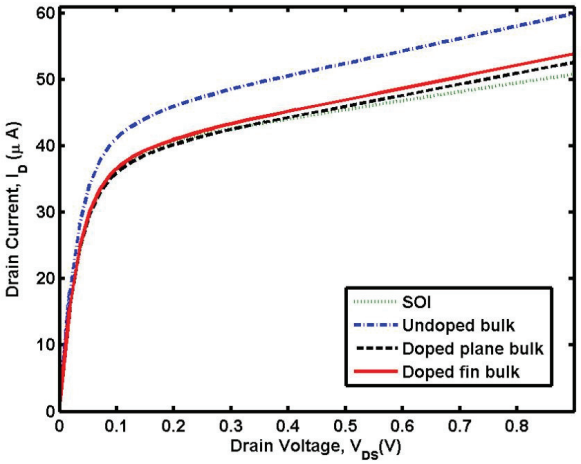


Fig. 5. Output characteristics for all examined structures (at V_{GS}= 0.9 V). Drain current and output conductance are higher for bulk FinFET structures.

V_{GS} = 0.9 V the channel is effectively wider than in the case of SOI FinFET due the inversion of the fin below the gate. Additionally, higher drain currents can be attributed to increased overdrive voltage due to lower threshold voltage. For that reason also, drain current and output conductance are higher for doped bulk FinFET structures. For the doped bulk structures the output conductance is approx. 16 µS, while for the undoped bulk it equals 19 µS and for SOI FinFET equals 13 µS.

Table I. Results of electrostatic analysis for all examined structures at V_{DS} = 0.9 V

Structure	SOI FinFET	Undoped Bulk FinFET	Doped Plane Bulk FinFET	Doped Fin Bulk FinFET
$V_{TH,LIN}$ [V]	0.191	0.135	0.226	0.219
I_{ON} [µA]	48.0	59.8	51.4	52.3
I_{OFF} [nA]	24.4	184.9	8.0	10.4
SS [mV/dec]	91	186	94	100
DIBL [mV/V]	139	211	132	143

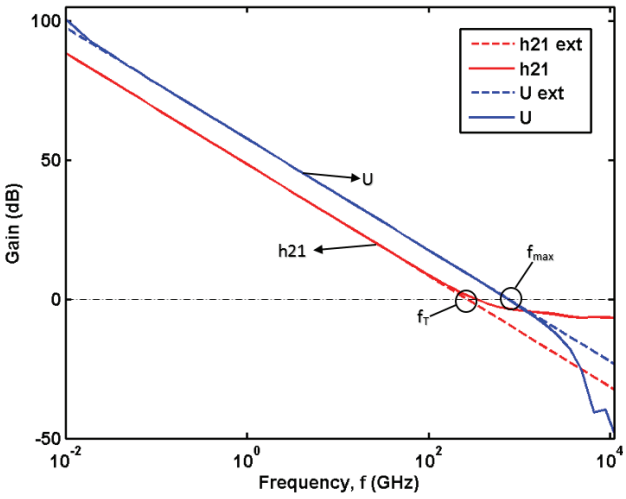


Fig. 6. Unilateral gain (U) and current gain (h_{21}) for SOI FinFET (V_{GS}= V_{DS}= 0.9 V). The f_{max} and f_T are extrapolated as frequencies where U and h_{21}, respectively, are equal 1 assuming a 20 dB/dec slope.

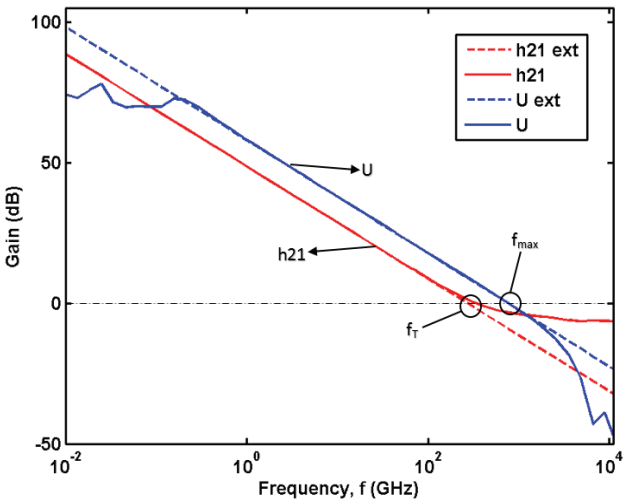


Fig. 7. Unilateral gain (U) and current gain (h_{21}) for plane-doped bulk FinFET (V_{GS}= V_{DS}= 0.9 V).

Table II. The cut-off and maximum oscillation frequency at V_{GS}= V_{DS}= 0.9 V for all examined structures

Structure	SOI FinFET	Undoped Bulk FinFET	Plane-doped Bulk FinFET	Fin-doped Bulk FinFET
f_T [GHz]	269.8	276.1	269.3	272.5
f_{max} [GHz]	776.5	713	744.4	760

B. RF Performance Analysis

In this section, the previously described SOI and bulk FinFETs are compared in terms of the cut-off (f_T) and maximum oscillation frequency (f_{max}). These two frequencies present the most relevant figures of merit for high frequency capabilities of MOS transistor [15]. The cut-off frequency can be determined by finding the frequency where the current gain equals 1, i.e. 0 dB. The current gain is defined by admittance parameters:

$$|h_{21}| = \frac{|y_{21}|}{|y_{11}|}, \qquad (1)$$

where y_{21} is the forward transfer admittance and y_{11} is the input admittance when output is short-circuited. The admittance parameters were extracted using small signal analysis in Sentaurus Sdevice. Using the definition of unilateral power gain (U), maximum oscillation frequency can be determined by finding the frequency where U equals 1, i.e. 0 dB. The unilateral power gain (U) is also defined using admittance parameters:

$$U = \frac{|y_{21} - y_{12}|^2}{4\left[\mathrm{Re}(y_{11})\mathrm{Re}(y_{22}) - \mathrm{Re}(y_{12})\mathrm{Re}(y_{21})\right]}, \qquad (2)$$

where y_{12} is the reverse transfer admittance and y_{22} is the output admittance when input is short-circuited. Figures 6 and 7 show unilateral power and current gains versus frequency for SOI and plane-doped bulk FinFET, respectively. Although U and h_{21} often drop at a rate much faster than 20 dB/decade at high frequencies, it is common to extrapolate f_{max} and f_T as frequencies where U and h_{21}, respectively, are equal 1 assuming a 20 dB/dec slope [15].

Extrapolated frequencies for all examined structures are listed in Table II. Cut-off frequencies are approx. the same, 270 GHz, for all the examined structures, whereas f_{max} differs significantly depending on the structure. Since the cut-off frequency is proportional to transconductance g_m and inversely proportional to gate capacitance C_g, we investigate the influence of these parameters on cut-off frequency.

Figure 8 shows transconductance versus gate voltage. The transconductance was extracted from transfer characteristic where drain voltage was fixed at 0.9 V and gate voltage was changed from 0 V to 0.9 V. It can be seen that the transconductance is the highest for the initial bulk FinFET which is a consequence of increased overdrive voltage due to lower threshold voltage for this structure. For higher gate voltages (V_{GS}> 0.5 V) in both doped bulk FinFETs channel is also formed in smaller part of fin below the gate influence, so the channel width is effectively wider compared with SOI FinFET which results with higher transconductance for both doped bulk FinFETs. Since PTS in both bulk doped structures slightly penetrates in the active fin, at lower voltages (V_{GS} < 0.5 V) channel width for both doped bulk FinFETs

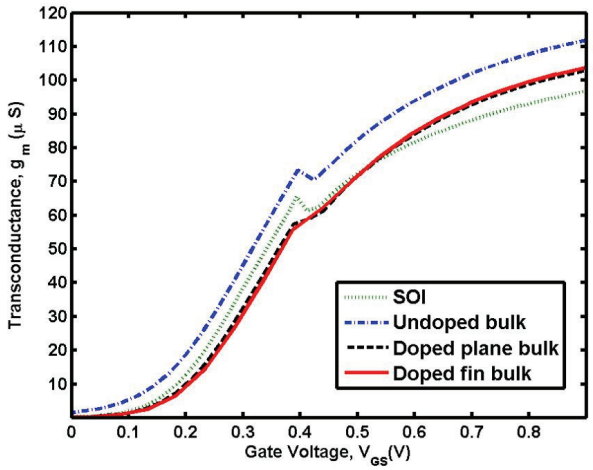


Fig. 8. Transconductance g_m versus gate voltage for all examined structures (V_{DS} = 0.9 V). As g_m is proportional to channel width, with smaller width we get smaller transconductance.

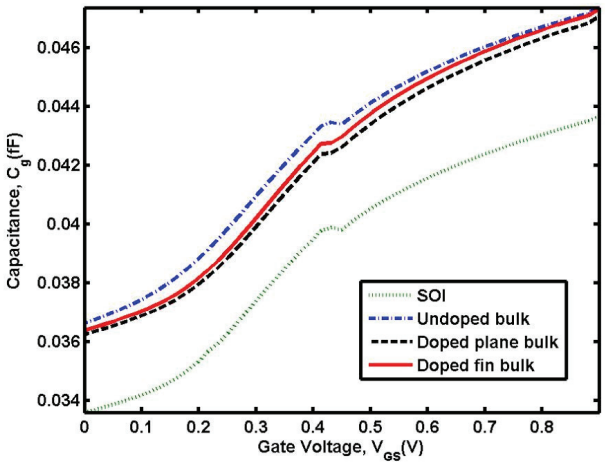


Fig. 9. Capacitance C_g versus gate voltage for all examined structures at 500 GHz (V_{DS} = 0.9 V). It is shown that capacitance is the lowest for SOI FinFET.

is narrower compared with SOI FinFET where fin is uniformly doped. As g_m is proportional to channel width, smaller width results in smaller transconductance. Figure 9 shows gate capacitance dependency of V_{GS} at 500 GHz. Capacitance was extracted using small signal analysis in Sentaurus Sdevice at the same bias as transconductance. It is shown that capacitance is the highest for undoped bulk FinFET. If we assume that at V_{GS}= 0.9 V the gate capacitance is mostly determined by gate oxide capacitance, then this capacitance has higher influence on total difference of C_g between bulk and SOI structures than capacitance toward substrate. Since the channel is effectively wider in case of bulk FinFETs, that determines their higher total gate capacitance. Also, if we compare capacitance towards substrate of SOI and bulk structures, we can conclude that bulk structures will have higher capacitance due higher dielectric constant of silicon than silicon oxide. In g_m-V_{GS} and C_{GS}-V_{GS} characteristics, kink effect was observed which is consequence of corner effect. Due to low body doping corner effect is not noticeable in the transfer characteristic [16].

Figures 10 and 11 show the dependence of f_T and f_{max} for different structures as a function of gate voltage, respectively. As expected, considering the values of C_g and g_m at V_{GS}= 0.9 V, the highest cut-off frequency is obtained for undoped bulk FinFET, and the lowest for SOI. With both bulk structures with implemented PTS in body region, satisfying results were achieved. Cut-off frequencies for both of them are almost equal to f_T of their SOI counterpart. Maximum oscillation frequency dependence on f_T and g_{ds} is approximated by

$$f_{\max} \sim \frac{f_T}{\sqrt{g_{ds}}}, \qquad (3)$$

where g_{ds} is the output conductance. Since the cut-off frequencies are approximately the same for all the examined structures, f_{max} is determined by the difference

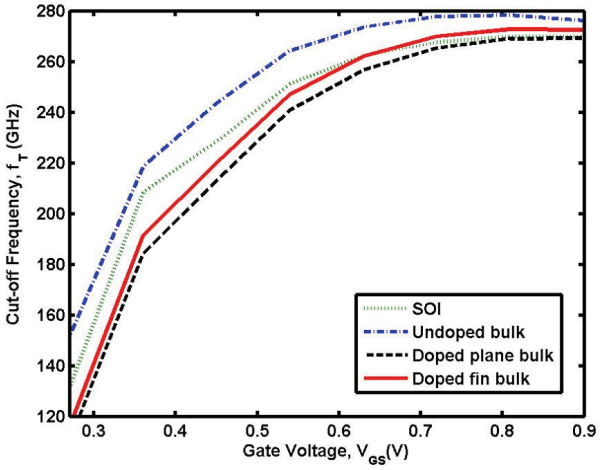


Fig. 10. Dependence of f_T on the gate voltage for all examined structures (V_{DS} = 0.9 V). For V_{GS} = 0.9 V, the highest cut-off frequency is obtained for undoped bulk FinFET, and the lowest for plane-doped bulk FinFET.

in g_{ds} for different structures. Therefore, due to the highest g_{ds}, undoped bulk FinFET exhibits the lowest f_{max} (713 GHz), while the SOI FinFET has the lowest g_{ds} and therefore highest f_{max} (776.5 GHz). The maximum oscillation frequency of plane-doped bulk FinFET is 744.4 GHz which is 4.5 % lower than SOI and 4.5 % higher than the initial bulk FinFET. On the other hand, fin-doped bulk FinFET shows slightly better results, with only 3 % lower f_{max} than SOI FinFET and 6.5 % higher f_{max} than the initial bulk structure. This slightly higher f_{max} in case of fin-doped bulk FinFET is good argument for using this type of bulk FinFET to achieve RF performance comparable to that of the SOI FinFET. However, the fin-doped type of FinFET is a more challenging solution from the fabrication point-of-view than plane-doped bulk FinFET. According to [1], [4], [7],

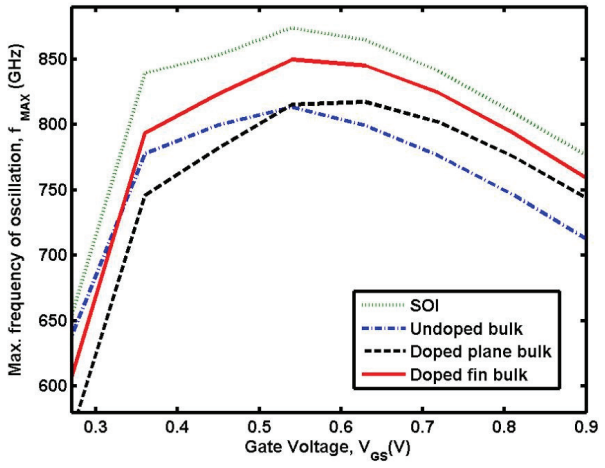


Fig. 11. Dependence of f_{max} on the gate voltage for all examined structures (V_{DS}= 0.9 V). For V_{GS}= 0.9 V, the highest f_{max} is obtained for SOI and the lowest for undoped bulk FinFET.

[17] with gate length downscaling the peak value of FOMs like f_T and f_{max} is increasing. In aforementioned literature, results of RF-FOMs for 12 nm-FinFETs are around 300 GHz which is approximately two times lower than results presented in this paper. Therefore, our results over-estimate the RF FOMs at the 20 nm gate length, most likely because high doping of source/drain extensions is assumed, resulting in significantly underestimated series resistances. In addition, since quantum effects were neglected, carrier mobility in our simulations is most likely significantly larger than expected in extremely-scaled structures. Finally, discrepancy with the measured data is partially attributed to omission of gate resistance since metal gates were used. Nevertheless, this study serves as a comparative analysis that indicates possible methods of improvement of RF performance in extremely-scaled bulk FinFETs.

IV. Conclusion

FinFETs with 20-nm gate length implemented on either bulk or SOI substrate were analyzed in terms of DC and RF performance by using 3D numerical device simulation. The initial undoped bulk FinFET has significantly higher current in the OFF state than its SOI counterpart, which demands a PTS doping in the inactive fin region in order to make bulk FinFET comparable with SOI in terms of DC performance. Bulk structures with two types of PTS doping were examined. For both bulk structures with PTS, approximately 20 times smaller current in the OFF state was achieved. Also, both doped bulk structures have shown similar RF performance as their SOI counterpart. Cut-off frequencies are roughly the same about 270 GHz, but bigger difference is seen in the f_{max}. The SOI FinFET has the highest f_{max} of 776.5 GHz, which is only 4.5 % and 3 % higher than in the plane-doped and fin-doped bulk FinFET, respectively. Therefore, considering the f_{max}, we can conclude that bulk structures with PTS exhibit significantly improved RF performance compared to undoped FinFETs, and are comparable with SOI FinFETs. The reported results for doped bulk FinFETs make a good argument for using them as future RF devices, even though the PTS doping is challenging from the fabrication point of view.

Acknowledgments

M. P. and T. S. acknowledge partial financial support from the University of Zagreb and the Ministry of Science, Education and Sports of the Republic of Croatia.

References

[1] V. Subramanian, A. Mercha, B. Parvias, M. Dehan, G. Groeseneken, W. Sansen, S, Decoutere, "Identifying the bottlenecks to the RF performance of FinFETs", 23rd International Conference on VLSI Design, pp. 111-116, 2010.

[2] T. Skotnicki, J. Hutchby, T. King, H. Wong, F. Boeuf, "The end of CMOS scaling: Toward the introduction of new materials and structural changes to improve MOSFET performance", IEEE Circuits Devices Mag., Vol. 21, No. 1, pp. 16-26, 2005.

[3] D. Hisamoto, W. –C. Lee, J. Kedzierski, H.Takeuchi, K.Asano, "FinFET: A self-aligned double-gate MOSFET scalable to 20 nm", IEEE Trans.Electron Devices, Vol.47, pp. 2320-2325, 2000.

[4] S. Nuttinick, B. Parvais, G. Curatola, A. Mercha, "Double Gate FinFETs as a CMOS Technology Downscaling option: An RF Perspective", IEEE Transactions on electron devices, Vol. 54, No. 2, pp. 279- 283, 2007.

[5] S.L.Tripathi, R. Mishra, R.A.Mishra, "Characteristic comparison of connected DG FINFET, TG FINFET and Independent Gate FINFET on 32 nm technology" 2nd International Conference on Power, Control and Embedded Systems, pp. 1-7, 2012.

[6] M.Poljak, V. Jovanović, T. Suligoj, "Improving bulk FinFET DC performance in comparison to SOI FinFET", Microelectronic Engineering, Vol. 86(10), pp. 2078-2085, 2009.

[7] J.C. Tinoco, S.S. Rodriguez, A. G. Martinez- Lopez, J. Alvarado. J. Raskin, "Impact of Extrinsic capacitances on FinFET RF performance", IEEE Transactions on microwave theory and techniques, Vol. 61, No.2, pp. 833- 840, 2013.

[8] M. Guillorn, J. Chang, A. Bryant, N. Fuller, O. Dokumaci, X. Wang, "FinFET performance advantage at 22 nm: An AC Perspective", IEEE Symp. VLSI Technology, pp. 12-13

[9] B. Parvais, M. Dehan, V. Subramanian, A. Mercha, K.T. San, M. Jurezak, " Analysis of FinFET parasitics for improved RF performance", Proc. Intl. SOI Conf., pp. 37- 38

[10] International Technology Roadmap for Semiconductors (ITRS), 2012 Edition, www.itrs.net

[11] Sentaurus User Guide, Synopsys Inc.

[12] M. Poljak, T. Suligoj, V. Jovanović, "Modeling study on carrier mobility in ultra-thin body FinFETs with circuit-level implications", Solid-State Electronics, Vol. 66, pp. 130-138, 2011.

[13] S.L.Tripathi, R. Mishra, V. Narendra, "Optimization of Pie-gate Bulk FinFET Structure", International Journal of Computer Applications, Vol. 59, No.2, pp.34- 39, 2012.

[14] T. Kanemura, T. Izumida, N. Aoki, M. Kondo, S.Ito, T. Enda, K. Okano, H. Kawasaki, A. Yagishita, A.Kaneko, S. Inaba, M. Nakamura, K.Ishimaru, K.Suguro, K. Eguchi, K.Ishiuchi, "Improvement of Drive Current in Bulk-FinFET using Full 3D Process/Device Simulations", SISPAD, pp. 131- 134, 2006.

[15] C.H. Doan, S. Emami, A.M. Niknejad, R.W. Brodersen, „ Millimeter –Wave CMOS Design", IEEE Journal of solid- state circuits, vol.40, no.1, pp. 144- 155, 2005.

[16] M. Poljak, V. Jovanović, T. Suligoj. "Suppression of corner effects in wide-channel triple-gate FinFETs", Microelectronic Engineering, vol.87, No. 2, pp. 192- 199, 2010.

[17] S. Lee, B. Jagannathan, S. Narashima, A. Chou, N. Zamdmer, J. Johnson, R. Willimas, L. Wagner, J. Kim, J. –O. Plouchart, J. Pekarik, S. Springer, G. Freeman, "Record RF performance of 45-nm SOI CMOS techonology", IEEE Int. Electron Device Meeting, pp. 255-258

Simulation Study of Aging in CMOS Binary Adders

Ting An, Hao Cai, Lirida Alves de Barros Naviner
Département Communications et Électronique
Institut Mines-Télécom, Télécom-ParisTech, LTCI-CNRS-UMR 5141
Paris CEDEX 13, 75634, France
Email:{ting.an, hao.cai, lirida.naviner}@telecom-paristech.fr

***Abstract*—Hot carrier injection (HCI) and negative bias temperature instability (NBTI) become dominant reliability issues in nanometer CMOS technology. These aging effects can induce additional delay which will be accumulated through logic gates and thus degrade the performance of the circuits. This paper discusses performance degradations induced by aging mechanisms in digital integrated circuits. We propose an aging-aware synthesis flow taking into account NBTI and HCI. This flow is demonstrated through the implementation of several architectures of adders using CMOS technology. The simulation results show that Kogge-Stone Adder (KSA) and SKlansky Adder (SKA) are the best solutions whether in terms of the complexity or the resistance to aging effects with induced delay degradation below 0.35%.**

I. Introduction

As essential arithmetic elements, adders are wildly used in digital integrated circuits (ICs), such as CORDIC and Finite Impulse Response (FIR) processors. They are also applied to build more complicated arithmetic elements in digital signal processors. Various binary addition architectures have been proposed over the last decades, in order to meet principal design specifications: speed, power and area [1].

The scaling down of CMOS technology causes reliability problems in ICs and systems [2]. NBTI and HCI are treated as main aging effects in sub-micron CMOS circuits [3]. These effects induce additional delay through many logic gates and thus degrade circuit performance. Aging effects have been intensively studied on logic gates (inverter, NAND, NOR) [4] [5]. Nevertheless, a few of researches focus on arithmetic operators. Siddiqua *et al.* reported NBTI-aware analysis on the adder, whereas only Kogge-Stone adder is studied [6].

Traditional digital IC's design tasks includes behavioral description (VHDL/Verilog), functional simulation, RTL (register transfer level) synthesis and post-layout simulation. In order to consider performance degradation due to scaling down, reliability should be included in this design flow. Since a non-critical path can become a critical path with the time progress due to NBTI effect [7], we particularly concentrate on the aging effects on critical path and near critical path of circuits [8].

In this paper, we propose a design flow that includes aging estimation taking into account both NBTI and HCI mechanisms. With the help of this flow, it is possible to aging-aware simulate the circuit netlist generated from RTL synthesis. Binary adders with 65 nm CMOS in different architectures are studied in order to demonstrate this flow. We show a comprehensive simulation study of aging effects in binary adders with 65 nm CMOS.

The paper is organized as follows. Section II reviews aging mechanisms in MOS transistors and introduces the adders architectures studied in this work. Section III presents the proposed synthesis flow considering aging effects. Section IV applies this flow to adders in different architectures. Section V outlines some conclusions and future work.

II. Preliminaries

A. Aging effects

NBTI and HCI are two major aging degradation effects in nanometer CMOS technology [9] [10]. Figure 1(a) and 1(b) show NBTI and HCI mechanisms in PMOS/NMOS type transistors. NBTI refers to the generation of positive oxide charge and interface traps at the Si-SiO_2 interface under negative gate bias to PMOS transistor, in particular at elevated temperature. HCI happens in NMOS type transistor, when hot electrons overcome the potential barrier and inject into gate oxide due to high lateral electric field near the drain.

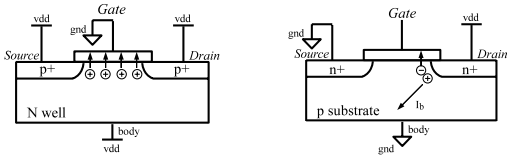


Fig. 1. HCI and NBTI mechanisms at CMOS device level: a. NBTI causes positive oxide charge and interface traps in PMOS type transistors, b. Hot carriers inject into the dielectric at the drain end of NMOS type transistors

As shown in Figure 2(a), HCI can influence sub-threshold swing coefficient (n_{fac}), intrinsic threshold voltage (v_{th0}), intrinsic mobility (μ_0), saturation velocity (v_{sat}), drain source resistance per width (r_{dsw}), subthreshold region DIBL coefficient (eta_0) and threshold voltage offset (v_{off}). Besides, the only BSIM4 [11] parameter affected by NBTI is v_{th0} of PMOS transistor (see Figure 2(b)). More severe v_{th0} degradation is observed when PMOS transistor work at high temperature ($150°C$) [12]. These physical parameter degradations can lead to circuit level degradation. In particular, this work focuses on delay degradation.

B. Digital adders

According to design requirements, different adder architectures can be applied to digital ICs design [16]. Five architectures are taken into account in this work. Ripple Carry Adder

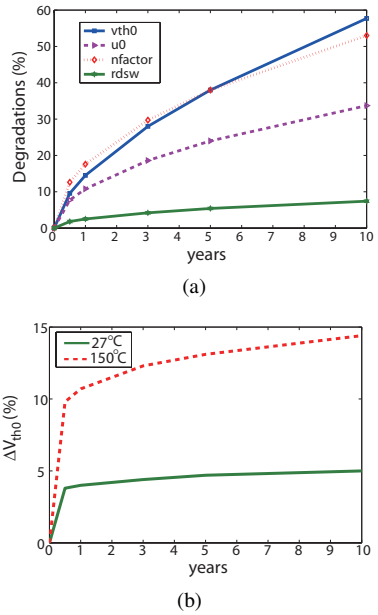


Fig. 2. HCI and NBTI induced physical parameters degradation in: a. NMOS and b. PMOS transistor

(RCA) is the simplest carry propagation adder leading to a low operation speed. Carry Select Adder (CSA) [13] accelerates the speed thanks to redundancy. Carry Lookahead Adder (CLA), also referred as Parallel-Prefix Adders (PPAs) is another fast adder suitable for high-speed application. We also explore some advanced PPAs like Kogge-Stone Adder (KSA) [14] and SKlansky Adder (SKA) [15]. Table I summarizes the characteristics of each architecture, according to area, delay, regularity (synthesis).

TABLE I. COMPARISONS OF DIFFERENT ADDER ARCHITECTURE [1]

Adder architecture	Area	Delay	Regularity
RCA	$O(N)$	$O(N)$	highest
CSA	$O(N)$	$O(N^{\frac{1}{2}})$	high
CLA	$O(N)$	$O(logN)$	medium
KSA	$O(NlogN)$	$O(logN)$	medium
SKA	$O(NlogN)$	$O(logN)$	medium

III. AGING-AWARE DESIGN FLOW

SPICE simulators are widely used to evaluate timing performance at gate level. Nevertheless, they are not suitable to directly simulate aging induced timing degradations in complex circuits due to low computation efficiency [17]. The proposed aging-aware design flow is based on the gate-level netlist generation of traditional synthesis, as illustrated in Figure 3. Traditional RTL synthesizers can produce a structural (i.e. logic or transistor level) description based on a HDL description for a given digital block.

According to design specifications, the *Behavioral modeling language* for digital circuits (e.g., VHDL, Verilog) is used to describe the functions of logic circuits. Their function correctness is verified by the *Functional simulation*. The *Library statement* includes necessary technology files. A gate-level netlist is generated by *RTL synthesis* and re-simulated by the *Aging simulator*.

In this work, we considered NBTI and HCI aging models available on Eldo simulator from Mentor Graphics [18]. With an initialization of aging setup (e.g., aging model selection, aging stress duration), a transient simulation is performed to evaluate the stress on each transistor [19]. Then, aged netlist which contains degraded information is generated and can be further applied to post-layout simulations. Finally, both ideal and aged synthesis report can be investigated by designers.

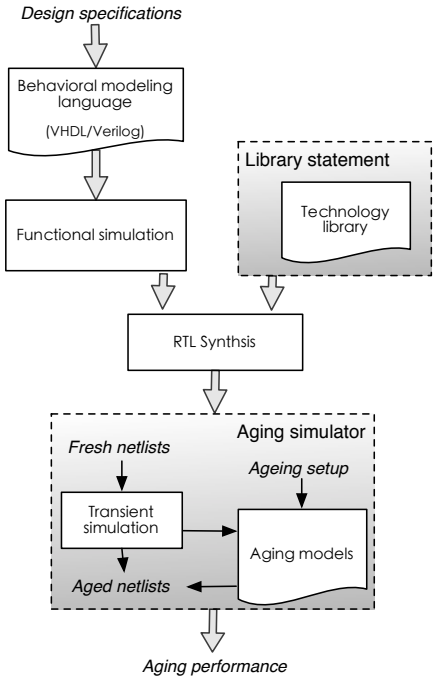


Fig. 3. Aging-aware design flow

IV. CASE STUDY: 65 NM ADDERS

This section presents the ASIC implementation results as well as the aging effect performance of different adders as mentioned in Section II-B. We used the 65 nm CMOS technology to demonstrate the proposed flow. Even the 65 nm technology is not highly susceptible to aging effects, our flow allows to explore how they affect adder circuits anyway. The simulation environment is characterized with room temperature of 27°C and 1.2 V supply voltage. The input duty cycle is set to 50%. Notice that the experimental results are obtained without the consideration of input probability. We conservatively assume that all the PMOS and NMOS transistors in the adders are affected equally by NBTI and HCI respectively.

A. Synthesis results

Table II shows the performance comparison, where ADP and PDP indicate relative area delay product and power delay product, respectively. Different word length adders are reported (4 bits, 8 bits and 16 bits). From this table, KSA and SKA

TABLE II. PERFORMANCE COMPARISON OF ADDERS

Adder Architecture	Area (μm^2)			Delay (ps)			Power			ADP (Normalized)			PDP (Normalized)			Dominant effect
	4	8	16	4	8	16	4	8	16	4	8	16	4	8	16	
RCA	49.9	99.8	199.7	524.8	992.4	1926.6	2.71	5.42	11.08	1.17	1.36	1.59	1.05	1.22	1.62	NBTI
CSA	96.7	193.4	386.9	372.7	600.7	1056.7	7.24	14.48	28.96	1.61	1.60	1.69	1.80	1.97	2.25	NBTI
CLA	64.5	130,0	257.9	348.2	651.4	1257.8	4.71	9.41	18.82	1.00	1.16	1.34	1.21	1.39	1.79	NBTI
KSA	64.0	160.2	402.5	350.4	532.5	601.2	3.87	9.15	21.96	1.00	1.17	1.00	1.00	1.10	1.00	HCI
SKA	57.7	132.1	302.6	416.6	550.2	819.6	3.59	8.03	17.75	1.07	1.00	1.03	1.10	1.00	1.10	HCI

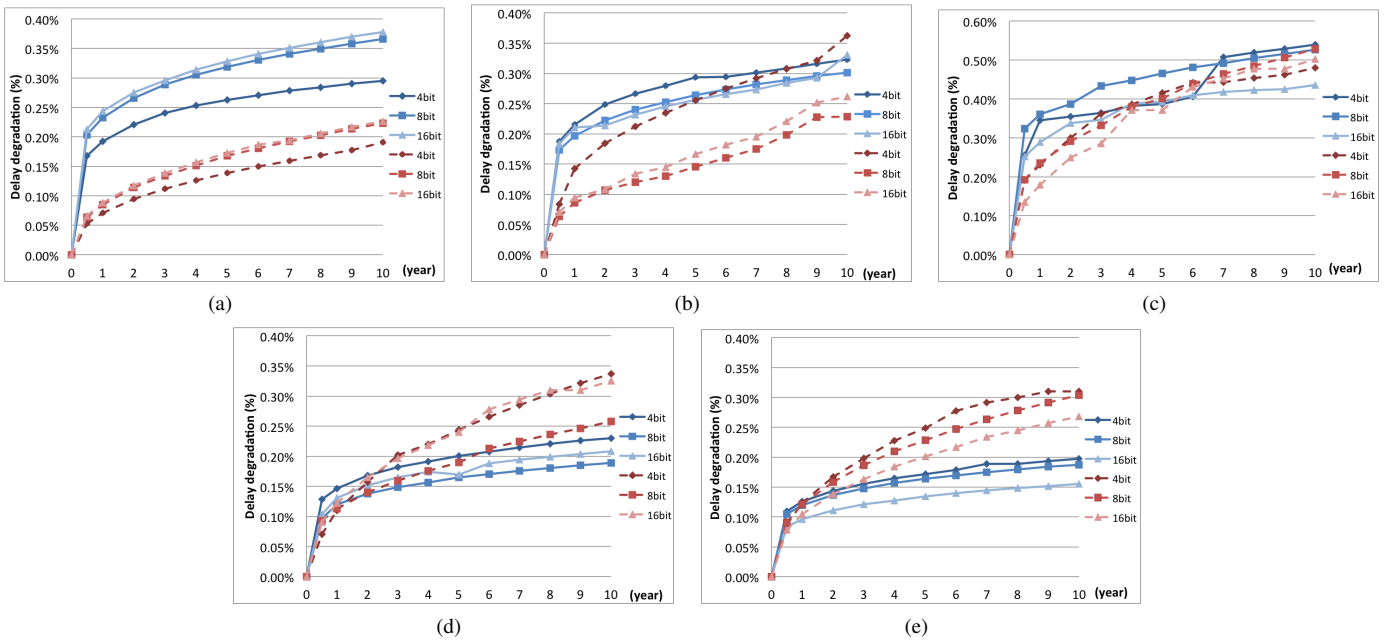


Fig. 5. NBTI and HCI Delay degradation over time of different adders, T=300K: (a) RCA (b) CSA (c) CLA (d) KSA (e) SKA ; (d)An example (8 bits-length KSA) to demonstrate the critical path changing with time under the HCI effect

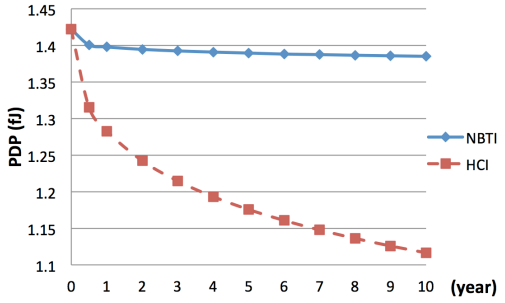


Fig. 4. Power delay product versus year of 4 bits-length RCA

are the best tradeoffs considering area, delay and power. It is noteworthy both adders have an interesting delay performance, especially in the case of long word length. Thus, they are referred to fast adders in this work.

As it can be seen from Table II, the impacts of NBTI and HCI exhibit different features according to adder architectures. Aging effects can lead to delay degradation and decrease the ADP product. Instead, they have positive influences on power and thus increase the PDP. This particularly true for HCI (see Fig. 4). We therefore concentrate on delay degradation. More details are shown in Figure 5, where blue lines stand for the delay degradation induced by NBTI and red lines for HCI. The impacts of NBTI and HCI on adders will be further discussed in the next subsection.

B. *Ageing consideration*

As shown in Figure 5, both NBTI and HCI induced additional delay are time-dependent and monotonic increasing. The dominant aging effect changes with adder architecture and depends on PMOS and NMOS partitions. The fast adders are more sensitive to HCI, while no significant difference between two effects is observed for CSA. NBTI is the dominant aging effect for RCA and CLA. Particularly, CLA's delay is severely degraded by NBTI (higher than 0.5% after 10 years stress). Remind that the experiments are performed at room temperature. NBTI degradation will get worse with temperature increasing [9], whereas HCI can not be intensified at high temperature [10].

Besides, we observe that the dominant aging effect varies with time in the case of faster adders. They become more sensible to HCI than NBTI after 2 years stress. The degradation speed caused by HCI is always faster than NBTI which makes the HCI as dominant effect in the later period, as expected according to Fig. 2.

Another interesting aspect is that the maximum delay

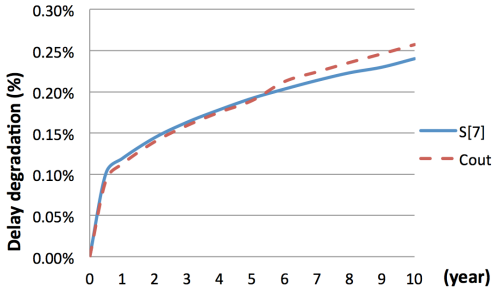


Fig. 6. An example (8 bits-length KSA) to demonstrate the critical path changing with time under the HCI effect

degradation is not always appearing at critical path. Taking CLA as example, the most significant degradation occurs on "carry out" bit which is near the critical path (MSB of the sum). As aforementioned, NBTI can lead to the possible critical path changing. Our experimental results prove that HCI may also result in this changing (see Figure 6). Let take a 8 bits KSA for instance, we notice the fresh circuit has bit "S[7]" as the critical path. However, the bit "Cout" is more sensible to HCI with the time progress, and will surpass the critical path after several years. Furthermore, the arise of delay do not always enhance the impact of aging effects. On the contrary, the impact could be reduced with increasing of delay. The 16 bits SKA suffers less degradation than the 4 bits SKA. This suggests that the aging effect is not a linear function of the impact on the individual logic gate.

Figure 7 depicts the performance comparison of different adders under aging effects. The observations are summarized as follows:

- Either NBTI or HCI, CLA is the most aging-sensitive adder among all architectures. For instance, a 8 bits-length CLA is thrice sensible to HCI than a 8 bits-length SKA. Moreover, CLA is much more vulnerable to HCI than others.
- SKA is the least NBTI-sensitive adder architectures (delay degradation$< 0.2\%$), while RCA is the most efficient architecture to resist HCI.
- Compared with the other adders, the fast adders can be the best trade-off considering both NBTI and HCI.
- The aging induced delay degradation in 65 nm CMOS adders is low compared with that related to other reliability effects (e.g., process variation). In other words, this observation demonstrates that 65 nm CMOS adders are reliable with respect to aging effects.

V. Conclusions

In this paper, we proposed an aging-aware design flow taking into account two major aging effects: NBTI and HCI. Five adder architectures were studied to demonstrate this flow using 65 nm CMOS technology. Adders were compared with respect to area, delay and timing performance under aging effect. The experimental results showed that the fast adders overcome both in hardware complexity and resistance to aging effects.The aging consideration enables to point out that the dominant effect strongly depends on architecture. The fast adders are more vulnerable to HCI than NBTI and while the opposite is true for low-speed adders. The difference delay degradation among studied adder architectures can be up to three times according to the circuit size and time stress. In addition, it is possible that the increasing of critical path reduce the impact of aging effect, an shown for SKA. Future work will focus on the experimental verification of obtained results as well as other reliability effects in digital circuits.

Acknowledgment

This work received the financial support from CATRENE Rely and ANR Robust FPGA projects.

References

[1] R. Zimmermann, "Binary adder architectures for cell-based VLSI and their synthesis," *Ph.D. dissertation, Swiss Federal Institute of Technolog (ETH) Zurich, Hartung-Gorre Verlag*, 1998.

[2] "International Technology Roadmap for Semiconductors (ITRS)", *Online: http://www.itrs.net/ Links/2009ITRS/Home2009.htm*, 2011.

[3] G. Gielen, P. D. Wit, E. Maricau, J. Loeckx, J. Martin-Martinez, B. Kaczer, G. Groeseneken, R. Rodriguez and M. Nafria, "Emerging yield and reliability challenges in nanometer CMOS technologies," *Proc. Design, Automation and Test in Europe (DATE)*, pp. 1322–1327, 2008.

[4] R. Vattikonda, W. Wang and Y. Cao, "Modeling and minimization of PMOS NBTI effect for robust nanometer design", *Proceedings of the 43rd annual Design Automation Conference*. ACM, 2006: 1047-1052.

[5] Wang W. "Circuit Aging in Scaled CMOS Design: Modeling, Simulation, and Prediction".*Ph.D. dissertation, Arizona State University*, 2008.

[6] T. Siddiqua, S. Gurumurthi, M. R. Stan, "Modeling and analyzing NBTI in the presence of Process Variation", *IEEE 12th International Symposium on Quality Electronic Design (ISQED)*, pp. 1–8, 2011.

[7] W. Wang, S. Yang, S. Bhardwaj, R. Vattikonda, S. Vrudhula, F. Liu and Y. Cao. "The impact of NBTI on the performance of combinational and sequential circuits," *Proceedings of the 44th annual Design Automation Conference,* ACM, 2007: 364-369.

[8] R. Baranowski, A. Cook, M. Imhof, C. Liu and H.-J. Wunderlich. "Synthesis of Workload Monitors for On-Line Stress Prediction," *IEEE Symp. Defect and Fault Tolerance (DFT)*, 2013.

[9] J. KO and S. CM, "Negative bias stress of MOS devices at high electric fields and degradation of MOS device," *Journal of Applied Physics*, vol. 48, no.5, pp. 2004–2014, 1977.

[10] C. Hu, S. C. Tam, F. Hsu, P. Ko, T. Chan, and K. W. Terrill, "Hot-electron-induced MOSFET degradation–model, monitor, and improvement," *IEEE Tran. on Electron Devices*, vol. 32, no.2, pp. 375–385, 1985.

[11] [Online] *http:www device.eecs.berkeley.edubsimbsim4.htm*. BSIM4 homepage, version 4.7, 2012.

[12] , H. Cai and H. Petit and J.-F. Naviner, "A Hierarchical Reliability Simulation Methodology for AMS Integrated Circuits and Systems", *Journal of Low Power Electron*, vol. 8, no.5, pp. 697–705,2012.

[13] O. J. Bedrij," Carry-select adder," *Electronic Computers, IRE Transactions on*, no 3, p. 340-346, 1962.

[14] P. Kogge and H. Stone, "A Parallel Algorithm for the Efficient Solution of a General Class of Recurrence Equations," *IEEE Transactions on Computers*, 1973, C-22, 783-791.

[15] , J."An evaluation of several two-summand binary adders," *Electronic Computers, IRE Transactions on* ,no 2, p. 213-226, 1960.

[16] B. Parhami, "Computer arithmetic: algorithms and hardware designs,"*Oxford University Press*, 2000.

[17] D. Lorenz, M. Barke and U. Schlichtmann, "Efficiently analyzing the impact of aging effects on large integrated circuits," *Microelectronics Reliability Journal*, vol. 52, no.8, pp. 1546–1552, 2012.

[18] Mentor Graphics Corporation, "Eldo user manual," 1999.

[19] H. Cai, H. Petit, and J.-F. Naviner, "Reliability Aware Design of Low Power Continuous-Time Sigma-Delta Modulator," *Microelectronics Reliability Journal*, vol. 51, no.9-11, pp. 1449–1453, 2011.

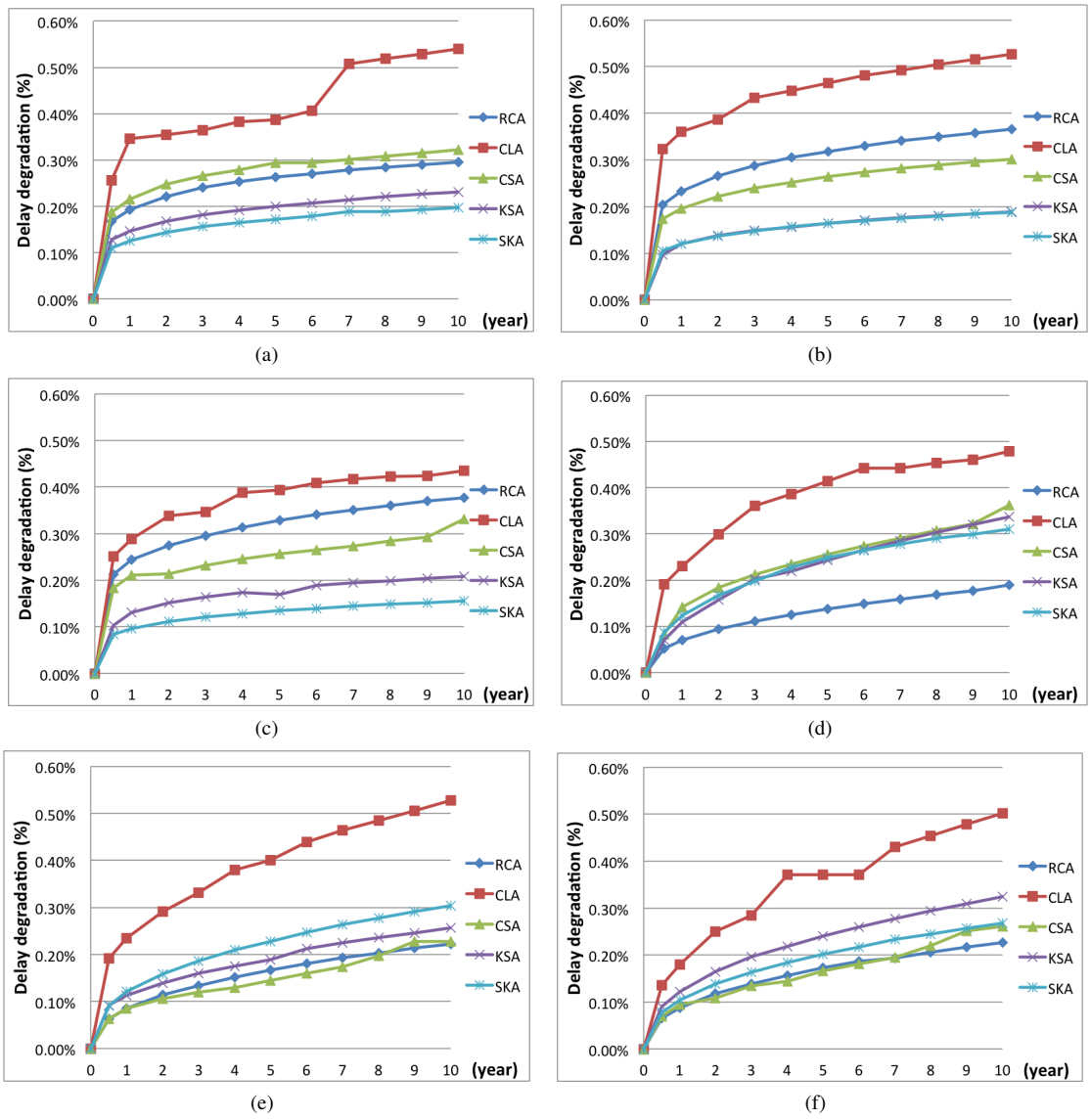


Fig. 7. Comparison of adders' performance on delay degradation over different bit length (a) 4 Bits length, NBTI (b) 8 Bits length, NBTI (c) 16 Bits length, NBTI (d) 4 Bits length, HCI (e) 8 Bits length, HCI (e) 16 Bits length, HCI

Analytical method for reliability assessment of concurrent checking circuits under multiple faults

Ting An, Kaikai Liu, Lirida Alves de Barros Naviner
Institut Mines-Telecom, Telecom ParisTech, CNRS LTCI, Paris, France
Email: {ting.an, kaikai.liu, lirida.naviner}@telecom-paristech.fr

***Abstract*—Reliability issues due to transient faults have increased with CMOS scaling and become an important concern for deep submicron technologies. Concurrent Error Detection (CED) scheme has been widely used against transient faults under the assumption of single fault and/or fault-free checking parts. In this work, we propose an analytical method in order to assess CED circuit reliability under more realistic hypothesis. In other words, we take into account the occurrence of multiple faults and fault-prone checking parts. This method allows to demonstrate the efficiency of CED schemes. The computational requirements for such an assessment are reduced by progressive analysis of the overall circuit through conditional probabilities. The proposed solution has been demonstrated on classical CED schemes.**

I. INTRODUCTION

With the shrinking of CMOS technology, electronic Integrated Circuits (ICs) are expected to suffer low reliability due to permanent and transient faults [1]. Permanent faults arise from the parametric variability in materials and fabrication processes. Transient faults mainly arise from the sensibility to external environment (radiation-related effects, low voltage, etc) and are important sources of soft errors.

Among existing fault-tolerant techniques, *Concurrent Error Detection* (CED) techniques are widely applied to detect errors occurring during normal circuit operation. This technique is particularly effective for the detection of soft errors which cannot be predicted in advance by test. Various CED schemes have been presented in last decades [2]–[5].

Traditionally, CED schemes are characterized under the hypothesis of single fault [6], [7]. However, the influence of multiple faults is no more negligible as circuits scale down to nanometer dimensions [8]. Multiple faults analysis of CED schemes have been only reported in few literature and most of them assume that faults do not affect fault-check parts [9]. Vasconcelos *et al.* proposed an analysis method based on fault injection and functional simulation approach [10]. The method assume that faults can occur independently on all components in CED circuits. Nevertheless, the time complexity of this approach is exponential with the number of binary inputs and the number of gates in circuit.

In this paper, we present an analytical solution for accurate reliability assessment of CED circuits. The solution deals with multiple faults, wherever they occur on, including fault-check parts.

The paper is organized as follows: Section II recalls basics on reliability analysis with probability matrices and CED schemes. Section III describes the proposed method. Section IV analyses the performance of the proposed solution. Finally, conclusions of this work are outlined in Section V.

II. BACKGROUND

A. Basics on PTM

The *Probabilistic Transfer Matrices* (PTMs) approach for reliability assessment is based on matrix representation of error probabilities [11].

Consider a logic block b with input x and output y, where $x \in \{x_0, x_1, \cdots, x_i, \cdots, x_{2^n-1}\}$ and $y \in \{y_0, y_1, \cdots, y_j, \cdots, y_{2^m-1}\}$, and both of them have binary representations. For example, x_{15} is noted as "1111" when n is "4". The PTM of block b, denoted PTM_b, has $2^n \times 2^m$ elements and each (i, j) element is the probability of getting output $y = y_j$ given the occurrence of input $x = x_i$, noted $p(j|i)$. In the case of an ideal (i.e. fault free) block, the PTM contains only zeros and ones and is referred as ITM (Ideal Transfer Matrix).

Fig. 1 shows the PTM of an OR logic gate with error probability equals to $1 - q$, where q represents the probability of getting a correct output. We can note that ITM is defined for $q = 1$. To calculate PTM of a whole circuit C, we combine PTMs of its basic blocks using matrix products and tensor products [11]. The reliability of a block or a circuit is directly

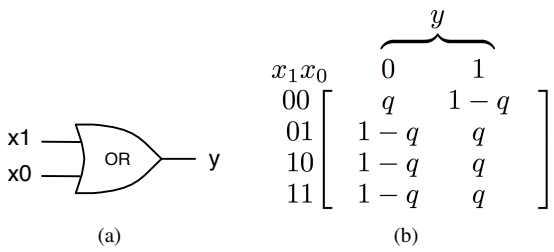


Fig. 1. (a) Logic gate OR and (b) its corresponding PTM.

extracted from the respective PTM and ITM, according to (1), where $p(i)$ denotes the probability that input x is x_i.

$$R = \sum_{ITM(i,j)=1} p(j|i) \cdot p(i) \tag{1}$$

The expression above gives the sum of probabilities corresponding to correct outputs. That is, the element in position (i, j) of PTM matrix will contribute to the sum only if the respective (i, j) element of the ITM has a value 1. Since the PTM gives the exact probabilities of getting correct outputs, this expression provides accurate reliability assessment.

B. Basics on CED

The general scheme of a CED structure is depicted in Fig. 2. It considers a main (target) *Function* F. A *Function Predictor* F_p predicts some special characteristics z of F's output y and a *Function Checker* F_c checks whether these characteristics are satisfied by y. This checking results in an one-bit output e. In the remainder of this paper, it is assumed that $e = 0$ when no error is detected and $e = 1$ otherwise. This is solely a convention that does not restrict the scope of the analysis. The output w of a such CED circuit is the concatenation of the outputs e and y: $w = [e, y]$.

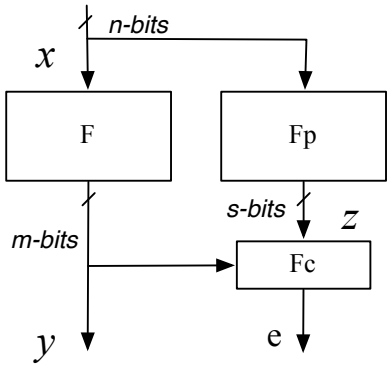


Fig. 2. Concurrent Error Detection circuit

C. Functional reliability of a CED circuit

Functional reliability can be defined as the ability of a system or component to perform its required functions under stated conditions for a specified period of time, even in presence of faults [12]. Actually, we can state that the functional reliability of a CED circuit shown in Fig. 2 is related to the probability of producing only the events E_1 or E_2 defined as follows:

Definition 1: The event E_1 is the occurrence of a correct output y combined with an output e indicating no error detected.

Definition 2: The event E_2 is the occurrence of an incorrect output y combined with an output e indicating error detected.

Indeed, the events E_1 and E_2 are mutually exclusive. Therefore, the functional reliability can be expressed as:

$$R = p(E_1) + p(E_2) \tag{2}$$

III. CED RELIABILITY ASSESSMENT

A. CED reliability assessment with PTM

In this section we show that R_{CED} can be obtained by using PTM principle.

First, consider rewriting the expression (2) to show the probability of occurrence of inputs and outputs as defined in section II-A. The expression (3) is produced by doing this, where $p(i)$ stands for the probability of $x = x_i$ and $p(j \in E_1|i) + p(j \in E_2|i)$ is related to a proper functioning of the CED for $x = x_i$. In other words, it represents the probability of having a proper $[e, y]_j$ given that the input is x_i.

$$\begin{aligned} R &= \sum_j \sum_i p(i)\left[p(j \in E_1|i)) + p(j \in E_2|i)\right] \\ &= \sum_j \sum_i p(i) \cdot PTM_{CED}(i,j) \cdot ITM_{CED}(i,j) \end{aligned} \tag{3}$$

In order to illustrate this idea, let's take an OR gate as Function F. The corresponding ITM_{CED} is presented in Fig. 3. We remark that left half of matrix ITM_{CED} (columns for $e = 0$) is the ITM_{OR}, while the right half (columns for $e = 1$) is the logic complement of ITM_{OR}.

$$\begin{array}{c} \\ 00 \\ 01 \\ 10 \\ 11 \end{array} \begin{array}{c} \overbrace{}^{e=0}\ \overbrace{}^{e=1} \\ \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \end{bmatrix} \end{array} \begin{array}{c} \\ 00\ 01\ 10\ 11 \\ \\ \\ \\ \end{array} = [ITM_{OR}, \overline{ITM_{OR}}]$$

Fig. 3. ITM_{CED} for an OR gate as function circuit F.

This can be generalized in a straightforward way for any CED circuit. Therefore, assuming F with n-bits input and m-bits output, the functional reliability can be expressed as:

$$\begin{aligned} R = &\sum_{j=0}^{(2^m-1)} \sum_{i=0}^{(2^n-1)} p(i) \cdot PTM_{CED}(i,j) \cdot ITM_F(i,j) \\ &+ \sum_{j=2^n}^{(2^{m+1}-1)} \sum_{i=0}^{(2^n-1)} p(i) \cdot PTM_{CED}(i,j) \cdot \overline{ITM_F(i,j)} \end{aligned} \tag{4}$$

B. CED reliability assessment: alternative approach

The following paragraphs present an alternative approach for estimating R. The objective is to reduce the computational complexity. For this, we define the events A and B:

Definition 3: The event A is the occurrence of a correct circuit output y.

Definition 4: The event B is the occurrence of a checker output e indicating no error.

Let $p(y_j^c)$ be as the probability of $y_j \in A$. According to the joint probability theory, the probability of E_1 is represented in (5), where $\coprod$ depicts the disjoint union.

$$\begin{aligned} p(A \cap B) &= p\Big(\coprod_j y_j^c \cap B\Big) \\ &= \sum_j \big(p(y_j^c) \cap B\big) \\ &= \sum_j p(y_j^c) \cdot p\big(B|y_j^c\big) \end{aligned} \tag{5}$$

We can notice the following equalities, where $p(z_k)$ is the probability that $z = z_k$.

$$p(y_j^c) \cdot p\big(B|y_j^c\big) = \sum_k p(y_j^c) \cdot p(B|y_j^c \cap z_k) \cdot p(z_k|y_j^c) \quad (6)$$

We now analyze the dependency of z with respect to y_j^c. Based on the law of conditional probability and the law of total probability, we get (7) and (8), respectively.

$$p(z|y_j^c) = \frac{p(z \cap y_j^c)}{p(y_j^c)} \quad (7)$$

$$p(y_j^c) = \sum_i p(x_i) \cdot p(y_j^c|x_i) \quad (8)$$

Since faults are considered to occur independently on F or F_p, y and z are independent for a given x_i:

$$\begin{aligned} & p(z \cap y|x_i) = p(z|x_i) \cdot p(y|x_i) \\ & \Rightarrow p(z \cap y) = \sum_i p(x_i) \cdot p(z|x_i) \cdot p(y|x_i) \end{aligned} \quad (9)$$

Applying (9) into (7) we have :

$$\begin{aligned} p(z|y_j^c) &= \frac{\sum_i p(x_i) \cdot p(y_j^c|x_i) \cdot p(z|x_i)}{p(y_j^c)} \\ &= \sum_i p(y_j^c|x_i) \cdot \frac{p(x_i)}{p(y_j^c)} \cdot p(z|x_i) \end{aligned} \quad (10)$$

We rewrite (5) by replacing $p(z_k|y_j^c)$ in (6) with that given in (10), and thus express the probability of E_1 as shown in equation (11).

$$\begin{aligned} p(A \cap B) = \sum_j \Big(& \sum_k p\left(B|y_j^c \cap z_k\right) \\ & \cdot \sum_i \big(p(y_j^c|x_i) \cdot p(z_k|x_i) \cdot p(x_i)\big) \Big) \end{aligned} \quad (11)$$

The probabilities in the equation (11) can be obtained from PTMs of F, F_p and F_c as the following:

$$p\left(B|y_j^c \cap z_k\right) = PTM_{F_c}((m \times j + k), 0) \quad (12)$$

$$p(y_j^c|x_i) = PTM_F(i,j) \cdot ITM_F(i,j) \quad (13)$$

$$p(z_k|x_i) = PTM_{F_p}(i,k) \quad (14)$$

Applying these equalities to (11) allows us to rewrite E_1 in (15). In an equivalent manner, we can get E_2 as given in (16). We notice that only PTMs of sub-circuits are required instead of that of the overall CED circuit.

IV. Performance analysis

A. Complexity evaluation

This section focuses on the computational complexity of methods proposed in previous section. Suppose PTMs and ITMs of sub-circuits F, F_p and F_c are already available. Meanwhile, denote $\oplus$ as the computational complexity for adding two numbers and $\otimes$ as the computational complexity for multiplying two numbers.

In the case of method described in III-A, the complexity is that of obtaining the global matrix PTM_{CED} and adding the appropriate elements in accordance with the equation (4).

On the basis of PTMs approach, the PTM of two blocks b_1 and b_2 in cascade results from the inner product of PTM_{b_1} and PTM_{b_2}. In the case of parallel structure, the resulting PTM comes from the tensor product of PTM_{b_1} and PTM_{b_2}.

Figure 4 shows the partitioning of the CED circuit in order to calculate the overall PTM_{CED}. Utilizing the rules above on the blocks in a level u leads to the values T_u in Table I. Similarly, the complexities related to the cascade involving different levels are shown in Table II.

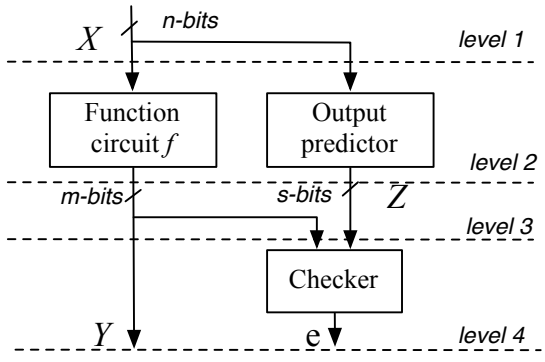


Fig. 4. Level partition of CED circuit

TABLE I. Computational complexity to obtain PTM for a given level.

T_1	$(2 \times 2^2)^n \times \otimes = 2^{3n} \times \otimes$
T_2	$2^{n+m} \times 2^{n+s} \times \otimes = 2^{2n+m+s} \times \otimes$
T_3	$(2 \times 2^2)^m \times (2^2)^s \times \otimes = 2^{3m+2s} \times \otimes$
T_4	$(2^2)^m \times (2^{m+s} \times 2) \times \otimes = 2^{3m+s+1} \times \otimes$

$$p(E_1) = \sum_{j=0}^{2^m-1} \Big(\sum_{k=0}^{2^s-1} PTM_{F_c}((m \times j + k), 0) \cdot \sum_{i=0}^{2^n-1} \big(PTM_F(i,j) \cdot PTM_{F_p}(i,k) \cdot ITM_F(i,j) \cdot p(i)\big) \Big) \quad (15)$$

$$p(E_2) = \sum_{j=0}^{2^m-1} \Big(\sum_{k=0}^{2^s-1} PTM_{F_c}((m \times j + k), 1) \cdot \sum_{i=0}^{2^n-1} \big(PTM_F(i,j) \cdot PTM_{F_p}(i,k) \cdot \overline{ITM_F}(i,j) \cdot p(i)\big) \Big) \quad (16)$$

TABLE II. COMPUTATIONAL COMPLEXITY TO OBTAIN PTM FOR A CASCADE OF LEVELS

$T_{1\to 2}$	$2^{n+2n+(m+s)}(\oplus+\otimes)$ $= 2^{3n+m+s}(\oplus+\otimes)$
$T_{1\to 3}$	$2^{2n+(m+s)+(2m+s)}(\oplus+\otimes)$ $= 2^{2n+3m+2s}(\oplus+\otimes)$
$T_{1\to 4}$	$2^{(m+s)+(2m+s)+(m+1)}(\oplus+\otimes)$ $= 2^{4m+2s+1}(\oplus+\otimes)$

The total complexity T_a thus is:

$$T_a = 2^{n+m+1}(2\times\otimes+\oplus)-\oplus+\sum_u T_u+\sum_v T_{1\to v} \quad (18)$$

In the case of method described in III-B, the total complexity comes straight from the computation of (15) and (16):

$$T_b = 2^{m+n+s+1}(2^2\times\otimes+\oplus)-\oplus \quad (19)$$

We can easily conclude that $T_b \ll T_a$, which confirms a significant gain in computation efficiency.

B. Experimental results

The proposed methodologies have been implemented with Matlab [13], while Modelsim [14] has been used as the simulation platform for a fault injection approach [10]. Experimental results were performed on a Linux workstation with a 2GHz AMD Athlon microprocessor and 2GB RAM.

Remind that in fault injection approach, the result accuracy can be improved by increasing the maximum number of considered simultaneous faults, but the computational complexity would be dramatically increased as well. In order to optimize the procedure of simulations, we assume the maximum number of simultaneous faults to be 3 in the fault injection approach

We consider three different CED schemes adopted in a ripple carry adder (RCA) operator: an identical duplication adder (noted as RCA_Dup), a diverse duplication adder (noted as RCA_Div_Dup) and a parity checking adder (noted as RCA_Parity).

Table III shows the run-time comparison. It is noteworthy that conventional PTMs based method is impracticable for RCA_Dup and RCA_Div_Dup due to excessive memory requirements. The obtained results verifies the remarkable performance of the solution described in III-B, particularly for identical duplication circuit.

TABLE III. TIME CONSUMPTIONS (S)

Circuit type	Evaluation method			Gate
	Fault injection [10]	Basic PTM	Alternative	Nb.
RCA_Dup	7398.1	NA	18.6	86
RCA_Div_Dup	6345.7	NA	19.3	78
RCA_Parity	367.2	50.3	4.7	32

V. CONCLUSION

This work dealt with the reliability assessment of CED circuits under multiple faults. We proposed an accurate analytical solution based on PTMs approach. We firstly presented how to apply PTMs to assess the functional reliability. Afterwords, we proposed an alternative approach to reduces the analysis computational complexity.

The computational efficiency of the alternative solution has been mathematically demonstrated. Arithmetic operators were considered as case studies. The experimental results confirmed that the proposed alternative solution outperforms the basic PTMs based method as well as the fault injection approach.

Future work will concentrate on exploring different CED schemes through the proposed method.

VI. ACKNOWLEDGMENT

The authors would like to thank Laboratory LTCI-CNRS (UMR 5141) and the financial support from CATRENE Rely and ANR Robust FPGA projects.

REFERENCES

[1] M. Stanisavljević, A. Schmid, and Y. Leblebici, *Reliability of Nanoscale Circuits and Systems: Methodologies and Circuit Architectures*, 2011.

[2] G. Kraft and W. Toy, "Microprogrammed control and reliable design of small computers," *Prentice-Hall Inc., Englewood Cliffs, N. J., 1981, 428*, 1981.

[3] M. Nicolaidis, R. Duarte, S. Manich, and J. Figueras, "Fault-secure parity prediction arithmetic operators," *IEEE Design & Test of Computers*, vol. 14, no. 2, pp. 60–71, 1997.

[4] M. Nicolaidis, "Carry checking/parity prediction adders and alus," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 11, no. 1, pp. 121–128, 2003.

[5] D. P. Vasudevan, P. K. Lala, and J. P. Parkerson, "Self-checking carry-select adder design based on two-rail encoding," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 54, no. 12, pp. 2696–2705, 2007.

[6] S. Mitra and E. J. McCluskey, "Which concurrent error detection scheme to choose?" in *Proceedings of the International Test Conference 2000*, 2000, pp. 985–994.

[7] S. Bayat-Sarmadi and M. A. Hasan, "Concurrent error detection in finite-field arithmetic operations using pipelined and systolic architectures," *IEEE Transactions on Computers*, vol. 58, no. 11, pp. 1553–1567, 2009.

[8] C. Rusu, A. Bougerol, L. Anghel, C. Weulerse, N. Buard, S. Benhammadi, N. Renaud, G. Hubert, F. Wrobel, T. Carriere *et al.*, "Multiple event transient induced by nuclear reactions in CMOS logic cells," in *13th IEEE International On-Line Testing Symposium, 2007. IOLTS 07.*, 2007, pp. 137–145.

[9] X. Guo and R. Karri, "Invariance-based concurrent error detection for advanced encryption standard," in *Proceedings of the 49th Annual Design Automation Conference*, 2012, pp. 573–578.

[10] M. C. de Vasconcelos, D. T. Franco, L. A. Naviner, and J.-F. Naviner, "Relevant metrics for evaluation of concurrent error detection schemes," *Microelectronics Reliability*, vol. 48, no. 8, pp. 1601–1603, 2008.

[11] K. Patel, J. Hayes, and I. Markov, "Evaluating circuit reliability under probabilistic gate-level fault models," in *Proceedings of the International Workshop on Logic and Synthesis*, 2003, pp. 59–64.

[12] W. Kuo, V. R. Prasad, F. A. Tillman, and C. L. Hwang, *Optimal Reliability Design: Fundamentals and Applications*. Cambridge University Press, 1999.

[13] Matlab. (2013). [Online]. Available: http://www.matlab.com/

[14] Model. (2013). [Online]. Available: http://www.mentor.com/

Synthesis of CMOS Multiplier Structures Using Multifunctional Circuits

C. Popa
Faculty of Electronics, Telecommunications and Information Technology, Bucharest, Romania
cosmin_popa@yahoo.com

Abstract - The paper will present new high accuracy multiplier circuits, focused on the implementation of original techniques for improving the linearity of the proposed structures. The new approach of designing high precision multipliers using multifunctional cores presents the important advantage of allowing a facile reconfiguration of the designed circuits, the multiplier core being able to implement a multitude of additional circuit functions: amplifying, squaring and square-rooting or simulating a positive and negative equivalent resistance. The multiplier structures are designed for low-voltage low-power operation (the supply voltage is $\pm 1.8V$ for implementing in $0.18\mu m$ CMOS technology and the current consumption is smaller than $50\mu A$ for any proposed multiplier).

Keywords - Analog signal processing, multiplier circuits, linearity error, multifunctional cores, translinear circuits

I. INTRODUCTION

Computational circuits and particularly multipliers are important building blocks for telecommunication applications, medical equipments, hearing devices and disk drives [1]-[4]. Different mathematical principles represent the functional cores for designing multiplier structures with increased linearity and either currents or voltages as input variables [5]-[7].

II. THEORETICAL ANALYSIS

In order to extend the possibility of generation of a multitude of additional circuit functions, the proposed multiplier circuits are designed using two original implementations of multifunctional cores - MFC (computational blocks with output variables containing a linear combination of elementary circuit functions). This new original approach of designing computational circuits using multifunctional cores presents the important advantage of generating large number of circuit functions using the same core (amplifying, multiplying or simulating a positive or negative equivalent resistance). Thus, the possibility of reconfiguration of these structures is strongly increased. This original design method facilitates the implementation of field programmable analog arrays (FPAAs) with improved performances and reduced design costs.

It will be presented a group of original implementations of multiplier circuits, the main goal being the improvement of their linearity for an extended range of the input voltage amplitudes. The operation of the proposed circuits is based on an original principle of operation and on two original implementations of multifunctional cores. In order to improve the circuits' frequency response, only MOS transistors biased in saturation will be used.

A. Principle of operation

The principle of operation for the proposed multiplier circuits is based on the compensation of the nonlieairities introduced by two complementary mathematical functions. As in CMOS technology it is very simple to implement the squaring function using MOS transistors biased in saturation region, the complementary functions are chosen to be squaring and square-rooting functions. The squaring dependencies are generated using a classical MOS differential amplifier, while the square-rooting functions are obtained considering the difference between two gate-source voltages of MOS transistors biased at different drain currents.
The mathematical relations that model the squaring dependencies are:

$$I_{O1} = \frac{K}{2}\left(V_{GS1} - V_T\right)^2 \tag{1}$$

$$I_{O2} = \frac{K}{2}\left(V_{GS2} - V_T\right)^2 \tag{2}$$

and:

$$V_{GS1} - V_{GS2} = V_3 - V_4 \tag{3}$$

resulting:

$$\sqrt{I_{O1}} - \sqrt{I_{O2}} = \sqrt{\frac{K}{2}}\left(V_3 - V_4\right) \tag{4}$$

The implementation of a square-root dependence of two output currents, I_{OUT1} and I_{OUT2}, on two biasing currents, I_{O1} and I_{O2}, respectively, exploits a translinear loop containing two gate-source voltages of MOS transistors biased at different drain currents. The difference of these gate-source voltages is forced to be equal with a multiple of a $V_1 - V_2$ differential input voltage:

$$V_{GS}\left(I_{OUT1}\right) - V_{GS}\left(I_{O1}\right) = a\left(V_1 - V_2\right) \tag{5}$$

a representing a constant that depends on the particular implementation of the circuit that uses this principle of operation. Imposing a biasing in saturation of all MOS transistors from the circuit, it results the following relation:

$$\sqrt{I_{OUT1}} = \sqrt{I_{O1}} + a\sqrt{\frac{K}{2}}(V_1 - V_2) \quad (6)$$

So, I_{OUT1} current will be expressed as:

$$I_{OUT1} = I_{O1} + a\sqrt{2KI_{O1}}(V_1 - V_2) + \frac{a^2K}{2}(V_1 - V_2)^2 \quad (7)$$

Similarly, another I_{OUT2} current will have a square-root dependence on I_{O2} biasing current:

$$I_{OUT2} = I_{O2} + a\sqrt{2KI_{O2}}(V_1 - V_2) + \frac{a^2K}{2}(V_1 - V_2)^2 \quad (8)$$

In order to obtain the multiplying function between two differential input voltages, the output current must have a particular linear dependence on the previous currents:

$$I_{OUT} = (I_{OUT1} + I_{O2}) - (I_{OUT2} + I_{O1}) \quad (9)$$

resulting:

$$I_{OUT} = a\sqrt{2K}\left(\sqrt{I_{O1}} - \sqrt{I_{O2}}\right)(V_1 - V_2) \quad (10)$$

Replacing (4) in (10), it results that I_{OUT} current will have the following expression:

$$I_{OUT} = aK(V_1 - V_2)(V_3 - V_4) \quad (11)$$

B. Multiplier architectures using the new principle of operation

The proposed multiplier circuits are based on multifunctional cores that are able to implement the required mathematical relations for obtaining the multiplying function. Additionally, these multifunctional cores allow the generation of a multitude of circuit functions: amplifying, squaring or simulating an active resistance. The proposed multifunctional core (MFC I) is presented in Figure 1. The M1-M2 differential input stage is biased using M3 transistor, the drain current of M2 transistor being forced to be equal to I_{O1} current.

The functional relation implemented by the circuit is:

$$V_{GS}(I_{OUT1}) - V_{GS}(I_{O1}) = V_1 - V_2 \quad (12)$$

So, the I_{OUT1} output current will be expressed by (7), for $a = 1$. The first proposed implementation of a multiplier circuit (Figure 2) uses two multifunctional cores from Figure 1.

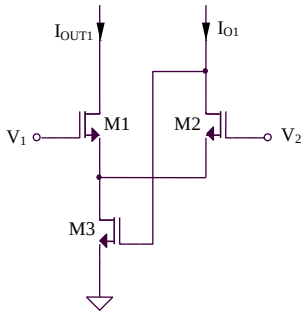


Figure1. The multifunctional core (MFC I)

The first multifunctional core, M1-M3, is biased at I_{O1} current and generates the I_{OUT1} output current. In order to extract I_{O1} current, M4 transistor is used. The expressions of I_{OUT1} and I_{OUT2} output currents are:

$$I_{OUT1,2} = I_{O1,2} + \sqrt{2KI_{O1,2}}(V_1 - V_2) + \frac{K}{2}(V_1 - V_2)^2 \quad (13)$$

As a result of two cross-connections, the output current of the multiplier circuit can be expressed using (9), resulting:

$$I_{OUT} = \sqrt{2K}\left(\sqrt{I_{O1}} - \sqrt{I_{O2}}\right)(V_1 - V_2) \quad (14)$$

The I_{O1} and I_{O2} currents are generated by the M9-M10 differential stage, the relation between them being given by (4). So, the output current of the multiplier circuit presented in Figure 2 will be proportional with the product of two differential input voltages:

$$I_{OUT} = K(V_1 - V_2)(V_3 - V_4) \quad (15)$$

The aspect ratios of MOS transistors are shown in Table 1.

TABLE 1. THE ASPECT RATIOS OF MOST RANSISTORS

M	M1	M2	M3	M4	M5	M6
W/L	4.5/0.9	4.5/0.9	9/0.9	4.5/0.9	4.5/0.9	4.5/0.9

M	M7	M8	M9	M10	M11	M12
W/L	4.5/0.9	9/0.9	9/0.9	9/0.9	4.5/0.9	4.5/0.9

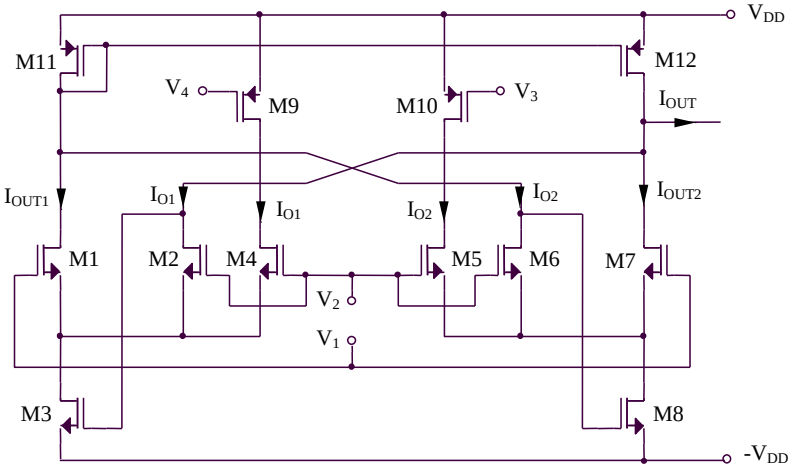


Figure 2. Multiplier circuit (I) using MFC I

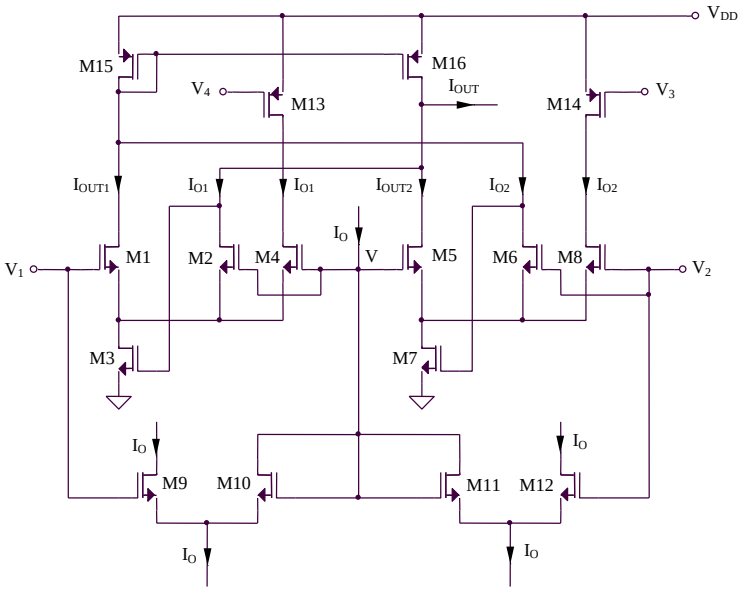


Figure 3. Multiplier circuit (II) using MFC I

The same multifunctional core from Figure 1 can be re-used for implementing another original multiplier circuit (Figure 3). Similar multifunctional cores M1-M3 and M5-M7 are used for realizing the required dependencies (7) and (8) of I_{OUT1} and I_{OUT2} output currents on I_{O1} and I_{O2} currents, respectively. The M4 and M8 transistors are used for connecting these cores with the M13-M14 differential amplifier that allows to obtain relation (4). The M9-M12 group computes the arithmetical mean of V_1 and V_2 input potentials, $V=(V_1+V_2)/2$. For the M1-M3 multifunctional core, the relation between the gate-source voltages can be expressed as follows:

$$V_{GS}(I_{OUT1})-V_{GS}(I_{O1})=V_1-V=\frac{V_1-V_2}{2} \qquad (16)$$

So, comparing with (5), it results that $a=1/2$, so the expression of the output current is obtained from (11):

$$I_{OUT}=\frac{K}{2}(V_1-V_2)(V_3-V_4) \qquad (17)$$

The aspect ratios of MOS transistors are shown in Table 2.

TABLE 2. THE ASPECT RATIOS OF MOS TRANSISTORS

M	M1	M2	M3	M4	M5	M6
W/L	4.5/0.9	4.5/0.9	9/0.9	4.5/0.9	4.5/0.9	4.5/0.9

M	M7	M8	M9	M10	M11	M12
W/L	9/0.9	4.5/0.9	4.5/0.9	4.5/0.9	4.5/0.9	4.5/0.9

M	M13	M14	M15	M16
W/L	9/0.9	9/0.9	9/0.9	9/0.9

The second proposed multifunctional core (MFC II) is presented in Figure 4.

The M1 and M3 transistors are biased at I_{O1} current, while M2 and M4 transistors have the same drain current, equal with I_{OUT1}.

The relation between their gate-source voltages is:

$$2V_{GS}(I_{OUT1})-2V_{GS}(I_{O1})=V_1-V_2 \qquad (18)$$

Relation (18) is similar with (5), so I_{OUT1} output current will be expressed by (7) (for $a=1/2$).

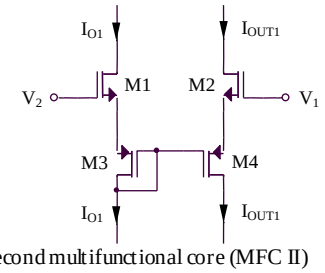


Figure 4. The second multifunctional core (MFC II)

The original proposed implementation of a multiplier circuit (Figure 5) uses two multifunctional cores from Figure 4 (M1-M4 and M5-M8 groups in Figure 5). Using (7) and (8), the expressions of I_{OUT1} and I_{OUT2} output currents will be:

$$I_{OUT1}=I_{O1}+\sqrt{\frac{KI_{O1}}{2}}(V_1-V_2)+\frac{K}{8}(V_1-V_2)^2 \qquad (19)$$

and:

$$I_{OUT2}=I_{O2}+\sqrt{\frac{KI_{O2}}{2}}(V_1-V_2)+\frac{K}{8}(V_1-V_2)^2 \qquad (20)$$

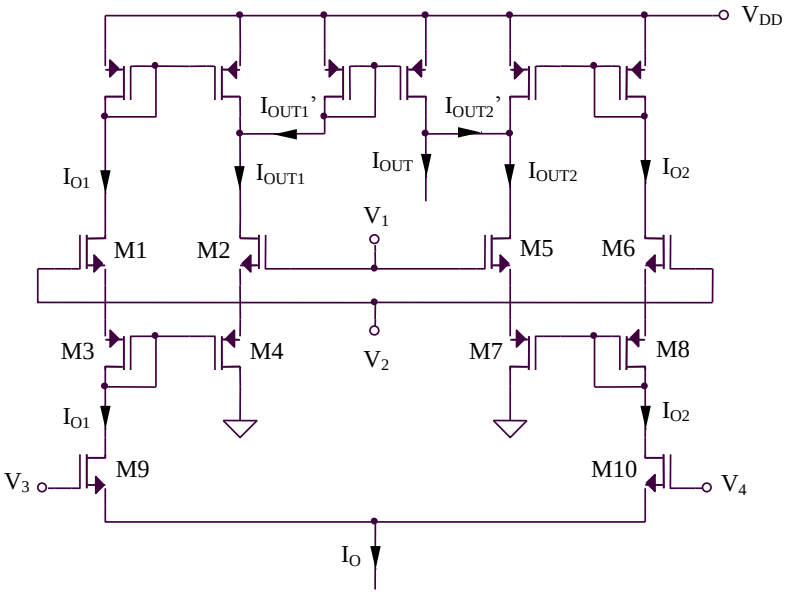


Figure 5. Multiplier circuit (III) using MFC II

As a result of the circuit connections, the output current of the multiplier can be expressed using (9), resulting:

$$I_{OUT} = \sqrt{\frac{K}{2}}\left(\sqrt{I_{O1}} - \sqrt{I_{O2}}\right)(V_1 - V_2) \qquad (21)$$

The I_{O1} and I_{O2} currents are generated by the M9-M10 differential stage, the relation between them being given by (4). So, the output current of the multiplier circuit presented in Figure 5 will be proportional with the product of two differential input voltages:

$$I_{OUT} = \frac{K}{2}(V_1 - V_2)(V_3 - V_4) \qquad (22)$$

III. CONCLUSIONS

There were presented original implementations of multiplier circuits, the main goal of designing these structures being represented by the increasing of their linearity. The new proposed method for obtaining high accuracy multiplier circuits was based on the utilization of two functional cores that are able to generate a multitude of circuit functions. This original approach of designing computational structures has the very important advantage of multifunctionality, the proposed multiplier circuits, implemented using multifunctional cores, allowing a facile reconfiguration. As the design and implementation costs were mainly related to the improving of the circuit core performances, the costs associated with the implementation of one circuit function can be strongly reduced using this new approach. Comparing with similar designs that describe CMOS implementations of voltage multipliers, the proposed circuits present the advantage of a relatively simple implementation, correlated with a very good linearity.

REFERENCES

[1] R. Harjani, "A Low-Power CMOS VGA for 50 Mb/s Disk Drive Read Channels", in *IEEE Transactions on Circuits and Systems II*, Analog Digital Signal Processing, 1995, pp. 370-376.

[2] A. Motamed, C. Hwang, M. Ismail, "A Low-Voltage Low-Power Wide-Range CMOS Variable Gain Amplifier", in *IEEE Transactions on Circuits and Systems II*, Analog Digital Signal Processing, 1998, pp. 800-811.

[3] C. Popa, "Synthesis of computational structures for analog signal processing," Springer Publishing House, 449 pp., 2011

[4] C. Popa, "Superior-order curvature-correction techniques for voltage references," *Springer Publishing House*, 179 pp., 2009

[5] Szczepanski, S., Koziel, S., "1.2V low-power four-quadrant CMOS transconductance multiplier operating in saturation region", Circuits and Systems, 2004. ISCAS '04. Proceedings of the 2004 International Symposium on, Vol.1, 2004 , Pp. I - 1016-19 Vol.1

[6] Manolescu, A.M., Popa, C., "A 2.5GHz CMOS Mixer with Improved Linearity", Journal of Circuits, Systems and Computers, 2010

[7] Akshatha, B.C., Akshintala, V.K., "Low Voltage, Low Power, High Linearity, High Speed CMOS Voltage Mode Analog Multiplier", Emerging Trends in Engineering and Technology (ICETET), 2009 2nd International Conference on, 2009, Pp. 149 – 154

Improved Accuracy Current-Mode Analog Function Synthesizer

C. Popa
Faculty of Electronics, Telecommunications and Information Technology, Bucharest, Romania
cosmin_popa@yahoo.com

Abstract - An original analog function synthesizer circuit with increased accuracy will be presented, allowing to implement a multitude of important continuous mathematical functions. The accuracy of the proposed structure is excellent and the range of the input variable is strongly extended as a result of the third-order approximation of the implemented functions. The circuit performance is expected to be stable against process corners as all functions exclusively depend on current ratios. The great advantages of the increased modularity and controllability and of the associated reduced design costs per function represent an immediate consequence of the multiple functions realized by the proposed structure. The function synthesizer circuit is designed for implementing in $0.18\mu m$ CMOS technology and it is supplied at $1.8V$. The SPICE simulations confirm the estimated theoretical results, showing an approximating error smaller than 0.01% for the composing squaring block and smaller than 0.013% for the composing multiplier/divider block.

Keywords - analog signal processing, approximation function, computational circuits, multiplier/divider

I. INTRODUCTION

An original implementation of a current-mode analog function synthesizer circuit will be presented. The proposed structure is based on a particular approximation function, being able to generate any continuous mathematical function. It presents the important advantages of an improved accuracy and of excellent reconfiguration properties.

It exists in literature a relatively limited number of analog function synthesizers implemented in CMOS technology [1] - [6], most applications being dedicated exclusively to the design of exponential [7] - [9], squaring, square-rooting or Euclidean distance functions. In [1], the approximation of the implemented function can be implemented by adding the weighted output currents of a number of basic building blocks, built around a basic current squarer, and a constant current, the circuit presenting the disadvantage of a relatively large complexity. The circuit proposed in [2] is based on approximating the required function using the first three terms of its Taylor series expansion. The approximations can be implemented by adding the output currents of a weighted current square, a weighted current amplifier (or attenuator) and a constant current. The errors are relatively large and, additionally, the range of the input variable is restricted. It was proposed in [3] a structure for synthesis of analog exponential functions, based on approximating the exponential function using rational functions. The circuit presents the important limitations of realizing only the exponential function and of an implementation in bipolar technology.

II. MATHEMATICAL ANALYSIS OF THE APPROXIMATION FUNCTION

In order to implement an improved accuracy analog function synthesizer circuit, the proposed method is to use a general continuous approximation function, $g(x)$, having a Taylor series that can be made to match a multitude of continuous mathematical functions, $f(x)$. The original function $g(x)$ that can approximate many continuous mathematical functions $f(x)$ has the following expression:

$$g(x)=\frac{p^2}{q}\frac{x}{1-\frac{q}{p}x}+\left(n-\frac{p^2}{q}\right)x+m \qquad (1)$$

m, n, p and q being constant parameters, depending on the particular function that must be approximated. In order to evaluate the capability of $g(x)$ function to approximate a continuous function $f(x)$, the Taylor series expansion for $g(x)$ must be determined.

III. IMPLEMENTATION OF THE FUNCTION SYNTHESIZER CIRCUIT BASED ON THE PROPOSED APPROXIMATION FUNCTION

A. The block diagram of the function synthesizer circuit

The proposed block diagram of the original function synthesizer circuit is presented in Figure 1.

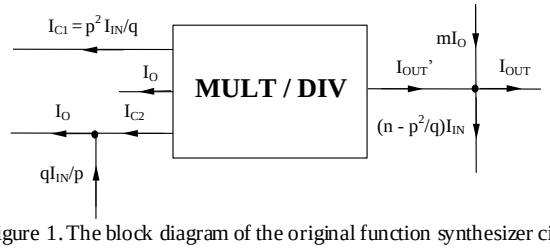


Figure 1. The block diagram of the original function synthesizer circuit

The core of the structure is represented by the "MULT / DIV" block, that must have the following expression of the output current I_{OUT}':

$$I_{OUT}' = I_O \frac{I_{C1}}{I_{C2}} \tag{2}$$

I_{C1} and I_{C2} represent the input currents of the "MULT / DIV" block, while I_O is a reference current. So:

$$I_{OUT}' = I_O \frac{(p^2/q)I_{IN}}{I_O - qI_{IN}/p} = I_O \frac{\frac{p^2}{q}\left(\frac{I_{IN}}{I_O}\right)}{1-\frac{q}{p}\left(\frac{I_{IN}}{I_O}\right)} \tag{3}$$

The output current of the circuit having the block diagram presented in Figure 1 will have the following expression:

$$I_{OUT} = I_{OUT}' - \left(n - \frac{p^2}{q}\right) I_{IN} + mI_O \tag{4}$$

Using the notation $x = I_{IN}/I_O$ and relation (1), it results that I_{OUT} current represents the third-order approximation of $f(x)$ function:

$$I_{OUT} = I_O g(x) = I_O g\left(\frac{I_{IN}}{I_O}\right) \cong I_O f(x) \tag{5}$$

B. The implementation of the "MULT / DIV" circuit

The proposed method for designing a current multiplier/divider circuit uses as circuit cores two current squaring circuits (Figure 2).

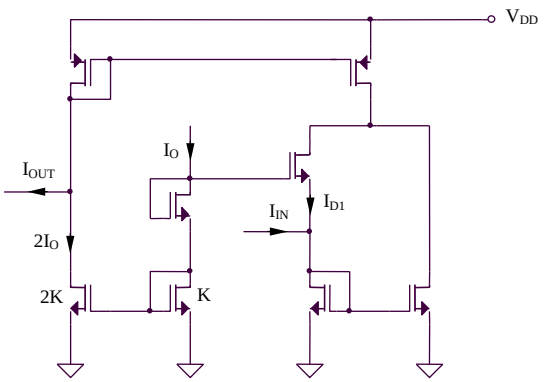


Figure 2. The implementation of the "SQ" block

For MOS transistors from Figure 2, the equation of the translinear loop can be expressed as follows:

$$2V_{GS}(I_O) = V_{GS}(I_{D1}) + V_{GS}(I_{D1} + I_{IN}) \tag{6}$$

The expression of the output current will be:

$$I_{OUT} = 2I_{D1} + I_{IN} - 2I_O = \frac{I_{IN}^2}{16I_O} \tag{7}$$

where:

$$I_{D1} = I_2 - (I_1 - I_O) + \frac{(I_1 - I_O)^2}{4I_2} \tag{8}$$

and:

$$I_{D2} = I_2 - (I_1 + I_O) + \frac{(I_1 + I_O)^2}{4I_2} \tag{9}$$

The original proposed implementation of the "MULT / DIV" block is presented in Figure 3. The expression of the output current for the circuit presented in Figure 3 will be:

$$I_{OUT} = I_{D2} - I_{D1} + 2I_O = I_O \frac{I_1}{I_2} \tag{10}$$

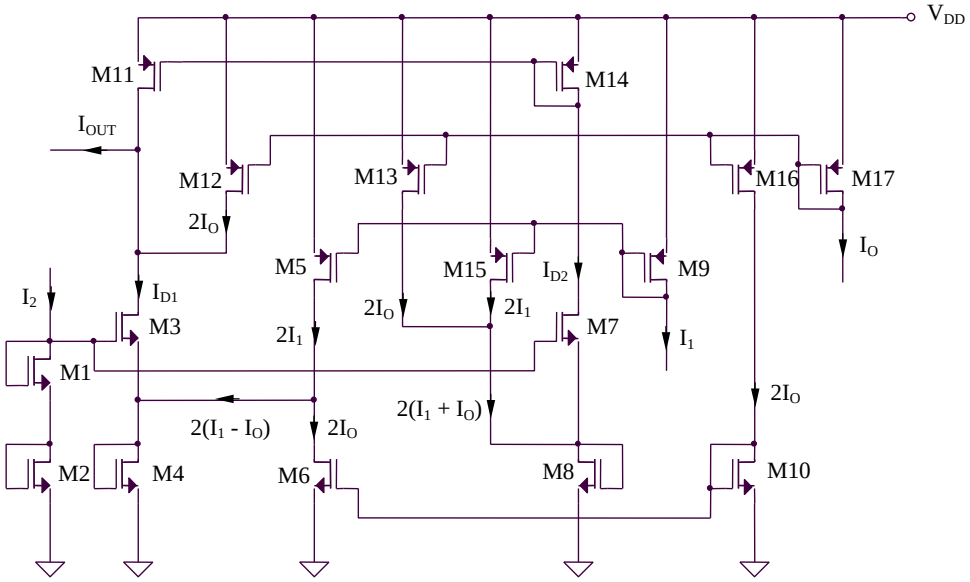


Figure 3. The implementation of the "MULT / DIV" block

The aspect ratios of MOS transistors are presented in Table 1.

TABLE 1. THE ASPECT RATIOS OF MOS TRANSISTORS FROM FIGURE 3

M	M1	M2	M3	M4	M5	M6	M7	M8	M9
W/L	4.5/ 0.9	4.5/ 0.9	4.5/ 0.9	9/ 0.9	4.5/ 0.9	4.5/ 0.9	4.5/ 0.9	4.5/ 0.9	4.5/ 0.9

M	M10	M11	M12	M13	M14	M15	M16	M17
W/L	4.5/ 0.9	4.5/ 0.9	4.5/ 0.9	4.5/ 0.9	4.5/ 0.9	2.25/ 0.9	4.5/ 0.9	4.5/ 0.9

The frequency range of the proposed multiplier/divider circuit is estimated to be very good, as a result of using exclusively MOS transistors biased in saturation region, correlated with a current-mode operation of the circuit.

IV. SIMULATIONS

The SPICE simulation $I_{OUT}(I_{IN})$ for the squaring circuit proposed in Figure 2 is presented in Figure 4. The I_O current is equal with $1mA$, while the range of I_{IN} current was chosen to be between 0 and $3mA$.

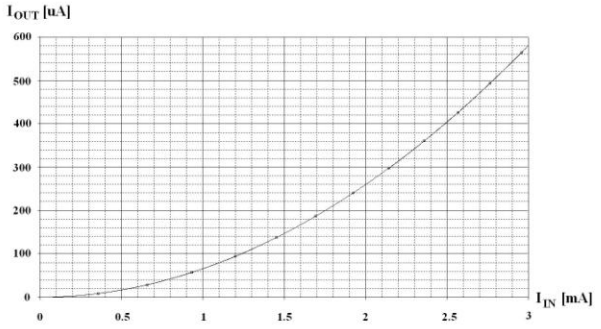


Figure 4. The SPICE simulation $I_{OUT}(I_{IN})$ for the squaring circuit

The simulated approximation error $\varepsilon_{SQ}(I_{IN})$ for the squaring circuit from Figure 2 is shown in Figure 5. The error is smaller than 0.01% for an extended range of the input current.

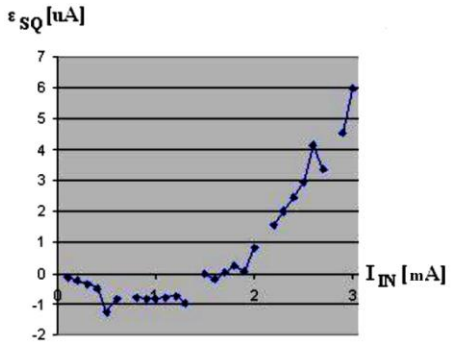


Figure 5. The simulated approximation error $\varepsilon_{SQ}(I_{IN})$ for the squaring circuit

The SPICE simulation $I_{OUT}(I_1)$ for the multiplier/divider circuit proposed in Figure 3 is presented in Figure 6. The I_O and I_2 currents have the following values: $I_O = 20\mu A$ and $I_2 = 20\mu A$, while the range of I_1 current was chosen to be between 0 and $20\mu A$.

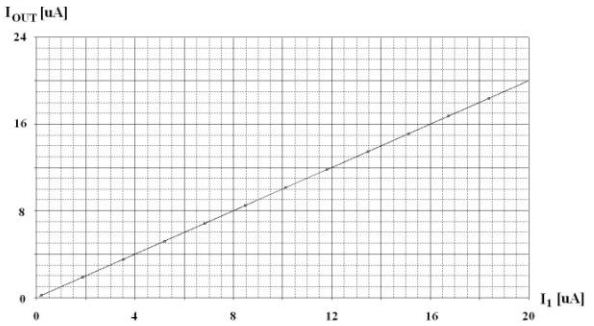


Figure 6. The SPICE simulation $I_{OUT}(I_1)$ for the multiplier circuit

The simulated linearity error of $I_{OUT}(I_1)$ characteristic for the multiplier/divider circuit from Figure 3 is shown in Figure 7. The linearity error is smaller than 0.013% for $I_1 = 20\mu A$.

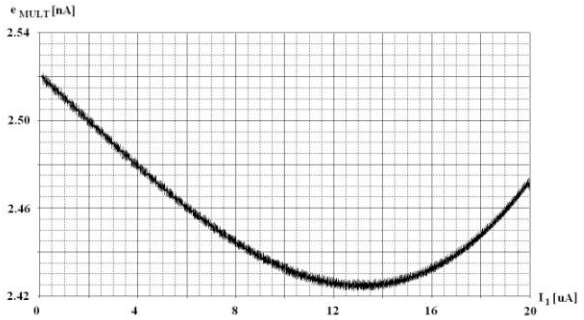


Figure 7. The simulated linearity error of $I_{OUT}(I_1)$ characteristic for the multiplier/divider circuit

For the same multiplier/divider circuit proposed in Figure 3, a transient analysis was performed. The I_O current is a sinusoidal current with an amplitude of $40\mu A$ and a frequency equal with $1kHz$, while I_1 current is a sinusoidal current having an amplitude of $0.5mA$ and a frequency of $30kHz$. The I_2 current is a continuous current, equal with $1mA$. The SPICE simulation of the output current is presented in Figure 8.

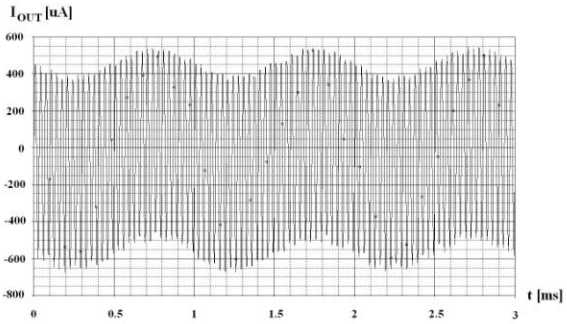


Figure 8. The SPICE simulation $I_{OUT}(t)$ for the multiplier/divider

The SPICE simulation $I_{OUT}(I_2)$ for the multiplier/divider circuit proposed in Figure 3 is

presented in Figure 9. The I_O and I_1 currents have the following values: $I_O = 20\mu A$ and $I_1 = 50\mu A$, while the range of I_2 current was chosen to be between 0 and $4mA$.

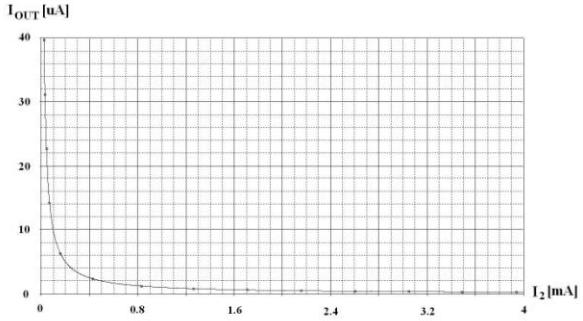


Figure 9. The SPICE simulation $I_{OUT}(I_2)$ for the multiplier circuit

A comparison between the simulated and the theoretical estimated results for the proposed multiplier/divider circuit is shown in Table 2.

TABLE 2. COMPARISON BETWEEN THE SIMULATED AND THE THEORETICAL ESTIMATED RESULTS FOR THE MULTIPLIER/DIVIDER CIRCUIT

I_2 (mA)	$I_{OUT\ th.}$ (μA)	$I_{OUT\ sim}$ (μA)
0.2	5.0000	4.9836
0.4	2.5000	2.4958
0.6	1.6667	1.6582
0.8	1.2500	1.2455
1.0	1.0000	0.9983
1.2	0.8333	0.8292
1.4	0.7143	0.7087
1.6	0.6250	0.5903
1.8	0.5555	0.5502
2.0	0.5000	0.4964
2.2	0.4545	0.4504
2.4	0.4167	0.4132
2.6	0.3846	0.3802
2.8	0.3571	0.3533
3.0	0.3333	0.3292
3.2	0.3125	0.3085
3.4	0.2941	0.2892
3.6	0.2778	0.2734
3.8	0.2632	0.2584
4.0	0.2500	0.2454

V. CONCLUSIONS

It was proposed, analyzed and implemented an original analog function synthesizer circuit The original structure can implement any continuous mathematical function, the circuit accuracy being excellent. Because of the third-order approximation of the computational structure, the range of input variables is relatively large comparing with previously reported similar circuits. An important additional improvement of the structure accuracy was obtained using a current-mode operation and providing an independence of the circuit performances on technological errors. The possibility of realizing a multitude of circuit functions is based on the increased modularity and controllability of the proposed function synthesizer circuit. The supply voltage was $1.8V$, while the implementing technology was $0.18\mu m$ CMOS. The theoretical estimated results are confirmed by SPICE simulations, resulting an approximating error smaller than 0.01% for the composing squaring block and smaller than 0.013% for the composing multiplier/divider block. Comparing with similar designs, the proposed multiplier/divider structure has better performances related to its linearity and to the maximal range of the input currents.

REFERENCES

[1] Cosmin Popa, "Current-Mode Analog Nonlinear Function Synthesizer Structures", Springer Publishing House, 2013, ISBN 978-3319010342

[2] Cosmin Popa, "High-Accuracy Function Synthesizer Circuit with Applications in Signal Processing", EURASIP Journal on Advanced in Signal Processing, 2012

[3] "A translinear current-mode programmable analog exponential function synthesizer", Abuelma'atti, M.T., Microelectronics, 1999. ICM '99. The Eleventh International Conference on, 2000, pp. 209 - 212

[4] Abuelma'atti, M.T., Al-Yahia, N.M., "An improved universal CMOS current-mode analog function synthesizer: Performance analysis", Circuits and Systems, 2008. APCCAS 2008. IEEE Asia Pacific Conference on, 2008, pp. 1196 - 1199

[5] Fattaruso, J.W., Meyer, R.G., "MOS analog function synthesis", Solid-State Circuits, IEEE Journal of, Vol. 22, 1987, pp. 1056 - 1063

[6] Jong-Kug Seon, Charlot, J.-J., "Design and applications of precise analog computational circuits", Circuits and Systems, 1999. 42nd Midwest Symposium on, Vol. 1, 1999, pp.275 - 278 vol. 1

[7] Jia Li, Fengyi Huang, Xinrong Hu, Xusheng Tang, "A 1GHz, 68dB CMOS variable gain amplifier with an exponential-function circuit", Signals Systems and Electronics (ISSSE), 2010 International Symposium on, Vol.2, 2010, pp. 1 - 4

[8] Raikos, G., Vlassis, S., "Low-voltage CMOS voltage squarer", Electronics, Circuits, and Systems, 2009. ICECS 2009. 16th IEEE International Conference on, 2009, pp.159 - 162

[9] Psychalinos, C., Vlassis, "A systematic design procedure for square-root-domain circuits based on the signal flow graph approach", S., Circuits and Systems I: Fundamental Theory and Applications, IEEE Transactions on, Vol. 49, Issue: 12, Publication, 2002, pp. 1702 – 1712

Analysis and optimization of a SAR ADC with Attenuation Capacitor

S. Brenna, A. Bonfanti, A. Abba, F. Caponio, A. L. Lacaita
Dip. di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Milano, Italy
email: stefano.brenna@polimi.it

***Abstract*—The conventional binary weighted array SAR ADC is the common topology adopted to achieve high efficiency conversion, i.e. with less than 10 fJ/conversion-step, even requiring extra effort to design and simulate full custom sub-fF capacitors. This paper presents the design and the optimization of an asynchronous SAR ADC with attenuation capacitor achieving an efficiency similar to conventional binary weighted array converters but adopting standard MiM capacitors. A monotonic switching algorithm further reduces the capacitive array consumption while an asynchronous and fully-differential dynamic logic minimizes the digital power consumption. A 10-bit prototype has been fabricated in a 0.13-μm CMOS technology. At 0.5-V supply and 200-kSps sampling frequency, the ADC achieves a SNDR of 52.6 dB, an ENOB of 8.45, and a power consumption of 420 nW, corresponding to a figure-of-merit (FOM) of 6 fJ/conversion-step. This efficiency is comparable to the best results published so far and it's the lowest among ADCs in 130-nm or less scaled technology. The ADC core occupies an area of only 0.045 mm^2.**

***Index Terms*—ADC, analog-to-digital conversion, nonlinearity, mismatch, asynchronous logic, successive approximation register.**

I. Introduction

In energy-limited applications, such as wireless sensor nodes, implantable medical devices or portable amusements, the adoption of ultra-low power circuits is mandatory in order to extent the system battery lifetime. Analog-to-digital converters (ADCs) featuring moderate sampling rate ($0.01 - 1$ Msps) and resolution (8-10 bit) are key components in such devices. Among different converter architectures, successive approximation analog-to-digital converter (SA-ADC) is the best choice due to its good trade-off among power efficiency, conversion accuracy and design complexity. In such converters, the primary sources of power consumption are the digital control circuit and the capacitive DAC array. While the digital power consumption benefits on the technology advancement, the power consumption due to the capacitive array is limited by the capacitor mismatch that is almost technology-independent. For this reason, a great number of DAC topologies and switching algorithms have been proposed in order to reduce DAC power consumption without penalty in terms of accuracy [1]–[3]. The latest trend is to rely on the high linearity properties of the conventional binary weighted (CBW) array adopting full-custom unit capacitance in the sub-fF range [4]–[6]. In fact, the minimum capacitor value supplied by a general-purpose design-kit is much larger than necessary to meet the linearity requirements, resulting in a considerably large array capacitance and hence in a high switching power. Thus, this approach requires extra-efforts to model and design the highly-matched unit capacitance or an error correction technique, the latter increasing area and circuit complexity. The purpose of this work is to demonstrate that a binary weighted with attenuation capacitor (BWA) converter [1] can achieve efficiency well below 10 fJ/conversion-step and a remarkable compactness, without requiring the design, modeling and accurate simulations of custom capacitors. The DAC power consumption is further reduced by applying to the BWA array an efficient switching procedure as the monotonic algorithm proposed in [3]. Moreover, an asynchronous and fully-differential dynamic logic is proposed. Thanks to the decreased number of transistors and the differential architecture, the digital power consumption is reduced helping to improve the converter efficiency. The proposed 10-bit SAR converter, shown in Fig. 1, has been integrated in $0.13 - \mu$m CMOS technology. At 0.5-V supply, the ADC achieves an efficiency of 6 fJ/conversion-step, in line with the best conventional binary weighted topologies but adopting as unit element a standard MiM capacitance of 34 fF.

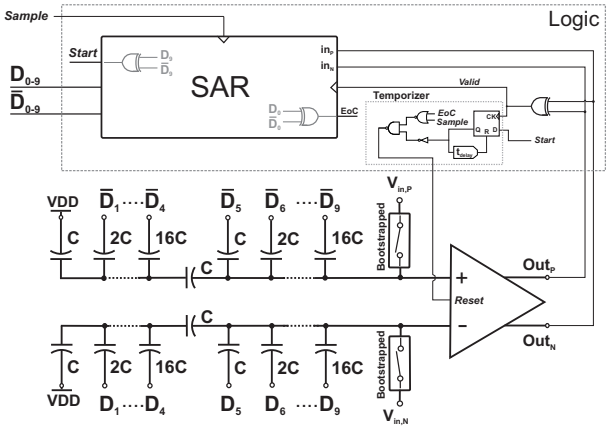


Fig. 1: Schematic of the proposed 10-bit converter.

The paper is organized as follows. Section II is devoted to a comparison between the traditional CBW and the bridge-capacitor DAC architecture in terms of power consumption and linearity due to statistical mismatch. The role of the parasitic capacitances in the BWA topology is also highlighted. Section III describes in detail the implementation of the switched capacitor network, of the asynchronous dynamic logic and of the dynamic comparator, while measurement results are shown in Section IV. Finally, conclusions are drawn in Section V.

II. Comparison between CBW and BWA topologies

The capacitive array typically serves as both sampling capacitance and feedback DAC, its linearity usually limiting the overall converter performance. The single-ended conventional N-bit binary weighted array is depicted in Fig. 2(a), where C_u is the unit capacitance. The main alternative to this

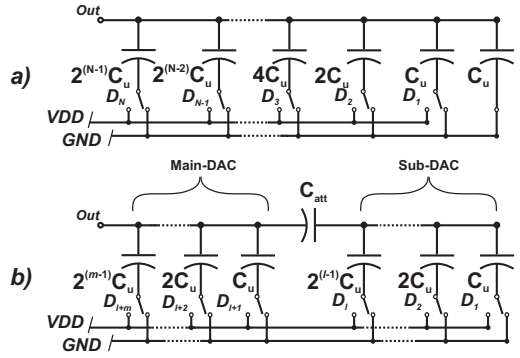


Fig. 2: N-bit CBW (a) and $(m+l)$-bit BWA (b) capacitive DAC.

TABLE I: COMPARISON OF THE CBW AND BWA ARRAYS

	CBW	BWA
C_{tot}	$2^N \cdot C_u$	$\left(2 \cdot (2^{\frac{N}{2}} - 1) + 1\right) \cdot C_u$
$\sigma_{DNL,max}$	$2^{\frac{N}{2}} \cdot \sigma\left(\frac{\Delta C}{C_u}\right)$	$2^{\frac{3N}{4}} \cdot \sigma\left(\frac{\Delta C}{C_u}\right)$
$\sigma_{INL,max}$	$2^{\frac{N}{2}-1} \cdot \sigma\left(\frac{\Delta C}{C_u}\right)$	$2^{\frac{3N}{4}-1} \cdot \sigma\left(\frac{\Delta C}{C_u}\right)$

structure is the binary weighted with attenuation (or bridge) capacitor array [1], shown Fig. 2(b). It features an attenuation capacitor, C_{att}, in order to divide the array into two binary weighted sub-arrays: a main-DAC and a sub-DAC of m and l capacitors, respectively. In particular, since this topology is intended to reduce the overall capacitance, we'll refer to the BWA architecture with equal main- and sub-DACs (i.e. m=l=$N/2$) and $C_{att} \cong C_u$, which has been shown to be the most energy efficient among possible combinations [2]. In this section, the impact of the capacitance mismatch and parasitics on the performance of both CBW and BWA arrays is accurately handled.

A. Capacitive mismatch analysis

The effect of capacitive mismatch in the DAC array is to degrade the linearity of the overall converter. The nonlinearity is usually quantified in term of differential nonlinearity (DNL), integral nonlinearity (INL) and effective number of bits (ENOB). However, while a precise formulation of the relationship between capacitive mismatch and ENOB is still lacking, the maximum standard deviations of DNL ($\sigma_{DNL,max}$) and INL ($\sigma_{INL,max}$) have been analytically derived as function of the unit capacitance relative standard deviation $\sigma\left(\frac{\Delta C}{C_u}\right)$ for the most common adopted capacitive arrays [2]. Table I shows the expressions of the overall capacitance, together with $\sigma_{DNL,max}$ and $\sigma_{INL,max}$, as function of $\sigma\left(\frac{\Delta C}{C_u}\right)$ [2], [7], for a N-bit BWA and CBW single-ended array. The standard deviation of the unit capacitor can be expressed in terms of Pelgrom mismatch coefficient, k_c, and specific capacitance, c_{spec}, being

$$\sigma\left(\frac{\Delta C}{C_u}\right) = k_c \cdot \sqrt{\frac{c_{spec}}{2C_u}}, \tag{1}$$

where the factor 2 takes into account that ΔC is referred to the single capacitance with respect to the nominal value. Considering the same number of bits and the same unit capacitance for both the topologies, the single-ended BWA array features an overall capacitance which is approximately a factor $2^{\frac{N}{2}-1}$ lower than in the CBW architecture.

Unfortunately, the effect of mismatch on static nonlinearity is a factor $2^{\frac{N}{4}}$ larger in the BWA than in the CBW array.

TABLE II: CAPACITANCE COMPARISON

	k_c $[\% \cdot \mu m]$	c_{spec} $[fF/\mu m^2]$	$k_c^2 c_{spec}$ $[\%^2 \cdot fF]$
Custom MoM (lateral) [10]	4	0.192	3.07
Custom MoM (1 layer) [11]	53	0.12	337
Custom MoM (2 layers) [11]	32	0.24	242
Custom MoM (lateral) [6]	0.5	0.25	0.1
Standard MiM (130 nM UMC)	0.95	1	0.9
Standard MiM (65 nM ST)	0.5	5	1.25

Thus, to achieve the same $\sigma_{DNL,max}$ the unit capacitance in the BWA array has to be a factor $2^{\frac{N}{2}}$ larger than in the conventional array. In this case, the overall array capacitance of the BWA network is approximately twice that of a CBW array, independently on the number of bits. In [2] the average switching energy for the CBW and the BWA topologies is analytically derived. In fact, when sized to have the same $\sigma_{DNL,max}$, the average switching energy consumption of the conventional array is a factor 1.91 lower than in the BWA topology.

Now, let us suppose to size a 10-bit capacitive array of a SA-ADC adopting a technology with k_c=$1\% \cdot \mu m$ and c_{spec}= $1fF/\mu m^2$. In order to achieve $3\sigma_{DNL,max}$<0.5 [5], [8], a unit capacitance of about 59 and 1.8 fF is needed for the BWA and CBW array, respectively. Moreover, the latest trend is to shrink further the value of the unit capacitance to reduce as much as possible area and power consumption [3], [6], [9], relying on the fact that ENOB is not compromised even with larger $\sigma_{DNL,max}$ [4]. Since capacitance smaller than 10 fF are not available among standard design-kit MiM and poly capacitors, the CBW topology requires a custom design of the array capacitors and an extra effort for their characterization, which needs dedicated CAD tools [6]. Table II shows a comparison of recent published custom capacitors with MiM capacitances from the 130-nm UMC and 65-nm ST-Microelectronics design-kit in terms of both Pelgrom coefficient k_c and specific capacitance c_{spec}. It's quite evident that custom capacitors can feature significantly worse matching properties with respect to standard MiM capacitances. The only exception is the custom lateral MoM capacitances presented in [5], which shows even better matching properties but having required a sophisticated simulation tool.

Thus, the unit capacitance in the BWA and CBW arrays that assures the same nonlinearity has to be sized taking into account the different k_c and c_{spec} parameters of custom and standard capacitors, leading to a ratio between the overall array capacitances of

$$\frac{C_{tot,BWA}}{C_{tot,CBW}} \cong 2\frac{\left(k_c^2 c_{spec}\right)_{standard}}{\left(k_c^2 c_{spec}\right)_{custom}}. \tag{2}$$

The product $k_c^2 c_{spec}$ for all the available custom capacitors as well for the standard MiM capacitances is also shown in Table II. The adoption of the custom capacitors proposed in [10], [11] results in a larger capacitance, and thus power consumption, of the CBW array than of the BWA network featuring standard MiMs. Only with the custom capacitance design in [6] the CBW array overwhelms the BWA architecture. In addition, Table II shows that custom capacitors typically present a specific capacitance much lower than those achievable with standard MiMs, even by a factor of 10,

resulting inefficient in terms of area and making more critical the array layout.

B. Effect of parasitic capacitances

Let us now discuss the impact of the parasitic capacitances on the linearity performance of the two considered DAC topologies. It is commonly assumed [2] that the parasitic capacitance connected to the top-plate nodes can degrade the linearity performance of the BWA converter, even if constant and voltage independent, while in the conventional array it only causes gain error without affecting the linearity. Let us try to be more quantitative expressing the voltage at the output node of the BWA array in Fig. 2(b) with $m{=}l{=}N/2$ and $C_{att}{=}C_u$. By indicating with C_{main} and C_{sub} the overall capacitance of the main- and the sub-DAC, respectively, and with $C_{par,main}$ and $C_{par,sub}$ the parasitic capacitances at the top-plate node of the corresponding array, the analog output voltage results

$$V_{out} \cong \left[\frac{\sum_{i=\frac{N}{2}+1}^{N} D_i \cdot C_i}{C_{main} + C_{par,main}} + AR \cdot \frac{\sum_{i=1}^{\frac{N}{2}} D_i \cdot C_i}{C_{sub} + C_{par,sub}} \right] V_{DD}, \tag{3}$$

C_i being the capacitance associated to the bit D_i and AR the attenuation ratio

$$AR \cong \frac{C_u}{C_{main} + C_{par,main}}. \tag{4}$$

Equations (3) and (4) show that only the parasitics $C_{par,sub}$ affects the linearity, since its effect on the value of the DAC output voltage is not constant for different input signals, while $C_{par,main}$ only causes gain error. In particular, $C_{par,sub}$ is responsible of a deterministic pattern of the DNL, and hence of the INL. The differential nonlinearity shows a peak every $2^{\frac{N}{2}}$ codes whose amplitude is

$$DNL_{peak} \cong \frac{\left(2^N - 2^{\frac{N}{2}}\right) C_{par,sub} + C_u}{2^N C_u}. \tag{5}$$

Eq. (5) highlights that $C_{par,sub}$ has to be lower than C_u in order to assure a monotonic behavior of the converter, i.e $DNL{<}1$. In order to verify the accuracy of the proposed analysis, a transistor-level simulation has been performed on a 10-bit single-ended BWA converter with a unit capacitance of 100-fF and a $C_{sub,par}$ of 50 fF. The simulation results are shown in Fig. 3. As expected, the DNL has a periodic pattern featuring a peak every 32, i.e $2^{\frac{N}{2}}$, codes of about 0.48 LSB, close to 0.49 as predicted by (5), while the INL is the range -0.5/0.5. The simulated ENOB is 9.85, suggesting that the parasitic capacitance at the top-plate of the sub-DAC can be as large as C_u without determining a severe ENOB drop.

III. Circuit design

The scheme of the proposed 10-bit AD converter is shown in Fig. 1. In order to achieve a better common-mode noise rejection and less distortion, a fully-differential topology has been adopted.

A. Capacitive array and switching algorithm

Being the converter fully-differential, the capacitive DAC is realized by means of two 10-bit binary weighted arrays with attenuation capacitor, one per branch. The fully-differential

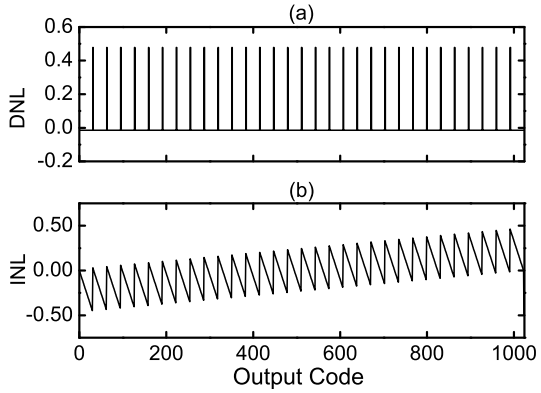


Fig. 3: Simulated DNL (a) and INL (b) for a 10-bit BWA DAC featuring C_u=100 fF and $C_{par,sub}$=50 fF.

structure allows to reduce $\sigma_{DNL,max}$ and $\sigma_{INL,max}$ by a factor of $\sqrt{2}$ with respect to the single-ended counterpart featuring the same unit capacitance [8], i.e. the same nonlinearity is obtained with half the unit capacitance. In order to further reduce the array power consumption, an efficient switching procedure, as the monotonic algorithm [3], has been applied to the capacitive DAC. In fact, the conventional trial-and-error search procedure [2], even if simple and intuitive, is not energy efficient, especially when unsuccessful trials occur.

The proposed capacitive array samples the differential input signal directly on the top of the two main-DACs via bootstrapped switches, which allow low-voltage operation, with the bottom-plates connected to the positive power supply, V_{DD}. After the switches are turned off, the first comparison is done without switching any array capacitance. According to the comparator output, the largest capacitor on the main-DAC corresponding to the positive input signal is switched to ground while the the other one remains at V_{DD}. The ADC repeats the procedure until the LSB is decided. For each cycle, only a capacitor is switched reducing the charge transfer and thus the array power consumption. Note that the proposed BWA array with monotonic switching scheme has the same overall capacitance of the original fully-differential BWA architecture, i.e. about $2^{\frac{N}{2}+2}C_u$. The monotonic switching scheme determines a significant efficiency improvement, reducing the average switching energy from 81.5 $C_u V_{DD}^2$ to roughly 32 $C_u V_{DD}^2$, i.e. by a factor of 2.5 with respect to traditional switching approaches.

Another advantage of the monotonic switching algorithm is that it reduces the effect of capacitor mismatch on the non-linearity of the converter. It can be verified following the same procedure adopted in [12] that the effect of mismatch on both $\sigma_{DNL,max}$ and $\sigma_{INL,max}$ is reduced by a factor $\sqrt{2}$ with respect to the traditional algorithm. This allows to adopt half the unit capacitance without impairing the linearity of the DAC, thus further reducing the array power consumption.

Based on statistical simulations, the unit capacitance was set to 34 fF, close to the technology minimum of 17 fF. Also the attenuation capacitance was set equal to C_u. With this value, the effect of capacitor mismatch is expected to keep the $\sigma_{DNL,max}$ lower than 0.5 LSB, while the total capacitance is 4.28 pF. A particular care was dedicated to the minimization of the sub-DAC top-plate parasitic capacitance, which was limited to about 15 fF, according to parasitic extraction tools.

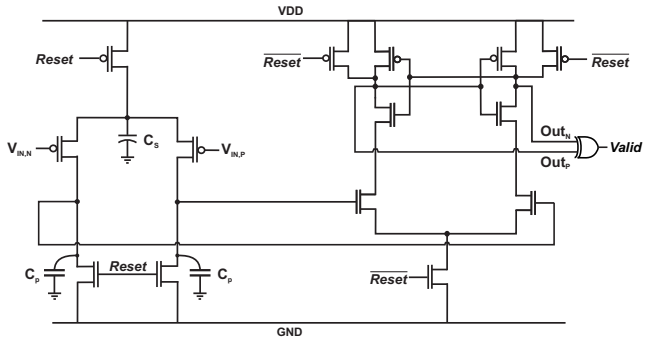


Fig. 4: Schematic of the dynamic comparator.

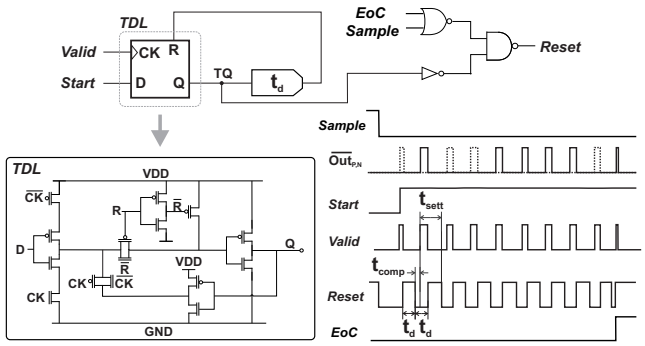


Fig. 5: Logic temporizer and timing diagram.

B. Dynamic Comparator

A two-stage dynamic comparator, shown in Fig. 4, has been adopted since it does not consume static current, thus being suitable for energy efficient design. It consists of a first stage similar to the one adopted in [9] followed by a differential latch. Since the monotonic algorithm makes the common-mode input voltage varying from $V_{DD}/2$ to 0 along the conversion cycle [3], the input stage features a PMOS diffetential pair. The operation is determined by the $Reset$ signal, which is generated by the logic circuit (see Fig. 1). Before the comparison takes part, the first stage output nodes are pre-charged low by a positive $Reset$ signal. Its falling edge stops the pre- charging phase and starts the amplification of the input signal. In fact, as $Reset$ becomes low, a current starts to flow into the differential pair charging the parasitic capacitances C_p at the drain nodes of the input transistors. The voltage on the two capacitors increases at a different speed, the difference depending on the input signal, $V_{in,p}$-$V_{in,n}$. As the first stage output nodes approximates the threshold voltage of the second stage input transistors, the latch starts to amplify the signal until the positive feedback takes over providing a rail-to-rail differential output. Consequently, the $Valid$ signal is pulled high to enable the asynchronous control logic. According to simulation, the first stage differential gain is about 5, high enough to make the noise of the second stage negligible. Thus, the equivalent input noise of the comparator is mainly determined by the input differential pair, its transistor working in weak-inversion region to maximize the comparator efficiency. The equivalent integrated input noise of the comparator is approximately [9]

$$v_{n,rms} \cong \sqrt{\frac{kT}{C_p}} \cdot \sqrt{\frac{8kT}{qV_T}}, \tag{6}$$

V_T being the threshold voltage of the second stage input transistors. In order to make this noise negligible with respect to the LSB for $V_{DD} = 0.5$ V, i.e. $(2 \cdot 0.5V)/2^N \cong 1$ mV, the parasitic capacitance has to be larger than approximately 3.3 fF. In the proposed comparator, the two parasitic capacitances are about 15 fF.

The comparator also determines the converter offset, which is mainly set by the first stage. The second stage contributes with both a static and dynamic offset [13]. However, its contribution to the comparator offset is negligible since if follows a gain stage. Even with a capacitance mismatch of 1 fF at the latch outputs, the input-referred offset is lower than 500 μV, i.e. half the LSB, while the static offset is lower than 250 μV rms. Moreover, the offset due to the second stage does not depend on the input signal, thus not affecting the converter linearity. As far as the first stage offset contribution, the main problem is that it depends on the common mode input voltage, which decreases during the conversion cycle due to the monotonic switching procedure. This dependence lead to distortion [3], [14]. In fact, the offset voltage of the comparator can be expressed as

$$V_{os} \cong \Delta V_T + \frac{V_{GS} - V_T}{2}\left(\frac{\Delta\beta}{\beta}\right), \tag{7}$$

where ΔV_T is the threshold voltage offset of the differential pair transistors, $V_{GS} - V_T$ is the effective voltage of the input pair and $\Delta\beta$ is the overall conductivity mismatch between the input transistors. In particular, the second term depends on the common mode voltage degrading the converter linearity. The simplest way to reduce its effect is to force these transistors to work in subthreshold region, thus minimizing the term $(V_{GS} - V_T)/2$, and to increase their area, slightly degrading the comparator power consumption performance.

Finally, a capacitor $C_s \cong 100$ fF has been added to the source of the first stage input transistors. Its effect is to reduce the kickback noise due to the falling edge of the $Reset$ signal. In fact, the PMOS current generator of the first stage must be sized large enough to guarantee an adequate current during the comparison. This can result in a capacitive coupling between its gate, its drain, and the capacitive arrays, which are connected to the comparator input nodes. In such conditions, at the first $Reset$ falling edge a charge is injected into the two DACs which determines a voltage drop at the top-plate of each array. If the two comparator input nodes are at the same voltage, the same amount of charge is induced on the two arrays since the two gate-source capacitances of input transistors are equal. However, when a differential signal is applied to the comparator, the two capacitances are different. This induces a signal- dependent charge on the two DACs, thus causing nonlinearity. The effect of the tail capacitance C_S is to reduce the amount of injected charge.

C. Asynchronous Logic

The SAR logic generates the necessary commands to control the comparator and the capacitive DAC. In order to reduce its power consumption, an asynchronous dynamic logic has been designed. By using a dynamic logic, less transistors are needed to implement the same functionality, while being asynchronous it requires only a low-speed sampling clock instead of an oversampled clock, thereby saving power.

The timing is assured by a logic temporizer (see Fig. 5) implemented using a dynamic latch (TDL) with a delayed

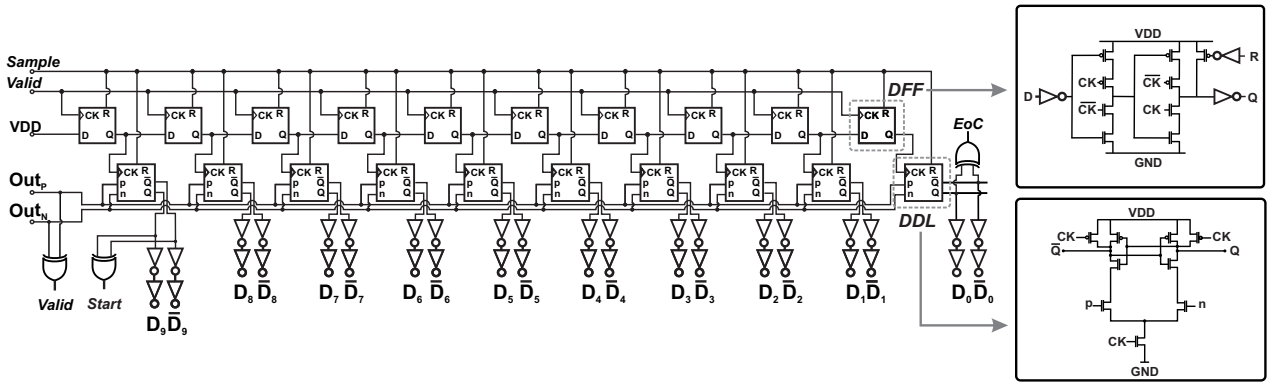


Fig. 6: Schematic of the asynchronous logic. The insets show the dynamic differential latch (DDL) and the dynamic flip-flop (DFF).

feedback loop. Its function is to enable the comparator, wait for its decision and then reset it for a time sufficient to assure the settling of the DAC voltage. Its working principle is briefly described in the following. At the end of the sampling phase, i.e. at the falling edge of the $Sample$ signal, the comparator evaluates the most significant bit (MSB) and enables the temporizer through the $Valid$ signal, which marks the end of each successful comparison. Only when the bottom plates of the MSB capacitors are switched, the $Start$ signal becomes high and remains in this state till the end of the conversion. Its rising edge triggers the first transition of the TDL output (TQ), causing the reset of the comparator ($Reset$=1) for a time $t_d \cong 75$ ns fixed by the delay unit and forcing $Valid$ to zero. At the end of the reset phase, the comparator evaluates the MSB-1 bit and sets high the $Valid$ signal. However, the comparator is not resetted till the falling edge of TQ has completed the feedback path, i.e. after a time t_d. Thus, the reset signal resembles a square wave with a period of $2t_d$. Since the comparator takes approximately $t_c \cong 10$ ns (see Fig. 5) for a decision, a time of $t_{sett} = 2t_d - t_{comp} \cong 55$ ns is left to the logic circuit and to the array to switch and settle, respectively, during each bit evaluation phase. When also the least significant bit (LSB) has been evaluated, the EoC rises and the conversion stops till the end of the successive sampling phase keeping the comparator in the reset state.

Fig. 6 shows the logic circuit implementing the successive approximation algorithm. It features a first row of dynamic flip-flops (DFFs) and a second row of dynamic differential latches (DDLs). The design goal was to minimizing the capacitive load of the most active signal lines, i.e. the $Valid$ signal and the outputs of flip-flops, latches and comparator, in order to reduce the power consumption. The DFFs show a C^2MOS structure, with the clock pin connected to 2 n-gates and 2 p-gates. Thus, each DFF output is loaded by 2 n-gates and 3 p-gates, while the DDL outputs are directly connected to a minimum area inverter that drives the corresponding array capacitors. Moreover, since the monotonic algorithm requires the bottom plates of all capacitors to remain at V_{DD} during the sampling phase, a key choice was to implement the DDL as a differential latch (Fig. 6), making possible to use a single set of 10 elements shared by both arrays, instead of 10 elements per each array. In fact, the outputs of the DDLs are kept high till the positive edge trigger performed by the related DFF output, then they switch according to the state of the comparator decision. Finally, the total differential capacitance loading the comparator is equal to 10x2 n-gates (plus 4 n- and 4 p-gates due to the XOR gate generating the $Valid$ signal).

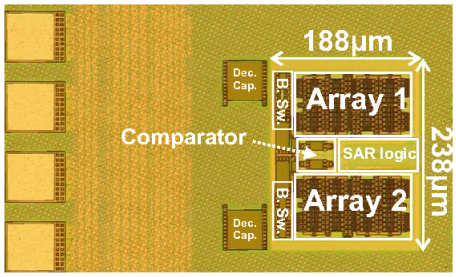


Fig. 7: Die photo of the proposed ADC.

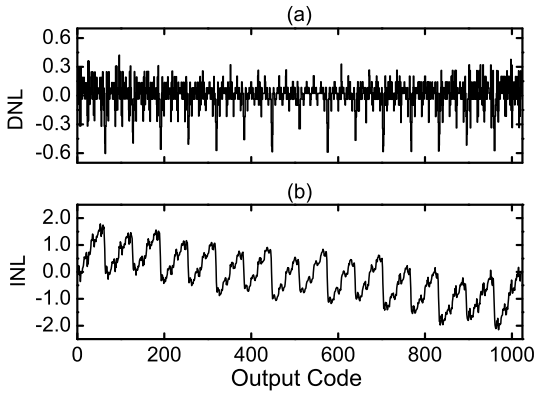


Fig. 8: Measured DNL and INL.

IV. Measurement Results

The ADC has been fabricated using a 2P8M 0.13-μm CMOS technology featuring 1.11-fF/μm^2 MiM capacitors. The converter occupies an area of 188 μm x 238 μm, as shown in Fig. 7. The performance of the ADC have been measured at 0.5-V supply for 8 samples. The average measured results are summarized in Table III.

The static performance of the ADC in terms of DNL and INL are shown in Fig. 8. The measured DNL and INL are -0.4/0.5 and -2/2 LSBs, respectively. Since each test chip shows a similar pattern and since the effect of the comparator has been accurately minimized, the DNL/INL performance are mainly due to the stray capacitances induced by the layout routing. This effect is well captured by accurate post-layout simulations.

Fig. 9 shows the output spectra at 0.5-V supply, 200-kSps sampling rate and for an input sinewave at 5.13-kHz and 96.48-kHz frequency, i.e. well below and slightly below the

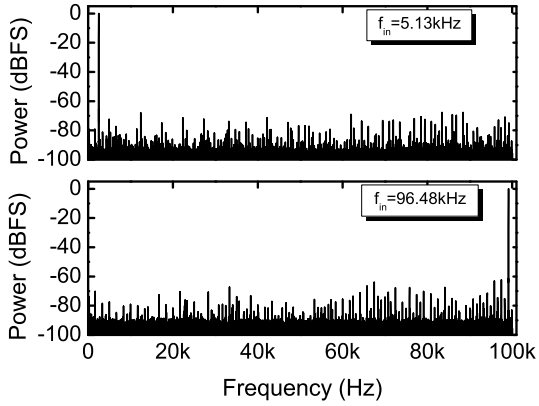


Fig. 9: Measured spectrum at 200-kSps sampling frequency with an input sine-wave at 5.13-kHz and 96.4-kHz frequency.

TABLE III: COMPARISON WITH THE-STATE-OF-THE-ART

	[1]	[9]	[3]	[6]	[8]	this work
Topology	BWA	CBW	CBW	CBW	CBW	BWA
Tech. (nm)	180	65	130	65	130	130
Unit cap. (fF)	120	0.5	4.8	0.25	13.5	34
Res. (bit)	10	10	10	10	10	10
Area (mm^2)	0.24	0.026	0.052	0.076	0.19	0.045
Voltage (V)	1	1	1.2	0.6	0.4†	0.5
Freq. (MSps)	0.1	1	50	0.04	1	0.2
ENOB (bit)	9.4	8.7	9.18	9.4	9.1	8.45
Power (μW)	3.8	1.9	826	0.072	0.052	0.42
FOM (fJ/step)	56	4.4	29	2.2	94.5	6

† A dual supply scheme is adopted: AVDD=1V, DVDD=0.4V.

Nyquist frequency. At low frequency, the average measured SNDR and SFDR are 52.6 and 67.5 dB, respectively. The resultant ENOB is 8.48 and its standard deviation considering the 8 tested samples is limited to 0.04. When the input frequency is increased near the Nyquist rate, the measured SNDR and SFDR drop to 50.8 and 62.1 dB, respectively.

The overall power consumption at 200-kSps sampling rate is 420 nW. Based on post-layout simulations, 50% of the power is due to the logic, 35% to the comparator and only 15% to the array. The leakage current is about 15 nA, thus becoming significant only at lower sampling rates.

To compare the proposed ADC with other works featuring different sampling rates and resolutions, the Figure-of-Merit (FOM) proposed in [15],

$$FOM = \frac{P_{diss}}{2^{ENOB} \cdot f_{sample}}, \tag{8}$$

is adopted. The FOM of the proposed ADC at the nominal supply of 0.5 V and for a 200-kSps sampling rate is 6 fJ/conversion-step. Table III shows that the proposed DAC well compares in terms of efficiency to the best recent published works, without adopting a CBW topology and sub-fF custom capacitors, and even being integrated in a older technology. It's worth noting that, since 85% of the power consumption is due to the digital circuits (asynchronous logic and comparator), the adoption of a more scaled technology would reduce the power consumption improving the efficiency of the converter. If compared to the converters implemented in 0.13-μm or less scaled technology, the proposed ADC is by far the best in terms of FOM. Finally, note that the proposed ADC favorably compares with the other converters in terms of area, with the exception of the work in [9] that adopts high-density custom capacitors and a conventional binary weighted array. The ADC proposed in [8] features a CBW array with 2 standard MIM capacitors connected in series in order to decrease the unit capacitance, but determining a large waste of area and an efficiency that is not so high as in other CBW realizations.

V. CONCLUSIONS

This paper presents a high efficient SA-ADC in 130-nm CMOS technology. It has been designed adopting a binary-weighted with attenuation capacitor array featuring a linearity and a total capacitance similar to a conventional binary weighted array but without requiring critical full-custom sub-fF capacitors. The design and the layout of the array has been accurately optimized in order to reduce the parasitic capacitances at the top-plate node of the sub-DAC which degrade the linearity. Moreover, an efficient switching scheme has been adopted in order to further reduce its power consumption. Finally, an asynchronous and fully-differential dynamic logic decreases the transistor count minimizing the digital power consumption. A prototype has been integrated and successfully tested showing an efficiency comparable to the state-of-the art converters even if realized in a less-scaled technology.

REFERENCES

[1] A. Agnes, E. Bonizzoni, and F. Maloberti, "A 9.4-ENOB 1V 3.8μW 100kS/s SAR ADC with time-domain comparator," *Dig. Tech. Papers Int. Solid State Circuits Conf.*, vol. 37, no. 2, pp. 246–610, Feb. 2008.

[2] M. Saberi *et al.*, "Analysis of power consumption and linearity in capacitive digital-to-analog converters used in successive approximation ADCs," *IEEE Trans. Circuits Syst. I: Reg. Paper*, vol. 58, no. 7, pp. 1736–1747, Aug. 2011.

[3] C. C. Liu *et al.*, "A 10-bit 50-MS/s SAR ADC with a monotonic capacitor switching procedure," *IEEE J. of Solid State Circuits*, vol. 45, no. 4, pp. 731–740, Apr. 2010.

[4] J. A. Fredenburg and M. P. Flynn, "Statistical analysis of ENOB and yield in binary weighted ADCs and DACs with random element mismatch," *IEEE Trans. Circuits Syst. I: Reg. Paper*, vol. 59, no. 7, pp. 1396–1408, Jul. 2012.

[5] P. Harpe *et al.*, "A 26 μW 8 bit 10 MS/s asynchronous SAR ADC for low energy radios," *IEEE J. of Solid State Circuits*, vol. 46, no. 7, pp. 1585–1595, Jul. 2012.

[6] P. Harpe, E. Cantatore, G. Haller, and B. Murmann, "A 2.2/2.7fJ/conversion-step 10/12b 40kS/s SAR ADC with data-driven noise reduction," *Dig. Tech. Papers Int. Solid State Circuits Conf.*, pp. 270–271, Feb. 2013.

[7] A. Chandrakasana and B. Ginsburg, "An energy-efficient charge recycling approach for a SAR converter with capacitive DAC," *Proc. Int. Symp. on Circ. and Syst. (ISCAS)*, pp. 184–187, May 2005.

[8] D. Zhang, A. Bhide, and A. Alvandpour, "A 53-nW 9.1-ENOB 1-kS/s SAR ADC in 0.13-μm CMOS for medical implant devices," *IEEE J. Solid State Circuits*, vol. 47, no. 7, pp. 1585–1593, Jul. 2012.

[9] M. van Elzakker *et al.*, "10-bit charge-redistribution ADC consuming 1.9μW at 1 MS/s," *IEEE J. of Solid State Circuits*, vol. 45, no. 5, pp. 1007–1015, May 2010.

[10] V. Tripath and B. Murmann, "Mismatch characterization of small metal fringe capacitors," *Proc. Cust. Int. Circ. Conf. (CICC)*, pp. 1–4, Sept. 2013.

[11] A. Abusleme, A. Dragone, G. Haller, and B. Murmann, "Mismatch of lateral field metal-oxide-metal capacitors in 180nm CMOS process," *IEEE Electronic Letters*, vol. 48, pp. 286–287, Mar. 2012.

[12] A. Chandrakasan and B. Ginsburg, "500-MS/s 5-bit ADC in 65-nm CMOS with split capacitor array DAC," *IEEE J. Solid State Circuits*, vol. 42, no. 4, pp. 739–747, Apr. 2007.

[13] A. Nikoozadeh and B. Murmann, "An analysis of latch comparator offset due load capacitor mismatch," *IEEE Trans. Circuits Syst. II: Exp. Briefs*, vol. 53, no. 12, pp. 1398–1402, Dec. 2006.

[14] H. Hong and G. Lee, "A 65-fJ/conversion-step 0.9-V 200-kS/s rail-to-rail 8-bit successive approximation ADC," *IEEE J. Solid State Circuits*, vol. 42, no. 10, pp. 2161–2168, Oct. 2007.

[15] R. H. Walden, "Analog to digital converter survey and analysis," *IEEE J. on Selected Areas in Communications*, vol. 17, pp. 539–550, Apr. 1999.

A Simulation and Modeling Environment for the Analysis and Design of Charge Redistribution DACs used in SAR ADCs

S. Brenna[1], A. Bonetti[2], A. Bonfanti [1], A. L. Lacaita[1]
[1] Dip. di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Milano, Italy
[2]ams International AG, Rapperswil, Switzerland
email: stefano.brenna@polimi.it, andrea.bonetti@ams.com

***Abstract*—The optimal design of SAR ADCs requires the accurate estimate of nonlinearity and parasitic effects in the feedback charge-redistribution DAC. Since the effects of both mismatch and stray capacitances depend on the specific array topology, complex calculations, custom modeling and heavy simulations in common circuit design environments are often required. This paper presents a novel MATLAB-based numerical tool to assist the design of classic, split and with attenuation capacitor binary weighted capacitive array topologies with an even number of bits from 6 to 14. The tool allows to perform both parametric and statistical simulations taking into account capacitive mismatch and parasitic capacitances in order to compute both differential- (DNL) and integral nonlinearity (INL). SNDR and ENoB degradation due to static non-linear effects is also estimated. An excellent agreement with the results obtained by the available circuit simulators (e.g. Cadence Spectre) is shown but featuring up to 10^4 shorter simulation time.**

***Index Terms*—Analog-to-digital conversion, numerical simulations, charge redistribution successive approximation registers.**

I. Introduction

In power-limited applications, such as portable instruments, wireless sensor networks or implantable biomedical systems, the use of ultra low-power analog-to-digital converters (ADCs) extends the battery life and limits heating phenomena. Due to their energy efficiency, successive approximation register (SAR) converters represent the best choice when moderate resolution and speed are required. The performance requirements of such converters are represented by static metrics, such as differential-non-linearity (DNL) and integral-non-linearity (INL), and by dynamic metrics, such as signal-to-noise-plus-distortion ratio (SNDR) and effective-number-of-bits (ENoB). All of these parameters mainly depend on the quality of the capacitive array, which implements the feedback charge-redistribution (CR) digital-to-analog converter (DAC). In the last decade, several novel CR-DAC topologies have been proposed for the design of SAR ADCs with outstanding efficiency [1], [2] and adopted in various integrated systems demonstrating their suitability for general purpose applications [3] [4]. The capacitive DAC plays a dominant role in degrading the overall performance of a SAR ADC as long as it is affected by mismatch and parasitic capacitances. Mismatch effects on DNL and INL can be considered statistical and they are well known for the most common array topologies [5]. Differently, the parasitic capacitance contribution is deterministic and strongly depends on the array architecture.

However, precise relationships that quantify the effects of both capacitor mismatch and parasitic capacitances on static and dynamic metrics still lacks. Thus, the designer has to size the DAC unit capacitance handling complex equations to target the desired requirements and quantify the converter nonlinearities running long transient simulations. Unfortunately, in common circuit design and simulation environments, like Cadence, such simulations are quite time-consuming and require heavily post-processing of the raw data, limiting the design time of the converter. Moreover, statistical simulations, like MonteCarlo analyses, are quite impractical since they require a large number of runs, at least 100, to achieve confident results.

This paper presents a MATLAB-based numerical tool to assist the design of the capacitive DAC adopted in SAR ADCs. The proposed tool implements the model of the three most common DAC topologies, such as the Classic Binary Weighted (CBW) [6], the Split Binary Weighted (SBW) [7] and the Binary Weighted with Attenuation Capacitor (BWA) [8] array, both single-ended and fully-differential, and allows to choose the switching algorithm between the conventional [6] and the monotonic scheme [9]. The numerical models take into account all possible mismatch and parasitic effects, giving the possibility to estimate by means of both statistical and parametric simulations the converter DNL, INL, SNDR, ENoB and power consumption.

The paper is organized as follows. In Section II, a brief overview of the array topologies is given, resuming the effects of mismatch and parasitic capacitances on both DNL and INL. Section III describes how the different array models are implemented and simulated with the proposed tool. Section IV shows the most significant results, which are in agreement with Cadence Spectre analyses but achieved with a considerable time saving. Finally, conclusions are drawn in Section V.

II. Converter topologies

In a generic N-bit capacitive DAC, as the one shown in Fig. 1, each digital code is associated with a particular configuration of the bottom switches that produces the corresponding output voltage. In an equivalent SAR ADC implemented by means of such a DAC in the feedback path, this voltage marks an input transition level between two adjacent digital codes. It derives that the conversion accuracy depends on the actual size of each capacitive bank. As a consequence of the technologic mismatch and parasitics, the capacitance of each block can differ from its nominal value and can be in general expressed as

$$C_i = 2^{i-1}C_u + C_{par,i} + \sum_{j=1}^{2^{i-1}} \delta_j \qquad i = 1, ..., N \qquad (1)$$

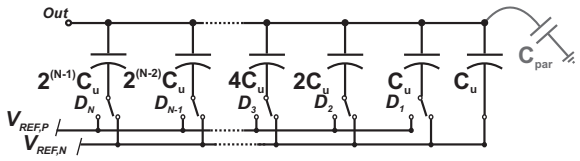


Fig. 1: Schematic of a N-bit CBW array.

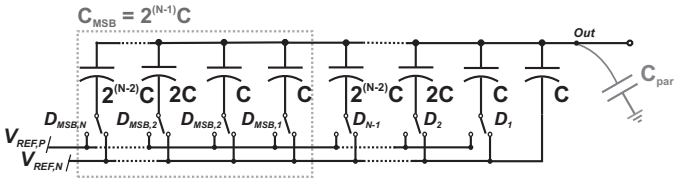


Fig. 2: Schematic of a N-bit SBW array.

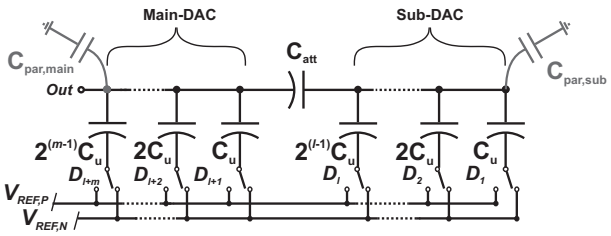


Fig. 3: Schematic of a $(m+l)$-bit BWA array.

where $C_{par,i}$ is the parasitic capacitance related to i^{th} capacitive block and δ_j is the mismatch equivalent capacitance affecting the unit capacitor. The effect of the parasitic capacitances $C_{par,i}$ can be considered deterministic, depending on layout inaccuracies, capacitor geometry and wirings. On the contrary, the capacitor mismatch can be modeled as a Gaussian distribution of the unit capacitor value with a mean equal to the nominal capacitance, C_u, and a standard deviation equal to [10]

$$\sigma_C = \frac{C_u k_c}{\sqrt{2A}} = k_c \cdot \sqrt{\frac{c_{spec} \cdot C_u}{2}}, \tag{2}$$

k_c, C_u, A and c_{spec} being the Pelgrom mismatch coefficient, the unit capacitance, the area and the specific capacitance, respectively. In presence of mismatch or parasitics, the DAC output voltage levels differ from their ideal values being not equally spaced and determining a nonlinearity in the ADC input-output characteristic. Another important parasitic in the charge-redistribution array is the stray capacitance connected to top-plate node of the array to a fixed voltage node, e.g. C_{par} in Fig. 1. This parasitic capacitance causes a gain error in the converter characteristic but in some ADC topologies [5] it can also affect the converter nonlinearity.

A. Conventional Binary Weighted Array (CBW)

The classic binary weighted array [6], depicted in Fig. 1, is the most common topology, its linearity being affected only by the mismatch and parasitic capacitances connected between the top- and the bottom-plate nodes of each capacitance block, C_i. The statistical DNL and INL depend on the technology mismatch and vary with the output code. In a single-ended configuration, the maximum standard deviation of the DNL occurs in correspondence of the converter mid-code, being [5], [7]

$$\sigma_{DNL,CBW} = 2^{\frac{N}{2}} \cdot \frac{\sigma_C}{C_u}, \tag{3}$$

while the correspondent INL maximum standard deviation is

$$\sigma_{INL,CBW} = 2^{\frac{N}{2}-1} \cdot \frac{\sigma_C}{C_u}. \tag{4}$$

In a fully-differential configuration, these results have to be divided by a factor $\sqrt{2}$ [11].

In this array topology, the parasitic capacitance related to the top-plate node, C_{par}, causes a difference between two adjacent output voltage levels to be smaller than the nominal value. This turns into a gain error without affecting the converter linearity.

B. Split Binary Weighted Array (SBW)

The split DAC topology [12] is shown in Fig. 2. It consists of a binary weighted array where the MSB capacitor implements a binary weighted sub-array that perfectly reflects the structure of the remaining banks.

This DAC topology features an improved switching efficiency [12] and also a reduced impact of the capacitors mismatch. In fact, the maximum standard deviation of DNL and INL, which still occurs at the mid-code, is a factor $\sqrt{2}$ lower than in the CBW array topology [7] being

$$\sigma_{DNL,SBW} = 2^{\frac{N-1}{2}} \cdot \frac{\sigma_C}{C_u} \tag{5}$$

$$\sigma_{INL,SBW} = 2^{\frac{N-2}{2}} \cdot \frac{\sigma_C}{C_u}. \tag{6}$$

These relations are referred to a single-ended configuration while a fully-differential topology is a further $\sqrt{2}$-factor less sensitive to mismatch. Moreover, as in the CBW array, only the parasitics connected between top- and bottom-plate nodes of each array capacitor limit the converter nonlinearity.

C. Binary Weighted Array with Attenuation Capacitor (BWA)

A single-ended BWA array [13] is shown in Fig. 3. An attenuation capacitor C_{att} divides the DAC into two binary weighted arrays, main- and sub-DAC, of m and l capacitors, respectively. In this work, we'll refer to the architecture shown in Fig. 3, with the same number of bits in the main- and sub-DAC (i.e. $m = l = N/2$) and $C_{att} = C_u$, which has been shown to be the most energy efficient [5] among all the possible choices. For the same unit capacitance, this topology is more sensitive to capacitor mismatch with respect to CBW and SBW arrays because of the reduction of the overall DAC capacitance. In fact, the maximum σ_{DNL} and σ_{INL} are

$$\sigma_{DNL,BWA} = 2^{\frac{3N}{4}} \cdot \frac{\sigma_C}{C_u} \tag{7}$$

$$\sigma_{INL,BWA} = 2^{\frac{3N}{4}-1} \cdot \frac{\sigma_C}{C_u} \tag{8}$$

which are a factor $2^{\frac{N}{4}}$ larger than in the conventional topology. As far as the parasitic effect, the linearity is degraded also by the stray capacitance connected to the top-plate node of the sub-DAC, $C_{par,sub}$ in Fig. 3, which determines a dependence on the input code of the DAC output voltage, while the parasitic capacitance connected to the main-DAC, $C_{par,main}$, only affects the converter gain [5].

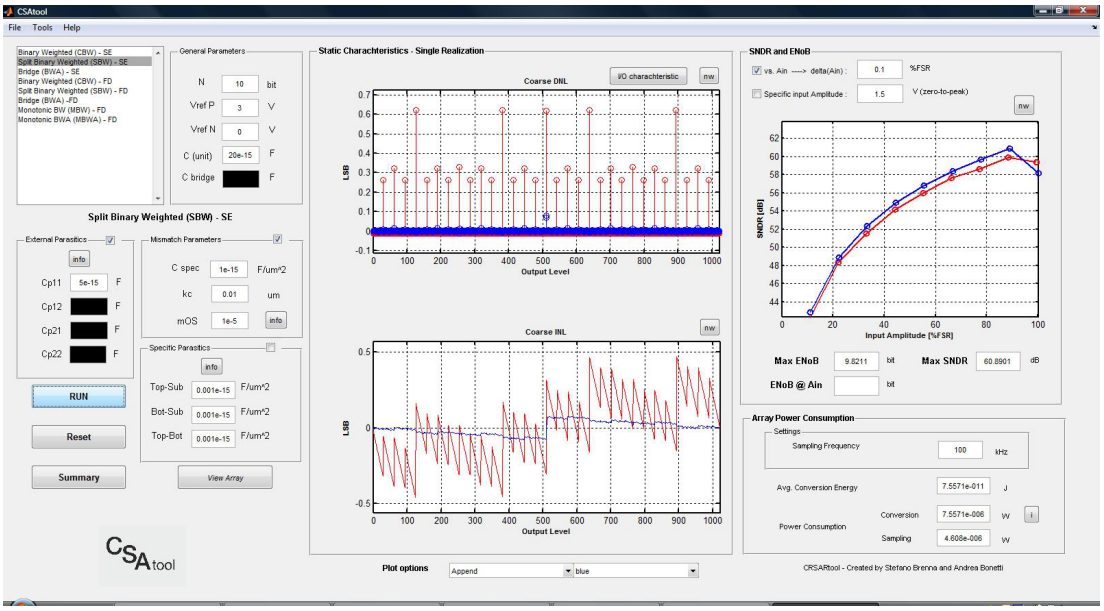


Fig. 4: Screenshot of the Graphic User Interface.

III. CSA MATLAB Simulation Tool

The proposed simulation platform (CSAtool, Charge-redistribution SAR ADC tool) is implemented in MATLAB and allows to evaluate the linearity, either static and dynamic, and the array power consumption of each array topology. Both statistical and deterministic effects due to capacitor mismatch and parasitics can be taken into account once evaluating the ADC performance. The tool provides a graphic user interface (GUI) which is shown in Fig. 4. A behavioral model is defined for each array topology and for a number of bits from 6 to 14. The model does not simply implement the known equations that estimate the nonlinearity (maximum standard deviation of DNL and INL) and the average power consumption of the different topologies. It reproduces the behavior of a specific circuit architecture, which is described by functional capacitive blocks, including the effects of mismatch and parasitics (see Fig. 5). The simulations can be performed with a high degree of customization, giving the possibility to accurately evaluate the impact of mismatch, through both single or multiple statistical runs, and also of the parasitics of each specific array capacitor, thus representing a suitable alternative to Cadence MonteCarlo and post-layout simulations. This section is devoted to the description of the different converter models that are implemented in the proposed tool.

A. CBW model

In a generic SAR ADC, the analog-to-digital conversion is effectively performed by comparing the input analog voltage signal with subsequent voltage levels generated by the capacitive DAC. For a N-bit converter, the input signal is compared with N successive DAC output levels. In a conventional binary weighted topology, the DAC output voltage at each conversion step can be written as

$$DAC_{out,CBW} = FSR \cdot H \tag{9}$$

where FSR is the full scale range of the converter and H, as shown in Fig. 5, is the scalar product

$$H = \frac{1}{C_{tot} + C_{par}} \times \bar{C} \times \bar{D}' \tag{10}$$

where C_{tot} is the total capacitance of the array, C_{par} the parasitic capacitance shown in Fig. 1, $\bar{C}$ the vector of the array capacitances C_i (see (1)) and $\bar{D}$ the vector of the digital word updated at each conversion cycle,

$$\bar{C} = \left[\begin{array}{ccc} C_1 & \dots & C_N \end{array} \right] \tag{11}$$

$$\bar{D} = \left[\begin{array}{ccc} D_1 & \dots & D_N \end{array} \right] \tag{12}$$

The digital word $\bar{D}$, which encodes the DAC output levels at each conversion step, is determined by the adopted switching algorithm. By means of (10), (11) and (12), it is possible to compute the ADC conversion characteristic.

B. SBW model

The simple model described in the previous section can be extended to the SBW architecture of Fig. 2. The MSB capacitor is implemented as a sub-array and the switching scheme differs from the conventional algorithm [12]. Thus, the DAC output voltage can be expressed as

$$DAC_{out,SBW} = FSR \cdot \left(H_{MSB} + H_{1,MSB-1} \right), \tag{13}$$

H_{MSB} and $H_{1,MSB-1}$ being coefficients related to the MSB and the residual capacitance array, respectively,

$$H_{MSB} = \frac{1}{C_{tot} + C_{par}} \times \bar{C}_{MSB} \times \bar{D}'_{MSB} \tag{14}$$

$$H_{1,MSB-1} = \frac{1}{C_{tot} + C_{par}} \times \bar{C}_{1,MSB-1} \times \bar{D}'_{1,MSB-1}. \tag{15}$$

Thus, the conversion voltage level is set by two different N-bit words, $\bar{D}_{MSB}$ and $\bar{D}_{1,MSB-1}$, and two vectors of capacitances, $\bar{C}_{MSB}$ and $\bar{C}_{1,MSB-1}$, related to the MSB sub-array and to the residual array, respectively.

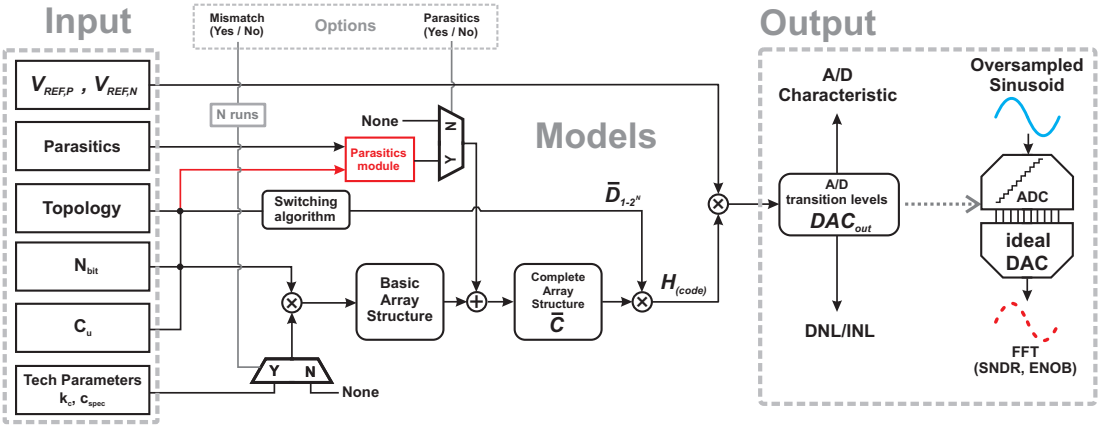


Fig. 5: CSAtool block diagram.

C. BWA model

As for as the BWA topology concerns, two equal capacitive arrays must be considered: the main-DAC, which is connected to the MSB switches, and the sub-DAC, related to the least significant bits. Let us indicate as $C_{tot,main}$ and $C_{tot,sub}$ to the overall capacitance of the main-DAC and of the sub-DAC, respectively, and as $C_{par,main}$ and $C_{par,sub}$ the parasitic capacitance at the top-plate node of the corresponding DAC (see Fig. 3). Due to the presence of the attenuation capacitor, C_{att}, the sub-DAC contribution to the overall DAC output voltage is reduced by an attenuation factor AR,

$$AR = \frac{C_{att}}{C_{tot,main} + C_{par,main} + C_{att}}, \tag{16}$$

$C_{tot,main}$ being the total capacitance of the ideal main-DAC, while $C_{par,main}$ is the parasitic capacitance connected to the top-plate node of the main array. Thus, the DAC output in the BWA topology is evaluated as

$$DAC_{out,BWA} = FSR \cdot (H_{main} + AR{\cdot}H_{sub}), \tag{17}$$

where H_{main} and H_{sub} are coefficients related to the main- and sub-DAC,

$$H_{main} = \frac{1}{C_{tot,main} + C_{par,main} + C_{att}} \times \bar{C}_{main} \times \bar{D}'_{main} \tag{18}$$

$$H_{sub} = \frac{1}{C_{tot,sub} + C_{par,sub} + C_{att}} \times \bar{C}_{sub} \times \bar{D}'_{sub}. \tag{19}$$

where $\bar{C}_{main}$, $\bar{C}_{sub}$, $\bar{D}_{main}$ and $\bar{D}_{sub}$ are the capacitance and digital output code vectors related to the main- and the sub-DAC,

$$\bar{C}_{main} = \left[\; C_{\frac{N}{2}+1} \quad \dots \quad C_N \; \right] \tag{20}$$

$$\bar{C}_{sub} = \left[\; C_1 \quad \dots \quad C_{\frac{N}{2}} \; \right] \tag{21}$$

$$\bar{D}_{main} = \left[\; D_{\frac{N}{2}+1} \quad \dots \quad D_N \; \right] \tag{22}$$

$$\bar{D}_{sub} = \left[\; D_1 \quad \dots \quad D_{\frac{N}{2}} \; \right]. \tag{23}$$

IV. Simulation Results

The CSAtool implements in MATLAB the three above-mentioned converter topologies by considering each array capacitor as a composition of unit elements, C_u, each with its own variability that depends on technologic parameters, i.e. k_c and c_{spec}. This simple modeling approach allows to accurately take into account the effects of mismatch and stray capacitances, which affect the converter linearity. The tool allows to run both nominal and statistical simulations in order to accurately evaluate the DNL and INL taking into account at the same time capacitor mismatch and parasitic effects. The tool also allows to estimate the degradation of both SNDR and ENoB due the converter nonlinearity. These two parameters are evaluated by feeding the real ADC with an oversampled sinewave of variable amplitude and re-converting the digital output signal to an analog waveform by means of an ideal DAC. Hence, the signal-to-noise ratio and the effective number of bits are evaluated by performing a FFT on the DAC output. Both SNDR and ENoB can be estimated while considering both deterministic parasitic effects and the contribute of statistical capacitor mismatch on multiple runs. Thus, CSAtool allows to obtain results accurate enough to be compared with post-layout and MonteCarlo simulations in Cadence Virtuoso environment but with a large improvement in terms of simulation time. In this section, the simulation results for the three converter topologies are shown and compared to both analytic expressions and Cadence results in terms of accuracy and simulation time. The presented results are related to converters featuring a 20-fF MiM unit capacitor with a specific capacitance of $1fF/\mu m^2$ and a Pelgrom coefficient of $1\% \cdot \mu m$.

A. Static metrics

The estimate of parasitic contribution to static non-linearity has been evaluated by means of the CSAtool and compared with Cadence Virtuoso simulation results. A single-ended SAR ADC has been designed in Cadence Virtuoso adopting a VerilogA description for the logic circuit and the comparator, while the feedback DAC has been implemented with ideal capacitors. Arbitrary parasitic capacitances have been added between the main capacitor top- and bottom-plate nodes in order to estimate their effect on the converter nonlinearity. Hence, the input-output characteristic has been evaluated by means of a transient simulation applying a full-scale ramp as input signal. To reduce the simulation time, only the converter input and output values have been saved. The strobe and the sampling periods have been chosen short enough to guarantee

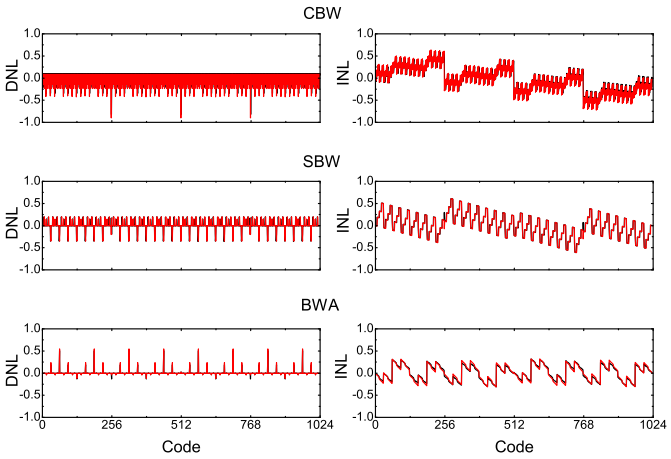


Fig. 6: Comparison between DNL and INL estimated by Cadence simulations (black lines) and by CSAtool (red lines) for a given pattern of parasitic capacitances and for each converter topology.

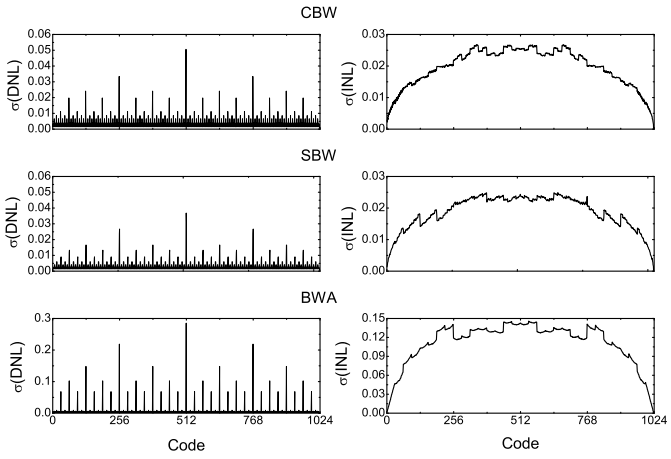


Fig. 7: Standard deviation of DNL and INL as a function of the output code for a CBW, SBW and BWA 10-bit converter.

at least 100 points per each conversion level, thus keeping the systematic error on the DNL below 1%. Fig. 6 shows the comparison between the DNL and INL characteristics obtained by CSAtool and Cadence Spectre simulations for the three converter topologies affected by parasitics extracted in Assura. The error between the CSAtool and Cadence results is never larger than 0.05 LSB, confirming the good accuracy of the implemented converter models.

Fig. 7 shows the results obtained by the CSAtool once the capacitor mismatch has been considered in terms of standard deviation of DNL and INL. Since a MonteCarlo simulation performed in the Cadence environment requires at least 100 runs to achieve confident results, thus being impractical, the effect of capacitor mismatch has been compared to the analytic expressions by using (3), (4), (5) and (8). Table I compares the maximum values of DNL/INL standard deviation, i.e. σ_{DNL} and σ_{INL}, computed by the CSAtool running 100 simulations and the estimates obtained by the analytic expressions. The error between estimates and CSAtool results is always lower than 0.005 LSB.

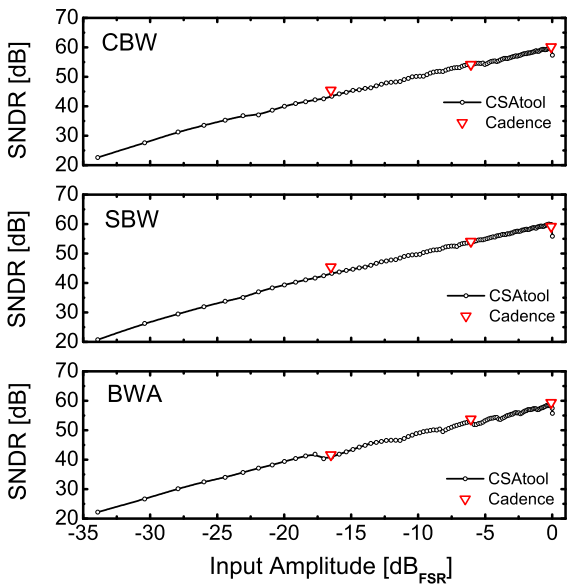


Fig. 8: SNDR as function of the input signal amplitude for the three converter topologies.

B. Dynamic metrics

The dynamic metrics have been computed in the CSAtool for the three single-ended converter topologies (with added parasitic capacitances) and compared to Cadence simulation results. The Cadence test-bench is similar to the one described in the previous section. The input signal is a sinewave at 1-kHz frequency but featuring a variable amplitude, while the sampling rate has been set to 30 kHz.

Fig. 8 shows the SNDR evaluated by means of Cadence Spectre simulations and CSAtool as function of the input signal amplitude (referred to the full scale range) for the three converter topologies. The maximum discrepancy between the Cadence and CSAtool results is always lower than 0.45 dB. However, it's worth pointing out that the CSAtool allows to easily compute the SNDR vs. input amplitude curve, while the same simulations in Cadence takes a long time allowing to compute only few points of the dynamic characteristic. This can results in a not correct evaluation of the SNDR peak and thus of the ENoB.

C. Power consumption

The models implemented in the CSAtool allow to evaluate the array switching energy as function of the output code.

Fig. 9 shows the average switching energies as function of the number of bits for the three considered single-ended arrays evaluated by means of the analytic expressions [5],

$$E_{ave,CBW} \cong 0.66 \cdot 2^N \left[C_u \left(V_{REF,P} - V_{REF,N} \right)^2 \right] \quad (24)$$

$$E_{ave,SBW} \cong 0.41 \cdot 2^N \left[C_u \left(V_{REF,P} - V_{REF,N} \right)^2 \right] \quad (25)$$

$$E_{ave,BWA} \cong 1.25 \cdot \left[C_u \left(V_{REF,P} - V_{REF,N} \right)^2 \right]. \quad (26)$$

The average energy evaluated with the CSAtool for the same arrays is also shown in Fig. 9 as symbols, showing a good agreement with the theoretical estimates.

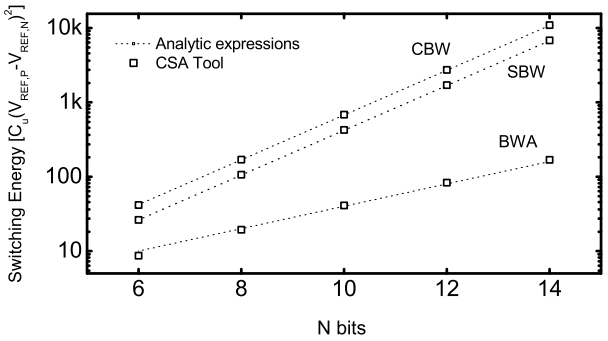


Fig. 9: Average switching energy for the three converter topologies. Dashed lines refer to analytic equations, symbols refer to CSAtool results.

TABLE I: ESTIMATES OF $\sigma_{DNL,max}$ AND $\sigma_{INL,max}$

	$\sigma_{DNL,max}$		$\sigma_{INL,max}$	
Topology	CSAtool	Equation	CSAtool	Equation
CBW	0.0502	0.0506	0.0249	0.0253
SBW	0.0351	0.0358	0.0252	0.0253
BWA	0.286	0.286	0.143	0.144

D. Simulation time

Table II reports a comparison between the simulation times needed to compute the static and dynamic metrics with the CSAtool and Cadence environment. The simulation times refer to a CBW converter with a number of bits from 6 to 14 and to a single simulation run. All the simulations have been performed with a 3-GHz Pentium Xeon featuring a 4-Gbyte main memory. For the same accuracy (i.e. DNL lower than 1%), the CSAtool features an improvement in terms of simulation time up to about 10^4. Finally, it's worth pointing out that the CSAtool allows to estimate the static nonlinearities and the dynamics metrics in presence of statistical mismatch and parasitic capacitance effect allowing to perform a MonteCarlo simulation over a large number of runs. Table III shows the simulation time for a MonteCarlo analysis of 100 runs. For a 14-bit converter, i.e. the worst case, the tool allows to compute the static and dynamic metrics in less than 10 minutes, while the same analysis cannot be performed in a reasonable amount of time in Cadence Virtuoso envirmonment.

V. CONCLUSIONS

In this work, a fast and accurate simulation tool for the analysis and design of charge redistribution SAR capacitive DACs and ADCs is presented. A graphic user interface MATLAB environment supports the implemented models that allow to simulate both technology mismatch and parasitic effects on linearity. The proposed tool overwhelms the conventional simulation method based on transient analyses in terms of computation time for a given accuracy and does not require a fine calibration of simulation parameters (points per conversion step, strobe period, etc.). In particular, the method is helpful every time a fast analysis and an accurate sizing of commonly-used converter architectures is needed. The proposed CSAtool can be downloaded with encrypted scripts from ftp://ftp.elet.polimi.it/outgoing/Stefano.Brenna.

TABLE II: SINGLE SIMULATION TIME

	Static metrics		Dynamic metrics	
Number of Bits	CSAtool	Cadence	CSAtool	Cadence
6	0.064s	625s	1.74s	$4{\cdot}10^3$s
8	0.087s	$2.5{\cdot}10^3$	1.99s	$4{\cdot}10^3$s
10	0.272s	10^4s	2.66s	$4{\cdot}10^3$s
12	1.005s	$4{\cdot}10^4$s	2.549s	$4{\cdot}10^3$s
14	3.997s	$1.6{\cdot}10^5$s	3.322s	$4{\cdot}10^3$s

TABLE III: MONTECARLO SIMULATION TIMES FOR 100 RUNS

Resolution	Static metrics	Dynamic metrics
6	2.13s	160s
8	6.491s	172s
10	25.25s	183s
12	100s	245s
14	425s	573s

REFERENCES

[1] M. van Elzakker *et al.*, "10-bit charge-redistribution ADC consuming 1.9μW at 1 MS/s," *IEEE J. of Solid State Circuits*, vol. 45, no. 5, pp. 1007–1015, May 2010.

[2] P. Harpe, E. Cantatore, G. Haller, and B. Murmann, "A 2.2/2.7fJ/conversion-step 10/12b 40kS/s SAR ADC with data-driven noise reduction," *Dig. Tech. Papers Int. Solid State Circuits Conf.*, pp. 270–271, Feb. 2013.

[3] W. S. Liew, X. Zou, and Y. Lian, "A 0.5V 1.13-μW/channel neural recording interface with digital multiplexing scheme," *Proceedings of ESSCIRC*, pp. 219–222, Sept. 2011.

[4] H. Gao *et al.*, "HermesE: A 96-channel full data rate direct neural interface in 0.13-μm CMOS," *IEEE J. of Solid-State Circuits*, vol. 47, April 2012.

[5] M. Saberi, R. Lotfi, K. Mafinezhad, and W. Serdjin, "Analysis of power consumption and linearity in capacitive digital-to-analog converters used in successive approximation ADCs," *IEEE Trans. Circuits Syst. I: Reg. Paper*, vol. 58, no. 7, pp. 1736–1747, Aug. 2011.

[6] J. McCreary and P. Gray, "All-MOS charge redistribution analog-to-digital conversion tecqniques. I," *IEEE J. of Solid State Circuits*, vol. SC-10, no. 6, pp. 371–379, Dec. 1975.

[7] A. Chandrakasan and B. Ginsburg, "500-MS/s 5-bit ADC in 65-nm CMOS with split capacitor array DAC," *IEEE J. Solid State Circuits*, vol. 42, no. 4, pp. 739–747, Apr. 2007.

[8] A. Agnes, E. Bonizzoni, and F. Maloberti, "Design of an ultra-low power SA-ADC with medium/high resolution and speed," *Proc. Int. Symp. on Circ. and Syst. (ISCAS)*, pp. 1–4, 2008.

[9] C. C. Liu, S. J. Chang, G. Y. Huang, and Y. Z. Lin, "A 10-bit 50-MS/s SAR ADC with a monotonic capacitor switching procedure," *IEEE J. of Solid State Circuits*, vol. 45, no. 4, pp. 731–740, Apr. 2010.

[10] S. Haenzsche, S. Henker, and R. Shcuffny, "Modeling of capacitor mismatch and non-linearity effects in charge redistribution SAR ADCs," *Proceedings of the Mixed Design of Integrated Circuits and Systems (MIXDES)*, pp. 300–305, June 2010.

[11] D. Zhang, A. Bhide, and A. Alvandpour, "A 53-nW 9.1-ENOB 1-kS/s SAR ADC in 0.13-μm CMOS for medical implant devices," *IEEE J. Solid State Circuits*, vol. 47, no. 7, pp. 1585–1593, Jul. 2012.

[12] A. Chandrakasana and B. Ginsburg, "An energy-efficient charge recycling approach for a SAR converter with capacitive DAC," *Proc. Int. Symp. on Circ. and Syst. (ISCAS)*, pp. 184–187, May 2005.

[13] A. Agnes, E. Bonizzoni, and F. Maloberti, "A 9.4-ENOB 1V 3.8μW 100kS/s SAR ADC with time-domain comparator," *Dig. Tech. Papers Int. Solid State Circuits Conf.*, vol. 37, no. 2, pp. 246–610, Feb. 2008.

High-speed Energy-efficient 5:2 Compressor

Ardalan Najafi, Somayeh Timarchi
Dep. of Electrical and Computer Engineering
Shahid Beheshti University
Tehran, Iran
Ar.najafi@mail.sbu.ac.ir S_timarchi@sbu.ac.ir

Amir Najafi
Dep. of Electrical, Computer and Biomedical Engineering
Qazvin Branch, Islamic Azad University
Qazvin, Iran
Amir.najafi@qiau.ac.ir

***Abstract*— Multipliers are important components that dictate the overall arithmetic circuits' performance. The most critical components of multipliers are compressors. In this paper, a new 5:2 compressor architecture based on changing some internal equations is proposed. In addition, using an efficient full-adder (FA) block is considered to have a high-speed compressor. The number of transistors in the proposed design is less than the best existing 5:2 compressor architectures. Three 5:2 compressors are considered for comparison. The proposed architecture is compared with the best existing designs presented in the state-of-the-art literature in terms of power, delay and area. Architectures are simulated in 90-nm CMOS technology under 1 V supply voltage. The simulation results show that the proposed compressor improves power and delay by 24.59% and 18.54% respectively, compared to two of the best existing architectures. In addition, voltage scaling and temperature analysis show that the proposed architecture outperforms the other designs from power-delay product (PDP) point of view in comparison to the aforementioned designs.**

Keywords—5:2 compressor; multiplier; high-performance arithmetic circuit; low-power design

I. INTRODUCTION

Rapid growth of using multipliers has attracted lots of researchers' attention into making high-performance multipliers to consume less power and operate faster. Multiplication process includes three steps: 1) partial product generation; 2) partial product reduction; 3) final addition with carry propagating. The second step makes the worst-case delay, consumes the main part of power, and occupies the high fraction of silicon area. To decrease the latency of this step, compressors have been widely employed. Therefore, designing a low-power and high-speed compressor is an important issue that should be raised to have a proper multiplication and subsequently a fast arithmetic computation.

5:2 compressors are widely used as building blocks of multipliers [1, 2]. Lots of architectures for 5:2 compressors have been proposed in the literature [3-6]. By a vast research on these structures, it has been revealed that the structures presented in [4] and [6] have better performance than others. In [4], the authors have proposed a new 5:2 compressor. They have simulated their proposed compressor by CMOS XOR-XNOR modules and two different multiplexer implementations. In [6], the authors have proposed a new structure based on a new realization of 5:2 compressors' truth table.

In this paper, we propose a new 5:2 compressor by applying some changes to the structure proposed in [4] to have a high-speed and low-power compressor. We use Carry Generator (CGEN) modules to produce C_{out1} and C_{out2} signals. We also make use of a high-performance CMOS FA to produce main output signals.

The paper is organized as follows: in Section II, we describe two architectures proposed in [4] and [6]. In Section III, we explore our proposed compressor. In Section IV, the simulation results and comparison are detailed. Finally, in Section V we conclude the paper.

II. BACKGROUND

Every 5:2 compressor has seven inputs and four outputs. Five inputs are primary inputs and the rest are two input carries which receive their values from the previous stage of one bit lower in significance. All the seven inputs, as well as output Sum bit have the same weight. The other three output bits weight one bit higher order.

A 5:2 compressor with five primary inputs X_1, X_2, X_3, X_4, X_5 and two output bits Sum and Carry along with carry input bits, C_{in1} and C_{in2}, and carry output bits, C_{out1} and C_{out2}, is governed by the following equation:

$$x_1 + x_2 + x_3 + x_4 + x_5 + C_{in1} + C_{in2} = Sum + 2\left(Carrry + C_{out1} + C_{out2}\right) \quad (1)$$

Various structures for 5:2 compressors are available in the literature. The simplest implementation of 5:2 compressor is obtained by cascading three full-adders in a hierarchical structure. The structure has a critical path delay of 6Δ where Δ refers to delay of XOR-XNOR, XOR, MUX or CGEN module, as their critical delay difference is trivial.

The structure of 5:2 compressor with critical path delay of 4Δ has been proposed in [7]. A modified structure has been proposed in [8], which has a critical path delay of 5Δ. In [3, 4] and [6] authors have proposed architectures with a critical path delay of 4Δ.

As shown in [4, 6], designs presented in these papers have better performance than others. The results of these two papers urged us to use these structures for our survey. These two designs are introduced below.

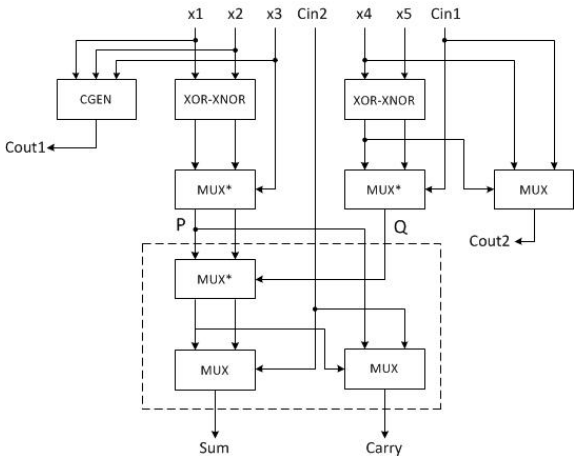


Fig. 1. 5:2 compressor architecture proposed in [4]-Arch-07

A. *The compressor proposed in [4]*

Since CMOS implementation of MUX performs better than a XOR gate in terms of power and delay [9], in [4] the authors have proposed a 5:2 compressor based on using multiplexers in place of XOR gates as shown in Fig. 1. In CMOS implementation of the MUX and XOR-XNOR blocks, as shown in Fig. 2(b) and Fig. 5(a), outputs and their complements are generated. But, complement outputs are not being utilized in architectures proposed in [3, 7, 8, 10]. If the output of multiplexer is used as select bit of another multiplexer, it can be used efficiently, and an extra stage to compute the negation in multiplexer structure can be saved. The authors used CMOS-CGEN in their architecture to produce C_{out1} signal. Fig. 3 shows the CMOS implementation of CGEN module. The architecture proposed in [4] will be called ***Arch-07*** in the rest of this paper and is depicted in Fig. 1.

Arch-07 is one of the most efficient architectures that exist for 5:2 compressors and has been used in [2, 11, 12].

The authors performed simulation by using CMOS multiplexer and XOR-XNOR modules which are depicted in Fig. 2(b) and Fig. 5(a), respectively. In addition, Fig. 2(a) has been used in Arch-07 as the MUX block of intermediate stages.

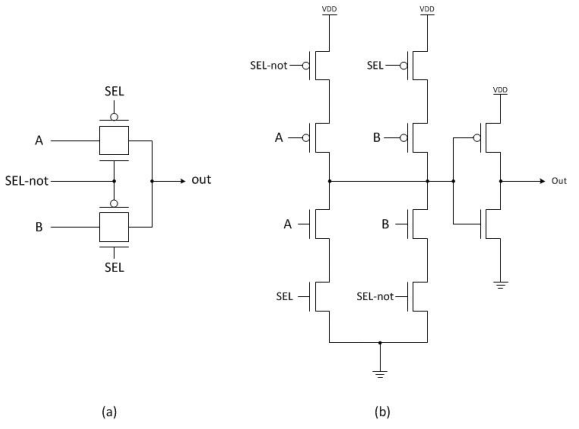


Fig. 2. (a) Transmission Gate implementation of multiplexer (b) CMOS implementation of multiplexer

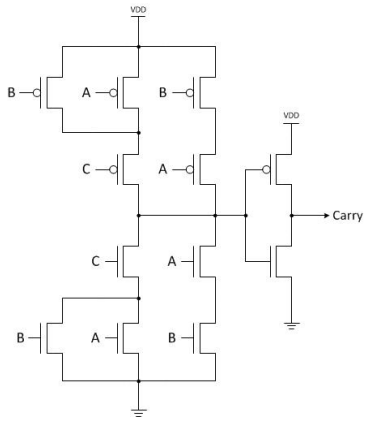


Fig. 3. CMOS implementation of Carry Generator module

B. *The compressor proposed in [6]*

In [6], authors have proposed architectures with a critical path delay of 4Δ. Pass-transistor logics and static CMOS logics have been used based on a new understanding of 5:2 compressors' truth table. This architecture is depicted in Fig. 4 and will be called ***Arch-13*** in the rest of this paper. CMOS-CGEN modules have been used to produce C_{out1} and C_{out2} signals, in this architecture. This block is also a main block of 7-2 compressor proposed in [13].

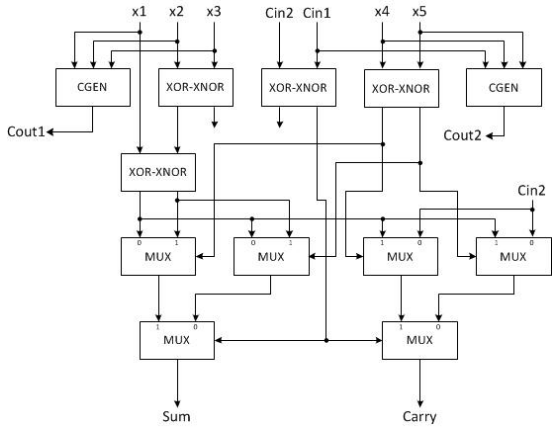


Fig. 4. 5:2 compressor proposed in [6]-Arch-13

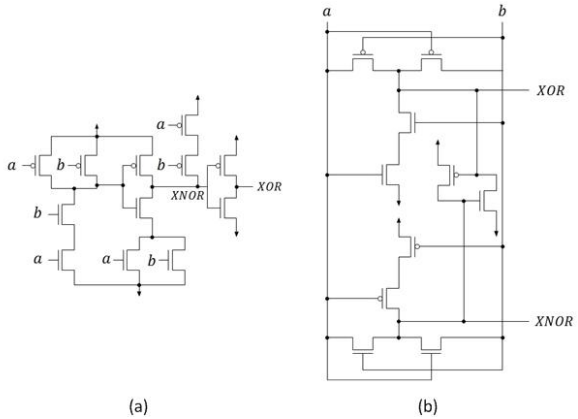


Fig. 5. (a) CMOS implementation of XOR-XNOR module (b) XOR-XNOR module proposed in [3]

In Arch-13, XOR-XNOR modules of Fig. 5(b) are used to feed complementary inputs of MUXs. In addition MUX modules of Fig. 2(a) have been used in this architecture. The authors have claimed power and latency reduction by decreasing middle stage capacitances.

III. PROPOSED COMPRESSOR

In this section the proposed architecture is introduced. Two improvement approaches are used to propose the new 5:2 compressor architecture. First, by a closer look at dashed box of Fig. 1, it represents the functionality of a conventional FA and can be replaced by variety of FAs presented in the literature. This replacement is expected to lead to considerable speed improvement, due to 34% faster operation of CMOS FA in comparison to two cascaded CMOS XOR gates, as it is explained in [14]. Therefore, the CMOS FA presented in [15] is used in our proposed 5:2 compressor architecture.

To further improvement, we make some changes to internal equations of the 5:2 compressor to eliminate final Not gates of the CMOS FA. By doing so, we could have reduced power dissipation as well as improved operational speed. To achieve this goal, we have to use XNOR gates instead of XORs of the second stage of the architecture. Hence, we propose the architecture which is shown in Fig. 6. This design uses 82 transistors in its architecture (i.e. 6 transistors less than Arch-13). The proposed design will be called ***New Arch*** in the rest of this paper.

In this architecture, FA-not is CMOS FA which it's final Not gates have been eliminated. This module is shown if Fig. 7.

As shown in Fig. 6, CGEN modules have been used to produce C_{out1} and C_{out2} output signals. The CMOS-CGEN which has been shown in Fig. 3 is used in the proposed architecture. In addition, outputs of the XOR gates have been fed to inputs of the XNOR gates, as it can be seen in Fig. 3. This way, outputs of the XNOR gates are negation of what it was before for conventional 5:2 compressors and by replacing a FA-not instead of a FA we can have valid Sum and Carry signals.

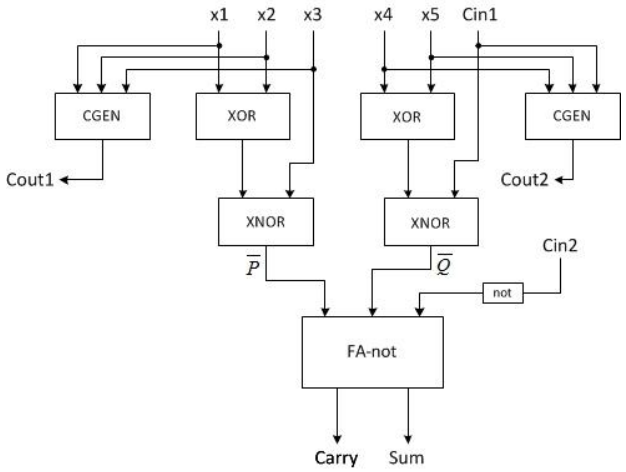


Fig. 6. Proposed 5:2 compressor design-New Arch

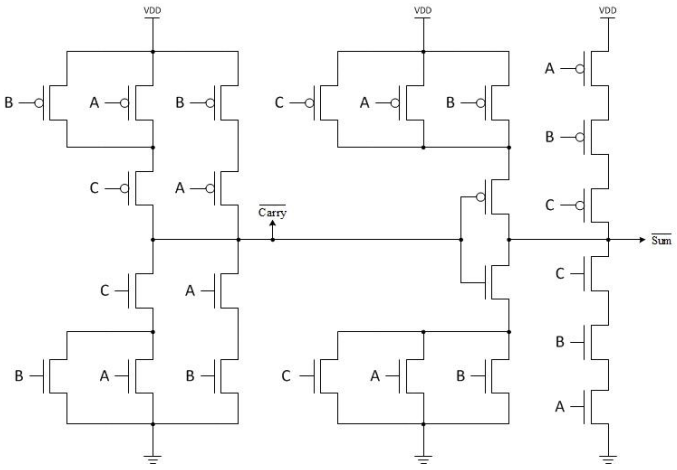


Fig. 7. CMOS implementation of FA-not cell

We know the following equations:

$$\begin{aligned} \bar{x} \oplus \bar{y} &= x \oplus y \\ \overline{x \oplus \bar{y}} &= x \oplus y \end{aligned} \tag{2}$$

Therefore, (3) and (4) proves the accuracy of the mentioned methods.

$$\begin{aligned} \overline{\left(\overline{P} \oplus \overline{Q}\right) \oplus \overline{C_{in2}}} &= \left(P \oplus Q\right) \oplus \overline{C_{in2}} \\ &= P \oplus Q \oplus C_{in2} \end{aligned} \tag{3}$$

$$\begin{aligned} \overline{\left(\overline{P} + \overline{Q}\right) \cdot \overline{C_{in2}} + \overline{P}\overline{Q}} &= PQ + PC_{in2} + QC_{in2} \\ &= \left(P + Q\right) \cdot C_{in2} + PQ \end{aligned} \tag{4}$$

It is worth mentioning that $\overline{P}$ and $\overline{Q}$ are nodes which represent XNOR gates outputs of the proposed architecture and have been shown in Fig. 6. In addition, they are logical complement of P and Q nodes, respectively, which have been shown in Fig. 1.

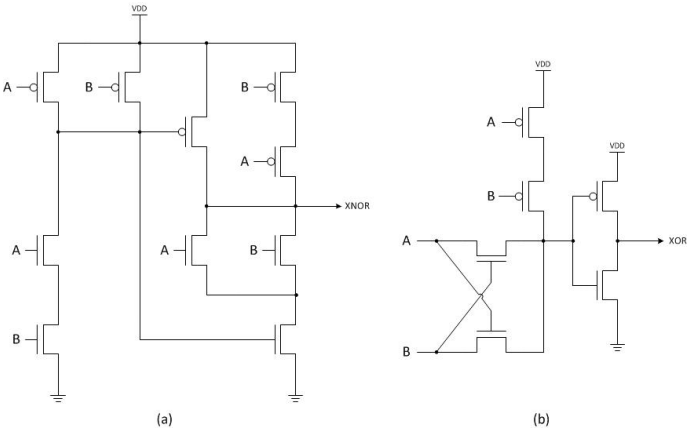


Fig. 8. (a) CMOS implementation of XNOR gate (b) XOR gate using pass-transistor logic presented in [16]

The logic equations governing the proposed 5:2 compressor architecture are detailed in (5)-(8):

$$Sum = x_1 \oplus x_2 \oplus x_3 \oplus x_4 \oplus x_5 \oplus C_{in1} \oplus C_{in2} \tag{5}$$

$$C_{out1} = (x_1 + x_2) \cdot x_3 + x_1 x_2 \tag{6}$$

$$C_{out2} = (x_4 + x_5) \cdot C_{in1} + x_4 x_5 \tag{7}$$

$$\begin{aligned} Carry = &\left((x_1 \oplus x_2 \oplus x_3) + (x_4 \oplus x_5 \oplus C_{in1})\right) \cdot C_{in2} \\ &+ (x_1 \oplus x_2 \oplus x_3) \cdot (x_4 \oplus x_5 \oplus C_{in1}) \end{aligned} \tag{8}$$

The XOR and XNOR gates which have been used in the proposed architecture are shown in Fig. 8.

IV. SIMULATION RESULTS AND COMPARISON

A. Simulation environment

The simulations have been performed by HSPICE software. All the simulations are targeted for TSMC 90-nm technology. The circuits under test are simulated at 1 V supply voltage. It should be noted that all inputs are driven by frequency of 200MHz.

The simulation environment for 5:2 compressor is shown in Fig. 9. Three cascaded compressors are used due to the fact that the critical path of some data patterns may cross neighboring compressors in the same stage of CSA tree. The dashed line in Fig. 9 indicates the most probable scenario of such potential critical paths. The delay is measured from the earliest input signal reaching 50% of the supply voltage to the latest output signal reaching 50% of the supply voltage for each input cycle. The worst case is the largest delay among all input data.

B. Results and comparisons

The proposed architecture is compared to two of the best existing 5:2 compressor architectures in terms of power, delay and area. Therefore, three different 5:2 compressor designs are simulated: Arch-07, Arch-13 and the proposed architecture.

The simulation results for 1V supply voltage at ambient temperature of 25°C are tabulated in Table 1. Bold numbers represent the best ones among the three architectures in each figure of merits. In addition, the last raw of the table shows the improvement percentage of our proposed architecture in comparison to the best of the existing ones.

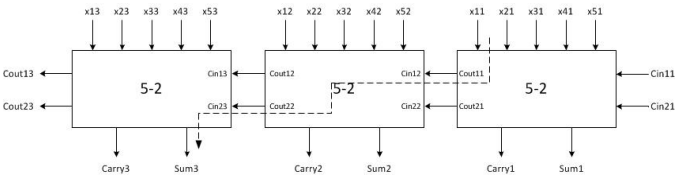


Fig. 9. Simulation environment

TABLE I. SIMULATION RESULTS IN 25°C WITH THE 1V

	Power [µw]	*Delay [ps]*	*Power-delay product [fj]*	*Transistor count*
Arch-07	6.18	228.3	1.41	90
Arch-13	6.66	210.3	1.4	88
New Arch	**4.66**	**168.2**	**0.78**	**82**
Improvement percentage	24.59%	18.44%	44.28%	6.82%

As it can be seen in the table, New Arch reduces power by 24.59% in comparison to Arch-07. In addition, New Arch reduces delay, PDP and transistor count by 18.44%, 44.28% and 6.82%, respectively, in comparison to Arch-13.

We have also performed voltage scaling and temperature analysis. For voltage scaling analysis, 5:2 compressors have been simulated at temperature of 25°C under supply voltages ranging from 0.8V to 1.2V. Fig. 10 shows PDP versus supply voltage of the three 5:2 compressor architectures.

For temperature analysis, the architectures have been simulated in the wide temperature range of 20-120°C with 1V supply voltage. Fig. 11 shows PDP versus temperature of the three 5:2 compressor architectures

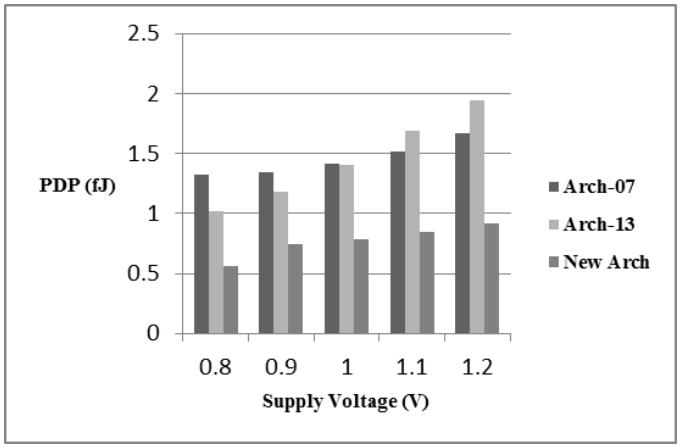


Fig. 10. Power-delay product Vs. supply voltage

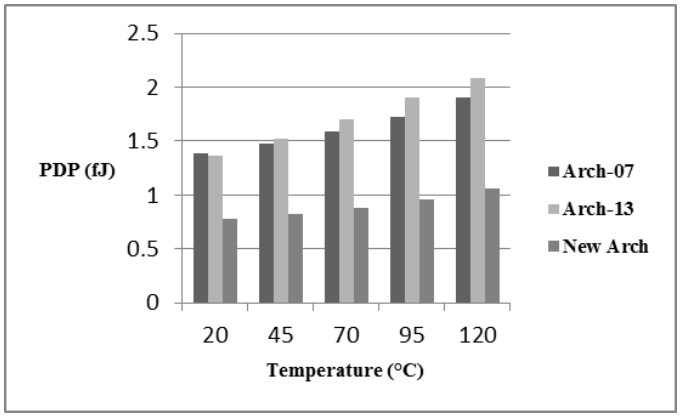


Fig. 11. Power-delay product Vs. temperature

Fig. 10 and Fig. 11 show that the proposed architecture outperforms the other two designs from energy efficiency point of view. These figures as well as Table 1 show up to 45% improvement in PDP of the proposed architecture in comparison to the best existing ones.

V. CONCLUSION

In this paper, we proposed a new 5:2 compressor architecture by changing some internal equations. The new compressor is proposed by presenting a module named: FA-not. This module is obtained by eliminating final Not gates of a CMOS FA block. In addition, CMOS-CGENs are used to produce C_{out1} and C_{out2} signals. By applying these methods, we have come up with a 5:2 compressor which uses 82 transistors.

Simulation and synthesis are performed by HSPICE software; and all inputs are driven by the frequency of 200MHz. It is concluded that among the three simulated architectures, the proposed architecture consumes the lowest power, operates faster and consequently is the most energy efficient one. The proposed architecture, in term of PDP, shows 44.28% improvement compared to the best existing architecture.

REFERENCES

[1] O. Kwon, K. Nowka, and E. E. Swartzlander, "A 16-Bit by 16-Bit MAC Design Using Fast 5:3 Compressor Cells," *The Journal of VLSI Signal Processing*, vol. 31, pp. 77-89, 2002.

[2] R. Modugu, C. Minsu, and N. Park, "A fast low-power modulo 2^n+1 multiplier design," in *Instrumentation and Measurement Technology Conference, 2009. I2MTC '09. IEEE*, 2009, pp. 951-956.

[3] C. H. Chang, J. Gu, and M. Zhang, "Ultra low-voltage low-power CMOS 4-2 and 5-2 compressors for fast arithmetic circuits," *IEEE Transactions onCircuits and Systems*, vol. 51, pp. 1985-1997, 2004.

[4] S. Veeramachaneni, K. M. Krishna, L. Avinash, S. R. Puppala, and M. B. Srinivas, "Novel Architectures for High-Speed and Low-Power 3-2, 4-2 and 5-2 Compressors," in *Proc. of 20th Int. Conf. on VLSI Design*, 2007, pp. 324-329.

[5] R. Menon and D. Radhakrishnan, "High performance 5 : 2 compressor architectures," *Circuits, Devices and Systems, IEE Proceedings -*, vol. 153, pp. 447-452, 2006.

[6] M. Tohidi, M. Mousazadeh, S. Akbari, K. Hadidi, and A. Khoei, "CMOS implementation of a new high speed, glitch-free 5-2 compressor for fast arithmetic operations," in *Mixed Design of Integrated Circuits and Systems (MIXDES), 2013 Proceedings of the 20th International Conference*, 2013, pp. 204-208.

[7] O. Kwon, K. Novka, and E. E. Swartzlander, "A 16-bit × 16-bit MAC design using fast 5:2 compressor," in *IEEE Int. Conf. Application-Specific Systems, Architectures, Processors*, 2000, pp. 235-243.

[8] K. Prasad and K. K. Parhi, "Low-power 4-2 and 5-2 compressors," in *Proc. of the 35th Asilomar Conf. on Signals, Systems and Computers*, 2001, pp. 129-133.

[9] R. Zimmermann and W. Fichtner, "Low-Power Logic Styles: CMOS Versus Pass-Transistor Logic," *IEEE J. Solid-State Circuits*, vol. 32, pp. 1079-1090, 1997.

[10] G. Jiangmin and C. Chip-Hong, "Low voltage, low power (5:2) compressor cell for fast arithmetic circuits," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on*, 2003, pp. II-661-4 vol.2.

[11] C. Vinoth, V. S. K. Bhaaskaran, B. Brindha, S. Sakthikumaran, V. Kavinilavu, B. Bhaskar, M. Kanagasabapathy, and B. Sharath, "A novel low power and high speed Wallace tree multiplier for RISC processor," in *Electronics Computer Technology (ICECT), 2011 3rd International Conference on*, 2011, pp. 330-334.

[12] R. Shende, P. Zode, and P. Zode, "Efficient design 2^k-1 binary to residue converter," in *Devices, Circuits and Systems (ICDCS), 2012 International Conference on*, 2012, pp. 482-485.

[13] M. Rouholamini, O. Kavehie, A. P. Mirbaha, S. J. Jasbi, and K. Navi, "A New Design for 7:2 Compressors," in *Proc. IEEE/ACS Int. Conf. on Computer Systems and Applications*, 2007, pp. 474-478.

[14] A. Pishvaie, G. Jaberipur, and A. Jahanian, "Improved CMOS (4;2) compressor designs for parallel multipliers," *Computers & Electrical Engineering*, vol. 38, pp. 1703-1716, 2012.

[15] N. Weste and K. Eshraghian, *Principles of CMOS VLSI design*, 2nd ed.: Addison Wesley, 1993.

[16] S. Nishizawa, T. Ishihara, and H. Onodera, "Analysis and comparison of XOR cell structures for low voltage circuit design," in *Quality Electronic Design (ISQED), 2013 14th International Symposium on*, 2013, pp. 703-708.

An electrochemical sensor ASIC for agriculture applications

Tan Kong Yew[1], Yuzman Yusoff[1], Lam Kien Sieng[1], Hanif Che Lah[1], Hasmayadi Majid[1], Noor Shelida[1]
[1] MIMOS Berhad, Integrated Circuit Design Department, Kuala Lumpur, Malaysia
kongyew.tan@mimos.my

Abstract – This paper presents a fully integrated application specific integrated circuit (ASIC) for the readout of multiple electrochemical sensor signals. It consists of an analog router that routes readout circuits to various sensors, a 10-bit successive approximation register analog to digital converter (SAR ADC), two digital to analog converters (DAC), a multiplexer and I2C/SPI serial interfaces for connection to micro-controller (MCU). The ASIC is fabricated in a 0.35-µm CMOS technology, occupies an area of 1.5 × 3.0 mm^2, and consumes a current of 3.2 mA from a 3.3 V supply. All calibration and configurations are set using the I2C/SPI serial interface connections.

I. INTRODUCTION

In agriculture applications, low cost, ease of use, reliability and functional flexibility[1][2] are essential considerations for design. To measure sensor signals of different types, an adaptive universal signal conditioner has been designed. Integrating a micro-controller (MCU) with this universal sensor interface ASIC combines analogue and digital processing capability and allows customization to different applications. Such an arrangement enhances performance while simultaneously reduces cost and power when considered as a system.

In this paper, a sensor ASIC that integrates a 16-channel reconfigurable interface, a 10-bit SAR ADC and two 10-bit DAC is presented. The sensor interface performs signal readout for current and voltage output sensors, providing a unified interface for heterogeneous mix of sensors. The ASIC supports serial interfaces I2C/SPI that permits communication with an MCU that sets configuration, calibration, control and sensor data readout and processing. This multi-sensor interface can be re-configured to satisfy the readout and signal processing needs of many different sensor applications, providing flexibility and lowering costs.

This paper is divided into three main sections. The system architecture and circuit implementation are explained in Approach and Methods section. The Results and Discussion section presents simulation results and physical layout. Finally, the conclusions are drawn in the last section.

II. APPROACH AND METHODS

The architecture as shown in Fig. 1 consists of an analog router which connects 24 inputs to 4 different input multiplexers depending on sensor type (2/3/4 port sensors). This allows as many as 6 sensors (max. 4 port) to be connected simultaneously in addition to provision for external sensor readout through open ports as shown in Fig. 2. The input multiplexers routes the sensor signals to the respective readout circuits before connecting to a 16:1 multiplexer which output to an external filter.

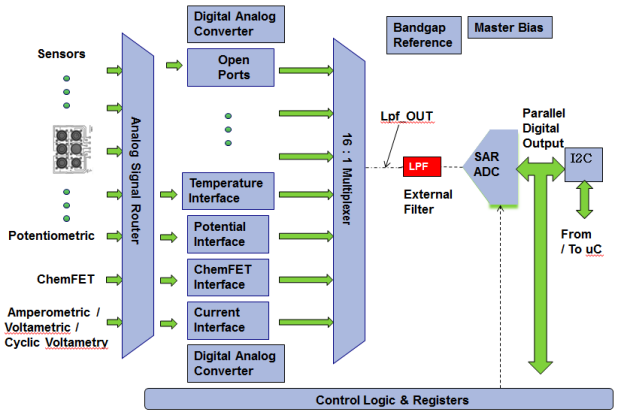


Fig. 1 – Architecture of sensor ASIC

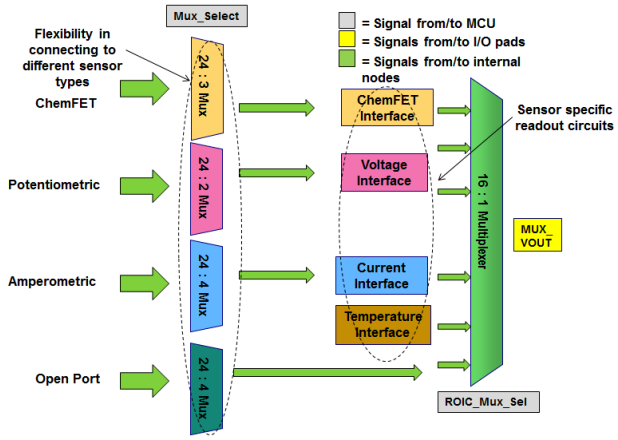


Fig. 2 – Analog signal router connection block diagram

Signals output from external filter then passes through an ADC to be digitized and serialized. All the control

sequence are controlled and configured through I2C interface by external micro-controller. Outputs from the two 10-bit DAC connect to all sensor readout circuits and also to the SAR ADC, allowing it to be used as reference signals for calibration and to set offsets if needed.

A. ELECTROCHEMICAL SENSOR READOUT CIRCUITS

The three readout integrated circuit (ROIC) designed is for chemical field effect transistors (ChemFET), current (Amperometric) and voltage (Potentiometric) sensors. ChemFET [2] sensors are field effect transistors with gates replaced by an intrinsic electrolyte layer covered by a dedicated ion selective membrane. It is used for measurement of a specific ion mole activity from a dissolved solution composing of other ions. The readout circuit [3] for the designed ChemFET sensor interface is depicted in Fig. 4, where provisions are made for various settings of input sources, gain, dc shift and biasing.

The Amperometric readout circuit [4] shown in Fig. 5, is configurable for two, three or four electrode type current sensors. A trans-impedance amplifier converts the current input to voltage where its gain is selectable before being converted to digital output by ADC later. The sensor bias can be set either through internal DAC (programmable) or external pin.

The Potentiometric readout circuit designed, shown in Fig. 6, is a conventional three op-amp type with programmable settings for gain and offset.

B. DATA CONVERTERS

Both the analog to digital converter (ADC), Fig.7 and digital to analog converter (DAC), Fig.8 designs have 10 bit resolution and uses resistor string DAC sub-block in order to minimize area. Two DAC's are designed for use to set direct current (DC) offsets and provide reference signals for calibration. The outputs from the DAC can be selected to route to various readout circuits and to ADC.

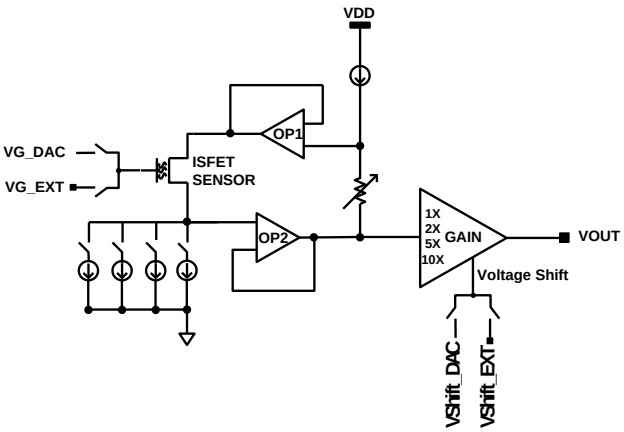


Fig. 4 – ChemFET sensor readout circuit

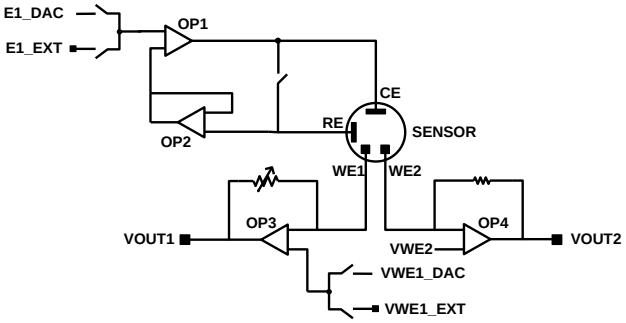


Fig. 5 - Amperometric sensor readout circuit

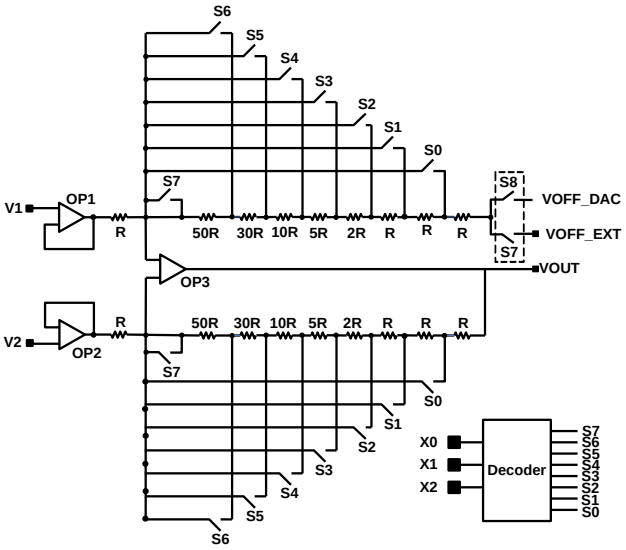


Fig. 6 - Potentiometric sensor readout circuit

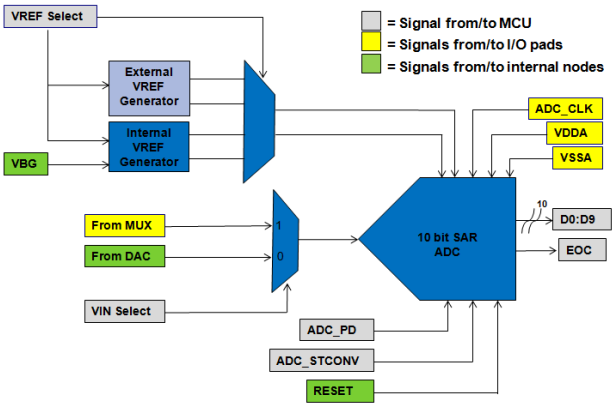


Fig. 7 - 10 bit Successive Approximation Analog Digital Converter

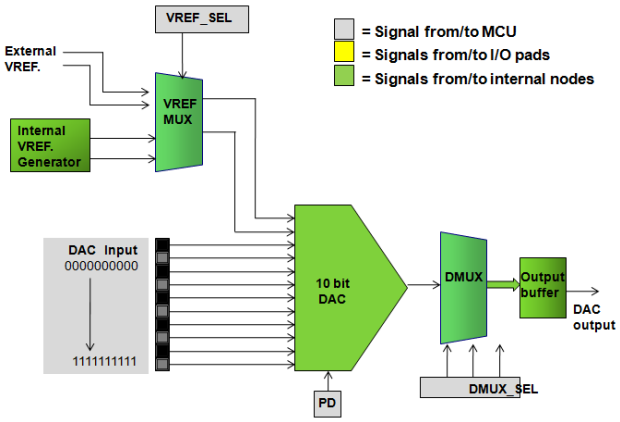


Fig. 8 - 10 bit Resistive string Digital to Analog converter

C. SERIAL INTERFACES

All data into and out of ADC, DAC and other ROIC blocks uses either I2C or SPI interface, Fig 9 to reduce pin count. The serial interfaces are also used for setting configuration data as well as command from external MCU.

The Serial Interface consists of START and STOP detection logic to identify the beginning and end of data packet. The Bit Counter and State Machine are used to keep track of operation and serially shift in or out data from or to the data base. These data are written into/read from the Register Bank, which are used to interface with analog circuitry.

III. RESULTS AND DISCUSSION

This circuit was simulated and taped out using 0.35um CMOS process technology. The layout of the sensor ASIC is shown in Fig. 10 while the measurement setup used for characterizing the taped-out chip is shown in Fig. 11, where in addition to bench equipments being connected, an I2C/SPI host adapter with its associated software is also interfaced in order to provide stimulus and read response through I2C/SPI protocol.

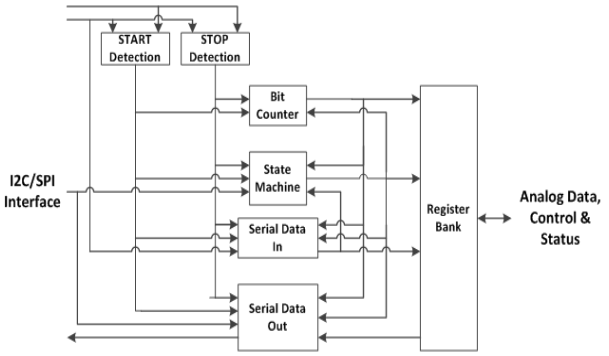


Fig. 9 - Serial Interface

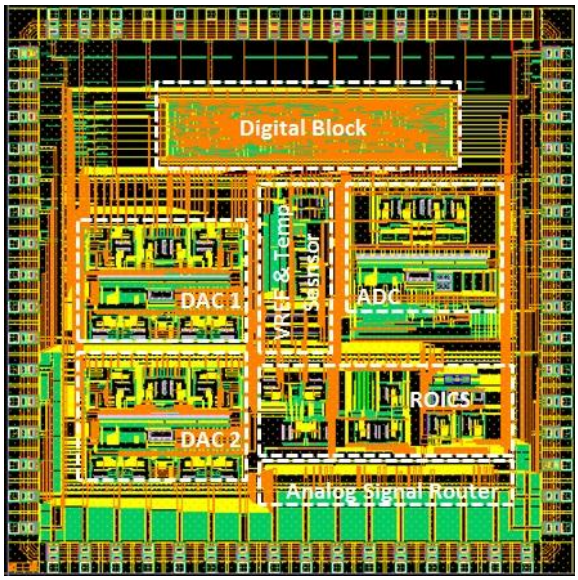


Fig. 10 - Layout implementation of sensor ASIC (5mm x 5mm) [6] ©2013 MIMOS Berhad

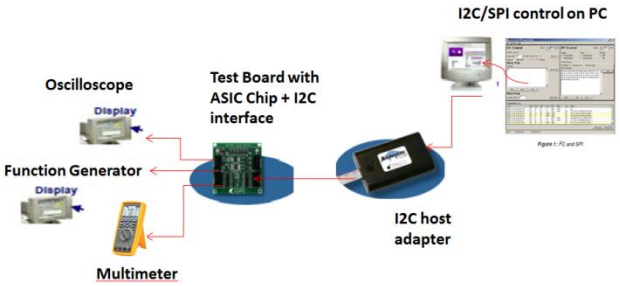


Fig. 11 – Test setup used for sensor ASIC characterization

Fig. 12 shows the functional test result for DAC block where three sets of ten consecutive codes, representing the lower, middle and upper DAC ranges were input through I2C interface and DAC ouput voltage measured at pin Lpf_OUT (refer to Fig. 1), which is the output of 16:1 multiplexer. The result (13 out of 14 chips tested) shows that transition from one code to the next consecutive code was correct, i.e. ~2mV (equivalent to one least significant bit, LSB) for reference of 2 volt for a 10 bit DAC.

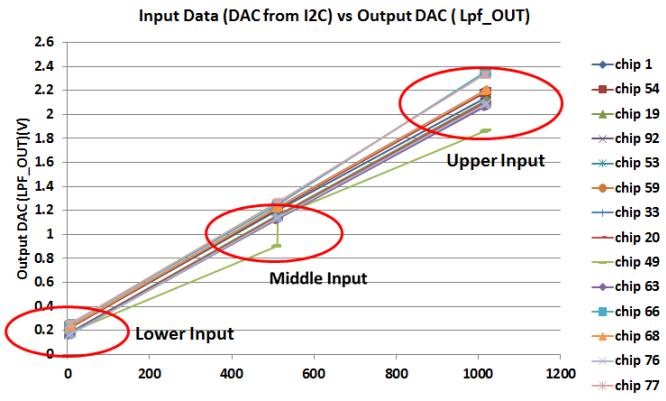


Fig. 12 – DAC block functional test result

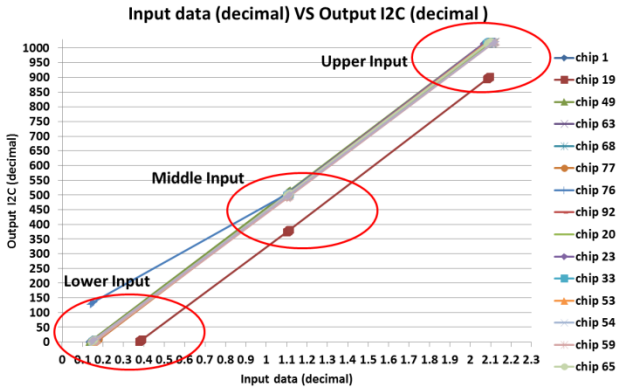


Fig. 13 – ADC block functional test result

The ADC block was functionally tested for three sets of ten consecutive codes (Fig. 13), representing the lower, middle and upper ADC ranges. In this case the ADC input was increased until ten consecutive codes from those ranges were exercised. Zoom-in results for the functional tests are depicted in Fig. 14, Fig. 15 and Fig. 16. The results show that the ADC increase by one code for approximately 2mV change in input for a 2V reference which corresponds to 1 LSB (least significant bit) and also the increase are linear over those ten codes.

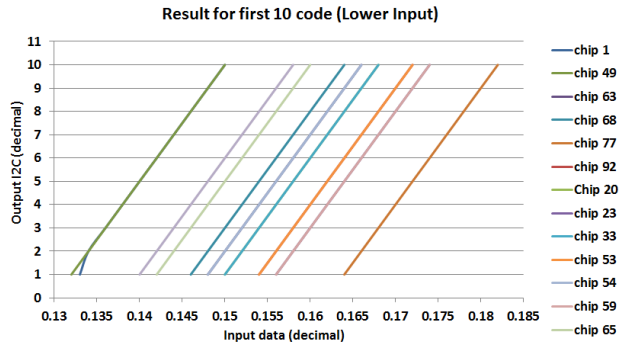


Fig. 14 – ADC block functional test result for first 10 lower code

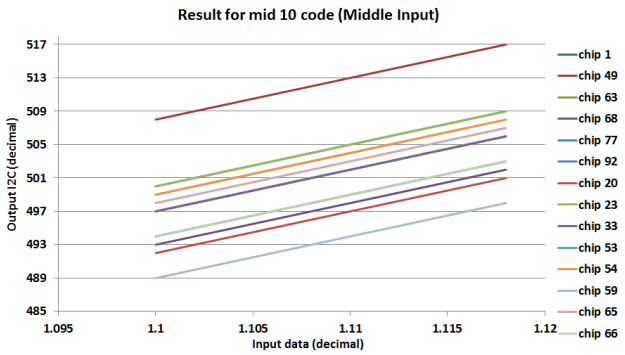


Fig. 15 – ADC block functional test result for first 10 middle code

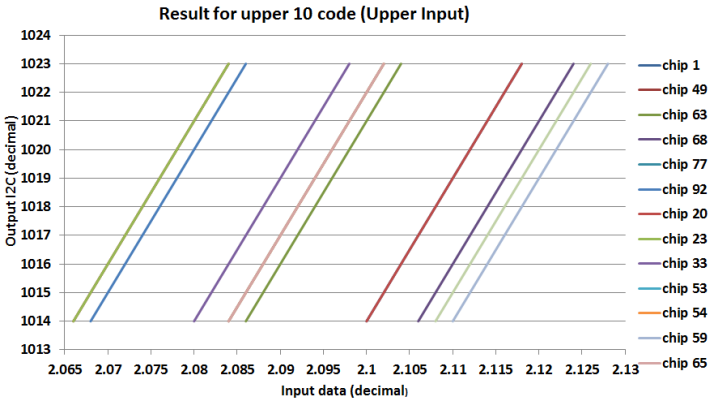
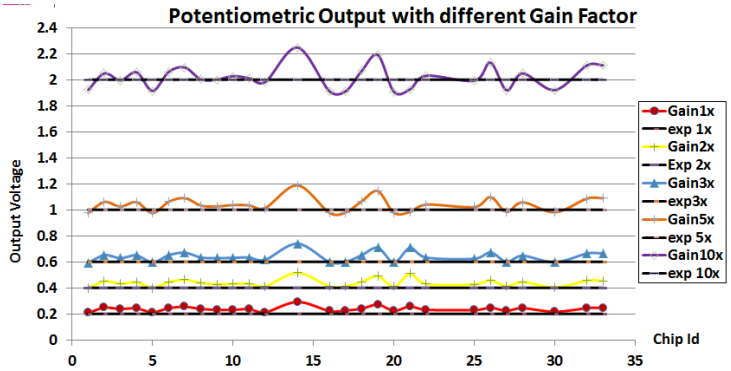


Fig. 16 – ADC block functional test result for first 10 upper code

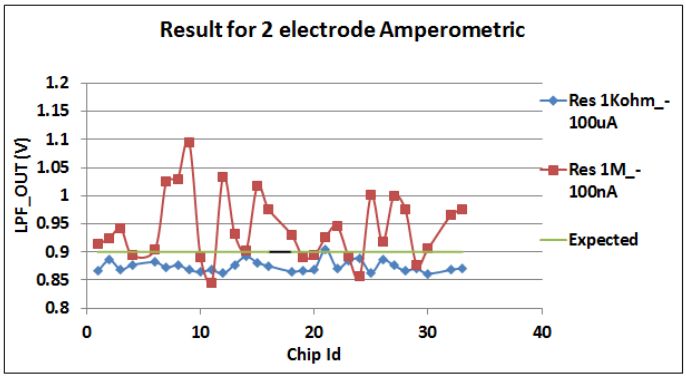


Output Voltage	Gain factor				
	1x	2x	3x	5x	10x
Min (V)	0.208	0.396	0.591	0.975	1.906
Max (V)	0.292	0.515	0.74	1.188	2.247
Average(V)	0.237	0.438	0.64	1.039	2.019

Note : exp refers to expected value from simulation data

Fig. 17 – Potentiometric ROIC block functional test result

For potentiometric readout circuit characterization, the input was set at 0.2V and gain set to 1x,2x,3x,5x and 10x respectively. The test result is as shown in Fig. 17 where the output voltage matches the expected value within 15% error for smaller gain values (1x,2x,3x) and within 25% error for larger gain values (5x,10x).



Voltage(V)	1K ohm	1M ohm
Min	0.8616	0.845
Max	0.9045	1.007
Average	0.8743	0.923

Note : Expected refers to parameter value from simulation data

Fig. 18 – Amperometric ROIC block functional test result

The Amperometric readout test result shown in Fig. 18 is for a 2 electrode configuration using external resistor of value 1 Kilo-ohm and 1 Mega-ohm for gain setting with bias current of 100 uA and 100 nA set respectively at WE1 (see Fig. 5). The gain-select using internal resistors was verified not working. The input voltage VWE1_EXT_AMP I set to 1V and the theoretical output voltage was calculated to be 0.9V. The result shows that with resistor of 1K-ohm the output was within 5% of expected value however using 1M-ohm resistor results in greater output error, as much as 22% from expected value.

Finally for the ChemFET sensor block (Fig. 4), a commercial ion sensitive field effect transistor (ISFET) sensor was used to test the readout circuit. ISFET sensor can detect pH of solution by the output voltage read out from ChemFET block. Fig. 19 and Fig. 20 shows the voltage output read-out using a drain source voltage (V_{DS}) bias of 1V and 0.5V respectively with bias current set at 100uA. The result shows that 18 of the 19 chips tested shows fairly constant output voltage for the different pH buffer solutions used i.e. ph4, ph7 and pH10.

Fig. 21 and Fig. 22 shows for those same drain source voltage and current bias, the output voltage change for each step of pH change. The result shows the sensitivity of the sensor together with readout circuit. This is in line with theoretical calculated value of ISFET sensor sensitivity of 55 mV/pH.

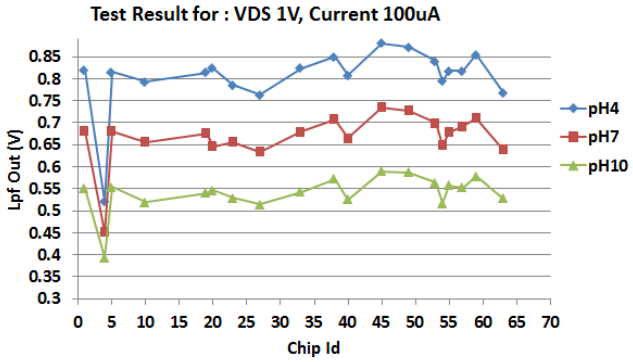


Fig. 19 – ChemFET ROIC block functional test result 1

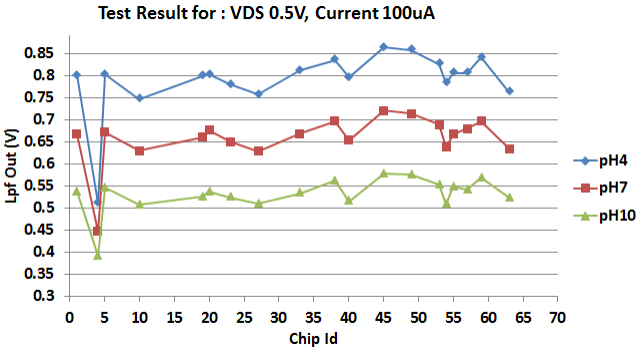


Fig. 20 – ChemFET ROIC block functional test result 2

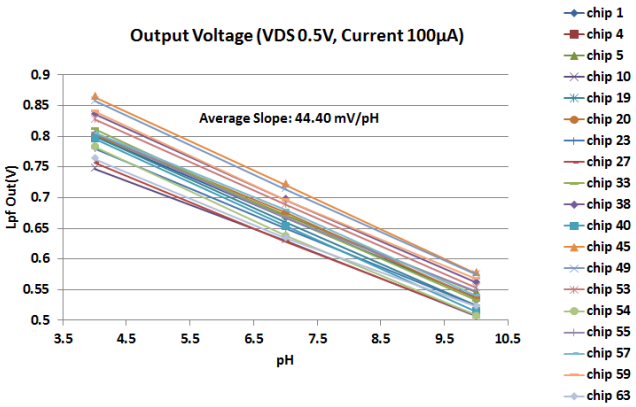


Fig. 21 – ChemFET ROIC block functional test result 3

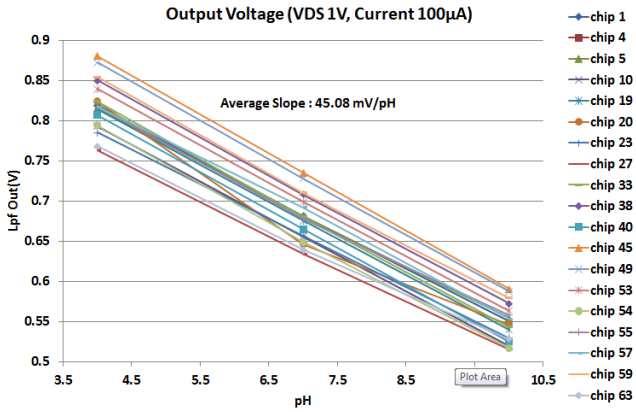


Fig. 22 – ChemFET ROIC block functional test result 4

So far, functional result has been shown for each individual block within the ASIC chip. Currently more extensive characterization work is still undergoing to check the linearity of the ADC, DAC and the interface with micro-controller.

IV. Conclusion

A sensor ASIC has been designed that caters to electrochemical sensor readout in agriculture applications with features of small area, low pin count, high re-configurability and power down options. The sensor ASIC has been fabricated and functionally tested to satisfy its performance criteria.

References

[1] M. Mailand, S. Getzlaff, D. Breitmeyer, R. Richter, U. Meyer, E. Boehme,A. Krasselt, "A Flexible, Programmable Sensor Signal Conditioning IC for High-Precision Smart Sensors, Semiconductor Conference Dresden (SCD), 2011

[2] Jinwen Xi, Chao Yang, Andrew Mason and Peixin Zhong "Adaptive Multi-Sensor Interface System-On-Chip", IEEE Sensors 2006, EXCO, Daegu, Korea / October 22~25, 2006

[3] K.Tukkineami, Study of CHEMFET interface electronics, MIXDES (2002)

[4] B. Palan, F.V. Santos, J.M. Karam, B. Courtois, M. Husak, New ISFET sensor interface circuit for biomedical applications, Sens. Actuators B 57 (1999) 63–68
[5] Wang, J. Analytical electrochemistry, Jonh Wiley & Sons: New York, 2nd, 2000
[6] MIMOS Copyright

Application of charge amplifiers with Polyvinylidene Fluoride materials

A.S. Marinov*, Or. P. Stanchev* and E. B. Bekov*
* Technical university of Varna/Department of electronics and microelectronics, Varna, Bulgaria
or.stanchev@gmail.com

Abstract - The current paper proposes an improved specialized charge amplifier as an interface for measurement based applications that utilize Polyvinylidene Fluoride (PVDF) material. As this material and its interfacing electronics are still being researched, most of the applications and topologies are in development. PVDF material characteristics are reviewed and set as a basis for the design of the proposed charge amplifier. Experimental results of the proposed circuitry are compared to conventional voltage amplifiers. The charge amplifier characteristics and benefits are presented and discussed. Methodology for design of charge amplifier, applied to several measurement applications that utilize PVDF, is discussed.

I. INTRODUCTION

Transducer materials convert one form of energy to another and are widely used in sensing applications. The piezoelectric polymer sensors are among the fastest growing technologies on the sensor market. There is a considerable number of applications where the piezoelectric polymer film could be a proper solution.

The polyvinylidene fluoride (PVDF) is a polarized fluoropolymer which has very high piezo-activity. It is also pyroelectric, producing electrical charge in response to changes in temperature.

The PVDF material is extruded into a thin film that is flexible, lightweight and tough plastic. It also can be directly attached to a structure without disturbing its mechanical motion. The piezo film properties as transducer include [1]:

- Wide frequency range from 10^{-3} Hz to 10^{8} Hz;
- Vast dynamic range from 10^{-8} to 10^{6} psi;
- Low acoustic impedance that is close to water, human tissue and adhesive systems;
- High elastic compliance;
- High voltage output that is 10 times higher than piezo ceramics for the same force input;
- High dielectric strength, withstanding strong fields up to 75 V/μm where most piezo ceramics depolarize;
- High mechanical strength and impact resistance;
- High stability, resisting moisture, most chemicals, oxidants and intense ultraviolet and nuclear radiation;
- Fabrication into unusual designs;
- Ability to be glued with commercial adhesives.

The piezo film has acoustic impedance that is very close to water and human tissue. This feature makes it suitable for medical applications. It is also used in the automotive industry and other areas for vibration sensing, impact sensing, accelerometers, audio and ultrasonic applications. Application in meteorological sensors is also being researched [2][3].

A major drawback is that the piezo film is relatively weak electromechanical transmitter when compared to ceramics, particularly in low frequency applications. Also the PVDF material is not recommended for use at temperatures above 100 °C. If the electrodes of the film sensor are exposed, became very sensitive to electromagnetic radiation. Therefore good shielding techniques should be used for high EMI/RFI environments.

The typical piezo film sensors are rectangular elements of piezo film with silver ink screen printed electrodes on both sides and protective coating. The construction is shown on Fig. 1. The PVDF film strip has thickness of 9, 28, 52 or 100 μm. Forces or stresses applied to the PVDF elements are considered in three main dimensions – length, width and thickness. The highest conversion is obtained when the mechanical force is applied to the PVDF element over its thickness, because of the very small cross-sectional area and thus relatively small

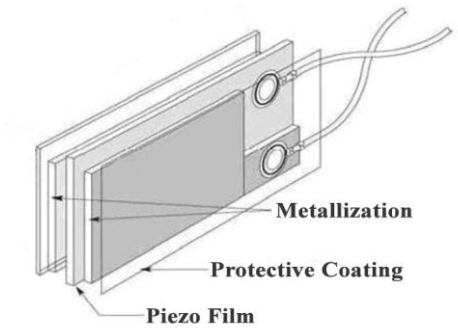


Figure 1. Construction of PVDF film sensor [1]

The carried out research is realized in the frames of the project BG051PO001-3.06-0005, Program 'Human Resources Development', Ministry of education, youth and science, Bulgaria.

longitudinal forces. They create very large stresses within the material. After applying mechanical force, electrical charge occurs on the surface of the PVDF element. This charge is directly proportional to the mechanical deformation and causes a voltage between the electrodes on the surface of the element.

The developed charge of the PVDF sensor elements for small stress is given by:

$$Q = A \cdot d_{3n} \cdot X_n, \text{ for } n = 1, 2 \text{ or } 3 \quad (1)$$

where Q is the developed charge, A is the conductive electrode area, d_{3n} is the piezoelectric coefficient, X_n is the applied stress and n is the axis of applied stress.

The output voltage of the PVDF sensor elements for small stress is given by:

$$V = g_{3n} \cdot X_n \cdot t, \text{ for } n = 1, 2 \text{ or } 3 \quad (2)$$

where V is the output voltage, g_{3n} is the piezoelectric coefficient, X_n is the applied stress, t is the film thickness and n is the axis of applied stress.

The equivalent circuit of piezo film sensor is shown on Fig. 2. It consists of capacitance and a voltage source. Fig. 2 shows the simplified equivalent circuit. The series capacitance represents the piezo film capacitance and the voltage source amplitude is equal to the open circuit voltage of the piezo film. This simplified equivalent circuit can be used in most cases except at very high frequencies that are close to the ultrasonic wave range.

II. USING AMPLIFIERS WITH POLYVINYLIDENE FLUORIDE

The Polyvinylidene Fluoride is a new material and in the technical literature there is limited information considering the material and the interface circuits for PVDF sensors. The interface circuits' manufacturers offer solutions that usually are not fully described or powered with single supply [6].

The presented research is focused on investigation of the interaction between PVDF sensor elements and interface circuits, especially the comparison between charge and voltage differential amplifiers.

A. Voltage Amplifiers

The piezo film sensor can be used as a voltage source when the information value is the output voltage, generated between the electrodes. As the primary output of the PVDF sensor is charge, the indirect voltage is obtained by placing a resistor in parallel to the output (Fig. 2). The parallel resistance and the piezo film capacitance form RC group and due to the frequency response of this network it acts like a high-pass filter. The cut-off frequency of the filter is given by:

$$f_c = 1 / (2 \cdot \pi \cdot R \cdot C) \quad (3)$$

For low frequency measurement the resistance has to be high enough to ensure that the cut-off frequency is below the operational one so that signal reduction is avoided.

When the PVDF sensor is used for vibration or low stress measurement, the voltage output has to be amplified. A typical one stage non-inverting amplifier configuration is proposed by the piezo film manufacturers to be used when high signal to noise ratios are desirable [1]. The non-inverting configuration has high input impedance that is desirable.

When the signal to noise ratios are low an instrumentation amplifier circuit is more appropriate in that case for precise measurements due to its very high common-mode rejection ratio and input impedance [4].

The advantage of the voltage amplifier can be seen when ambient temperature is considered. The voltage sensitivity variation over temperature is smaller than the charge sensitivity variation. Consequently, voltage amplifiers with piezo film exhibit less temperature dependence.

When the piezo film sensors are used with voltage amplifier the capacitance of the film and the connecting cables have adverse effect on the measurement network transfer function.

B. Charge amplifiers

The PVDF film sensors have capacitive nature and charge amplifiers can be used. Fig. 3 shows a basic charge amplifier structure, proposed by the piezo film manufacturers [1]. The gain of the given charge amplifier is the ratio between the piezo film sensor characteristic capacitance C and the feedback capacitance C_f. The output voltage of the charge amplifier is determined by the charge Q developed by the piezo film and the feedback capacitance C_f of the charge amplifier.

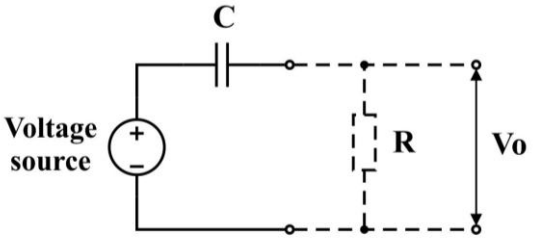


Figure 2. Equivalent circuit of piezo film sensor [1]

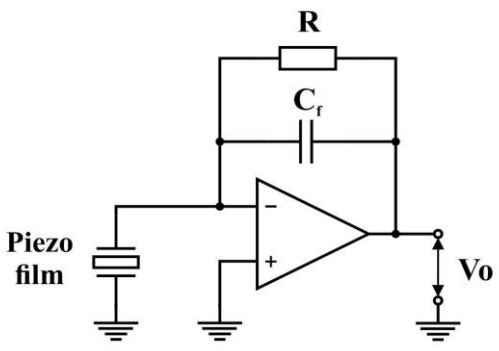


Figure 3. Basic charge amplifier [1]

The output voltage is given by:

$$V = - Q / C_f \quad (4)$$

The output voltage is very likely to drift. A proper compensation technique is placing a resistor in parallel to the feedback capacitor. This resistor creates a time constant which is independent from the film capacitance and can be precisely controlled. It determines a cut-off frequency that limits the input signal which is similar to the high-pass filter that is present in the piezo film voltage amplifier topology. The time constant for charge amplifier is given by the feedback capacitor C_f and the parallel resistor R:

$$\tau = R \cdot C_f \quad (5)$$

The piezo film is very sensitive to EMI disturbances. It also has strong pyro sensitivity for the thermal radiation in the range from 7 to 20 μm and care must be taken when designing low frequency mechanical sensors to avoid disturbances from ambient temperature changes. Common-mode rejection can be used to isolate either very low frequency mechanical strain from simultaneous pyro-effects and EMI.

The most critical part of the measurement network is the input resistance. The input resistance affects low frequency measurement capability and the signal amplitude. The charge amplifier requires operational amplifiers having high input impedance and low bias current so that operational amplifiers with JFET input stage transistors are appropriate. A high input resistance avoids charge leakage from the feedback capacitor and low bias currents prevent feedback capacitor from parasitic charging and discharging.

Considering different amplifier topologies a two stage differential dual supply charge amplifier is proposed for processing the signal from PVDF sensor element [5][6]. It has similar advantages in noise cancellation like the instrumentation amplifier. The first stage consists of input buffer charge amplifiers with low gain close to unity, high input impedance and high rejection ratio. The second stage consists of differential amplifier with high gain, providing the necessary output signal magnitude. A major problem with the two stage charge differential configuration is the matching condition between the feedback capacitors and resistors that could cause asymmetry.

III. Methodology for PVDF Charge Amplifier Design

The proposed methodology for design of charge amplifier for PVDF film sensor consists of the following steps:

- Selection of a proper sensor according to application specifics;
- Determination of the signal bandwidth and possible noise frequency (for example the 50 Hz grid supply voltage);
- Calculation of the maximum charge, generated by the piezo film sensor, due to the maximum applied force (1) and the piezo strain constants;
- Choosing a charge amplifier topology (one stage or two stage differential) due to the expected signal to noise ratio;
- Choosing proper high input impedance JFET operational amplifiers;
- Calculation of the feedback capacitor, depending on the amplifier topology and desired output signal magnitude (4);
- Calculation of the time constant (5), liming the signal lowest frequency;
- Calculation of the output low-pass filter, limiting the signal spectrum.

IV. Experimental Verification

A. Experimental Setup

The proposed voltage and charge amplifiers are compared experimentally. Both amplifiers have one stage and use high quality JFET OPA2141 operational amplifiers [8].

The experimental setup is shown on Fig. 4. It has the following principle of operation. A vibration generator applies simultaneously a mechanical stress on two identical PVDF film sensors, mounted under similar conditions. The developed charge of the first piezo film and the output voltage of the second piezo film are amplified via charge and voltage amplifiers. The amplifiers' outputs are filtered with low pass filters. The two measurement networks are independent, but operate under the same conditions. Measurements are taken with digital oscilloscope and processed with personal computer

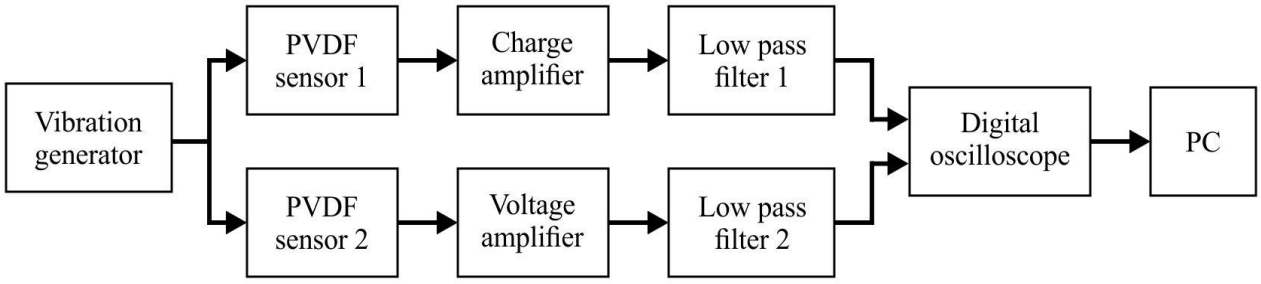


Figure 4. Experimental setup

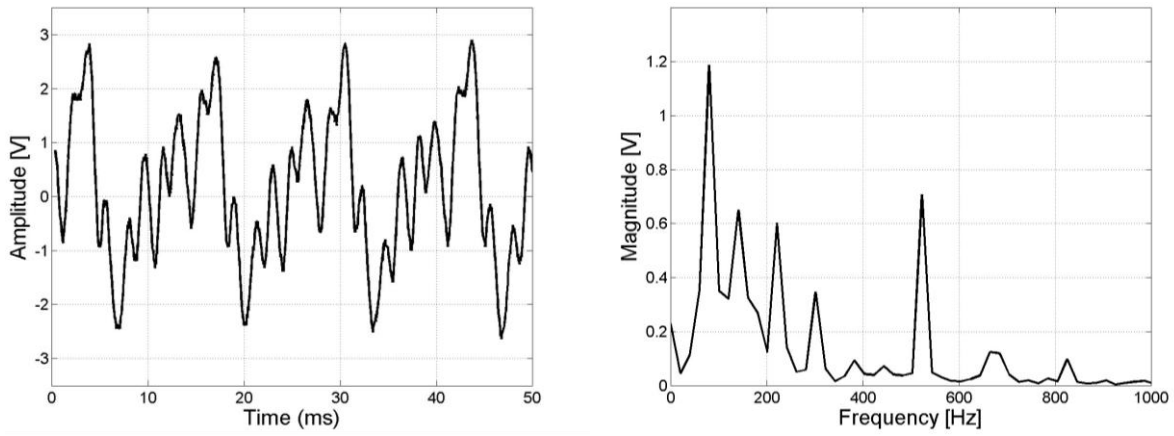


Figure 5. Charge amplifier output at periodic impact – time domain signal and frequency spectrum

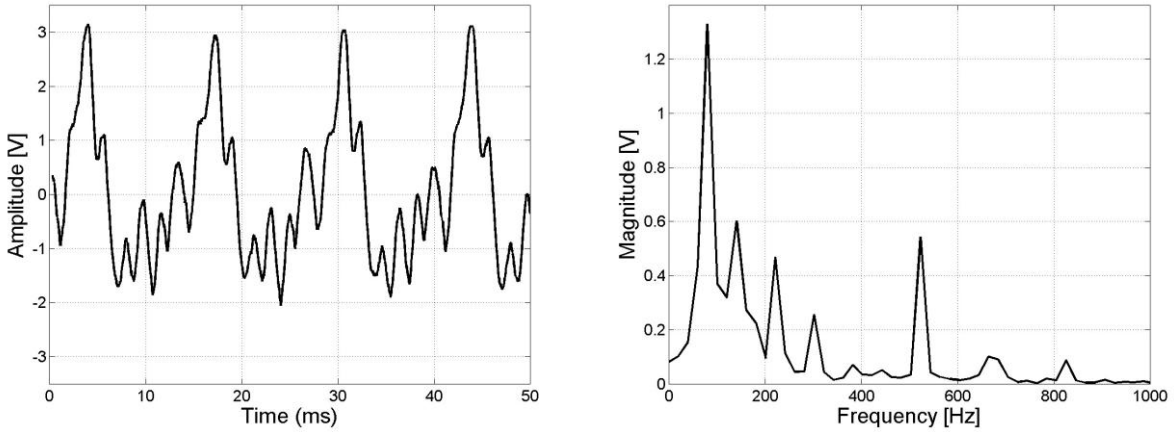


Figure 6. Voltage amplifier output at periodic impact – time domain signal and frequency spectrum

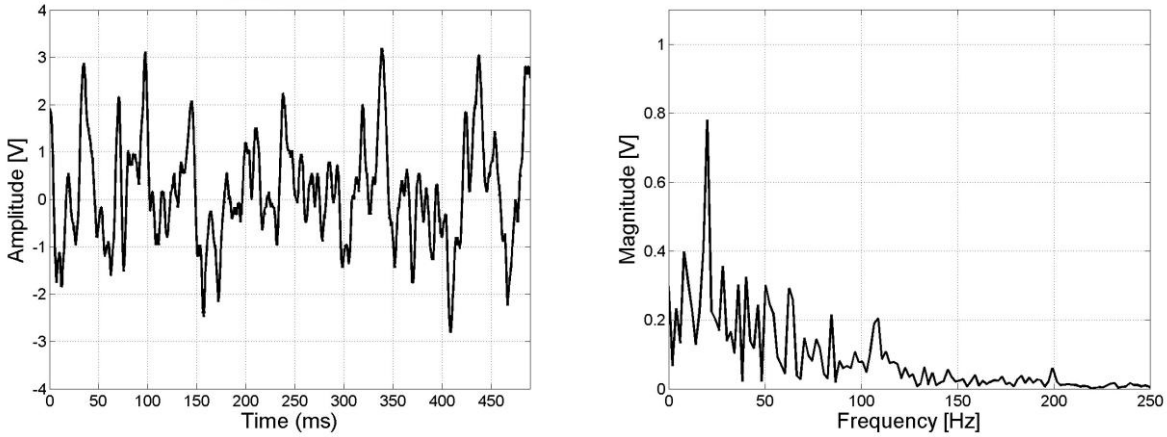


Figure 7. Charge amplifier output at aperiodic (random) impact – time domain signal and frequency spectrum

and specialized software for signal processing.

The used sensors are two LDT4-028K PVDF film strips with 28 μm film thickness and 11 nF capacitance [7].

The examined charge amplifier has a similar structure to the amplifier shown on Fig. 3. The feedback consists of 1.5 nF capacitor and 3.3 MΩ resistor with a gain of 7.33 and 5 ms time constant. The amplifier PCB layout is designed due to the specifics for charge amplifier input pins routing and shielding [1].

The examined voltage amplifier has a standard non-inverting single operational amplifier structure. A 56 kΩ feedback resistor and 56 kΩ input resistor are used with a gain of 2. Parallel to the input a 510 kΩ resistor is placed setting the voltage mode and giving a 28.3 Hz cut-off frequency.

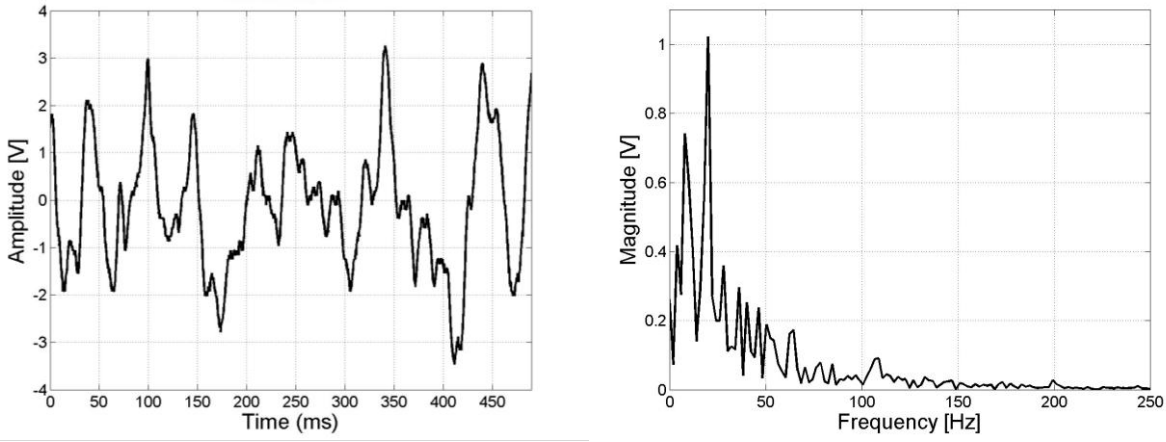


Figure 8. Voltage amplifier output at aperiodic (random) impact – time domain signal and frequency spectrum

The passive output low pass-filters limit the signal bandwidth to 400 kHz (matching the ADC input signal requirements).

Both amplifiers are powered with filtered DC dual supply that has extremely low voltage ripple as the charge amplifier is very sensitive to unstable power supply.

B. Experimental Results

The experiment is accomplished in two parts – in the first part a periodic vibration is applied to the piezo film sensors and in the second – a vibration with aperiodic (random) nature. The time domain signals and frequency spectra of the output signals of both amplifiers are shown on Fig. 5 and Fig. 6 from the first part and on Fig. 7 and Fig. 8 from the second part of the experiment.

The experimental results demonstrate that the charge amplifier provides better resolution of details in the output shape of the signal, especially when the piezo film sensor output has more oscillatory nature. The signal frequency spectrum consists of more high harmonics with bigger magnitude so that the signal carries more information for the primary excitation of the sensor.

V. CONLUSION

The Polyvinylidene Fluoride sensors are more and more used in important areas like medicine and automotive industry. Thus the measurement networks and interfaces for data processing of the sensors' outputs are constantly developed and improved.

The piezo film sensors have enhanced capacitive nature so that using charge amplifiers is recommended. It is experimentally proven that they provide better resolution of details in the output shape of the signal and in the frequency spectrum, when the piezo film sensor output has more oscillatory nature. A configuration for differential charge amplifier is discussed. A methodology for charge amplifier design is proposed and used for experimental amplifier design.

ACKNOWLEDGMENT

The carried out research is realized in the frames of the project BG051PO001-3.06-0005, Program 'Human Resources Development', Ministry of education, youth and science, Bulgaria.

REFERENCES

[1] Measurement Specialtiets, "Piezo Film Sensors Technical Manual," http://www.meas-spec.com, March 2008

[2] A. Marinov, E. Bekov, V. Valchev, "PVDF Based Wind Direction and Speed Sensor for Weather Assessment Relevant to Renewable Energy Generation," EPE-PEMC 2012 ECCE Europe, September, 4-6, 2012, Novi Sad, Serbia.

[3] E. Bekov, A. Marinov, V. Valchev, "PVDF based rain sensor for weather assessment relevant to renewable energy sources," PCIM Europe 2012, 8-10 May 2012, Nuremberg.

[4] J. Wilson, "Sensor Technology Handbook," Elsevier, USA, 2005

[5] M. Massarotto, A. Carlosena, A. J. Lopez-Martin, "Two-Stage Differential Charge and Transresistance Amplifiers," Instrumentation and Measurement, IEEE Transactions on vol. 57, issue 2, pp. 309-320, February 2008

[6] Maxim Integrated, "Amplifier Provides Signal Conditioning for Piezofilm Sensor," Application Note 1127, June 2002.

[7] Measurement Specialties, "LDT1-028K Piezo Sensor," http://www.meas-spec.com/

[8] Texas Instruments, www.ti.com

Design of Transdermal Drug Delivery System with PZT actuated micropump

T. Dolžan, D. Vrtačnik, D. Resnik, U. Aljančič, M. Možek, B. Pečar and S. Amon
University of Ljubljana, Faculty of Electrical Engineering,
Laboratory of Microsensor Structures and Electronics
Ljubljana 1000, SLOVENIA
tine.dolzan@fe.uni-lj.si

Abstract – Investigation of transdermal drug delivery system, based on silicon injector chip with silicon microneedles and PZT (lead zirconate titanate) actuated micropump, for drug delivery such as insulin is reported. First, basic skin properties and dosing requirements are reviewed. Next, the whole system and its components (injector chip, micropump, drug container) are discussed. Based on this, microneedles injector chip design is discussed and prototypes fabricated. Then, design of PZT actuated micropumps is discussed and prototypes fabricated. From measurements on fabricated micropumps, fluidic parameters are evaluated. Basic requirements for drug container, control electronics and power supply parameters are also presented. Finally, maximum drug delivery, based on measured prototype micropump parameters, through the prototype injector chip is determined.

I. INTRODUCTION

Microneedles transdermal drug delivery systems are recently gaining influence in various medical applications. When designing a transdermal drug delivery system, good knowledge of the skin is essential as its anatomy and properties will determine many important parameters of the whole system. Human skin acts as a natural barrier and protects the body from surroundings. When we want to deliver a drug through the skin, we have to break this barrier in some way. To do this efficiently, we have to know exactly how skin functions. Skin is composed of three layers (Figure 1), from top to bottom they are: stratum corneum, epidermis and dermis [1]. The most upper layer, stratum corneum, is a natural barrier which we have to break or remove to effectively deliver drug into the body. Stratum corneum is made of dead skin cells and is approximately 10 to 20 μm thick. Epidermis is made of keratinocytes and contains no blood vessels. It is approximately 50 to 150 μm thick [1]. Dermis is the bottom layer, which contains blood vessels and nerve endings. If we want to painlessly deliver a drug into the skin, we have to break stratum corneum and deliver the drug into dermis, from where it will diffuse into the blood vessels and then into the body.

Next we need to know what kind of drug will be injected in the body and what the required quantities are. This will determine the flow rate and drug container capacity. Viscosity of drug solution will also be an important issue for proper determination of flow rate needed.

In our case transdermal drug delivery system will be mainly used for delivery of insulin for basal needs. Standard insulin preparations contain 100 U/ml of active substance. Healthy human body daily produces on average 0,7 units of insulin per kg of body mass (0,5 – 1,0), that is about 50 units in the whole day for an average person with body mass of 70 kg. Approximately half of the produced insulin is used for covering basal metabolic needs. There are two types of diabetes, Type 1 and Type 2. Patients with Type 2 diabetes still have some own insulin and so need smaller doses, therefore transdermal drug delivery system should be able to deliver up to 0,30 ml of insulin preparation daily for Type 1 patients and less for patients with Type 2 diabetes. Based on those facts, drug delivery system should be capable to inject an appropriate amount of drug, depending on the patient's individual needs. It is desirable that the dose of the drug would be easily adjustable for every patient.

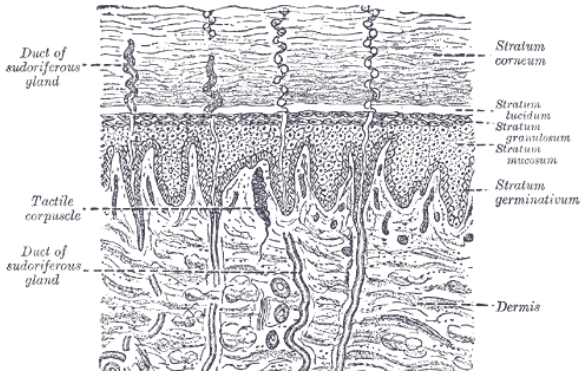


Figure 1: Skin cross-section [2].

II. MICRONEEDLES TRANSDERMAL DRUG DELIVERY SYSTEM BASICS

Main components of an advanced, autonomous microneedles drug delivery system usually consist of: injector chip with microneedles, micropump, drug container, control electronics and power supply. All components should be small enough to fit in the compact microhousing that could be attached on the skin.

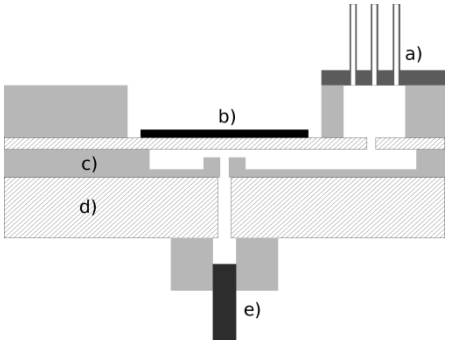


Figure 2: Assembled system: a) silicon micronedles array; b) PZT actuator on glass membrane; c) PDMS layer with micropump structure; d) support glass; e) fluidic connection

All components are assembled together as shown in Figure 2 (electronics and power supply are not presented), with smallest possible number of assembly steps.

III. SILICON MICRONEEDLE INJECTOR CHIP

Microneedles can be made from many different materials. Silicon is one of the standard MEMS materials, from which different types of precise 3D microneedle structures can be made. This is why silicon was selected for our microneedle injector chip material. As injector chip will be in contact with living tissues, silicon biocompatibility is also an important issue in this case [3].

Based on skin anatomy, microneedles must be long enough to pierce stratum corneum and deliver drug to the epidermis where capillaries are present to take and transport the drug into the body. On the other side, needles must not be too long to penetrate deeper into the tissue where nerves begin, to avoid pain. Combined, dermis and stratum corneum are typically around 110 µm thick. As the skin bends when needle is inserted by force, we need longer needles than the thickness of stratum corneum, but short enough not to pierce the dermis. Also based on findings of different authors, we concluded that the needles should be 150-200 µm long [4], [5], [6].

As we have an array of many needles, insertion force is distributed uniformly across the whole array. Insertion force should be small enough to enable simple insertion of the injector chip by hand. Therefore, to achieve small insertion force, microneedles must have sharp tips.

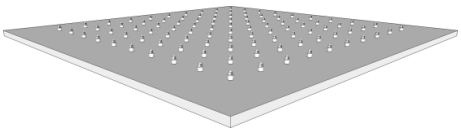


Figure 3: Design of 10x10 matrix injector chip.

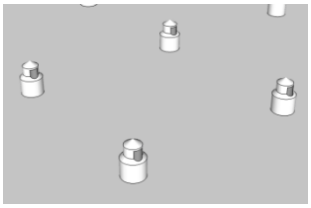


Figure 4: Detailed view of 10x10 microneedle array design, distance between needles is 1mm.



Figure 5: Detailed view of designed single microneedle with side opening.

Total delivered drug flow through all needles is directly dependent on needle bore and number of needles. Obviously, if we have smaller needle bore, we must add more needles to the array, to achieve equal flow rate.

For our first prototype we decided to fabricate 200 µm long needles with base diameter of 130µm and bore diameter of 50 µm. On injector chip there will be a 10x10 matrix of 100 needles (Figure 3). Figure 4 represents a detailed view of the injector chip. The needles will have sharp tips to guarantee minimal needed insertion force and side openings to prevent clogging after insertion (Figure 5).

The basic fabrication process for injector chip with microneedles consists in our case of proper interchanging silicon anisotropic and isotropic etching steps.

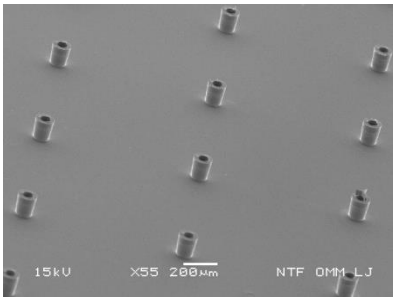


Figure 6: Fabricated prototype injector chip with hollow microneedles.

Recently we have successfully fabricated prototype injector chips, containing arrays of 10x10 microneedles, one type with 200 µm high solid sharp microneedles and second type with 200 µm high hollow microneedles with opening on the top. On Figure 6, fabricated injector chip with hollow microneedles is shown. Fabricated injector chips will enable us to set up experimental system for evaluating basic parameters of needle penetration and to investigate properties of injecting drug liquids through the skin into the body.

IV. PZT ACTUATED MICROPUMP

Micropump is the main active component of the system and must be capable of producing high enough pressure and liquid flow to achieve delivery of adequate drug dose. At the same time micropump should be small enough and as thin as possible. Piezoelectric actuated pump can fulfill all this requirements [7].

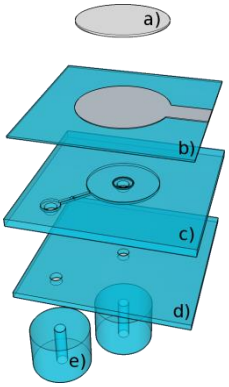


Figure 7: Designed micropump structure: a) PZT actuator; b) glass membrane with conductive glue; c) PDMS layer with micropump structure; d) support glass; e) prototype fluid connectors

Selected prototype micropump is in our case made from PDMS, which is cast into the silicon mold. Mold is created with our standard MEMS processing for the fabrication of two level structures. PDMS structure is afterwards bonded on support glass and glass membrane on the other side of the structure. PZT actuator with contacts is attached on a glass membrane as shown in Figure 7. All used materials in contact with drug solution are known to be biocompatible. More details on PDMS micropump design and fabrication are given in [8].

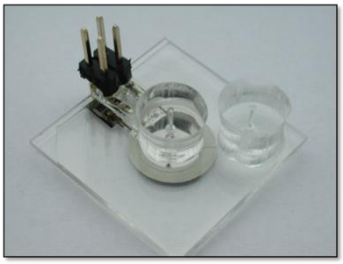


Figure 8: Assembled micropump prototype.

Assembled micropump prototype is shown on Figure 8, with two throttles achieves maximal backpressure of 450 mbar (50 kPa) and maximal flow rate of 2,4 ml per minute (144 µl/s). Micropump test actuation signal has amplitude of 250 V and frequency of 110 Hz.

V. DRUG CONTAINER, CONTROL ELECTRONICS AND POWER SUPPLY

Container capacity should be large enough to hold volume of drug solution. For example, based on a typical patient daily need (0,30 ml), container should have capacity of 1 ml for autonomy of three days, before refilling or changing. Because drug container is in contact with drug solution, it will be made from biocompatible materials. As drug container has only limited capacity, option to refill or exchange whole drug container without replacing entire system is under study.

Control electronics also presents some unique design challenge. For actuation of PZT element on micropump, rather high voltages are needed. Control system will also have to incorporate some time/dosage control and possibility of adjustment on demand. Because the system should be autonomous, small power consumption and size will be critical.

Power supply is mainly limited by small size of the system and will have to supply enough power for micropump actuation and electronics over a predetermined period of time. In addition to integrated battery, it is possible to include some sort of energy harvesting device. That could provide all or some extra power for the system, depending on actual system consumption and harvesting device power generation capacity.

VI. DISCUSSION

Pressure needed to inject drug solution in the tissue depends on many different factors, as are skin properties on site where needles are inserted, depth of insertion, clogging and others, so it is nearly impossible to asses all this analytically.

The basic microfluidics parameters of our problem are connected by the following equations [9]

$$\Delta P = \frac{128\mu LQ}{n\pi d^4}$$

$$\Delta P = aQ + b \qquad (1)$$

$$Q = \frac{100b\pi d^4}{128\mu L - \pi d^4 a}$$

The first equation gives pressure drop on injector chip at volumetric flow rate Q of fluid with dynamic viscosity μ, through needles with bore diameter d, channel length L and number of needles n. The second equation is our approximation for measured linear connection between micropump generated backpressure and flow rate. In the third equation, derived from first and second equation, maximum possible volumetric flow rate is given as a function of injector chip, micropump parameters and viscosity of used fluid.

For now, we can calculate volumetric flow rate when microneedles are not inserted into skin. This calculation will provide us the theoretical maximum possible volumetric flow rate through given microneedle array. For pressure difference we will use actual backpressure that our prototype micropump can achieve, that is 450 mbar. Maximum flow at zero backpressure is therefore 2,4 ml per minute. From measured data we can derive needed coefficients a and b in second equation of (1). As insulin solution contains more than 90% DI water, we will in our calculations as a first approximation use the viscosity of DI water.

Calculations based on given equations have revealed basic relations between critical parameters of delivery system. As a practical example, if we could deliver drug into the skin with this rate, only two delivery intervals in duration of approximately 5 seconds daily would be enough to deliver daily amount of 0,30 ml of drug solution to the patient with Type 1 diabetes.

VII. CONCLUSION

Investigation of a microneedles transdermal drug delivery system is reported. Design and fabrication of system basic components such as injector chip with silicon microneedles and PZT actuated micropump, for drug delivery such as insulin, is presented. Based on skin properties, microneedles injector chip design is discussed and prototype injector chips were fabricated. Design and fabrication of PZT actuated micropumps and evaluation of fluidic parameters for manufactured prototype micropumps is reported. Basic requirements for drug container, control electronics and power supply parameters are discussed. Finally, maximum drug liquid flow through the prototype injector chip, based on measured prototype micropump parameters, was determined.

ACKNOWLEDGMENT

Authors wish to thank ARRS (grant L2-4122, P2-0244) and Center of Excellence NAMASTE for their support.

REFERENCES

[1] N. Roxhed, A Fully Integrated Microneedle-based Transdermal Drug Delivery System, Stockholm, 2007.

[2] H. Gray, Anatomy of the Human Body, 1918.

[3] G. Voskerician, M. S. Shive, R. S. Shawgo, H. Recum, J. M. Anderson, M. J. Cima and R. Langer, "Biocompatibility and bio fouling of MEMS drugdelivery devices," *Biomaterials*, vol. 24, no. 11, p. 1959–1967, 2003.

[4] D. V. McAllsiter, F. Cros, S. P. Davis, L. M. Matta, M. P. Prausnittz and M. G. Allen, "Three-dimensional hollow microneedles and microtube arrays," in *The 10th Int. Conf. on Solid-state Sensors and Actuators*, 1999.

[5] S. P. Davis, W. Martanto, M. G. Allen and M. R. Prausnitz, "Hollow Metal Microneedles for Insulin Delivery to Diabetic Rats," *IEEE Transactions Biomedical Engineering*, vol. 52, no. 5, pp. 909-915, 2005.

[6] M. A. L. Teo, C. Shearwood, K. C. Ng, J. Lu and S. Moochhala, "In vitro and in vivo characterization of MEMS microneedles," *Biomedical Microdevices*, vol. 7, no. 1, pp. 47-52, 2005.

[7] B. Pečar, D. Vrtačnik, D. Resnik, M. Možek, U. Aljančič and S. Amon, "Bubble Tolerant Valveless Piezoelectric Micropump for Liquids and Gasses," in *MIDEM Conference Proceedings*, 2011.

[8] B. Pečar, D. Križaj, D. Vrtačnik, D. Resnik, M. Možek, U. Aljančič, T. Dolžan and S. Amon, "Numerical Simulation of STMT Piezoelectric Micropump operation," in *MIDEM Conference Proceedings*, 2012.

[9] B. J. Kirby, Micro- and Nanoscale Fluid Mechanics: Transport in Microfluidic Devices, Cambridge: Cambridge University Press, 2010.

Modern Monitoring System for Water Pollution by Petrochemical Products based on Optical Fibre Refractometer

J. Turán* and Ľ. Ovseník*
*Technical University of Košice/Department of Electronics and Multimedia Communications, Košice, Slovakia
jan.turan@tuke.sk, lubos.ovsenik@tuke.sk

Abstract - Modern industrial quality monitoring systems are based on optical fibre sensors, microelectronic sensory technology and miniaturized infrared spectroscopy systems based on low-power, low-cost, high-efficient and innovative hardware solutions. The proposed paper presents development work related to create remote control system for monitoring water pollution petrochemical products using optical fibre refractometer.

I. INTRODUCTION

The current European Water Framework Directive, implemented in the European Union (EU) is a structured programme for management of water resources through Integrated River Basin Management (IRBM), with the aim to achieve overall good quality of water by 2015. This is accomplished through careful monitoring of surface water and groundwater habitats in Europe and implementation of stepwise improvements gradually to attain a 'good ecological status' which is the ultimate goal [1].

Environmental water quality, also called ambient water quality, relates to water bodies such as lakes, rivers, and oceans. Water quality standards for surface waters vary significantly due to different environmental conditions, ecosystems, and intended human uses. Pollutants may take various forms and as a consequence, several types of continuous on-line measurement may be performed. Sensors may be installed to provide indicative measurements or specific parameters may be monitored when the potential pollutant is known. A typical water pollution monitoring system may therefore comprise a suite of instruments or sensors to cover various aspects of chemical and physical water quality [2,3].

Monitoring is a process by which the industry measures its emissions, releases, and/or parameters of its performances that are to provide information on the release of polluting materials and/or performance of control technologies. It provides secret or public information on the nature of pollutants emissions and/or the performance of purification plant. Industrial monitoring of emissions from the industry is performed by the purification plant` s operator, in accordance with appropriate, defined and amended programme for taking samples and in accordance with accepted measuring protocols (standards or analytical methods or evaluation methods).

Data on industrial monitoring, integral control and pollution prevention in Oil and Petrochemical Industry are important for environment pollution prevention. By defining the place of pollution we determine control and pollution reduction measures with the possibility of choosing the type of monitoring and the dynamic of measuring the pollutants. The designed modern industrial water pollution monitoring system presented in the paper is based on optical fibre refractometer technology based on low-power, low-cost, high-efficient and innovative hardware solutions, including WWW remote control. In Chapter 2 is described architecture of photonics multimedia web-based laboratory system. Chapter 3 contains design of interactive web controlled fibre optic refractometer instrument. Experiments and results are presented in Chapter 4. Conclusions are covered in Chapter 5.

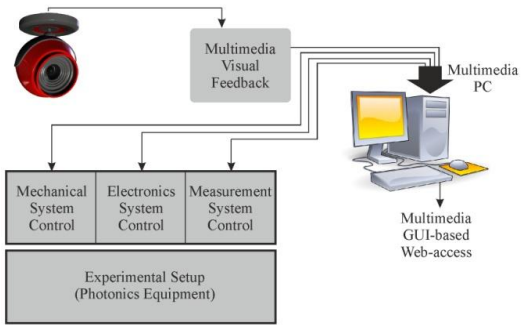


Figure 1. The architecture of monitoring system with optical fibre refractometer.

II. MONITORING SYSTEM WITH OPTICAL FIBRE REFRACTOMETER

The architecture of monitoring system with optical fibre refractometer is on Figure 1. The core of the system is experimental setup to realize chosen photonics equipment - Fibre Optic Refractometer. Experimental setup is controlled trough three data acquisition systems: a) Mechanical b) Electronics and c) Measurement, to provide a desired laboratory function of the photonics equipment and measurement of its parameters. The used simple web-camera assists for multimedia visual feedback in the telelaboratory (monitoring system).

Fibre optics sensors of various constructions are valuable devices for wide industrial applications [4,5]. Using fibre optics technology and micro-optics is possible to design innovative fibre optic index of refraction transducer that has unique properties. Basically it consists of input - output pair of simple multimode fibres that interrogate a small lens. The cone of light injected into the lens from the input fiber is internally reflected from inside surface of the lens and focused back into the output fibre. When the outer surface of the lens is in contact with a liquid, the attenuation of the light reaching the output fibre depends strongly on the refractive index of the liquid. Depending on the construction and design of such a transducer, this transducer may have a wide dynamic range. Its relative sensitivity, i.e. the ratio of fractional change of the optical intensity for a given change in index of refraction, is substantial, (of the order of 5 to 10) over a wide range of index, e.g. n = 1.3 to 1.6. For convenience, usually employed lenses have diameter from 5 to 1 mm. In principle, it should be possible to go to even smaller diameters, e.g. 250 to 300 microns - so that relatively small catheter type transducers are feasible. Since the basic design is quite simple it also is possible to produce extremely rugged transducer elements that are suitable for field/industrial applications [6,7,8].

This fibre optic small, probe type structure transducer elements, can be easily inserted into the top of liquid containers or through a simple fitting can be simple to insert in a flow line. It is essential to combine these very practical and versatile transducers capable of detecting small changes of index of refraction with "smart" data acquisition and signal processing technology to develop versatile and practical instruments. Refractometers frequently are used for medical, pharmaceutical, industrial fluid, chemical, petrochemical, plastic, food, and beverage industry applications. For example, they are used for measuring the concentrations of aqueous solutions: as the concentration or density of a solute increase, the refractive index increases proportionately. Included in such measurements is the percentage of sugar in fruits, soft drinks, canned syrups and other solutions. They also are used to determine the salinity of aquariums and of solutions used in food processing, the freezing point of coolants and deicing fluids, the charge status of acid batteries, the serum protein and urine specific gravity, etc.

Almost all currently available classical refractometers employ a prismatic element on which the liquid sample is placed. These instruments yield an output that is based on the degree of a light beam at the liquid-prism interface. In the simplest refractometers, the index is determined visually, via a lensed viewing tube, by observing where the output beam intersects a graduated scale. In more modern, digital type instruments, the degree of bending is measured automatically using a linear photodetector array [1,4]. Since the index of refraction is strongly dependent on temperature and also, to a lesser degree, on wavelength, these effects must be corrected for in designing and/or using such instruments. They are of major importance when one attempts to determine an index of refraction to 1 part in 10.000 or better, which is the desired sensitivity for a high quality, laboratory type instrument. Their importance can easily be seen in the case of water, for example, which has a temperature coefficient $\Delta n/\Delta T$ of 1,5 parts 10.000 per °C.

As with any other refractometer, any instrument that employs this fiber optic based transducer must be capable of correcting for the intrinsic temperature dependence of a liquid's index of refraction [4,9]. In addition, however, for an intensity type fiber optic sensor, other corrections and precautions must be taken especially if, as already mentioned, a precision of 1 part in 10.000 is to be attained. It will be necessary to employ low noise electronic circuitry and/or correct for photodetector dark current, especially at higher indices, where the output light intensity is strongly attenuated. It also will be necessary to correct for light source and photodetector temperature sensitivities and for any stray light that might affect the photodetector. In one sense, in terms of capabilities of today's microprocessor controlled "smart" sensor technology, it should be straightforward to design instruments that automatically "massage" the raw transducer data to correct them for each of these effects.

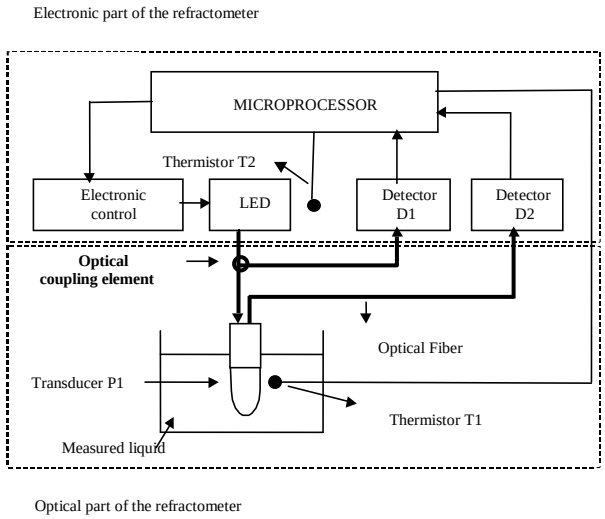


Figure 2. Block diagram outlining the design of a basic fiber optic refractometer.

Referring to the block diagram in Figure 2, the basic system, as presently conceived, consist of the following elements: a light emitting diode (LED) or semiconductor laser diode (LD) light source and its electronic driver/pulser; a monitor photodetector to determine the output light level of the use light source, a second photodetector to record the return light from the transducer; and a microprocessor to automatically control the system and process the data from the various elements. It will compute the ratio of the intensity of the transducer output for an unknown liquid and that recorded earlier for a standard liquid, e.g., water.

The microprocessor will then determine an index value, either by a comparison and interpolation process between this ratio and those in a calibration data lookup table, or by computation using a transducer response equation. In addition, the temperature of the liquid sample and of the source/detector module will be measured simultaneously with the index, to correct for their temperature dependence. The liquid sample temperature

will be determined using a thermistor, as indicated in Figure 2, or using a fiber optic temperature sensor in applications requiring an all dielectric transducer, e.g., for use in explosive or high voltage environments [10,11].

III. WEB-CONTROLLED MONITORING SYSTEM ARCHITECTURE

This equipment was developed with the appropriate Measured Liquid Magazine, Sensor Module Positioner, Control Servomotors, Heating Element, Sensor Cleaning Element, Visual Camera Feedback, Control and Communication Software (Figure 3) to create an interactive web controlled Fibre Optic Refractometer instrument. Sensor Module Positioner controls using servomotor the on (measured) and off (non-measured) position of the sensor head. Measured Liquid Magazine is based on the revolver system controlled by servomotor and provides the change of measured liquid (5 different liquids are possible one liquid is used for Sensor Cleaning Element if necessary – oil, petrochemical product measurements, etc.). Heating Element provide controlled heating of the measured liquid. Visual Camera Feedback was added to the system to provide students with the feeling to be virtually present at the measurement place and also as visual feedback for verify correct function of the mechanical parts of the instrument [11,12].

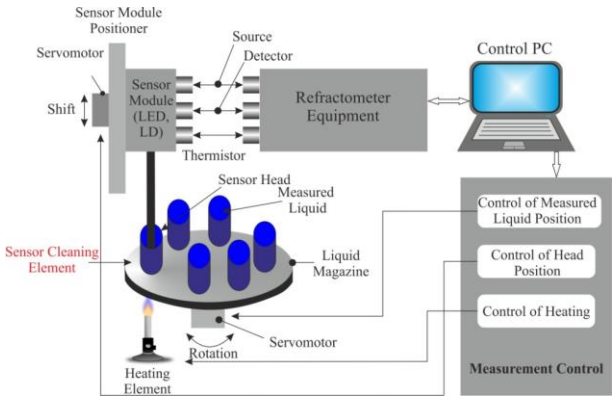


Figure 3. Web-controlled remote monitoring system design.

The complete system was set together with the two (Laboratory Refractometer Client Server and Refractometer Server) servers based on LabWindows. Operation of needs approval and authorized permission to be granted in advance through the local internet service provider. Therefore, and in step to avoid such complications the system was operated and tested using the internet network available on University campus.

IV. EXPERIMENTS AND RESULTS

Performance of the developed fibre optic refractometers was evaluated by various test measurements; this is here demonstrated by results of several classical experiments.

The first experiment was to measure the refractive index of various liquids. As field tests we used some calibration liquids of known index of refraction. They make it difficult to clean the sensor face however.

We recommend keeping the containers of the mixtures sealed when not in use since they tend to drift in index, though this is not much of a problem. We frequently make up relatively large sample, e.g., 100 to 200 ml, and then use one half for our measurements for while and after a week or two, take a reading in the other half that has been kept well sealed. That way we can determine if there has been any shift in the index of our measurement sample. The results of measured index of refraction are depicted on TABLE I (Food products) and TABLE II (Petrochemical products), and are in very good success as compared with results obtained by classical methods (e.g. the refractive index of water at 20°C is 1.3330 or propylene glycol at 20°C has a refractive index of 1.4312).

TABLE I. RESULT OF FOOD PRODUCTS MEASUREMENTS

Serial Number	**Refractive index of various liquids**		
	Liquid products	***Refractive index***	***Temperature (℃)***
1.	Water	1.3333	21
2.	Distilled water	1.3331	23
3.	Vinegar	1,3358	23
4.	Pepsi cola	1,3461	23
5.	Coca cola	1,3417	23
6.	Synthetic alcohol	1.3620	21
7.	99% alcohol	1,3611	23
8.	LACALUT Mouthwash	1,3572	23
9.	Rapeseed oil	1,4697	23
10.	Sunflower oil	1,4682	23
11.	Olive oil	1,4662	23

TABLE II. RESULT OF PETROCHEMICAL PRODUCTS MEASUREMENTS

Serial Number	**Refractive index of various liquids**		
	Liquid products	***Refractive index***	***Temperature (℃)***
1.	Propylen glycol	1.4268	21
2.	Engine oil MOL Arol 2T	1,4791	23
3.	Gear grease Madit PP80	1,4806	23
4.	Oil Mobil VS-200	1.4399	21
5.	Oil Mobil motor 5W-50	1.4678	21
6.	Brake fluid Castrol	1,4424	23
7.	Gasoline Natural 95 Slovnaft	1,4565	23
8.	Gasoline Natural 95 OMV	1,4572	23
9.	Diesel Slovnaft	1,4591	23
10.	Diesel OMV	1,4609	23

The second experiment was to measure the change in refractive index depending on the temperature of the liquid and third experiment was to measure the refractive index changes depending on the concentration of the liquid. The test measurement is the data on the index

refraction of propylene glycol/water mixture are good to use since it is simple to clean the sensor after dipping in the various solutions, since it is necessary that the sensor face be clean and dry to get a good reading in air and also before dipping in water to get a calibration/comparison reading. Even though the index varies with wavelength we take the water reading at 20°C to correspond to an index of 1.3330, and refer all other readings to this value. The dn/dt for water in the range 15 to 30°C is 0.0001 per degree °C, while that for 100% propylene glycol is 0.0003. Thus for glycol/water solutions one could assume a linear dependence of dn/dt, that is, for example assume dn/dt=0.0002 for a 50% solution. In all instruments with automatic temperature compensation, it is assumed that they are to be used with water solutions, and it is assumed that dn/dt=0.0001. We may use also propylene glycol, either 100%, taking its index to be 1.4312, or 100% plus a few mixtures. Additional results are measured for the second experimenter (change refractive index depending on the temperature):

1) *Engine oil MOL Arol 2T:*
   a) *Range: 20 - 75ºC*
   b) *dn/dt: - 0,00035/ºC*

2) *Rapeseed oil:*
   a) *Range: 20 - 75ºC*
   b) *dn/dt: - 0,00035/ºC*

3) *Olive oil:*
   a) *Range: 23 - 75ºC*
   b) *dn/dt: - 0,00004/ºC*

Additional results are measured for the third experimenter (changes refractive index depending on the concentration):

1) *Alcohol:*
   a) *Range: 10 – 99%*
   b) *dn/dc: 0,00032/%*

2) *Sugar solution (sugar in water):*
   a) *Range: 9 – 50%*
   b) *dn/dc: - 0,0018/%*

3) *Saline solution (salt in water):*
   a) *Range: 0 – 36%*
   b) *dn/dt: - 0,0013/%*

CONLUSIONS

Development work related to create remote control system for monitoring water pollution petrochemical products using optical fibre refractometer in the Applied Photonics distance education courseware was presented. The system modules structure enables its accommodation to solve remote access trough WWW to various photonics equipments for a relative large pool of students. The system was experimentally tested with module Fibre Optics Refractometer Instrument. Future work includes adding more modules to the system and designing more web-based equipments to enhance teaching and learning of subject Applied Photonics.

ACKNOWLEDGMENT

This work was supported by Cultural and Educational Grant Agency (KEGA) of the Ministry of Education, Science, Research and Sport of the Slovak Republic under the project no. „063TUKE-4/2013 - The Use of Remote Controlled Optical Fibre Refractometer in Teaching“ (50%), the Slovak Research and Development Agency under the contract no. “APVV-0025-12 - Mitigation of Stochastic Effect in High-Bitrate All Optical Networks” (50%).

REFERENCES

[1] N. N. AzizulHaq and G. A. A. A. Alawi, "Various Learning Courses Based on Digital Library Conceptualization," International Conference on Technology for Education 2010: T4E 2010, Mumbai, India, July 1-3, 2010, pp. 222-223.

[2] G. Govindan, S. G. Raj and D. Sastikumar, "Measurement of Refractive Index of Liquids Using Fiber Optic Displacement Sensors," Journal of American Science, vol. 5, no. 2, 2009, pp. 13-17.

[3] A. L. Gurashkin and et al., "Communication: High Speed Optical Investigations of a Character of Boiling-up Onset," Journal of Chemical Physics, vol. 136, no. 2, 2012, pp. 1-4.

[4] A. Leva and F. Donida, "Multifunctional Remote Laboratory for Education in Automatic Control: The CrAutoLab Experience," EEE Transactions on Industrial Electronics, vol. 55, no. 6, June 2008, pp. 2376-2385.

[5] G. Mudhana, K. S. Park, S. Y. Ryu and B. H. Lee, "Fiber-Optic Probe Based on a Bifunctional Lensed Photonic Crystal Fiber for Refractive Index Measurements of Liquids," IEEE Sensors Journal, vol. 11, no. 5, 2011, pp. 1178-1183.

[6] J. B. Rosolem, C. Floridia, A. A. Juriollo and E. W. Bezerra, "Comparative Analysis of Fiber Grating Versus Fiber Powering for Fiber Optical Sensing Applications," SBMO/IEEE MTT-S International Microwave and Optoelectronics Conference: IMOC 2009, Belem, Brazil, November 3-6, 2009, pp. 641-645.

[7] J. L. Santos, O. Frazãoa, J. M. Baptista,J P. A. S. Jorge, I. Dias, F. M. Araújo and L. A. Ferreira, "Optical Fibre Sensing Networks," SBMO/IEEE MTT-S International Microwave and Optoelectronics Conference: IMOC 2009, Belem, Brazil, November 3-6, 2009, pp. 290-298.

[8] M. Sheeba, M. Rajesh, C. P. G. Vllabhan, V. P. N. Nampoori and P. Radhakrishnan, "Fibre Optic Sensor for the Detection of Adulterant Traces in Coconut Oil," Measurement Science and Technology, vol. 16, no. 11, November 2005, pp. 2247-2250.

[9] W. Steingartner, D. Galinec, "The Rôle of Categorical Structures in Infinitesimal Calculus," Journal of Applied Mathematics and Computational Mechanics, vol. 12, no. 1, 2013, pp. 107-119.

[10] R. T. Takeishi and et al., "Flexfuel Vehicle Alcohol-Gasoline Blend Ratios Determination by Optical Fiber Sensoring," 19th International Conference on Optical Fibre Sensors, Proc. of SPIE vol. 7004, 2008, pp. 70044F1-70044F4.

[11] J. Turán, E. F. Carome and Ľ. Ovseník, “Fiber Optics Refractometer for Liquid Index of refraction Measurement,” TELSIKS-2001 International Conference on Telecommunication in Modern Satellite, Cable and Broadcasting Service, Niš, Serbia and Montenegro, May 19-23, 2001, vol. 2, pp. 489-492.

[12] S. Zadvomov and A. Sokolovsky, "An Electro-Optic Hybrid Methane Sensor," IEEE Canadian Conference on Electrical and Computer Engineering: CCECE 2008, Niagara Falls, ON, Canada, May 4-7, 2008, pp. 969-974.

Broadband Characterization of SMA Connectors by Measurements

Marko Magerl, Tvrtko Mandic, Adrijan Baric
University of Zagreb Faculty of Electrical Engineering and Computing
Zagreb, Croatia
Email: marko.magerl@fer.hr

***Abstract*—This paper presents the characterization of various types of SubMiniature version A (SMA) connectors. The characterization is performed by measurements in frequency and time domain. The SMA connectors are mounted on microstrip (MS) and conductor-backed coplanar waveguide (CPW-CB) manufactured on high-frequency (HF) laminates. The designed characteristic impedance of the transmission lines is 50 Ω and deviation from the designed characteristic impedance is measured. The measurement results suggest that for a given combination of the transmission line and SMA connector, the discontinuity in terms of characteristic impedance can be significantly improved by choosing the right connector type.**

***Index Terms*—SMA connector; microstrip; conductor-backed coplanar waveguide; measurement; characterization**

I. Introduction

The omnipresent interconnect used in RF measurements is the SubMiniature version A (SMA) connector. The SMA connectors are used to connect measurement equipment and they are usually de-embedded as they are considered an intrinsic part of the transmission line. Nevertheless, the de-embedding is feasible only in the useful frequency range of the SMA connector-to-transmission line transition where insertion and return losses are not prohibitively large.

The S-parameters and time domain reflectometry (TDR) measurements are commonly used to verify SMA connector-to-transmission line transition performance [1], [2]. This is often performed for single connector type only [3], [4] and in limited frequency range [5], [6]. In this work various SMA connector types are mounted on microstrip (MS) and conductor-backed coplanar waveguide (CPW-CB). The MS and CPW-CB transmission lines are manufactured on low-loss high-frequency (HF) laminates. The measurements are performed up to 18 GHz for the following SMA connector types: vertical through hole (TH), vertical surface mount (SM), end-launch (EL) and reusable end-launch (SW). Furthermore, TDR measurements are performed to extract the characteristic impedance profile of the connectors with the best performance.

This paper is structured as follows: Section II presents the measurement set-up, MS and CB-CPW transmission line parameters and geometrical parameters of SMA connector footprints. Section III compares the measurements performed on various types of SMA connectors mounted on MS and CPW-CB transmission lines. The conclusion is given in Section IV.

II. Measurement Set-up

A number of structures for characterization of SMA connectors are manufactured on two-layer HF laminates. The structures consist of two SMA connectors mounted on transmission lines of different type and length. Several examples are shown in Fig. 1.

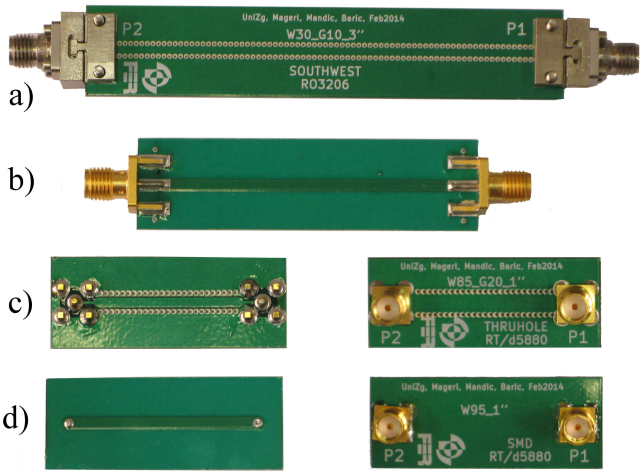


Fig. 1. Examples of the manufactured structures for characterizing SMA connectors: a) SW connectors mounted on a 3 inch CPW-CB line on RO3206 laminate (bottom layer view), b) EL connectors mounted on a 2 inch MS line on RO3206 laminate (top layer view), c) TH connectors mounted on a 1 inch CPW-CB line on RT/duroid 5880 laminate (top layer view on the left and bottom layer view on the right), d) SM connectors mounted on a 1 inch MS line on RT/duroid 5880 laminate (top layer view on the left and bottom layer view on the right).

A. Design of Transmission Lines

Two Rogers HF laminates are used: RT/duroid 5880 ($\varepsilon_r = 2.2$) and RO3206 ($\varepsilon_r = 6.15$). Parameters of the HF laminates most important for the design of transmission lines are summarized in Table I. The HF laminates have two metal layers.

Table II presents the dimensions of MS and CPW-CB transmission lines designed to have a 50 Ω characteristic impedance. The values are obtained using ADS LineCalc [7]. For a given HF laminate, the MS line characteristic impedance is defined only by the trace width, therefore the choice of

TABLE I
ROGERS HIGH FREQUENCY LAMINATE CHARACTERISTICS

	RT/duroid 5880	**RO3206**
substrate thickness	787.4 μm (31 mil)	1.27 mm (50 mil)
Cu thickness	17 μm	35 μm
ε_r	2.2	6.15
$tg\ \delta$	0.0004	0.0027

TABLE II
MICROSTRIP AND CONDUCTOR-BACKED COPLANAR WAVEGUIDE 50 Ω TRANSMISSION LINE DIMENSIONS

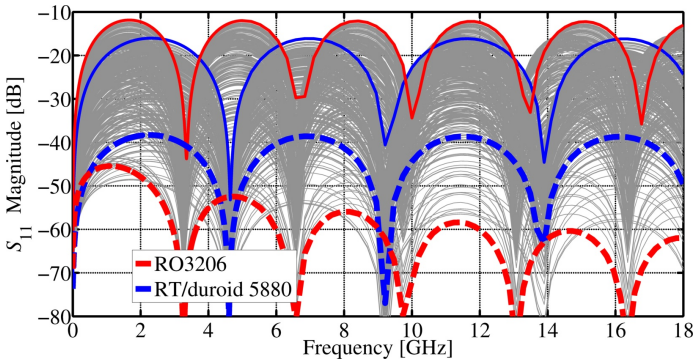
	RT/duroid 5880	**RO3206**
MS (w)	2.413 mm (95 mil)	2.032 mm (80 mil)
CPW-CB (w/g)	1.778 mm / 254 μm (70 mil / 10 mil)	762 μm / 254 μm (30 mil / 10 mil)
CPW-CB (w/g)	2.159 mm / 508 μm (85 mil / 20 mil)	1.27 mm / 508 μm (50 mil / 20 mil)

the MS line trace width is unique for a given characteristic impedance. However, the CPW-CB line characteristic impedance is defined by the trace width and the gap size between the trace and the ground plane. This allows design freedom to make CPW-CB lines more robust to the PCB manufacturing process tolerances. Monte Carlo simulations are performed in Agilent ADS [7] in order to analyze the influence of PCB manufacturing process tolerances on the characteristic impedance of CPW-CB transmission lines. A ±63.5 μm (±2.5 mil) tolerance is associated with the CPW-CB line trace width and gap size. Fig. 2 presents Monte Carlo simulation results for the return loss of 50 Ω CPW-CB lines with 3 different gap sizes. According to the simulation results, a gap size increase from 127 μm (5 mil) to 254 μm (10 mil) increases the worst-case return loss by ≈10 dB, while the gap size increase from 254 μm (10 mil) to 508 μm (20 mil) increases the worst-case return loss by only ≈5 dB. This indicates that CPW-CB lines with larger gap sizes have a reduced sensitivity to the variation of geometrical parameters. Based on the simulation results, the CPW-CB lines used in measurements are designed with gap sizes of 254 μm (10 mil) and 508 μm (20 mil), as indicated in Table II.

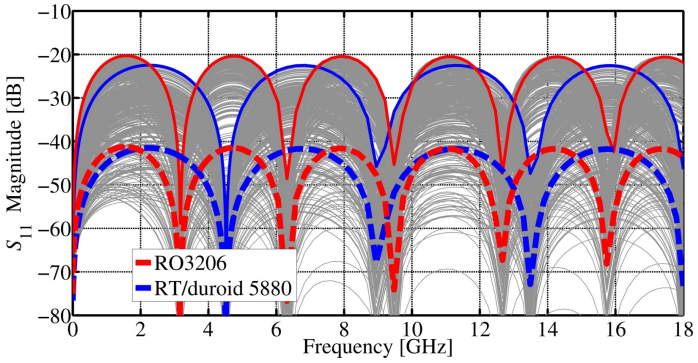
The CPW-CB ground planes placed on top and bottom layers are connected by vias positioned along the line length as shown in Fig. 1 c). The via drill diameter is equal to 0.45 mm (17.7 mil) and the spacing between vias is 1.016 mm (40 mil), which is less than $\lambda/8$ in the frequency range up to 18 GHz for both HF laminates. The geometrical parameters of vias are identical for all characterized structures.

B. SMA Connectors

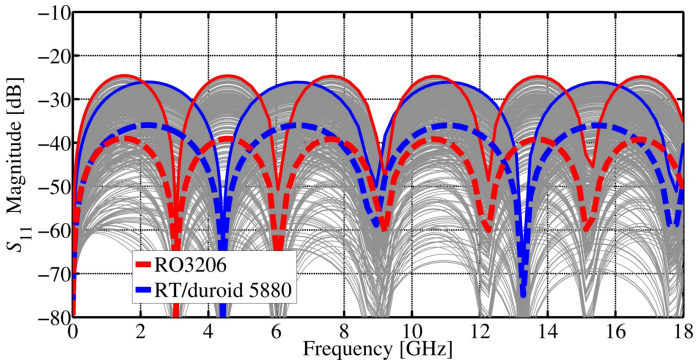
The SMA connector footprints are designed according to the manufacturer recommendations. The footprint drawings and design considerations are described in this section.



(a) 50 Ω CPW-CB: g = 127 μm (5 mil).
RO3206: w = 381 μm (15 mil), RT/duroid 5880: w = 1.27 mm (50 mil).



(b) 50 Ω CPW-CB: g = 254 μm (10 mil).
RO3206: w = 762 μm (30 mil), RT/duroid 5880: w = 1.778 mm (70 mil).



(c) 50 Ω CPW-CB: g = 508 μm (20 mil).
RO3206: w = 1.27 mm (50 mil), RT/duroid 5880: w = 2.159 mm (85 mil).

Fig. 2. Monte Carlo simulation results for the return loss of 50 Ω CPW-CB lines with 3 different gap sizes: a) 127 μm (5 mil), b) 254 μm (10 mil), c) 508 μm (20 mil). Nominal cases are represented by dashed lines and worst cases by solid lines. Statistical model: $g = g_{nom} + \Delta g$, $w = w_{nom} - \Delta g$, Δg = uniform(-2.5 mil, 2.5 mil), 250 trials.

1) Vertical connectors: Two vertical SMA connectors are used: through hole (TH) [8] and surface mount (SM) [9]. The connectors are mounted on the bottom layer of the PCB. Fig. 3 presents the top layer layout of the TH and SM vertical connector footprints. The reference planes and ground plane edges on the bottom layer are represented by dashed lines. A short fixture line extending from the signal pin of the connector to the reference plane is part of the TH and SM connector footprints. In the TH connector case, the fixture line has a width

of 508 µm (20 mil). The presented TH connector footprint has the same layout for all characterized structures, therefore the fixture line has a different characteristic impedance for the RO3206 laminate compared to the RT/duroid 5880 laminate.

The width of the SM connector fixture line is equal to the trace width of the transmission line extending from the footprint. The diameter of the antipad (dashed line in Fig. 3b) is recommended by the SM connector datasheet.

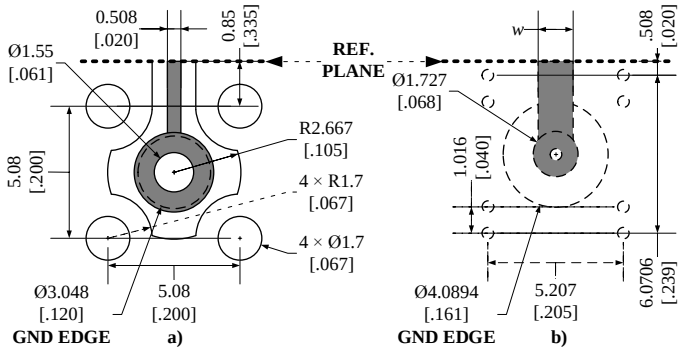


Fig. 3. Top layer layout of vertical connector footprints: a) TH connector and b) SM connector. Reference planes and ground plane edges on the bottom layer are represented by dashed lines.

2) End-Launch connectors: Two end-launch connectors are used: end-launch (EL) [10] and reusable end-launch (SW) [11]. Fig. 4 presents the top layer layout of the EL and SW end-launch connector footprints. The reference planes are represented by dashed lines, and the ground plane on the bottom layer is uniform. Both EL and SW connector footprints are inherently CPW-CB lines because of the ground pads next to and below the signal pad, as it can be seen in Fig. 4. In order to minimize signal reflections, the EL and SW connector footprints are designed as 50 Ω CPW-CB lines. The EL connector signal pad has the width w equal to the trace width of the transmission line extending from the footprint. The gap g between the signal pad and the ground pads of the EL connector footprint is calculated to result in a 50 Ω CPW-CB line. Since the EL connector is designed for laminates with a thickness of 1.575 mm (62 mil), it is mounted only on the RO3206 laminate.

The structures for the characterization of vertical and end-launch SMA connectors are designed as follows. A reference plane is defined for each SMA connector, as shown in Figs. 3 and 4, and 50 Ω transmission lines of different lengths are placed between the reference planes of two identical SMA connectors, as shown in Fig. 1. The transmission line lengths used on the RT/duroid 5880 laminate are: 25.4 mm (1 inch) and 50.8 mm (2 inch). Lengths used on the RO3206 laminate are: 25.4 mm (1 inch), 50.8 mm (2 inch) and 76.2 mm (3 inch). In order to have a ground plane that extends at least 3 trace widths on each side of the transmission line on all characterized structures, the width of all PCBs is equal to 7 times the maximum designed trace width (7 × 2.413 mm = 16.9 mm). The characterization of the proposed structures is performed by measuring two-port S-parameters using a vector network analyzer (VNA) in the frequency range from 10 MHz to 18 GHz, and by TDR measurements.

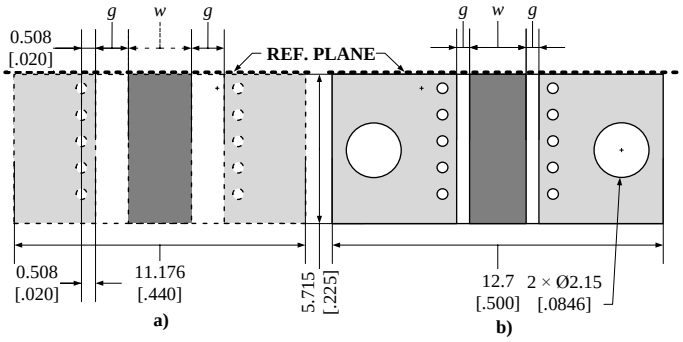
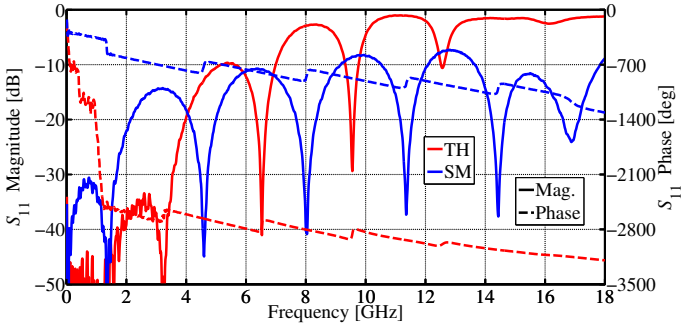
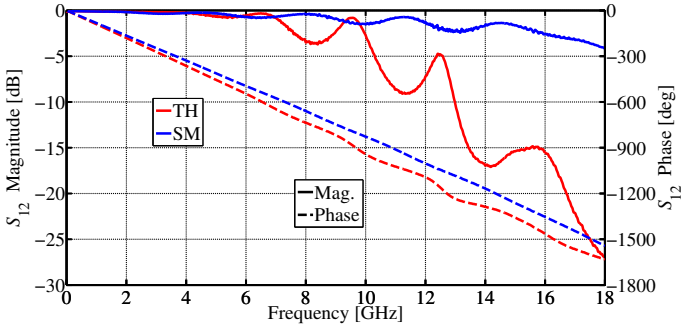


Fig. 4. Top layer layout of end-launch connector footprints: a) EL connector and b) SW connector. Reference planes are represented by dashed lines. The ground plane on the bottom layer is uniform.

III. Measurement Results and Discussion



(a) S11

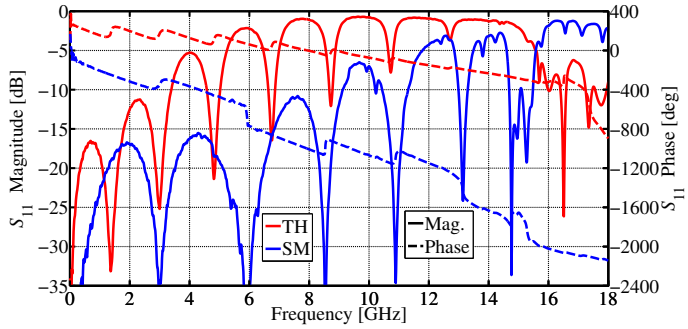


(b) S12

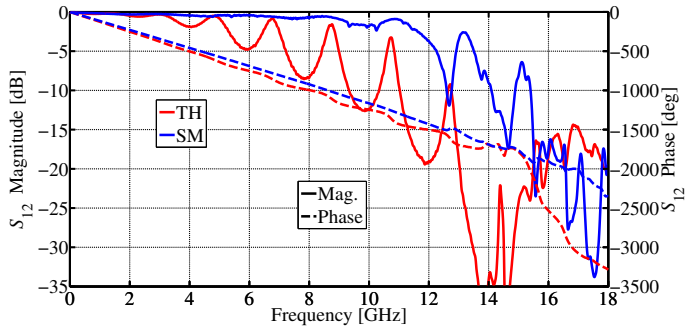
Fig. 5. S-parameter measurement results of TH and SM connectors mounted on MS lines (l = 25.4 mm (1 inch), w = 2.413 mm (95 mil)) on the RT/duroid 5880 laminate: a) return loss (S_{11}), b) insertion loss (S_{12}). Magnitude is represented by solid lines and phase by dashed lines.

Fig. 5 presents S-parameter measurement results of the TH and SM connectors mounted on MS lines (l = 25.4 mm (1 inch), w = 2.413 mm (95 mil)) on the RT/duroid 5880 laminate. The return loss (S_{11}) is shown in Fig. 5a and the insertion loss (S_{12}) is shown in Fig. 5b.

The TH connector has excellent matching with 35 dB return loss in the range up to 3 GHz, and 10 dB return loss in the

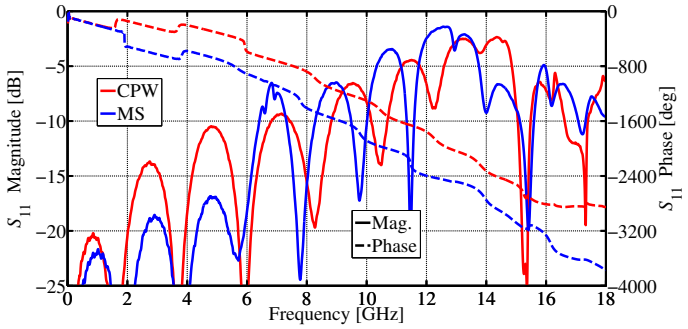


(a) S11

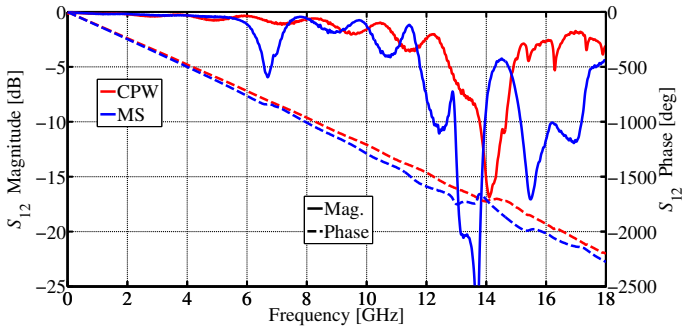


(b) S12

Fig. 6. S-parameter measurement results of TH and SM connectors mounted on CPW-CB lines (l = 25.4 mm (1 inch), w = 762 µm (30 mil), g = 254 µm (10 mil)) on the RO3206 laminate: a) return loss (S_{11}), b) insertion loss (S_{12}). Magnitude is represented by solid lines and phase by dashed lines.



(a) S11



(b) S12

Fig. 7. S-parameter measurement results of EL connectors mounted on CPW-CB line (l = 25.4 mm (1 inch), w = 1.27 mm (50 mil), g = 508 µm (20 mil) and MS line (l = 25.4 mm (1 inch), w = 2.032 mm (80 mil)) on the RO3206 laminate: a) return loss (S_{11}), b) insertion loss (S_{12}). Magnitude is represented by solid lines and phase by dashed lines.

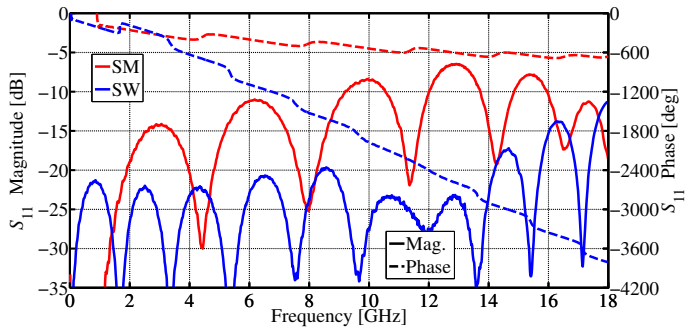
range up to 7 GHz. However, above 7.5 GHz the TH connector is poorly matched, partially because of the non-50 Ω fixture line in the TH connector footprint, shown in Fig. 3. The TH connector insertion loss 3 dB point is reached at 8 GHz. The SM connector has 10 dB return loss up to 9 GHz and 7.5 dB return loss in the entire measured frequency range. The SM connector insertion loss 3 dB point is reached at 17 GHz and is within 4 dB in the entire measured frequency range.

The TH connector shows better overall performance compared to the SM connector on the RT/duroid 5880 laminate in the frequency range up to 4 GHz. The S_{12} phase response is more linear for the SM connector than the TH connector in the frequency range above 7 GHz. It also shows that the SM connector is electrically shorter than the TH connector, which is consistent with the connector physical dimensions. Measurement results of the TH and SM connectors mounted on CPW-CB lines (w = 2.159 mm (85 mil), g = 508 µm (20 mil) and w = 1.778 mm (70 mil), g = 254 µm (10 mil)) on the RT/duroid 5880 laminate show performance similar to the results presented in Fig. 5.

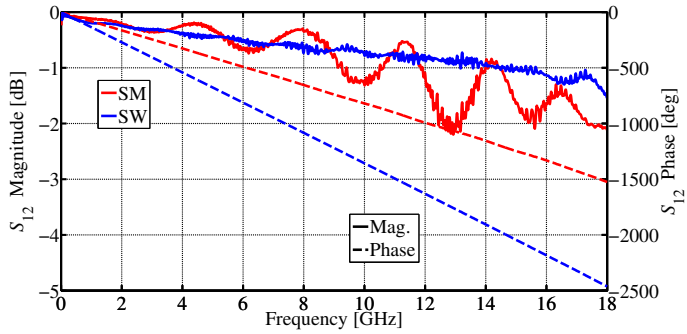
The TH and SM connectors are compared further for the RO3206 laminate. Fig. 6 presents S-parameter measurement results of the TH and SM connectors mounted on CPW-CB lines (l = 25.4 mm (1 inch), w = 762 µm (30 mil), g = 254 µm (10 mil)) on the RO3206 laminate. The return loss (S_{11}) is shown in Fig. 6a and the insertion loss (S_{12}) is shown in Fig. 6b. The TH connector has 10 dB return loss up to 3.3 GHz and the insertion loss 3 dB point at 5 GHz. The SM connector has 10 dB return loss up to 9 GHz and the insertion loss 3 dB point at 12 GHz.

The SM connector shows better overall performance than the TH connector on the RO3206 laminate in the frequency range up to 15.5 GHz. For frequencies above 15 GHz both connectors have more than 15 dB insertion loss. The S_{12} phase response for the SM connector has better linearity than the for TH connector. Measurement results of the TH and SM connectors mounted on MS lines (w = 2.032 mm (80 mil)) and CPW-CB lines (w = 1.27 mm (50 mil), g = 508 µm (20 mil)) on the RO3206 laminate show similar performance to the results presented in Fig. 6.

The overall performance of the TH and SM connectors on the RO3206 laminate is worse than on the RT/duroid 5880 laminate. The causes may be as follows. The RO3206 laminate is 482.6 µm (19 mil) thicker than the RT/duroid 5880 laminate and it has a ≈2.8 times higher dielectric constant. This results in smaller trace widths and gap sizes of 50 Ω transmission lines, which makes them more sensitive to PCB manufacturing process tolerances, as shown in Section II. Also, the RO3206 laminate has ≈7 times higher losses than the RT/duroid 5880 laminate, which results in higher insertion loss for the same line length. The TH and SM connectors show similar performance for both transmission line types regardless of the HF laminate used. However, this is not the case for the EL



(a) S11



(b) S12

Fig. 8. S-parameter measurement results of SW connectors mounted on CPW-CB line (l = 25.4 mm (1 inch), w = 762 μm (30 mil), g = 254 μm (10 mil) on the RO3206 laminate compared to the SM connector mounted on CPW-CB line (l = 25.4 mm (1 inch), w = 1.27 mm (50 mil), g = 508 μm (20 mil) on the RT/duroid 5880 laminate: a) return loss (S_{11}), b) insertion loss (S_{12}). Magnitude is represented by solid lines and phase by dashed lines.

connectors, as shown next.

Fig. 7 presents S-parameter measurement results of the EL connectors mounted on CPW-CB line (l = 25.4 mm (1 inch), w = 1.27 mm (50 mil), g = 508 μm (20 mil) and MS line (l = 25.4 mm (1 inch), w = 2.032 mm (80 mil)) on the RO3206 laminate. The return loss (S_{11}) is shown in Fig. 7a and the insertion loss (S_{12}) is shown in Fig. 7b. The MS line test case shows better overall performance compared to the CPW-CB line test case in the range up to 6 GHz. Although the EL connector is rated up to 18 GHz, both cases show more than 5 dB insertion loss in the range above 12 GHz. Further, a 6 dB dip in the insertion loss is observed for the MS line test case in the range between 6 GHz and 8 GHz, regardless of the MS line length used. This dip is therefore attributed to the EL connector performance.

The MS line test case has 10 dB return loss up to 7 GHz and the insertion loss 3 dB point at 6.4 GHz. The CPW-CB line test case has 10 dB return loss up to 6.8 GHz and the insertion loss 3 dB point at 12.5 GHz. A 17 dB dip in the insertion loss is observed for the CPW-CB test case in the range between 12 GHz to 15 GHz, which is probably due to the connector-to-transmission line transition. This transition may be improved by optimization using a 3D EM simulator. The S_{12} phase response shows better linearity for the CPW-

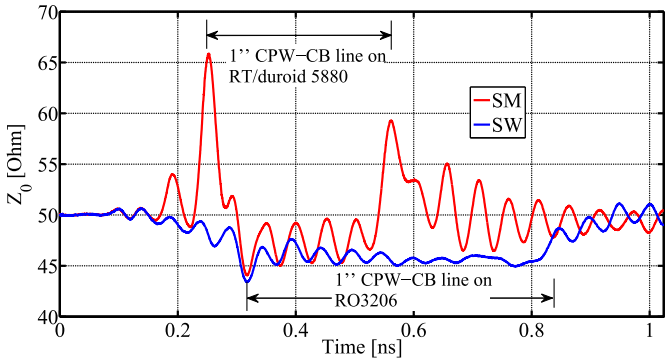


Fig. 9. Characteristic impedance profile of SW connectors mounted on CPW-CB line (l = 25.4 mm (1 inch), w = 762 μm (30 mil), g = 254 μm (10 mil)) and the SM connector mounted on CPW-CB line (l = 25.4 mm (1 inch), w = 1.27 mm (50 mil), g = 508 μm (20 mil)) on the RT/duroid laminate, obtained by TDR measurements. Other port of the transmission lines is terminated in 50 Ω.

CB line test case than for the MS line test case. Also, it shows that both structures have similar electrical length. This is an expected result because the same connector, transmission line length and HF laminate are used in both cases. Measurement results of the EL connectors mounted on CPW-CB line (w = 762 μm (30 mil), g = 254 μm (10 mil)) on the RO3206 laminate show similar performance to the results of the CPW-CB line test case presented in Fig. 6.

Reusable end-launch (SW) connectors show the best performance of all measured connectors. Fig. 8 compares S-parameter measurement results of the SW connectors mounted on CPW-CB line (l = 25.4 mm (1 inch), w = 762 μm (30 mil), g = 254 μm (10 mil) on the RO3206 laminate with with measurement results of the SM connector mounted on CPW-CB line (l = 25.4 mm (1 inch), w = 1.27 mm (50 mil), g = 508 μm (20 mil) on the RT/duroid laminate. The latter has the second best performance of all characterized structures. The return loss (S_{11}) is shown in Fig. 8a and the insertion loss (S_{12}) is shown in Fig. 8b.

The return loss of the SW connector is superior to all measured connectors, with 20 dB return loss up to 14 GHz and 10 dB return loss in the entire measured frequency range. The SW connector has far less ripple in the S_{12} magnitude response than the SM connector. Both connectors have insertion loss within 3 dB in the entire measured frequency range. Great linearity is observed in the S_{12} phase response for both connectors. The SW connector performance may be even further improved by designing a taper in the SW connector footprint, as described in [12].

Since they show the best performance, the SW and SM connectors are further compared in the time domain. TDR measurements are made using a 20 GHz sampling oscilloscope with TEK SD-24 sampling heads capable of launching signals with voltage swing of 250 mV and rise time of 30 ps. Fig. 9 presents the characteristic impedance profile extracted from the TDR measurements for the SW and SM connectors, which represents the time domain equivalent of the frequency domain measurements presented in Fig. 8. The transmission

line characteristic impedance can be observed from the 0.3 ns mark, and it equals ≈47.5 Ω for both test cases. The the transmission line in the SM connector test case is electrically shorter than in the SW connector test case due to a lower dielectric constant of the HF laminate used. The SM connector introduces a 15 Ω discontinuity in the characteristic impedance profile, while the SW connector does not significantly alter the characteristic impedance profile.

IV. Conclusion

The measurement results of various SMA connector types are presented, namely the vertical through hole (TH), the vertical surface mount (SM), the end-launch (EL) and the reusable end-launch (SW). The SMA connectors are mounted on 50 Ω MS and CPW-CB lines manufactured on two different HF laminates. The TH connector shows good return loss and insertion loss performance only in the frequency range up to several GHz. The SM connector mounted on CPW-CB line on the RT/duroid 5880 laminate shows the second best performance of all measured SMA connectors. Even though the EL connector is rated up to frequency of 18 GHz, due to the high insertion loss it has limited usability up to ≈13 GHz for the CPW-CB and up to 6 GHz for the MS line. The SW connector show the best performance of all measured SMA connectors, with excellent matching, low insertion loss and great S_{12} phase response linearity in the entire measured frequency range.

References

[1] J. Palecek, M. Vestenicky, P. Vestenicky, and J. Spalek, "Examination of SMA connector parameters," in *Intelligent Engineering Systems (INES), 2012 IEEE 16th International Conference on*, June 2012, pp. 259–263.

[2] S. McMorrow and A. Neves, "A hybrid measurement and electromagnetic field solver approach for the design of high-performance interconnects: An investigation of traces and SMA transitions," *DesignCon 2004 Conference Proceedings*, pp. 2–22, Jan. 2004.

[3] S. Wartenberg and Q. H. Liu, "A coaxial-to-microstrip transition for multilayer substrates," *Microwave Theory and Techniques, IEEE Transactions on*, vol. 52, no. 2, pp. 584–588, Feb. 2004.

[4] G. Antonini, A. C. Scogna, and O. A., "Equivalent circuit extraction for an sma connector," *Progress in Electromagnetic Research Symposium 2004*, pp. 61–64, March 2004.

[5] T. Mandic, R. Gillon, B. Nauwelaers, and A. Baric, "Characterizing the TEM cell electric and magnetic field coupling to PCB transmission lines," *Electromagnetic Compatibility, IEEE Transactions on*, vol. 54, no. 5, pp. 976–985, Oct. 2012.

[6] F. Vanhee, J. Catrysse, R. Gillon, and G. Gielen, "Scaleable equivalent circuit modelling of the E-field coupling to microstrips in the TEM cell," *Electromagnetic Compatibility - EMC Europe, 2008 International Symposium on*, pp. 1–6, Sep. 2008.

[7] *Advanced Design System 2013 Documentation Set*, Agilent Technologies, Inc., Santa Clara, US, 2013.

[8] Amphenol RF. (2014) 901-144-8RFX. [Online]. Available: http://www.amphenolrf.com/record.asp?N=0&sid=462D4880E17F&R=901-144-8RFX

[9] Lighthorse Technologies, Inc. (2014) LTI-SASF66GT. [Online]. Available: http://www.rfconnector.com/sma-surface-mount--jack-pcb-connector.html

[10] Emerson Connectivity, Inc. (2014) 142-0701-801. [Online]. Available: http://emersonconnectivity.com/OA_HTML/ibeCCtpItmDspRte.jsp?item=4253257

[11] Southwest Microwave, Inc. (2014) 292-04A-5. [Online]. Available: http://mpd.southwestmicrowave.com/products/product.php?need=endLaunch&item=2

[12] ——. (2011) The design & test of broadband launches up to 50 GHz on thin & thick substrates. [Online]. Available: http://mpd.southwestmicrowave.com/showImage.php?image=287

Application of TEM Cell in Measurement of Electromagnetic Fields of Electronic Modules

Krešimir Ivanuš* and Adrijan Barić**
* KONČAR–Electrical Engineering Institute, Zagreb, Croatia
** University of Zagreb, Faculty of Electrical Engineering and Computing, Zagreb, Croatia
kivanus@koncar-institut.hr, adrijan.baric@fer.hr

Abstract – The paper presents measured radiated emissions of an electronic module in the frequency range 150 kHz – 200 MHz using the TEM cell method. The TEM cell measurement method is explained and measurement results for different orientations of the electronic module are presented. An explanation for the differences in the measured radiated emissions for different module orientations is given.

I. INTRODUCTION

Measurement of radiated emissions of electronic device is essential part of EMC testing. This testing process should ensure proper operation of the measured equipment in complex environment in which electronic equipment must operate. In the environment, devices are exposed to the influence of other electromagnetic fields, on which they must be immune. On the other hand, to reduce the influence of a particular device on its environment and also to reduce the overall amount of electromagnetic pollution, the requirements are set to limit the emissions of electromagnetic fields.

The required immunity levels and allowed emission levels are defined for different types of equipment by appropriate standards. These standards prescribe the methods for immunity tests, as well as the methods for measurement of different types of emissions which electronic devices generate.

This work presents results of radiated disturbance measurement for analyzed electronic module (power supply module of a more complex modular system) in the TEM cell.

The TEM cell method [1]–[4] provides simple, affordable and reliable tool for the measurement of the radiated emissions of the electronic devices.

In the following chapters, firstly the analyzed electronic module is presented, and then the TEM cell and measuring method is explained. In the remainder of this paper the measurement results are presented and discussed, and then conclusions are given in Section V.

II. ANALYZED CIRCUIT

In this work a power supply module of a complex modular electronic system intended for industrial applications is analyzed. This system should be capable of working in highly complex environments, exposed to high levels of disturbances from the environment.

In order to reliably perform its operation, the system needs a quality power supply. The power supply module analyzed in this paper powers the system with quality and stable 5 V supply voltage, obtained from the 24 V input battery voltage. Battery voltage is usually very "dirty", exposed to severe disturbances, voltage drops and over-voltages. Therefore, the analyzed power supply module, besides the output voltage regulation, has to perform protection and electrical isolation. On the one hand, it must be resistant to interferences and protect powered devices from disturbances, but on the other hand, it must not pollute the environment with its emissions.

In this paper the emission of the radiated disturbances of the power supply module is analyzed. The power supply module is shown in Figure 1 and the block-diagram of its central component is presented in Figure 3.

Schematic of the power supply module is shown in Figure 2. The central component of the analyzed power supply circuit is the DC/DC converter (component A1 in Figure 2). It ensures galvanic separation of the input and output voltage and regulation of the output voltage. Additionally, it implements protection functions, like the protection from short-circuit on the output or protection from overheating.

As it can be seen on the block-diagram (Figure 3), the function of galvanic separation is achieved by means of inductive coupling between primary and secondary circuit

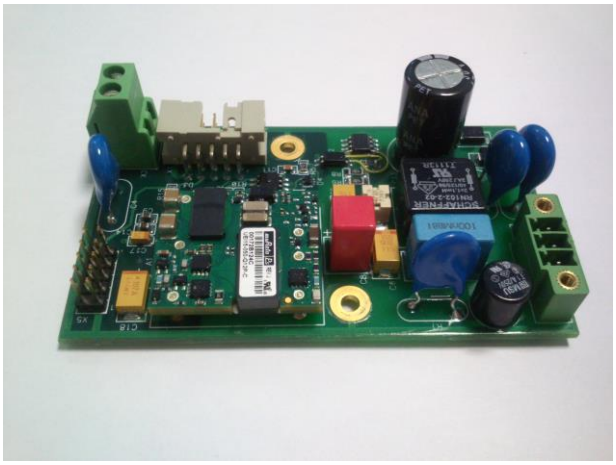


Figure 1. Picture of the analyzed power supply module.

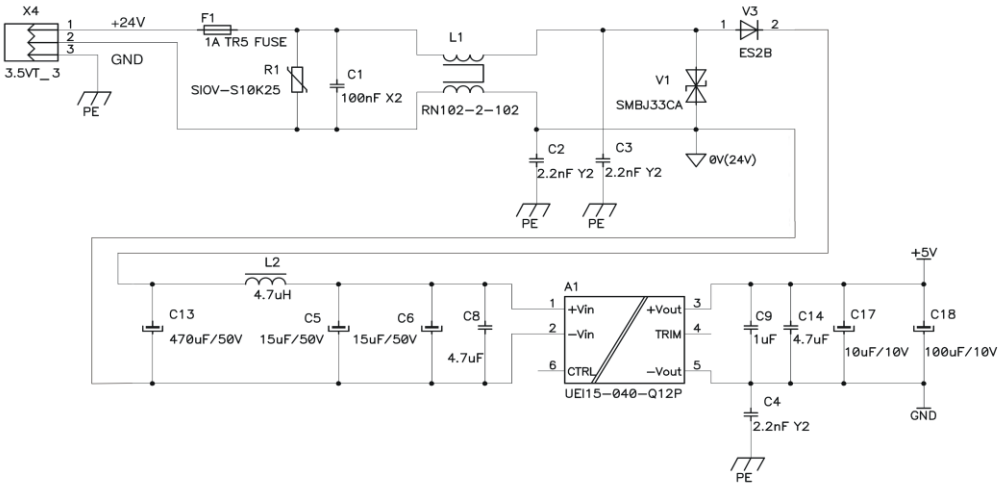


Figure 2. Schematic of the analyzed power supply module.

of the DC/DC converter. This inductive coupling between the primary and secondary side of the DC/DC converter is expected to be the main source of radiated disturbances.

The DC/DC converter used in this circuit is UEI15-050-Q12P-C [5] produced by Murata Power Solutions. The operating frequency of the DC/DC converter is 350 kHz.

The level of radiated disturbances is measured in the TEM cell by the modified method described in [6]–[8].

III. TEM CELL AND MEASUREMENT METHOD

TEM (Transverse ElectroMagnetic) cell is a rectangular coaxial waveguide in which electromagnetic waves propagate in transverse electromagnetic mode and by that create specific fields appropriate for different test procedures [9]. Mechanical design of coaxial waveguide ensures that the external conductor completely encloses the inner conductor (septum). Because of that, TEM cell does not emit electromagnetic energy that could affect nearby equipment. On the other hand, the inner of the waveguide is protected from external fields.

The principal cross-section of the TEM cell is shown in Figure 4. As it can be seen, the main part of the cell consists of the external conductor and septum, placed in the mid-point between upper and lower wall of the cell.

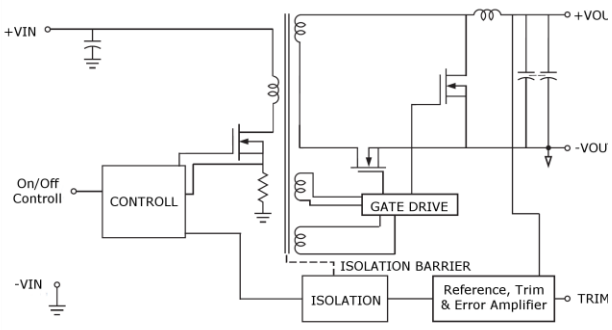


Figure 3. Block-diagram of the DC/DC converter.

In the external top wall there is an opening of 100 mm × 100 mm that enables access to the cell interior and placement of the equipment under test (EUT).

This opening is closed by adapter board on which EUT is placed and which provides all necessary connections between the EUT and outside world. The dimensions of the middle part of the TEM cell and tapered ends are selected to provide 50-Ω characteristic impedance through the whole cell [10].

At the ends of the TEM cell there are connectors that allow connection of external equipment. Generally, one port is used for connection of the measuring equipment: spectrum analyzer or disturbance generator, while the other port is terminated by a 50-Ω termination.

The IEC 61967-1 standard recommends placement of the EUT on the adapter board sized 100 mm × 100 mm. This adapter board performs multiple roles. On one hand it closes opening in the TEM cell external conductor and ensures reliable galvanic connection with it.

This is essential for preservation of a homogenous field within the TEM cell. On the other hand, this adapter board enables mounting of the EUT as well as all required electrical connections with the equipment outside of the TEM cell.

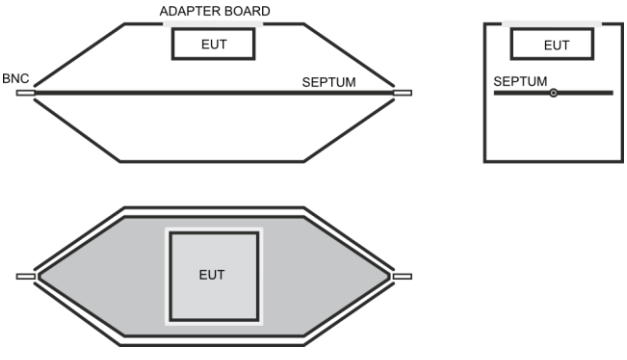


Figure 4. Principal cross-section of the TEM cell.

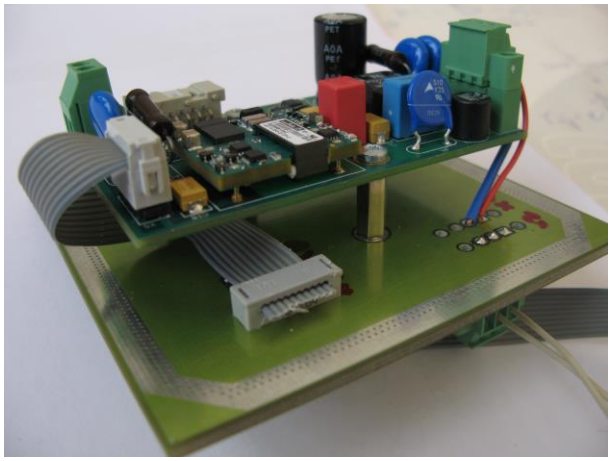


Figure 5. Power supply module mounted on the adapter board.

Because of its shape, the adapter board can be placed in TEM cell in four different orientations relative to the TEM cell port-to-port axis.

The mounting of the analyzed power supply module to the adapter board is shown in Figure 5. This way of mounting is deviating from the recommendations of the IEC 61967-1, since this method is intended for testing of integrated circuits. However this is a simple way how a module to be tested can be placed within a TEM cell.

To measure radiated emissions of the analyzed EUT, one port of the TEM cell is terminated by a 50-Ω termination, while the other port is connected to the spectrum analyzer Rohde&Schwartz FSP7 intended for measurements in the frequency range 9 kHz to 7 GHz. EUT is powered by a 24 V laboratory power supply, and the output is loaded by a variable resistor. The TEM cell used for measurement is FCC-TEM-JM2 [11], produced by Fischer Custom Communications, Inc.

IV. MEASUREMENT RESULTS

A series of measurements of the EUT is made. To facilitate the comparison of the measurement results, all measurements are performed with the same spectral analyzer settings: frequency range = 150 kHz – 200 MHz, reference level = -20 dBm, att = -10 dB, RBW = 3 kHz, VBW = 10 kHz and 8001 points.

As the adapter board can be placed in four different orientations, four sets of measurements are performed. For each orientation, seven measurements are performed:

1. EUT is placed within the TEM cell but without external wiring (supply and load);
2. external wiring is connected to the adapter board, but power supply is turned-off;
3. EUT is connected and power supply is turned-on. EUT is loaded with different loads: 0 mA (0% rated output current); 500 mA (25%), 1000 mA (50%), 1500 mA (75%) and 2000 mA (100%).

The output current is set by a variable resistor and during the whole measurement cycle the output current and voltage are monitored. The output current is kept constant during the whole measurement cycle within ±10 mA of the declared value.

The measurement results are presented below. To reduce the amount of unnecessary detail in representation of the results, the recorded spectrum is processed and spectrum envelope is calculated.

The measurement is started by placing the EUT mounted on the adapter board to the TEM cell. First the measurement is performed without connected wiring. Result is shown in Figure 6. It can be noticed that the spectrum is very similar in all four orientations.

This measurement represents the basic state of the TEM cell. It can be seen that the measured spectrum is practically flat – white noise. This is an expected result since the TEM cell is a closed metal box with the properties of the Faraday cage and there are no electromagnetic field sources within the TEM cell.

In the step 2 all necessary wiring is connected (power supply, load resistor, ampere meter and voltmeter), but the power supply is still turned-off. The result is shown in Figure 7.

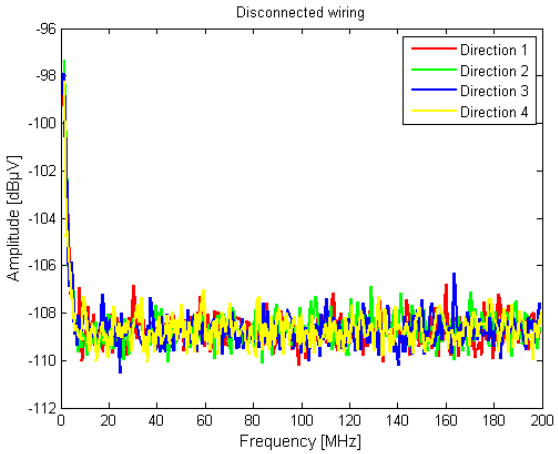


Figure 6. Spectrum in TEM cell with disconnected wiring, all four orientations.

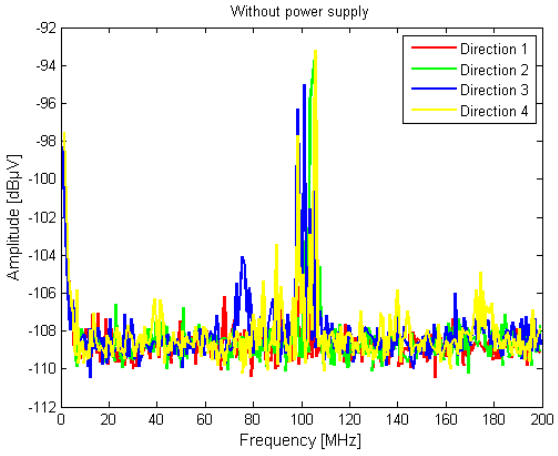


Figure 7. Spectrum in TEM cell with connected wiring, but without power supply, all four orientations.

The measured spectrum shows two distinctive spectral components, with amplitude of 14 dBµV above background noise level in the frequency range around 100 MHz. It can be concluded that this is the frequency range of FM radio, and that test wiring acts like antenna. This could be eliminated by placing the whole measurement system in a Faraday cage.

In the step 3 the EUT is powered and its output is loaded with different output loads. If the output load is 0, the DC/DC converter on the EUT is active, its control electronics is operational and the current through its internal coils is modulated. As expected, the spectral component at 350 kHz (DC/DC converter switching frequency) appears, as well as its higher harmonics, which is shown in Figure 8.

If this spectrum is analyzed in a wider range, up to 200 MHz (Figure 9), it can be seen that in the frequency range to 10 MHz the amplitude of radiated disturbances is falling, as expected. However, the amplitude is increasing again at the higher frequencies. At approx. 17 MHz there is a spectral component with the amplitude of 8 dBµV higher than the component at DC/DC converter switching frequency and the maximal amplitude is at approx. 115 MHz and it is 14 dBµV higher then amplitude on the switching frequency. Compared to the not-powered device case, the maximal increase in disturbance amplitude is 42 dBµV.

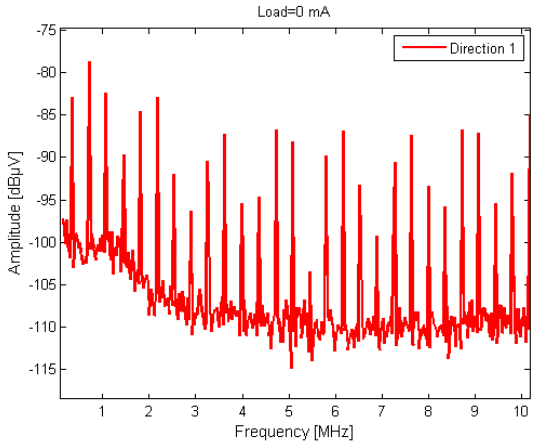


Figure 8. Spectrum in TEM cell in frequency range 150 kHz–10 MHz, with power supply, direction "1".

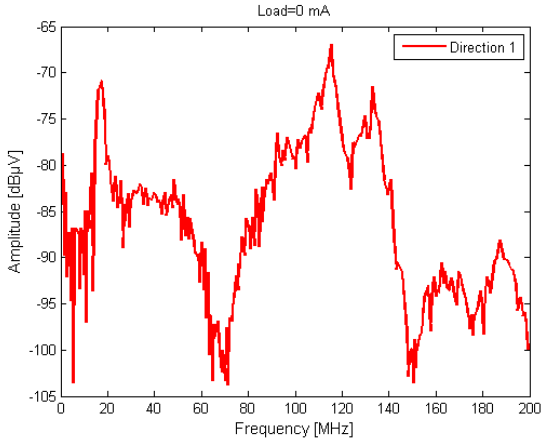


Figure 9. Spectrum amplitude envelope in frequency range 150 kHz–200 MHz, with power supply, direction "1".

If the measured spectrum amplitudes for all four directions of the adapter board are shown in the same image (Figure 10), it can be seen that on the same frequency, disturbance amplitudes of particular spectral components heavily differ, depending on the adapter board orientation. So, for instance, at 53.3 MHz the difference between direction "2" and direction "3" is larger than 20 dBµV.

In Figure 11 the spectrum amplitudes for all four directions and with 100% output load are shown.

Similarly as in the case of unloaded EUT, differences in the amplitude of radiated disturbances spectrum can be observed. But those differences are far smaller compared to unloaded EUT.

A more detailed analysis of the spectrum for orientation in direction "1" is shown in Figure 12. The spectrum is shown for different output loads. The load is changed in steps of 25 %, from 0 mA (0 %) to 2000 mA (100 % of the rated load).

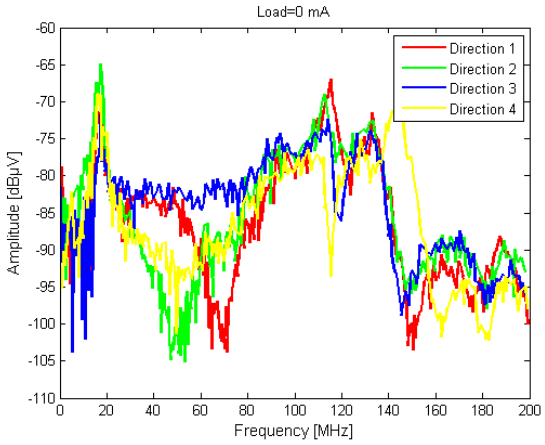


Figure 10. Spectrum amplitude envelope in frequency range 150 kHz–200 MHz, 0% load, all directions.

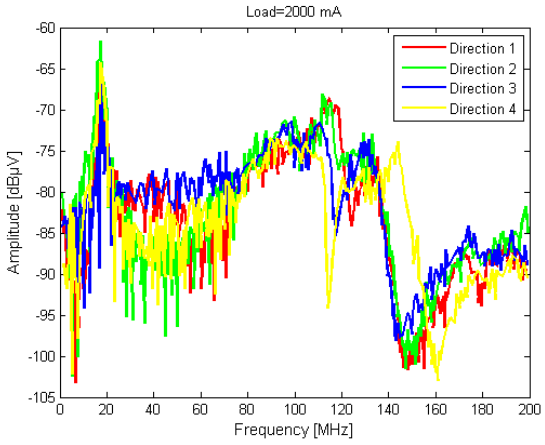


Figure 11. Spectrum amplitude envelope in frequency range 150 kHz–200 MHz, 100% load, all directions.

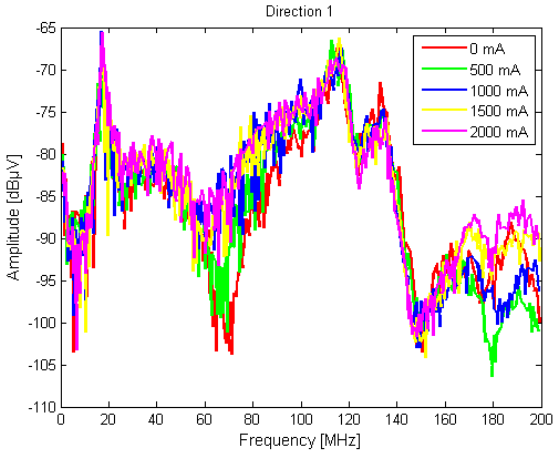


Figure 12. Comparison of the spectrum for the different output loads in direction "1".

By analyzing the measured spectrum it can be seen that, in most of the observed frequency spectra, with the increase of the load, i.e. the increase of the current through the DC/DC converter coil, the amplitude of the emitted electromagnetic radiation is increased. However, in some parts of the analyzed frequency spectrum this is not the case.

This behavior is probably the result of the operating mode of the output voltage regulator of the DC/DC converter.

Also, it is possible that these spectral components are related to the specific resonances of the TEM cell/module set-up or the resonances of the module PCB

V. CONCLUSION

The measurement method for measuring the emissions of electronic modules using the TEM cell is shown. The presented results show that this method enables relatively inexpensive, simple and reproducible procedure for measurements of electronic module emissions. An alternative to the presented method is to perform measurements in a semi-anechoic chamber, i.e. in Faraday cage with absorbers on the walls, which is much more expensive.

The measurement results show that the spectrum of the radiated noise of the analyzed power supply module is much richer than it could have been expected. Furthermore, the maximum emission is not at the switching frequency of the DC/DC converter (350 kHz), but at more than 300 times higher frequency, at 115 MHz.

The results also show the dependence between the measured emissions of the electromagnetic field of the power supply module and its load and orientation within the TEM cell. The measurements have shown that the amplitude of the radiated emissions does not grow uniformly with the increase of the output current. There are bands where the amplitude of the radiated emissions is the largest when the output of the power supply module is not loaded, i.e. the output current is zero.

In order to verify the presented measurement method and the results, it would be opportune to repeat all measurements in a semi-anechoic chamber and compare the results.

REFERENCES

[1] T. Mandic, R. Gillon,B. Nauwelaers, A. Baric, "Characterizing the TEM Cell Electric and Magnetic Field Coupling to PCB Transmission Lines", IEEE Transactions on Electromagnetic Compatibility, Volume:54, Issue: 5, pp. 976-985, Oct. 2012

[2] J.P. Muccioli, T.M. North, K.P. Slattery, "Predicting Module Level RF Emissions from IC Emissions Measurements using a 1 GHz TEM or GTEM Cell – a Review of Related Published Technical Papers", IEEE International Symposium on Electromagnetic Compatibility, EMC 2008., Detroit, MI, USA, 18-22 Aug. 2008.

[3] M. L. Crawford, "Generation of Standard EM Fields Using TEM Transmission Cells", IEEE Transactions On Electromagnetic Compatibility, vol. EMC-16, NO. 4, pp. 189-195, November 1974

[4] V. Kasturi,S. Deng, T. Hubing, D. Beetner, "Quantifying Electric and Magnetic Field Coupling from Integrated Circuits with TEM Cell Measurements", IEEE International Symposium on Electromagnetic Compatibility, EMC 2006., pp. 422-425 Portland, OR, USA, 14-18 Aug. 2006.

[5] Murata Power Solutions, "UEI15 Series, Isolated Wide Input Range 15-Watt DC/DC Converters", Datasheet, 2013.

[6] IEC 61967-1, Ed.1: "Integrated circuits - Measurement of electromagnetic emissions, 150 kHz to 1 GHz - Part 1: General conditions and definitions", International Electrotechnical Commission (*IEC*), 2002.

[7] IEC 61967-2, Ed.2: "Integrated circuits - Measurement of electromagnetic emissions, 150 kHz to 1 GHz - Part 2: Measurement of radiated emissions - TEM cell and wideband TEM cell method", International Electrotechnical Commission (*IEC*), 2005.

[8] IEC 62132-2: "Integrated Circuits, Measurement of Electromagnetic Immunity, 150 kHz to 1 GHz - Part 2: Measurement of Radiated Immunity – TEM-Cell and Wideband TEM-Cell Method", International Electrotechnical Commission (*IEC*), 2010.

[9] T. Mandić, "Modelling of integrated Circuit Packages and Electromagnetic Coupling to Interconnects", PhD Thesis, FER, Zagreb, 2013.

[10] IEEE Std 1309™-2005: "IEEE Standard for Calibration of Electromagnetic Field Sensors and Probes, Excluding Antennas, from 9 kHz to 40 GHz", IEEE Electromagnetic Compatibility Society, 2010.

[11] Fischer Custom Communications, "TEM Cells", Datasheet, 2009.

Model of Low-Noise, Small-Current-Measurement System Using MATLAB/Simulink Tools

Dejan Nagradić*, Krešimir Pardon** and Dražen Jurišić*
*University of Zagreb/Faculty of Electrical Engineering and Computing, Unska 3, HR-10 000 Zagreb, Croatia
**Systemcom Ltd., Kružičeva 4, HR-10 000 Zagreb, Croatia
dejan.nagradic@fer.hr; kresimir.pardon@systemcom.hr; drazen.jurisic@fer.hr

***Abstract*—In this paper the model of a system for measuring small currents, which generates the digital signal proportional to the input current, was designed. The analysed measuring system is the part of a larger microelectronic mixed-signal integrated chip (SoC – System on Chip), and it is located between the input sensor and the external digital device. Since the part that mostly contributes to the noise inside the whole measuring chain is at the input of the chain, the measurement method is analyzed to minimize the impact of noise. Because the noise is not definite, the simulation of measurement by the model was made using statistical principles. Performing multiple measurements of the input value (in our example it is current I_d) it is possible to calculate its mean value and standard deviation. It was shown that the measured quantity can be calculated from the mean value and that the standard deviation represents an amount of noise. Depending on the standard deviation it is needed to do either less or more sequential measurements to determine the reliable measured output digital value.**

Keywords: Small-currents measurement system, Low-noise systems, Sensors, Conversions into electrical quantity, Analog IC design, MATLAB/Simulink tools.

I. INTRODUCTION

Information processing devices that are placed between the sensor and the external circuit are often used in the practice. In this paper we present a model of one such system. The model is used for the functional analysis of a sub-circuit of a microelectronic integrated circuit, which is located between an input sensor and an external device for further digital data processing or value representing (e.g. microcontroller, LCD screen). The circuit under consideration generates a digital value proportional to the input analogue small current, considerably exposed to the noise (highly-sensitive, digital, integrated ampermeter) [1].

The considered circuit is convenient for a broad area of measurements and further analysis, using sensors where a non-electrical quantity (e.g. light, physical, chemical or electrochemical) has been transformed to electrical quantity (voltage or current). The application is very broad: bioindustry, medicine, ecology (water and air pollution), agriculture, food production, aviation, instrumentation, military and automobile industry, etc.

The MATLAB/Simulink model shows a principle to measure small input values. Using the model, we analyse the impact of the noise at the integrated measurement chain (realized as microelectronic integrated circuits on 180nm TSMC technology shown in Fig. 1) at the final output values [2].

II. NOISE

The main challenge in the system for measuring small input currents is noise. Noise is produced by random fluctuations of voltages and currents in the physical elements of electrical circuits, and as such represents a fundamental limitation in the precision of the measurements. In microelectronic systems there are several noise categories: internal noise (thermal, 1/f, shot, etc.), external noise (environment impact, interference etc.). Other effects include offset, finite open-loop gain and slew-rate of operational amplifiers, signal lines crosstalk, etc. [3].

Since in microelectronic devices and circuits it is very difficult to treat, quantize and measure the noise, an equivalent input RMS noise (root mean square noise) is used [4]. Basically, the equivalent input noise is a measure how much noise is added to the input current signal. Such representation of the noise at the input of the model does not have any impact on the functional or operational behaviour of the model and it is also independent of gain and current range settings. The value of all noise contributors in the system (all operational amplifiers, switches, sample & hold, current/voltage converter, etc.) is obtained from laboratory measurement as total RMS value at the output of the processed microelectronic chip. This combined RMS output noise value, also incorporating system's frequency behaviour, is then referred to the input of the model as the equivalent input noise. The effect of offsets of operational amplifiers is reduced by offset cancelation circuits.

Furthermore, this all gives ability using the model to run simulations of the integrated, complex microelectronic system.

III. MODEL IN MATLAB / SIMULINK

In the analysis of the model we will study the parts of the modelled measuring chain. As shown in Fig. 1, the integrated microelectronic measurement system consists of current-voltage converter (I/V converter), differential amplifier (diff. builder), programmable-gain amplifiers (gain chain), sample and hold circuit (S&H circuit) and analogue-to-digital converter (A/D converter).

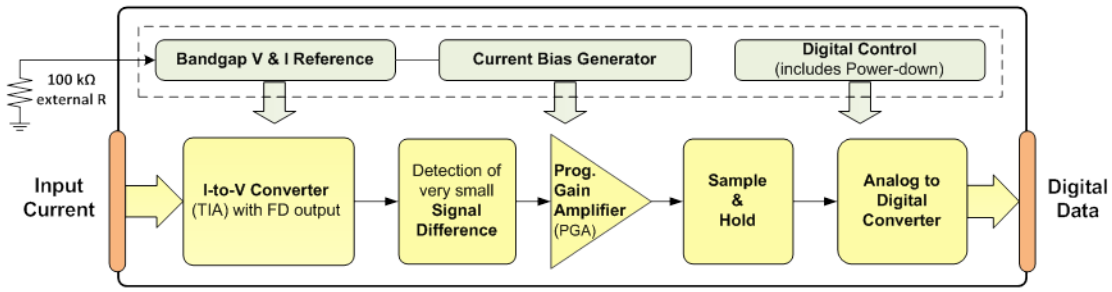


Figure 1. Block scheme of the measuring chain of low-noise, small-current-measurement system.

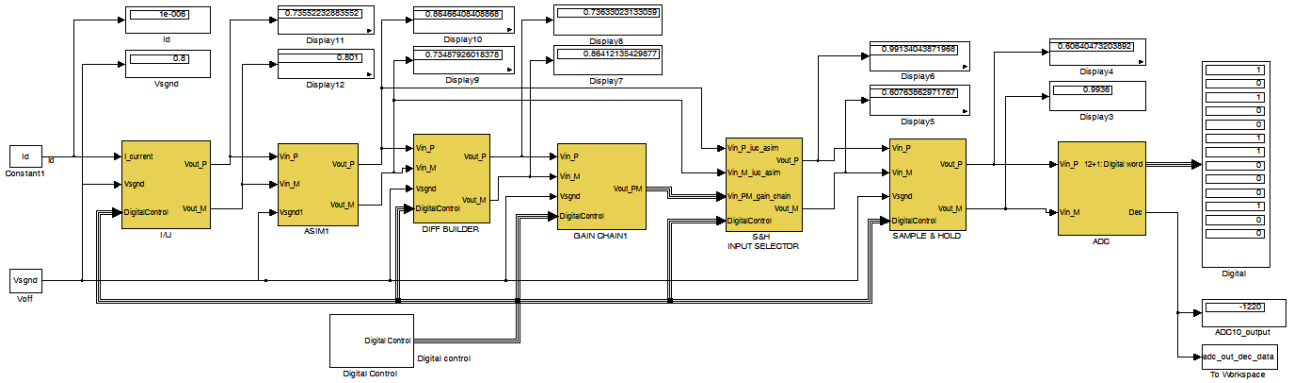


Figure 2. Model of the complete low-noise, small-current-measurement system in MATLAB/Simulink.

All blocks, except the A/D convert, are modelled with their own frequency behaviour. The model is built in MATLAB / Simulink tool [5] and is shown in Fig. 2. It is realized by 8 blocks, and additional blocks for measuring and displaying the output value. At the input there is a current source I_d, and a reference voltage V_{sgnd}. Each part of the system is described as follows.

A. *Current-to-voltage converter*

Current-to-voltage converter is the first block in the MATLAB/Simulink model and is shown in Fig. 3. Its purpose is to convert input current to voltage. It has three control signals: *az* (auto-zero), *cr_select* and *noise_select*.

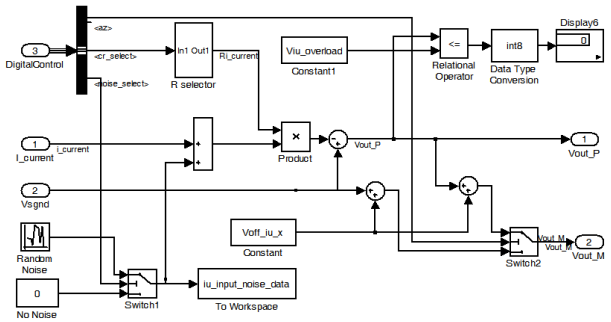


Figure 3. Model of the current-to-voltage converter.

The auto-zero signal controls the auto-zeroing process in order to eliminate offsets of operational amplifiers. When *az*=1, the block samples the combined voltage offset since this mode configures the circuit in the way it is equivalent to input current being zero (I_d =0). When *az*=0, the output has the voltage offset superposed to the useful signal. When those two output signals are subtracted, only the value of useful signal remains. The signal *cr_select* has 3 bits and controls the input current range by selecting the transresistance of the I/V converter, modelled as shown in Fig. 4. When *noise_select* is 1, the random noise is added to the input current. Once the noise is added, it propagates through the signal chain to the output of the model, so it is possible to perform simulations of measurements.

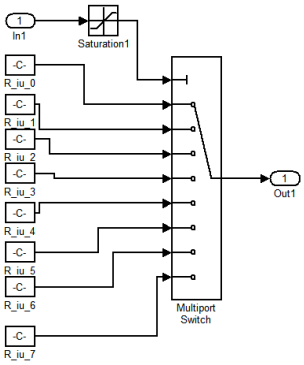


Figure 4. Reistance selector model in current-to-voltage converter.

B. *Symmetric converter*

The symmetric converter is shown in Fig. 5. The I/V converter and the symmetric converter operate as one block. The main purpose of the symmetric converter is to produce differential symmetrical signal. Thus, the current-to-voltage converter and the symmetrical converter together convert measured input current to a symmetric differential voltage signal since it is more robust to interference. As shown below, it is very important to reduce noise in the input stages, because they mostly contribute to the output total noise (see [6]).

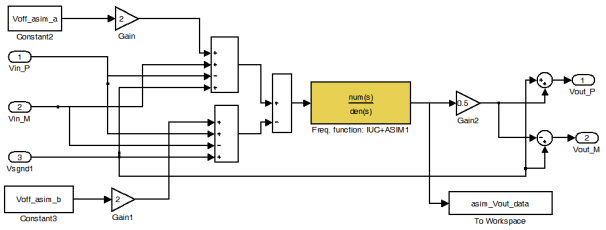


Figure 5. Model of the symmetric converter.

C. *Differentiator*

The next block in the chain is a differentiator and is shown in Fig. 6. The differentiator is the block where subtraction of the offset from useful signal from the I/V converter is done.

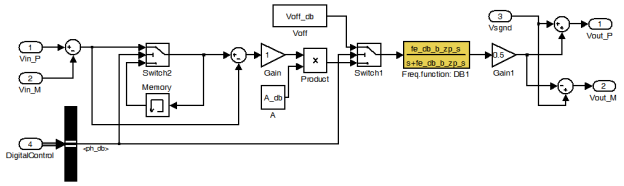


Figure 6. Model of the differentiator.

It operates in two phases, which are controlled by the signal *ph_db*, which serves to make a difference between the two input current sampling (I_d and I_d =0), i.e. the difference between the previous (t–1) state and current state (t).

$$\begin{aligned} &ph_db = 1, \quad \Delta V_{out} = V_{off} \\ &ph_db = 0, \quad \Delta V_{out} = -A\left[\Delta V_{in}(t-1) - \Delta V_{in}(t)\right] \end{aligned} \tag{1}$$

D. *Cascade amplifier*

The fourth block is a cascade amplifier, which consists of four-amplifier stage with bypass and gain control shown in Fig. 7. It determines the overall gain of the chain. Four amplifiers (instead of one) are used for practical reasons of circuit realization and also because of stability and impedance reduction. Since the overall gain of the integrated measurement system is specified before the designing and modelling the measurement system, the number of the gain stages is determined by achieving the gain and to have appropriate gain linearity.

The amplification of every amplifier can be programmed to be 3 or 6 and the maximum gain of the whole block is 6^4 = 1296. The multiplexer with bypass path is used if the desired stage gain is one. Therefore, it is possible to realize only the amplification which is the product of gains 1, 3 and 6 in 4 amplifiers. The final gain of the gain chain is gain multiplication of each stage, given by:

$$G_{final} = G_1 \cdot G_2 \cdot G_3 \cdot G_4 . \tag{2}$$

By setting the gains of the stages in the cascade amplifier, it is possible to optimize the quality of the measurements of small input currents.

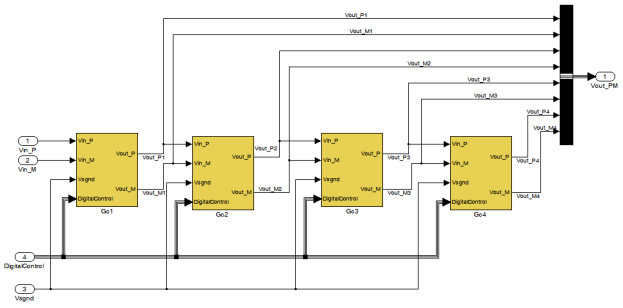


Figure 7. Model of the cascade amplifier.

As shown in Fig. 10 below, showing model of every amplifier in the cascade, each of these 4 amplifiers is controlled by 3 control signals *ph_gc*, *gain* and *ByPass* in the following way: when *ByPass*=0 and *gain*=1 then the gain is 6, when *gain*=0 then the gain is 3; when *ByPass*=1 then the gain is 1.

Each gain stage operates in two phases defined by the control signal *ph_gc*:

$$\begin{aligned} &A)\,\text{For } ByPass = 0 \\ &1)\; ph_gc = 1\ [\text{phase of samplig input signal } \Delta V_{in}(t)] \\ &\qquad \Delta V_{out} = V_{off} \\ &2)\; ph_gc = 0 \\ &\qquad \Delta V_{out} = -Gain\left[\Delta V_{in}(t-1) - \Delta V_{in}(t)\right] \\ &B)\,\text{For } ByPass = 1 \\ &\qquad \Delta V_{out} = \Delta V_{in}(t) \end{aligned} \tag{3}$$

The *noise factor* is a figure-of-merit for a circuit with respect to noise (see [6]–[8]), defined by:

$$F = \frac{S_i / N_i}{S_o / N_o} = \frac{N_o}{N_i} \cdot \frac{1}{A_a}, \tag{4}$$

where S_i and N_i are signal and noise powers at the input, respectively, while S_o and N_o are the same quantities at the output. A_a is system power gain.

If we consider a multistage amplifier (e.g. gain chain), where each stage has its corresponding noise factor F_1, F_2, etc., then the overall noise factor F of the cascade is given by the Friis formula (see [7], [8])

$$F = F_1 + \frac{F_2 - 1}{A_1} + \frac{F_3 - 1}{A_1 A_2} + \ldots + \frac{F_N - 1}{A_1 A_2 \cdot \ldots \cdot A_N}, \tag{5}$$

where A_1, A_2, …, A_N are the maximal power gains of each stage. From (5) and A_i>1; (i=1, 2, …, N) it follows that to minimize the noise factor F of the whole cascade structure, it is most important to have the first block with minimum noise factor F_1.

E. *Sample and Hold and Multiplexer block*

The sample and hold circuit is shown in Fig. 8. There is one control signal *ph_sh* that controls when the circuit takes the sample and when the circuit will let that sample propagate to the next block.

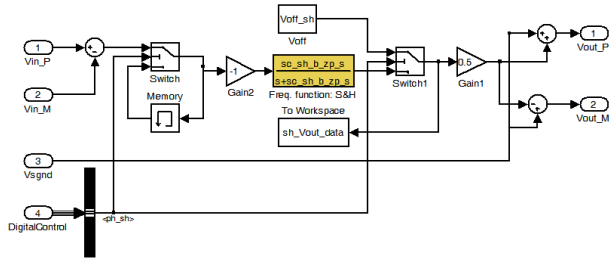


Figure 8. Sample and hold circuit.

The memory elements are capacitors that are charged and discharged, accordingly.

The multiplexer block, shown in Fig. 9, is controlled automatically depending on the state of bypass control bits, so user is not able to change or control block's functionality. The purpose of multiplexer block is to minimize impact of the amplifiers noise in the gain chain blocks in the case when gain = 1 by bypassing the particular gain stage or whole gain chain.

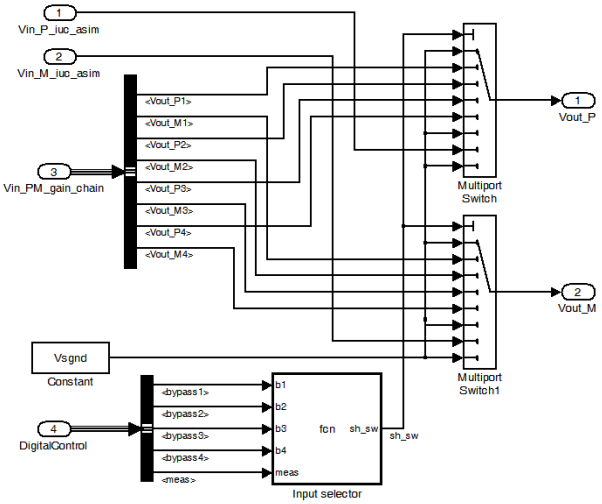


Figure 9. Multiplexer blok (Sample and Hold input selector).

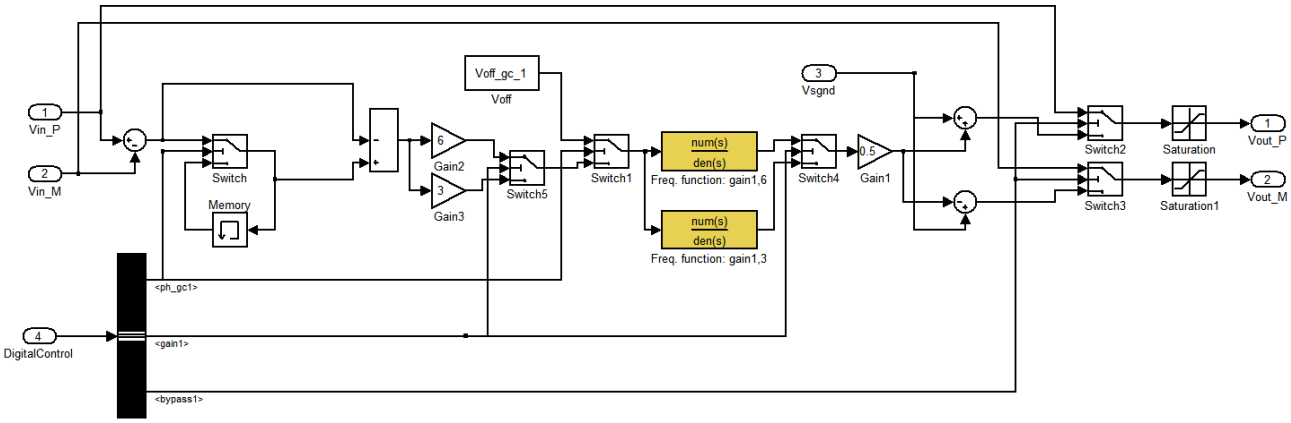


Figure 10. Model of every amplifier in the cascade.

TABLE I. SIMULATION RESULTS OF THE MODEL WITH VARIOUS NUMBER OF SAMPLES.

N	G=1 / I_d = 5.125E-6			G=54 / I_d = 9.4907E-8			G=324 / I_d = 1.5818E-8			G=648 / I_d = 7.909E-9			G=1296 / I_d = 3.9545E-9		
	StdDev	Error	Mean	StdDev	Error	Mean	StdDev	Error	Mean	StdDev	Error	Mean	StdDev	Error	Mean
5	2.191	0.048	–2081.6	131.354	2.873	–2089.6	825.152	18.874	1911.4	443.892	8.638	2640.8	2498.934	68.460	1332.4
10	1.581	0.035	–2081.5	139.828	3.044	–2110.3	558.366	11.038	2558.8	1039.873	22.731	2092.8	1950.097	44.260	1941.3
20	1.943	0.043	–2080.8	148.467	3.260	–2073.6	659.474	14.366	2107.3	1212.146	27.614	1926.8	2128.244	48.306	1941.1
50	2.143	0.047	–2081.2	133.334	2.930	–2071.3	678.909	15.133	2012.5	1417.575	30.998	2091.3	2203.085	49.143	1987.6
100	2.340	0.051	–2081.0	122.757	2.679	–2099.4	706.122	15.297	2130.8	1266.276	27.881	2062.8	2026.784	43.808	2140.4

F. Analog to digital converter

The last block in the chain is ideal analogue-to-digital converter and is shown in Fig. 11. It takes the voltage values stored in S/H circuit and converts them to a digital word. The digital word at the output has 13 bits, of which one bit defines the sign. The conversion is performed according to the formula

$$D_{Integer} = \frac{\Delta V_{in}}{V_{range}} \times 2^{N-1}, \qquad (6)$$

where V_{range} is the output voltage range, and N is the number of the bits (excluding the sign bit), and $D_{Integer}$ is integer value of binary output. Using MATLAB functions we convert measured values from the decimal to the binary system. Thus, we finally obtain a binary word displayed using 13 bits. Using a display in the model, the decimal value is shown too.

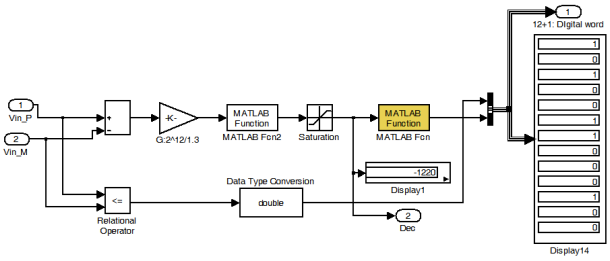


Figure 11. Analog to digital converter.

Finally, there also exist *digital control block* in the model, which contains all the signals with respective wave forms that control other blocks. The control waveforms are shown in Fig. 12.

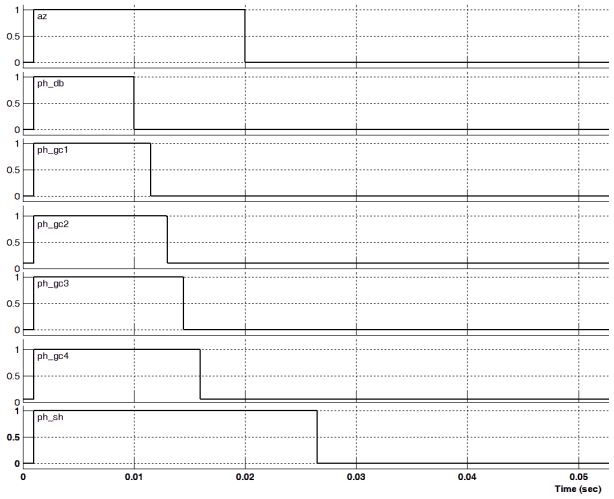


Figure 12. AZ and PH signals at control block.

IV. Operation Of The Model

Running the simulation of the system for small input currents measurement using the MATLAB/Simulink model is performed with *statistical* method of measurement, i.e. the statistical method will make the final calculation of measured quantity [4]. Small input current I_d is not uniquely determined because of the influence of the noise, but it is possible to calculate the mean value and deviation of a set of values obtained from multiple measurements (i.e. samples).

The optimized measurement means to determine the sufficiently large number of multiple measurements. By that, the method of measuring the output quantity with noise is to eliminate the noise by calculating the mean value and standard deviation. The measurement error and the quality of the resulting quantity (value of the mean) are determined by the number of measurements (*sample_size*), and the value of standard deviation. The standard deviation of the output quantity represents the value of RMS noise.

In the simulation of measurements, the automated multiple measurements are performed using a script in MATLAB [5]. As explained in the model description above, using signals *gain* and *ByPass* the gains in cascaded amplifier are determined. The simulations have been run for the gains of the whole system as G=1, 54, 324, 648 and the highest gain G=1296. The input current amount is used for simulation: I_d=5.124[μA]. The random noise generator is set at the input of the model adding the noise to the measured input current I_d.

The numbers (*sample_size*) of sequential running the simulation of the measurement are 5, 10, 20, 50 and 100 (the results are given in Table I). The mean values and standard deviations of the measurements are calculated using a script in MATLAB.

The mean value represents the measured value of the input current, whereas the standard deviation represents the value of the noise (RMS noise). This conclusion comes from the noise nature of a random signal with the normal (Gaussian) distribution of amplitudes and zero mean.

Consider the obtained mean values and standard deviations of the simulated measurements in Table I that vary as a function of gain and the number of measurements. In the case of higher gains (G=54, 325, etc.), the simulated input current at the input is divided by the total gain, in order to produce the same output digital value. In the columns "StdDev" standard deviations of the measurements are displayed and they represent the RMS value of the output noise. In the columns "Mean" the measured / simulated value of the signal (input current) is displayed in the digital form and is within the voltage range V_{range} defined by (6). The ± sign is dependent on the positive or negative gains of the amplifier stages, and it can be neglected (absolute value is used). It is shown that for higher gain the output RMS noise is also higher (higher standard deviation). With higher number of sequential measurements the obtained "Mean" values are nearer to the correct value of measured input current I_d.

V. Conlusion

In this paper we presented the model of a system for measuring small input currents. The model is built using MATLAB / Simulink and it is shown that it is possible to use such model to perform the simulation of the circuit solution that is the part of a mixed-signal IC chip. It is also shown that the noise is the limiting factor in measuring small input currents. The increased noise for higher gains produces the higher standard deviation. Given that the gains are large, and noise is also large, it is necessary to perform a larger number of sequential measurements of the same input value. According to the final value theorem (Gauss distribution), the calculated mean value (as the representative of measured input value) is reliable only after a sufficient number of measurements (see [4], [9]). The expected measurement error δ is given as:

$$\delta = \frac{\sigma}{\sqrt{N}}, \qquad (7)$$

where σ is the standard deviation and N the number of performed sequential measurement.

Therefore, if the standard deviation as a measure of noise is lower, the required number N of sequential measurements (sampling) is also lower. Conversely, if the standard deviation is higher (higher level of noise that comes with higher gain), to have the same level of measurement error, more measurements are needed to be performed and finally the correct mean value can be calculated, reliably representing the measured value (trusted value) [4].

Acknowledgment

This "diploma in industry" work has been possible by the cooperation between the Economic Council of the

Department of Electronic Systems and Information Processing, Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia and the industry represented by the company Systemcom Ltd. Zagreb, Croatia that developed the sensor interface chip described in the paper.

REFERENCES

[1] D. Nagradic. *Model of Low-Noise, Small-Current-Measurement System Using MATLAB/Simulink Tools*. M. Sc. Thesis, FER Zagreb, September 2013.

[2] Systemcom Ltd. *Current-Input Analogue Front End with 13-bit ADC. SC-I-AFE-180F110*, January 2013.

[3] R. Mancini. *Texas Instruments, Op Amps For Everyone, Design Reference*. Dallas, Texas, 2002.

[4] S. W. Smith. *The Scientist and Engineer's Guide to Digital Signal Processing*. *Chapter 2: Statistics, Probability and Noise*. San Diego, California Technical Publishing, 1997-1998.

[5] *MATLAB: Simulink*, MathWorks, Inc., Natick, MA, 2010.

[6] A. Santic. *Electronical instrumentation*. *Spot amplifiers and theirs properties: Noise of the Amplifier*. Zagreb, Školska knjiga, 1993.

[7] Texas Instruments. *Application Report, Noise Analysis in Operational Amplifier Circuits*. Dallas, Texas, 2007.

[8] J. Zurada and M. Bialko, "Noise and Dynamic Range of Active Filters with Operational Amplifiers," IEEE Trans. on Circ. and Sys., vol. CAS-22, no. 10, pp. 805-809, Oct. 1975.

[9] Jeongwook Koh. "*Low-Frequency-Noise Reduction Technique for Linear Analog CMOS IC's*", Ph.D. Thesis, Technische Elektroniks der Technischen Universität München, March 2005.

Algorithm for Fast Determination of LiFePO4 Battery Nominal Current

Igor Sirotić*, Viktor Šunde** and Željko Ban***
*,**Faculty of Electrical Engineering and Computing/Department of Electric Machines, Drives and Automation, Zagreb, Croatia
***Faculty of Electrical Engineering and Computing/Department of Control and Computing Engineering, Zagreb, Croatia
*igor.sirotic@fer.hr, *viktor.sunde@fer.hr, **zeljko.ban@fer.hr

Abstract - The development of production and increasing usage of light electric vehicles such as electric bicycles and electric scooters entails development of batteries, battery chargers and fast charging algorithms. Batteries with different nominal electrical characteristics and production technologies are currently in use. Among them, lithium - iron batteries are increasingly being used. The differences in basic characteristics of batteries and standardization issues require fast identification of various battery types and adapting of charging algorithms according to nominal current, voltage, capacity and state of charge.
This paper describes a method for determining the nominal current for unknown $LiFePO_4$ battery connected to the stationary filling station for charging electric bicycles. A prerequisite for determination of the rated current is identification of current-voltage curve at the moment of connection and identification of the number of cells. The algorithm for determining these parameters is given in the first part of the paper. The second part deals with determination of the nominal current for $LiFePO_4$ battery based on current-voltage response to the selected battery charging algorithm.

I. INTRODUCTION

Light electric vehicles (LEV) are more and more used for transportation of people and goods, especially in urban environments, tourist resorts and protected national or nature parks. Different kind of rechargeable battery types are used for main and auxiliary drives power supply, but lithium-ion based batteries are increasingly used due to their high energy density and compact size [1]. Other battery types, such as Lead-acid, NiCd and NiMH, are still used because of the significantly lower prices. Lithium iron phosphate batteries ($LiFePO_4$) are the safest and with longer cycle life than any other lithium-ion based battery. It has excellent thermal and chemical stability [2] as well so it is expected that $LiFePO_4$ battery will be widely used as energy storage element especially when it comes to portable electronics.

Necessity for light electric vehicles will increase in the near future so there is a need to develop infrastructure i.e. stationary filling stations network. Filling stations would enable user not to just recharge his vehicle battery at many different locations but also to recharge it fast. Main problem with it is that there is no standard present at the moment that defines battery type, nominal voltage, capacity and discharge rate of a battery that would be used in such a vehicle. Since charging has to be done as fast as possible an algorithm has to be developed to determine basic battery parameters and adjust charging algorithm according to them so that the battery would not be damaged or destroyed. Battery can be seriously damaged if its temperature rises above normal rated working temperature [3]. So battery charging algorithm has to be tuned in such a way to transfer most of its energy in form of battery capacity and as less energy possible in form of temperature. Pulse charge method reduces polarization and thus decreases battery temperature rise [4, 5]. Stationary filling stations will not be able to measure battery temperature since most of the LEVs does not have any kind of communication interface. So it is important to determine battery nominal current to tune pulse charging algorithm in a way that we can be sure that battery temperature will not exceed upper limit value.

II. EXPERIMENT SET UP

A. Battery

Batteries convert electrochemical energy to electrical energy with utilization of chemical reactions of oxidation and reduction. During the process, energy losses occur due to dipole polarization, ion migration through electrolyte, uneven ion concentration, accumulation of by products in the nearby of electrodes and electric resistance of electrodes. So, voltage-current relation depends mostly on chemical reaction parameters, not electrical. That is why there is no unique equivalent electrical model of a battery and they differ according to battery type and parameters which have to be observed and analyzed. Equivalent battery model that we use for battery voltage-current and voltage - open circuit voltage (OCV) relations is shown in Fig. 1. E_{m0} voltage source represents OCV of a battery. First RC parallel compound models battery voltage change as state-of-charge (SOC) changes and the second one models significant battery voltage rise when SOC of battery converges to 100%. Resistance R_0 models battery energy losses that linearly increase with load. Current source I_p models battery response when it is in the state of saturation and/or repletion by injecting current to the equivalent electrical model of a battery [6].

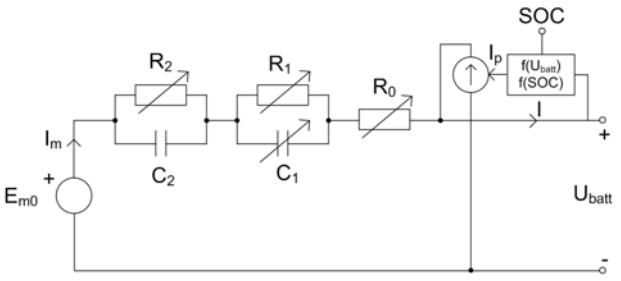


Fig 1. Equivalent electrical model of a battery

Battery terminal voltage equals:

$$U_{batt} = E_{m0} + U_{R0} + U_{R1} + U_{R2}, \quad (1)$$

where

$$U_{R0} = (I + I_p) \cdot R_0. \quad (2)$$

Total RC compound current is

$$i(t) = i_{Cx}(t) + i_{Rx}(t) = C_x \cdot dU_{Rx}/dt + U_{Rx}/R_x \quad (3)$$

so compound voltage can be expressed as

$$U_{Rx}(t) = (I + I_p) \cdot R_x - R_x C_x \int dU_{Rx}. \quad (4)$$

Thus, current-voltage relations of an equivalent model of a battery is stated as follows:

$$U_{batt}(t) = E_{m0} + (I + I_p) \cdot (R_0 + R_1 + R_2) - R_1 C_1 \int dU_{R1} - R_2 C_2 \int dU_{R2}. \quad (5)$$

Laboratory experiments were performed on the $LiFePO_4$ batteries that consists of cylindrical power cells. Batteries that were tested have different parameters and number of cells. Cells parameter specification is shown in Table I.

TABLE I. CELL PARAMETER SPECIFICATION

$LiFePO_4$ power cell		
	Battery 1 (Cell)	**Battery 2 (Cell)**
Parameter	***Value***	***Value***
Nominal capacity	10000mAh (0,5C)	15000mAh (0,5C)
Nominal voltage	3.2V	3.25V
Inter impedance	≤10mΩ	≤8mΩ
Maximum charge current	3C (30A)	2C (20A)
Maximum charge voltage	3.65 ± 0.05V	3.65 ± 0.05V
Discharge stop voltage	2.0V	2.2V
Work temperature - charge	0~45ºC	0~45ºC
Work temperature - discharge	-20~60ºC	-20~60ºC

B. *Control, Power and Measurement*

Block schematics of elements and connections used in experiment is shown in Fig. 2. Charging algorithm is developed using *Matlab* software which is in charge of controlling *Magna* current source. Communication is established using TCP/IP communication protocol and control using SCPI commands. *Magna* current source is connected directly to battery terminals by means of 16mm^2. To eliminate any errors due to voltage drop in power supply cables, output current source voltage is externally measured at the battery terminals.

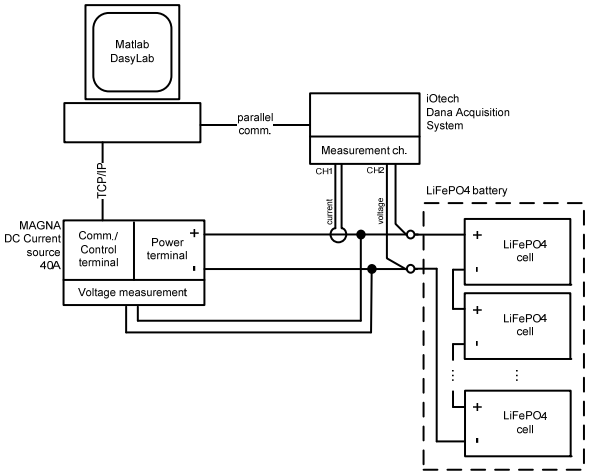


Fig. 2. Block schematics of experiment set up

iOtech WaveBook Data Acqusition System is used for recording battery voltage and current during charging process. *WaveBook* is configured using *DasyLab* software which is also used for data analyses.

III. OBSERVATION AND RESULTS

$LiFePO_4$ battery OCV - SOC relation can practically be divided in three areas depending on the SOC (as SOC changes from 0-100%): aperiodically, linear and exponential voltage increase area, Fig. 3.

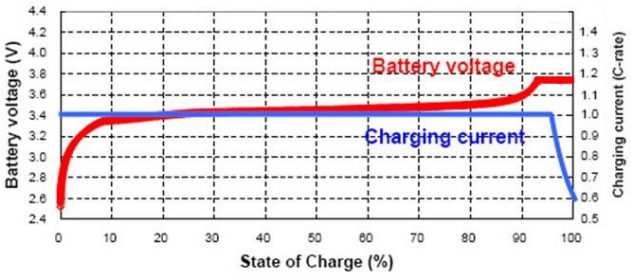


Fig. 3. OCV - SOC curve - aperiodical, linear and exponential voltage increase

While applying impulses of current to the battery we can observe the change of battery terminal voltage gradient to determine in which of those three areas of OCV - SOC curve battery momentarily is. Relation of initial voltage gradient in order to the SOC of the battery (charging current equals 0.5C) is shown in Fig. 4.

When the battery is in the exponential voltage increase area of OCV - SOC curve it basically means that

is already fully charged and no additional charging is required. Contrary to that, if the battery is in the area of aperiodical voltage increase, it means that the battery is almost completely empty.

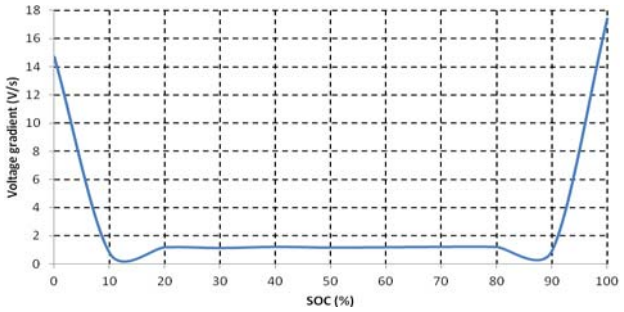


Fig. 4. Initial voltage gradient in relation to SOC (charging current = 0,5C)

To be able to make any conclusions, battery first needs to be charged until it reaches linear voltage increase area of OCV-SOC curve. Since $LiFePO_4$ battery accepts charge very well when it is discharged to low SOC values, transition to linear area can be done in under a minute.

When we reach this linear area and then apply charging current (up to 0.9C) battery cell terminal voltage value is certainly in the span of 3.2 and 3.4V. Uncertainty of 0.2V allows us to determine number of cells[1] of the battery if it has up to n = 16 cells. Because of all of the other uncertainties in the system, this number in practice is n = 14. Information of number of battery cells is important for nominal current determination because it is based on fixed terminal voltage levels of a battery cell. Voltage measured at the battery terminals is scaled down to single cell voltage level and that is one of entry values of the nominal current determination algorithm. Since number of cells determination is already done, we can be sure that the battery is in the linear area of OCV - SOC curve. Under this condition, when constant current is applied to the battery, its terminal voltage will at first start to aperiodically[2] increase up to a certain level. This takes some time, after which voltage continues its slow linear increase (Fig. 5.). Figure 5. also shows an example of how voltage gradient differs when different charging current is applied (at the same SOC).

Linear increase slope is so small that voltage is practically constant value until battery reaches SOC of nearly 100%. As the input current increases, voltage level increases accordingly, Fig. 6.

Experiment included charging of the batteries of different kind of nominal voltage and nominal parameters with following input current values: 0.5C, 0.8C, 1C, 1.2C, 1.5C and 2C. The results are shown in Table II.

[1]It is meant on number of cells connected in series within one branch of a battery.

[2]This aperiodical voltage increase mustn't be confused with aperiodical voltage increase at very low SOC. They differ by battery terminal voltage gradient behavior when charging current is applied. In this case we already know that the battery is in the linear area of OCV-SOC curve and no further check is required.

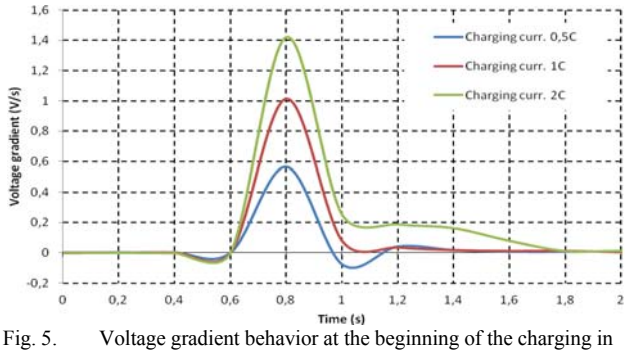


Fig. 5. Voltage gradient behavior at the beginning of the charging in relation to different charging currents at the same SOC

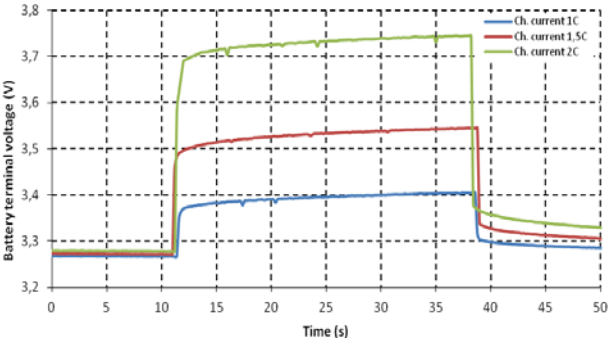


Fig. 6. Battery terminal voltage in relation to charging current

For every experiment that was conducted with defined current values, battery terminal voltage is within boundaries shown in Table II. Since battery terminal voltage rise is not stepwise, we also monitor battery terminal voltage gradient during its aperiodical increase. Most right column of the Table II. shows maximum gradient that appeared during experiments for specified current values.

TABLE II. EXPERIMENT RESULTS

Charging current	Battery terminal voltage span (V)	Battery terminal voltage gradient span (mV/s)
0.5C	3.22 - 3.28	≈ 0
0.8C	3.26 - 3.33	≤ 0.9
1.0C	3.35 -3.43	≤ 2.53
1.2C	3.41- 3.49	≤ 3.32
1.5C	3.55 - 3.65	≤ 4.44
2.0C	3.65 - 3.75	≤ 9.0

Having information of voltage levels and voltage gradient behavior allows us to determine nominal current of a battery up to a certain level of accuracy. This level of accuracy is good enough for setting initial parameters of pulse charging algorithm. The aim that we aspire to is to determine nominal current to a better accuracy level throughout pulse charge algorithm and then adjust parameters according to new input values. Adapting the pulse charge algorithm while actually charging the battery decreases necessary charging time.

IV. $LiFePO_4$ BATTERY NOMINAL CURRENT DETERMINATION ALGORITHM

Fig. 7. shows nominal current determination algorithm. Values of current, time and voltage gradients boundaries are set according to common batteries available in market for LEV-s and experience gained during the laboratory experiments.

At the very beginning of the algorithm process, small charging current is applied and it is increasing stepwise while voltage gradient is being monitored. If gradient reaches 100mV/s, its value is memorized and the process is repeated once more. In the case that gradient exceeds 100mV/s value again, it means that the battery is either in the state of aperiodical voltage increase or in the state of exponential voltage rise. We can differ these two states by comparison of two voltage gradient levels that have just been measured. If its value has risen it means that the battery is in the exponential area i.e. it is already fully charged and thus, charging process is finished. However, if it has reduced, battery is in aperiodical area and needs to be charged until it reaches linear voltage increase area. We consider that the battery has reached that area when the voltage gradient reduces under the value of 1mV/s. In the case that gradient value doesn't exceed 100mV/s at all, battery is already in the state of linear voltage increase so we can proceed to battery nominal current determination.

A 30 second pause follows to let the battery terminal voltage to come near to steady state value and then the determination begins. At first, charging current of 5 amps is applied to the battery terminals since it is the smallest expectable nominal current value of electric bike batteries. As algorithm proceeds, current is increasing in steps of 1 amp every 5 seconds. During that process voltage and voltage gradient are analyzed and compared to data from the database (Table II.). When voltage level and its gradient correspond to the values in the Table II., row 3 (charging current of 1.0C), temporary charging current is considered to be nominal.

Nominal current value is the output of this algorithm and the input of the pulse charging algorithm which is yet to be developed.

CONCLUSION

Experimental tests performed on LiFePO4 power cell batteries show recognizable behavior of battery terminal voltage in relation to input current. Thus, conclusions of number of battery cells and their nominal current can be made. A $LiFePO_4$ battery nominal current determination algorithm has been proposed. Determined nominal current is used as initial condition for adjusting the pulse charge algorithm. Pulse charge algorithm must be developed in such a way to adapt its initial parameters during charging process according to new data obtained out of constant measurements.

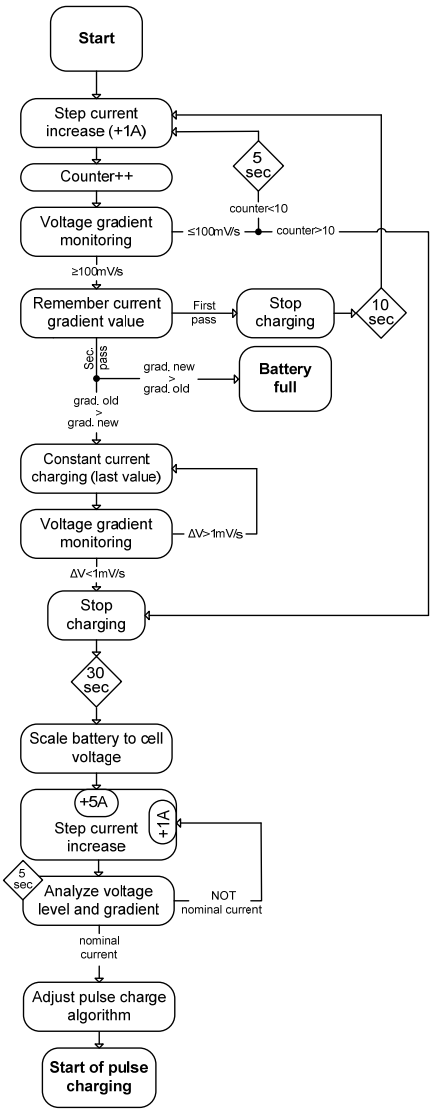


Fig. 7. Algorithm for LiFePO4 nominal current determination

REFERENCES

[1] K. M. Tsang, L. Sun, W. L. Chan, "Identification and Modeling of Lithium Ion Battery", Energy Conversion and Management, vol.51, pp. 2857-2862, 2010

[2] Chang Yoon Chun, Gab-Su Seo, Bo-Hyung Cho, Jonghoon Kim, "A Fast Stat-of-Charge Estimation Algorithm for $LiFePO_4$ Batteries Utilizing Extended Kalman Filter", ECCE Asia Downunder, Melbourne, Australia,June 2013, pp. 912-916

[3] Jiexun Liu, Dawei Gao, Jianhua Cao, "Study on the Effects of Temperature on $LiFePO_4$ Battery Life", Vehicle Power and Propulsion Conference, Seoul, Korea, October 2012, pp. 1436-1440

[4] A Chih-Chiang Hua, B. Zong-Wei Syue, "Charge and Discharge Characteristics of Lead-Acid Battery and $LiFePO_4$ Battery", International Power Electronics Conference, Sapporo, pp. 1478-1483, June 2010

[5] L. W. Yao, J.A. Aziz, "High Capacity $LiFePO_4$ Battery Model with Consideration of Nonlinear Capacity Effects", International Power Electronics and Motion Control Conference - ECCE Asia, Harbin, China, June 2012, pp. 182-187

[6] I. Grbeša, " Identification of model batteries for energy storage", diploma thesis, Faculty of Electrical Engineering and Computing University of Zagreb, 2012, Zagreb.

100 nA Power Management Circuit for Energy Harvesting Devices

D. Vinko and G. Horvat
University of Osijek, Faculty of Electrical Engineering, Department of Communication, Osijek, Croatia
davor.vinko@etfos.hr

Abstract – This paper presents a power management circuit suitable for energy harvesting devices. Energy harvesting methods provide very low instantaneous power which is not sufficient to continuously power a wireless sensor node. Presented power management circuit is designed to turn on the wireless sensor node when the voltage across storage capacitor reaches 5 V and turns it back off as the voltage drops to 4.5 V. In such operating conditions maximal current that power management circuit draws is under 100 nA. Laboratory model of a power management circuit is realized and the performance measurements for various inputs were conducted. Additionally, a mathematical model of the power management circuit is derived and compared with measurement results.

I. INTRODUCTION

Energy harvesting is gaining more popularity as an energy supply for low power electronics. Various applications in powering wireless sensor nodes [1]-[3] have been reported. Energy harvesting, also known as power scavenging, uses variety of methods to generate useful power. Most common methods of energy harvesting include harvesting energy from RF radiation, ambient light, vibration, small scale wind energy harvesting, thermal energy, and bioelectricity. All of these energy harvesting methods are able to provide low instantaneous power. Available instantaneous power ranges from 1 µW to 5 mW [4], [5].

Majority of the applications require higher instantaneous power. For example, wireless sensor nodes require up to 60 mW as peak power consumption while the RF part of wireless sensor node is active [6]. Since energy harvesting methods cannot be used as continuous power supplies, they can be used in applications where real-time response is not required.

Wireless sensor nodes can be used to monitor weather. It is sufficient for a wireless sensor node to send the data at a very low sampling rate (e.g. weather information; couple of times per hour). This means that a sensor node can collect and store harvested energy over a longer period of time, and use stored energy for a short active period for data transmission. For that an additional power management circuit is needed which monitors the amount of harvested (stored) energy and which turns on the sensor node when enough energy has been harvested.

Power management circuit acts as mediator between energy harvesting circuit and the load. Main design guideline for power management circuit is to reduce the circuit power consumption to a minimum. Significant percentage of energy harvesting methods can only provide instantaneous power in the range of few tens of µW, therefore the power management circuit must be designed with even lower power consumption. This ultra-low power consumption requirement refers to a period during which the device collects and stores energy from energy harvesting circuit.

Common energy harvesting circuit comprises of energy harvester followed by voltage multiplier, rectifier circuit and storage capacitor [7]. Voltage multiplier uses low amplitude AC voltage from energy harvester and generates higher DC voltage. The energy harvesting circuit (together with voltage multiplier) can be replaced and modeled with Thevenin voltage and resistance. Without loss of generality, Thevenin voltage can be set in the range of 10 V, while the Thevenin resistance must correspond to the reported instantaneous power available from various energy harvesting methods. For available power ranging from 1 µW to 5 mW, the Thevenin resistance ranges from 50 MΩ to 10 kΩ, respectively, giving maximal instantaneous DC current of energy harvesting circuit from 200 nA to 1 mA.

Load circuit is intended as a low power electronic device, such as wireless sensor node. With inherently nonlinear power consumption profile (power up, active operation, RF activity), the maximal power consumption is reported during RF activity [8] and is under 100 mW. To model a majority of available wireless sensor nodes, the load circuit can be modeled as a device which requires 5 V supply voltage and 50 mA continuous current, which gives an average power consumption of 250 mW [9].

Additionally, as the load circuit operates in cycles, device powered using power management circuit can be modeled by the energy required to complete a single operating cycle. Such modeling enables optimization of energy storage and power management, what will be discussed in more detail in following section of the paper.

This work was sponsored by J. J. Strossmayer University of Osijek under project "Passive Wireless Sensor Network".

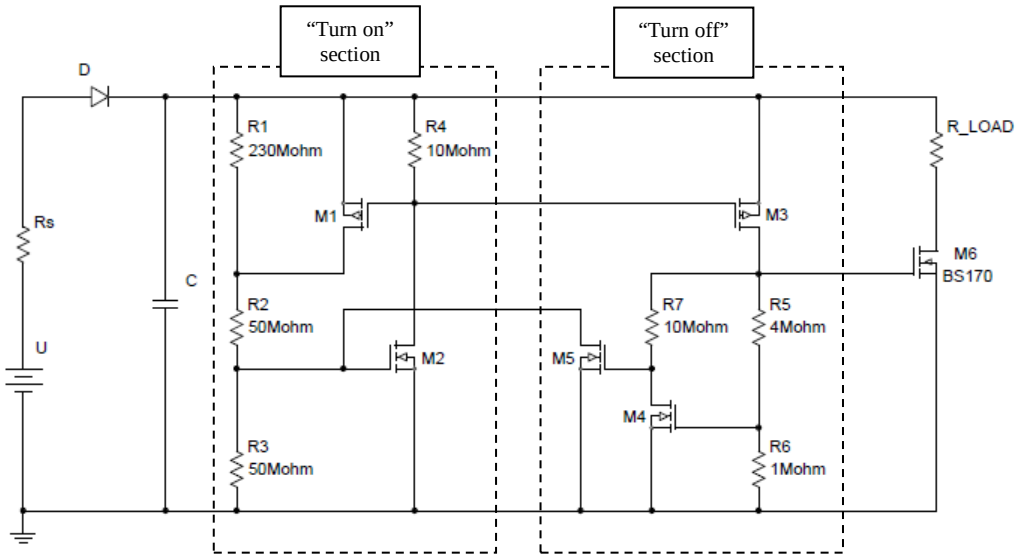


Figure 1. Schematic of the power management circuit

II. Power Management Circuit

The schematic of proposed power management circuit (PMC) is given in Fig. 1. Voltage source U and resistor R_S represent the energy harvesting circuit, while resistor $R_{_LOAD}$ represents a load circuit. Diode D serves as a rectifier and capacitor C is storage element which collects harvested energy.

There are two operating phases of power management circuit: the charging phase and the discharging phase. In charging phase the energy harvester is charging the capacitor C with disconnected load. In discharging phase, the energy stored in capacitor C is used to supply the load. This way an energy harvesting method with low current capability can be used to power a load with higher current demand.

The PMC is divided in two sections: "Turn on" section and "Turn off" section, which monitor charging and discharging phases, respectively. "Turn on" section uses voltage divider (resistors R_1, R_2 and R_3) to monitor the voltage across capacitor C during charging phase. High value resistors were used to minimize the power consumption during charging phase. This is important due to fact that the majority of energy harvesting methods is only able to provide instantaneous power in the range of few tens of µW [3].

When the voltage across capacitor C reaches "turn on" level, the transistor M_2 turns on and pulls gates of M_1 and M_3 to the ground, thus turning them on and starting the discharging phase. Transistors M_3 acts as a switch for "Turn off" section and also for the load circuit (represented with resistor $R_{_LOAD}$ on Fig. 1). Load can be driven directly by transistor M_3 or by additional transistor M_6 which can supply higher load current. Transistors M_1 to M_5 are CMOS transistors integrated in HCC4007UB integrated circuit and their maximal drain current is in the range of few mA. To test the PMC with loads that require up to 50 mA, the additional transistor BS170 (M_6 in Fig. 1) with maximal drain current of 500 mA is added.

With connected load circuit the charge stored in capacitor C discharges and voltage across capacitor drops. Voltage divider (R_5, R_6) in "Turn off" section keeps transistor M_4 turned on as long as voltage across capacitor C is above "turn off" level. When the voltage across capacitor C drop below "turn off" level, transistor M_4 turns off, causing transistor M_5 to turn on. M_5 then turns off transistor M_2, which turns off all transistors in PMC, and starts the charging phase again.

In applications such as wireless sensor networks, the energy harvesting is used to power the wireless sensor nodes. For each sensor node the energy required for one operating cycle can be measured. With known required energy the needed capacitance of capacitor C value can be calculated.

Energy consumption ΔE of a load circuit during discharge phase is equal to difference of energy stored in capacitor C at the beginning (E_{max}) and at the end (E_{min}) of discharging phase, (1).

$$\Delta E = E_{max} - E_{min}\,. \tag{1}$$

With capacitor C charged to a voltage level U, the energy stored in capacitor is equal to:

$$E = \frac{1}{2} C U^2 \tag{2}$$

and the energy consumption ΔE can be derived from maximal and minimal capacitor voltage levels (U_{max} and U_{min}) , which correspond to a beginning and the end of the discharging phase:

$$\Delta E = \frac{1}{2} C \left(U_{max}^{\ 2} - U_{min}^{\ 2} \right). \tag{3}$$

From (3), for known required energy per operating cycle, a needed capacitance value can be expressed as:

$$C = \frac{2\Delta E}{U_{max}^{\ 2} - U_{min}^{\ 2}}\,. \tag{4}$$

A. *Mathematical Model of charging process*

Charging process of proposed power management circuit is modeled by schematic given by Fig. 2.

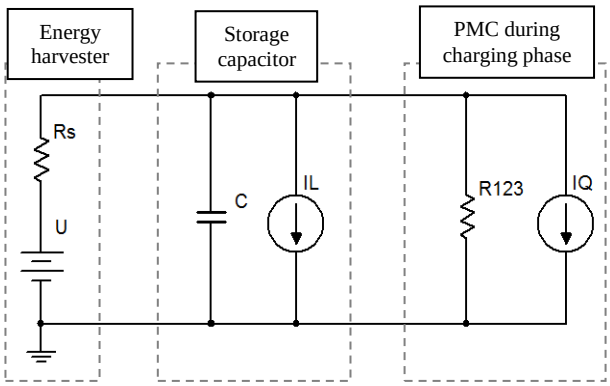


Figure 2. Charging process model

Energy harvester is modeled as voltage source U with series resistance R_S (equal to Fig. 1.). Leakage current of storage capacitor is modeled as current sink I_L. During charging phase the only active part of the power management circuit is resistive divider (R_1, R_2 and R_3) which is represented as their series equivalent R_{123}. Additional current sink (I_Q) was added to represent quiescent current found in CMOS transistors used in PMC (IC HCC4007UB).

Differential equation of the charging process equals to:

$$\frac{du_C}{dt}+u_C\frac{1}{C}\left(\frac{R_S+R_{123}}{R_S R_{123}}\right)=\frac{U-R_S\left(I_L+I_Q\right)}{R_S C}. \quad (5)$$

When solved, the expression for voltage across storage capacitor during charging phase is:

$$u_C=\left[U-R_S\left(I_L+I_Q\right)\right]\frac{R_{123}}{R_S+R_{123}}\left(1-e^{-\frac{R_S+R_{123}}{R_S R_{123}C}\cdot t}\right). \quad (6)$$

Time needed to charge a storage capacitor to a certain voltage level U_C is:

$$t_{U_C}=-\frac{R_S R_{123}C}{R_S+R_{123}}\ln\left(1-\frac{U_C}{U-R_S\left(I_L+I_Q\right)}\cdot\frac{R_S+R_{123}}{R_{123}}\right). \quad (7)$$

Unknown values of leakage current I_L and quiescent current I_Q can be calculated using measured time t_{Uc} needed to charge the storage capacitor C to a voltage level U_C:

$$I_L+I_Q=\frac{U}{R_S}-U_C\frac{R_S+R_{123}}{R_S R_{123}}\left(1-e^{-\frac{R_S+R_{123}}{R_S R_{123}C}\cdot t}\right)^{-1}. \quad (8)$$

Maximal achievable voltage across storage capacitor is then:

$$U_{C_{MAX}}=\left[U-R_S\left(I_L+I_Q\right)\right]\frac{R_{123}}{R_S+R_{123}}. \quad (9)$$

III. MEASUREMENT RESULTS

Experimental verification of presented power management circuit is performed on laboratory model using discrete components. The power supply U and R_S (Figs. 1. and 2.) represents energy harvester. Reported energy harvesting methods [4], [5], [7] significantly vary in output voltage (and output current) they can provide. Also, various approached are used to ensure voltage levels sufficient to power wireless sensor nodes. Depending on voltage levels available from energy harvester some use voltage multipliers (lower voltage levels from energy harvester) and some voltage down converters (higher voltage levels from energy harvester). Therefore, without any loss of generality, a power supply (U and R_S) used in measurements can be set to a fixed value of U = 10 V, while the series resistance R_S is varied during measurements to provide wider range of energy harvester output current. With R_S ranging from 10 kΩ to 50 MΩ a maximal output current of energy harvester ranges from 1 mA to 200 nA, respectively. Then the maximal instantaneous power that energy harvester can provide ranges from 5 mW to 1 µW, which is in concordance with reported energy harvesting methods [4], [5].

Values of resistive dividers (R_1, R_2, R_3 and R_5, R_6) were chosen as such that the load circuit turns on when the voltage across capacitor reaches 5 V, and turns back off when the voltage drops to 4.5 V. This can be seen in Fig. 3 which shows the voltage across storage capacitor (upper trace) and voltage across load circuit (bottom trace).

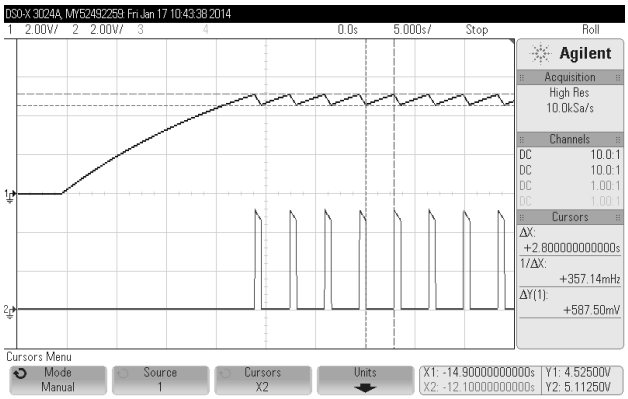


Figure 3. Voltage across storage capacitor and load circuit

Measurement shown in Fig. 3 is done with storage capacitor C = 470 µF, resistor R_S = 50 kΩ and resistive load $R_{_LOAD}$ = 10 kΩ. With measured turn on and turn off voltage levels equal to 5.1125 V and 4.525 V, respectively, the energy delivered to resistive load per operating cycle is equal to 1.3 mJ (3). This amount of energy is sufficient to power one operating cycle of wireless sensor node [6], [9].

With resistor R_S = 50 kΩ the measured charging time between two operating cycles of PMC is 2.8 seconds. With higher values of R_S the charging time is increasing, due to lower current available from energy harvester. To find a minimal current from energy harvester sufficient to charge storage capacitor and power a resistive load in repetitive cycles the following measurement is conducted. Resistor R_S values were varied from 10 kΩ to 5 MΩ for the PMC with storage capacitor C = 470 µF, and the

charging time between cycles is measured. Besides charging time, it is important to measure the "on" and "off" voltage levels so that the measurement results can be compared with mathematical model of the presented circuit (7).

Results are shown in Fig. 4. Due to high capacitor value (C = 470 µF) and low charging current, the charging time is rather large and is in the range of several minutes (for higher values of R_S). Since the charging time is measured by oscilloscope (Agilent DSO-X 3024A) where maximal time interval which can be measured is 500 s, minimal operating current could not be measured.

Therefore, to reach minimal operating current, additional measurements were performed with higher values of R_S and lower capacitance of storage capacitor C = 1 µF. Fig. 4 also shows calculated charging time values using mathematical model (5)-(8). Parameters in mathematical model also include leakage current I_L and quiescent current I_Q whose values are unknown. In order to extract correct I_L+I_Q value from measurements, the following expression is derived from (7):

$$I_L + I_Q = \frac{U}{R_S} - \frac{R_S + R_{123}}{R_S R_{123}} \cdot \frac{U_{on} - U_{off} e^{-\Delta t \frac{R_S + R_{123}}{R_S R_{123} C}}}{1 - e^{-\Delta t \frac{R_S + R_{123}}{R_S R_{123} C}}}. \quad (10)$$

With higher R_S (lower input current) the impact of leakage and quiescent currents is more emphasized. R_S value is increased up to a point where the storage capacitor could not reach the turn on voltage. The measurement results for the highest value of R_S at which the PMC could still operate were then used to calculate I_L+I_Q value. This point is marked by an arrow in Fig. 4. Using measured values of charging time, "turn on" and "turn off" voltage levels, the value of I_L+I_Q = 61.87 nA is calculated from (10).

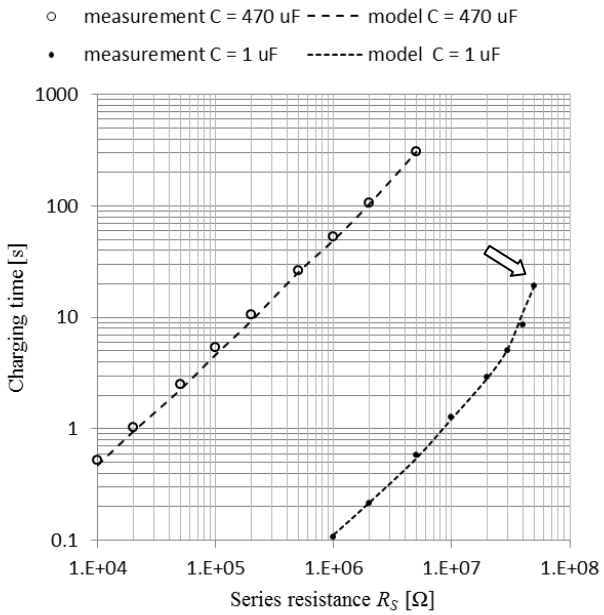


Figure 4. Comparison of measured and calculated values of storage capacitor charging time for different values of resistance R_S

Component datasheets provide wide range for both leakage and quiescent current values. Leakage current value for tantalum capacitor ranges from 0.01 nA/µFV to 10 nA/µFV, with typical value being 0.1 nA/µFV. For a presented PMC with maximal voltage across storage capacitor of 5 V and capacitance of 1 µF, the leakage current ranges from 0.05 nA to 50 nA , with typical value of 0.5 nA. Quiescent current for HCC4007UB integrated circuit ranges from 10 nA (typical value) to 250 nA (maximal value). The calculated value of I_L+I_Q = 61.87 nA is well within values provided by component manufacturers.

Together with current drawn by voltage divider R_{123}, the overall current demand of PMC during charging phase is well under 100 nA. This is in concordance with value of R_S = 50 MΩ at which the PMC could still operate. Namely, with turn on voltage level U_{on} = 5 V, the maximal current from energy harvester at the end of the charging phase is 100 nA.

Such ultra-low current consumption (together with storage capacitor of 470 µF) ensures that ripple voltage from voltage multiplier/rectifier (in paper represented by DC voltage source U) stays under 100 µV. With 0.5 V difference between turn on and turn off voltage level, ripple voltage has almost no effect on the performance of the PMC.

IV. CONLUSION

Presented power management circuit is designed to act as a biasing circuit between energy harvester and wireless sensor node. The harvested energy is collected over a longer period of time, and when sufficient energy is stored, it is used to power a wireless sensor node for a short time interval. Measurements on laboratory model of power management circuit are performed and results agree with proposed mathematical model. With operating voltage level of 5 V, maximal current that power management circuit draws is under 100 nA. This makes it suitable for usage with almost all energy harvesting methods, since it can operate with 1 µW of input power.

The future work will address the issue of rather high quiescent current in more detail. The aim is to reduce current consumption of power management circuit even further by implementing devices with lower quiescent current.

REFERENCES

[1] N. Srivastava, "Challenges of Next-Generation Wireless Sensor Networks and its impact on Society," Journal of Telecommunications, Volume 1, Issue 1, pp128-133, February 2010.

[2] M. Belleville, H. Fanet, P. Fiorini, P. Nicole, M. J. M. Pelgrom, Christian Piguet, R. Hahn, Chris Van Hoof, Ruud J. M. Vullers, Marco Tartagni, Eugenio Cantatore, "Energy autonomous sensor systems: Towards a ubiquitous sensor technology," Microelectronics Journal 41(11): 740-745 (2010)

[3] M. Marzencki, Y. Ammar, S. Basrour, "Integrated power harvesting system including a MEMS generator and a power management circuit," Sensors and Actuators A, 145-146 (2008) 363-370.

[4] M. Belleville, H. Fanet, P. Fiorini, P. Nicole, M.J.M. Pelgrom, C. Piguet, R. Hahn, C. Van Hoof, R. Vullers, M. Tartagni, E. Cantatore, "Energy autonomous sensor systems: state and perspectives of a ubiquitous sensor technology," Proceedings of the third International Workshop on Advances in sensors and Interfaces, IWASI 2009, pp. 134–138, June 2009.

[5] M. Belleville, E. Cantatore, H. Fanet, P. Fiorini, P. Nicole, M.J.M. Pelgrom, C. Piguet, R. Hahn, C. Van Hoof, R. Vullers, M. Tartagni, "Energy autonomous systems: future trends in devices, technology, and systems, Report," CATRENE Working Group on Energy Autonomous Systems, 2008, Table 3, p. 20.

[6] Aggeliki S. Prayati, Christos D. Antonopoulos, Tsenka Stoyanova, Christos Koulamas, George D. Papadopoulos, "A modeling approach on the TelosB WSN platform power consumption," Journal of Systems and Software 83(8): 1355-1363 (2010)

[7] A. Harbd, "Energy harvesting: State-of-the-art," Journal of Renewable Energy, No 36, pp. 2641-2654, 2011.

[8] H. Yu and Q. Yue, "Indoor Light Energy Harvesting System for Energy-aware Wireless Sensor Node," Proceedings of 2012 International Conference on Future Energy, Environment, and Materials, 16, pp. 1027-1032, 2012.

[9] G. Horvat, D. Sostaric, D. Zagar, "Power consumption and RF propagation analysis on ZigBee XBee modules for ATPC," 35th International Conference on Telecommunications and Signal Processing, TSP 2012, Prague, Czech Republic, pp. 222-226, July 3-4, 2012.

Evaluation of the operation of Depletion-mode SiC Power JFET in DC-DC converter applications

Josip Bačmaga[1], Kristijan Bene[1], Branimir Pejčinović[2], Adrijan Barić[1]
[1]University of Zagreb, Faculty of Electrical Engineering and Computing; [2]Portland State University
josip.bacmaga@fer.hr

***Abstract*—This paper presents the system for the evaluation of operation of a depletion-mode silicon carbide (SiC) power junction field-effect transistor (JFET). The main part of the system is a dc-dc step-down converter which represents realistic operating conditions for the switching devices in a synchronous buck configuration. In order to test the importance of the dead-time value on the operation and efficiency of the synchronous buck converter, a precise two-channel time-delay pulse signal generator is developed and its operation is described. The ability to precisely regulate control signal parameters of the high-side FET and low-side FET (switching voltage and current, operating frequency, duty cycle, dead-times, etc.) is needed in order to fully characterise SiC power switches. The functionality of the complete system is verified by measurements performed under various operating conditions.**

***Index Terms*—SiC JFET, buck converter, time-delay signal generation**

I. INTRODUCTION

In switched-mode power supply systems (SMPS), switching devices are submitted to a broad range of switching frequencies as well as voltage and current levels at their terminals. Although silicon (Si) power switching devices have been used in the industry for many years, their switching behaviour is far from being ideal. The two most disadvantageous effects, high voltage drop during on- and high leakage current during off-state, significantly increase power losses, i.e. reduce the highest achievable power conversion efficiency. The most important advantages of SiC power devices over silicon ones include capability of sustaining high voltages and currents, operation at higher frequencies, higher thermal conductivity and, consecutively, lower on-resistance [1].

In order to realize a SiC power switch having the best possible characteristics, the status of the developed SiC power devices is being frequently reviewed [2]–[5]. It has been shown that SiC MOSFET would be the device of choice as soon as the SiO_2/SiC interface and reliability issues are solved. On the other hand, operation of the JFETs is not dependent on the quality of gate dielectrics, but relies primarily on pn-junction performance [2]. In this way, SiC JFET designs are still more interesting for power switching applications, because of the higher quality of the conduction channel, achievable low on-resistance, high-frequency operation and high-voltage endurance.

The normally-on property of a depletion-mode JFET counts as a main drawback when this device is used in a dc-dc converter. However, the enhancement-mode (normally-off) JFETs require a narrow and relatively low doped channel to ensure the normally-off property which disturbs the on-state performance [6]. Moreover, the normally-off SiC JFET requires a more complex gate driver circuit in terms of a significantly larger gate voltage that is required to fully turn-on the device [7], [8]. Therefore, in order to completely exploit the advantages of the SiC transistor technology, the focus of the SMPS design is set on the depletion-mode (normally-on) SiC power JFETs. It is possible to turn a normally-on behaviour of the SiC JFETs into a normally-off by using a „Baliga-pair“ (Si MOSFET/SiC JFET cascode switch) [1]. The behaviour of „Baliga-pair“ is well known and studied in depth [7], [9]–[11]. It was shown that this cascode is suitable for operation of the normally-on SiC JFET, at the expense of increase in total on-resistance. In order to reduce the on-resistance and increase the switching speed, the driving circuit should be capable of providing high voltage and current to the cascode switch.

Section II presents an overview of the developed evaluation system based on a synchronous buck dc-dc converter. Section III presents the developed dual-channel precise time-delay signal generator used to estimate an influence of the driving signal parameters on the operation of the device under test (DUT). The functionality of the

This work was supported by the Agency for Innovation by Science and Technology (Agentschap voor Innovative door Wetenschap en Technologie – IWT, Flanders, Belgium) through project INVENT (Integrated voltage regulator using novel topologies and devices), project number 58686-2012.

complete system is verified by the measurement results given in Section IV. The conclusion is given in Section V.

II. The SiC JFET Evaluation System

In order to characterise the behaviour of the SiC JFET as a power switch, a synchronous buck converter capable of high input voltages and currents and very fast switching speeds is developed. The system is mainly based on the Semisouth SSJHB12R085-1 evaluation board [12]. As a device under test (DUT) which is to be characterised and used to prove the functionality of the evaluation system, the SJDP120R085 SiC JFET in a TO247 package produced by Semisouth Laboratories is chosen. It has a breakdown voltage of 1200 V and an on-state resistance of 0,085 Ω. The pinch-off voltage equals -5 V while the reverse gate breakdown voltage is -15 V [13]. The DUT is placed in a „Baliga-pair“ configuration together with the Infineon's BSC059N04LS power MOSFET. This MOSFET has an on-resistance of 5,9 mΩ and is capable of sustaining a drain current of 73 A and a drain-source voltage of 40 V. A gate-source threshold is 1,5 V and a gate-source breakdown voltage equals ±20 V.

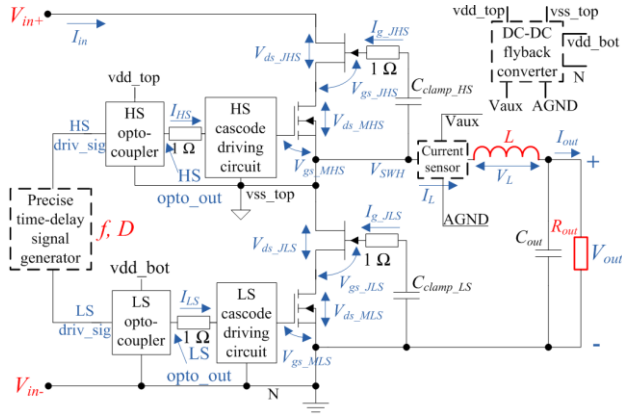


Fig. 1. Simplified circuit diagram of the SiC JFET evaluation system.

The simplified circuit diagram of the developed system is shown in Fig. 1. The central part of the system is the synchronous buck converter with corresponding driving circuit. The synchronous dc-dc converter contains high-side (HS) and low-side (LS) SiC JFET power switch, in the aforementioned cascode configuration, which operation is to be evaluated. The traditional cascode switch circuit exhibits some disadvantages like limited control of the switching speed [14], [15], so the capacitor-clamped cascode is used. The capacitors C_{clamp_HS} and C_{clamp_LS} store the Miller displacement charge during turn-off and then recycle this charge at turn-on to quickly charge the SiC JFET input capacitance for a fast turn-on transient [12], [14]. Both driving signals, generated by the precise time-delay signal generator, are optically coupled to the cascode driving circuit. The externally added inductor L and load resistor R_{out} (Fig. 1), as well as the input voltage V_{in}, switching frequency f and duty cycle D, are chosen according to the voltage and current valuesapplied to the DUT, i.e. the SiC power JFET. The flyback-based dc-dc converter providing three separate outputs of 15 V is used to supply the HS and LS driving circuit as well as the current sensor placed on the evaluation board and used to record the waveform of the inductor current. The measurable voltages and currents are marked in Fig. 1.

III. The Precise Time-Delay Signal Generator

In order to synchronously drive the cascode switches, the drive signal generator is implemented as a dual-channel precise time-delay square wave generator. Apart from the signal frequency, one channel is controlled separately from the other. It is required that the generator be able to produce signal frequencies of up to 10 MHz and duty cycle values up to 50%. The most relevant requirement of the developed generator is the ability to precisely adjust duty cycle and time-delay of the generated signal.

The block diagram of the developed pulse generator is presented in Fig. 2. It can be observed that every channel has its own block for a pulse-width modulation (PWM) and time-delay regulation while they share the signal from the local oscillator.The buffer stages, drawn as the blocks with right-point triangle in Fig 2., are placed in between every stage of the pulse generator in order to prevent the input impedance of the following stage from affecting the preceeding stage.

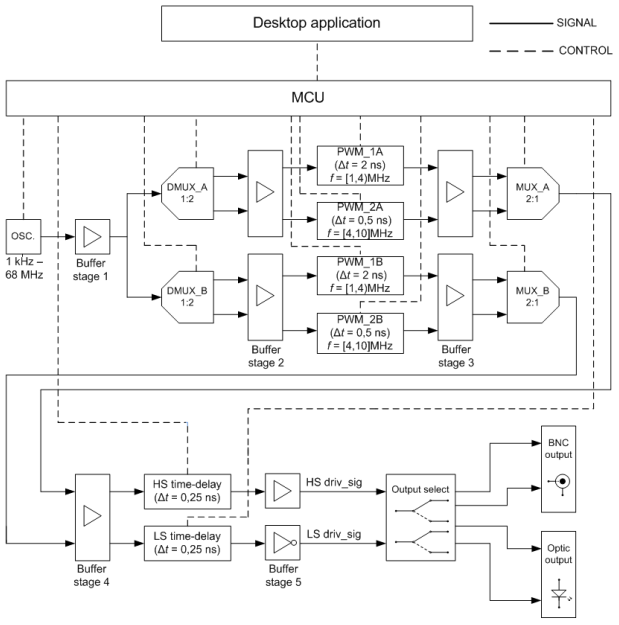


Fig. 2. Block diagram of the signal generator.

Both PWM and time-delay stage are realized using the 8-bit programmable timing element with fixed time-step value. The pulse width of every channel is determined using the time-step of either 2 ns or 0,5 ns. The highest achieveable pulse width for the former one is

$(2^8–1)\cdot 2$ ns = 510 ns while the maximum pulse width for the later one equals 127,5 ns. In a low frequency range (f < 4 MHz), the 2ns-PWM block is used to obtain the demanded duty cycle values while for higher frequencies the 0,5ns-PWM block is active. The frequency „border" of 4 MHz is chosen for the purpose of satisfying the trade-off between achieving the duty cycle values up to 50% and keeping the high pulse-width resolution. This border can easily be changed, either to higher or lower values, in order to obtain higher duty cycle values or to achieve more precise pulse-width adjustment. The time-delay stage uses the same timing element, but with a time-step of 0,25 ns for the whole freqency range of interest.

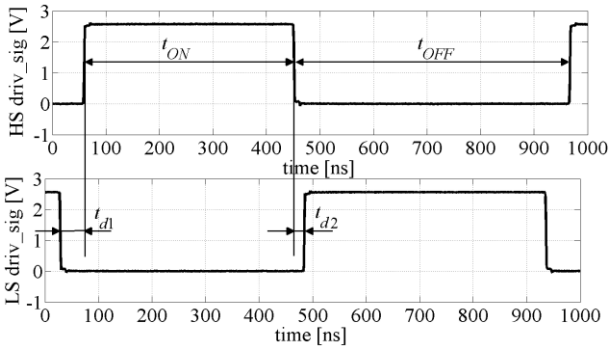


Fig. 3. The HS and LS driving signal for the SiC JFET evaluation board (f = 1,1 MHz; t_{ON} = 380 ns; t_{OFF} = 520 ns; t_{d1} = 27 ns; t_{d2} = 46 ns).

The HS and LS driving signals for the evaluation board are presented in Fig. 3. The waveforms are recorded using the Agilent MSO7034B oscilloscope and probe configuration as follows: 1:1 ratio, DC coupling, bandwidth limit: OFF, input impedance: 50 Ω. Two dead-times, t_{d1} and t_{d2}, are tuned using the time-delay and PWM property of the pulse generator, respectively. The t_{d1} increases with differential delay between HS and LS driving signal, while t_{d2} depends on the pulse width of either HS or LS driving signal.

The performance of the PWM and time-delay stage is presented in Fig. 4. The precision of adjusting the pulse width of the LS driving signal at lower (f < 4 MHz) and higher frequencies (4 MHz ≤ f < 10 MHz) is shown. The property of the pulse generator to adjust the pulse-width with a time-step of 2 ns at lower frequencies and 0,5 ns at higher frequenices, when one-step increment is applied, can also be observed. The bottom waveform presents the ability of the time-delay stage to tune the differential delay between generated signals with a time-step of 0,25 ns. Due to the maximum of 255 time-steps, the highest achievable delay equals 63,75 ns. Both generated pulse signals are brought to the evaluation board via the 50 Ω BNC cables. It is also possible to use the optical transmission to drive the SiC JFET power switches, but this part of the signal generator is not yet tested.

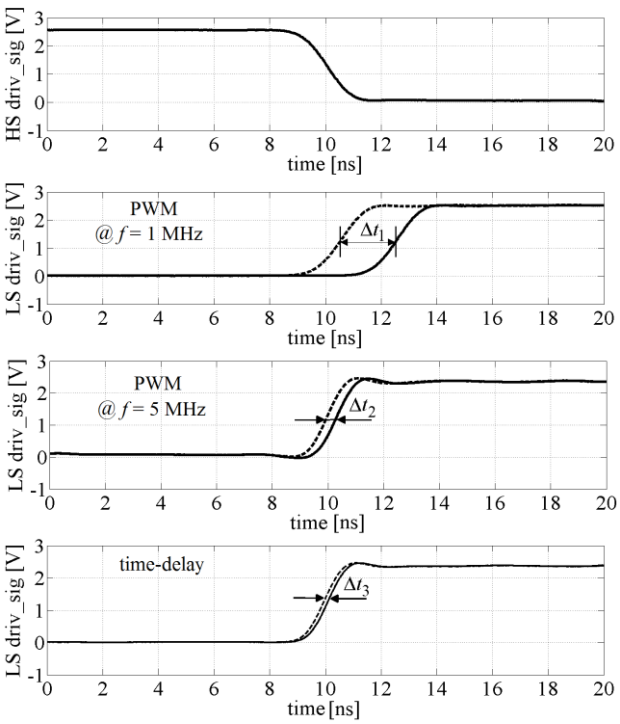


Fig. 4. Measured resolution of the PWM and time-delay stage – before (dashed line) and after (solid line) a one time-step increment is applied (Δt_1 = 2 ns; Δt_2 = 0,5 ns; Δt_3 = 0,25ns).

Since digital PWM and time-delay devices are constrained by their bit-resolution, the precision of adjusting the duty cycle of generated signals directly depends on the signal frequency. Furthermore, the operation of both PWM and time-delay stage is limited by the maximum amount of 255 time-steps.

IV. Measurement Results

A. Transient analysis

In order to prove the functionality of the developed evaluation system, the time-domain waveforms at the relevant test points from Fig. 1 are recorded. The chosen signal and system parameters used to verify the performance of the evaluation system are: f = 200 kHz; t_{d1} = t_{d2} = 180 ns (D = 0,43); V_{out} = 3,3 V; I_{out} = 2 A (R_{out} ≈ 1,67 Ω); L = 6 µH, C_{out} = 10 µF.

The relationship among the generated signals, optocoupler outputs and gate-source voltages of HS and LS driving MOSFETs is presented in Fig. 5. The optocoupler inverts the driving signal and provides the peak-peak output voltage equal to its supply voltage. The cascode driving circuit, proposed in [12], acts as a negative bias network by lowering the dc value of the driving signal in order to increase the turn-off time of the cascode switch. The dead-times t_{d1} and t_{d2} should be as short as possible, yet long enough to avoid the shoot-through effect.

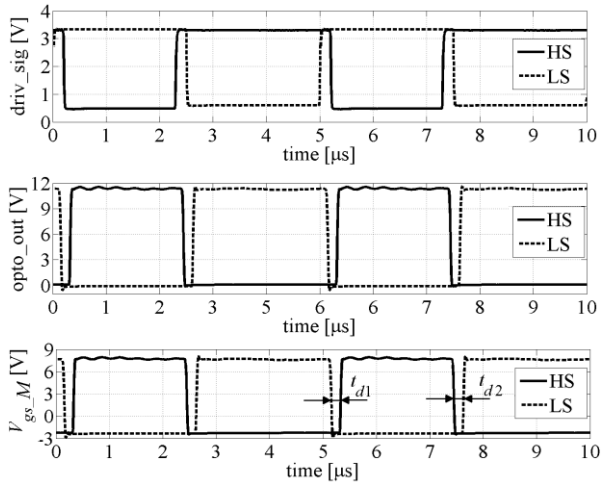


Fig. 5. Operation of the optocoupler and driving circuit (f = 200 kHz; t_{d1} = t_{d2} = 180 ns).

The operation of the low-side cascode switch is verified by the measured voltage and current waveforms shown in Fig. 6. The high-side cascode switch could be analysed in the same way, but the measurement procedure is slightly more complicated due to the difference between the high-side reference potential and ground potential.

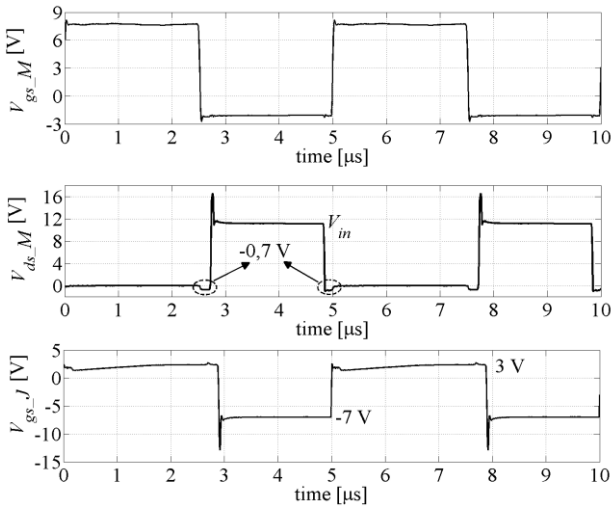


Fig. 6. Operation of the low-side cascode switch.

During the off-state, the drain-source voltage V_{ds_M} of the driving MOSFET equals input voltage V_{in} while, during the conduction mode, the voltage drop is negligible. The voltage drop of approximately -0,7 V, present during both dead-times, is caused by the negative current flowing through the MOSFET body diode and the channel of the JFET due to its normally-on structure. The voltage spike, observed at the beginning of the turn-off process of the LS cascode switch, is caused by the decrease of the reverse cascode current and parasitic inductance between MOSFET and JFET terminals. At the end of the turn-off process, the voltage spike is limited by the junction potential of the body diode.Finally, the gate-source voltage of the normally-on SiC JFET v_{gs_J} is positive (3 V) during the conduction mode in order to reduce the on-resistance. The blocking mode is ensured by the v_{gs_J} of -7 V so the normally-off property is established.

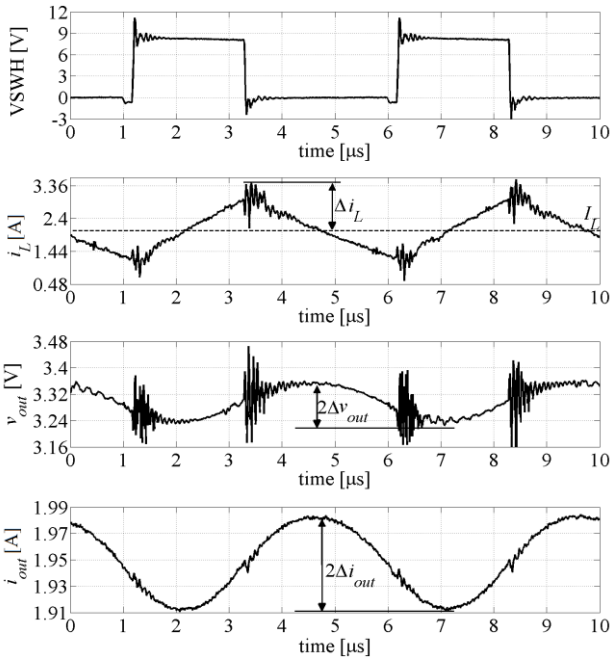


Fig. 7. Operation of the dc-dc evaluation system ($\Delta i_L \approx 1{,}3$ A; $\Delta v_{out} \approx 60$ mV; $\Delta i_{out} \approx 35$ mA).

The operation of the complete evaluation system is presented in Fig. 7. The voltage at the switching node VSWH contains parasitic oscillations. The ringing frequency of approximately 20 MHz is determined by the LC resonance in the current loop formed by the input decoupling capacitor, parasitic loop inductance and both high-side (on-state – R_{on}) and low-side (off-state – C_{oss}) cascode switch [16]. The waveform of the inductor current $i_L(t)$ is obtained form the current transducer LEM LTS 15-NP. The driving signal and system parameters, chosen to verify the operation of the complete system, ensure that the system operates in the conduction mode (CCM) as it is shown in Fig. 7. Since only an ac component of the inductor current I_L flows through the output capacitor C_{out}, i.e. $i_c(t)$=$i_{L_ac}(t)$, the output voltage is given by

$$v_{out}(t) = v_c(t) = \frac{1}{C}\int i_{L_ac}(t)dt. \qquad (1)$$

When the $i_{L_ac}(t)$ is positive, the C_{out} is being charged. Therefore, between the two zero-crossings of the $i_{L_ac}(t)$, the $v_{out}(t)$ changes between its minimum and maximum value. The output voltage $v_{out}(t)$ contains an undesired ripple because of the incomplete attenuation of the switching harmonics by the low-pass output LC filter. This voltage ripple causes the ripple of the output current $i_{out}(t)$ through the load R_{out}. The waveform of the output current $i_{out}(t)$ is recorded using the Agilent N2783B current probe with a transfer ratio of 0,1 V/A.

B. Evaluation of the SiC JFET power switch

The operation of the SiC JFET in a buck converter application is evaluated by measuring the converter efficiency and output voltage ripple under various operating conditions. The values of non-varied parameters are shown in Table I. The focus is set on the evaluation of the SiC JFET in a low-voltage buck converter under five test cases (TC) as described by Table I. Moreover, every test case is additionaly analysed by varying the load resistance R_{out}.

TABLE I.

TEST CASES – SIGNAL AND SYSTEM PARAMETERS.

Set	**f_{SW}, MHz**	**V_{out}, V**	**L, µH**	**t_{d1}, ns**	**t_{d2}, ns**
	0,4 / 1	3,3	6	50	50
Varied	**TC1**	**TC2**	**TC3**	**TC4**	**TC5**
	f_{SW}	V_{out}	L	t_{d1}	t_{d2}

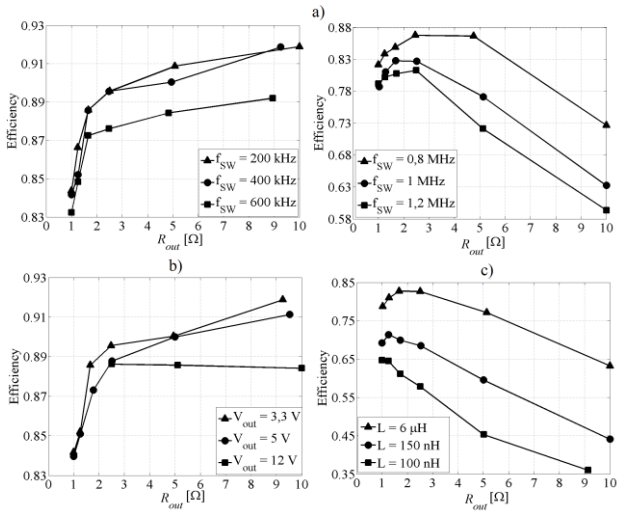


Fig. 8. a) TC1: V_{out} = 3,3 V; b) TC2: f_{SW} = 400 kHz; c) TC3: V_{out} = 3,3 V; f_{SW} = 1 MHz .

The measured power conversion efficiency of a synchronous buck converter as a function of the load resistance and switching frequency as a parameter is shown in Fig 8a. The curves in Fig. 8a (f_{SW} < 0,8 MHz) could be approximated by the function $R_{out}/(a+R_{out})$ where constant a represents the resistance of the inductor wire [15]. For high values of f_{SW}, the efficiency decreases for R_{out} > 2,5 Ω because the converter operates in a discontinous conduction mode (DCM). Due to the fixed t_{d1} and t_{d2}, an increase of the f_{SW} causes the decrease of the duty cycle. This leads to the higher switching losses as well as the lower efficiency. However, the SiC JFET/Si MOSFET cascode switch in low-voltage applications is able to operate with power coversion efficiencies higher than 0,8 even for the output voltages up to 12 V as seen in Fig. 8b. The operation in the DCM is more obvious in Fig. 8c. The f_{SW} of 1 MHz was chosen for a better comparison of the obtained results. For higher R_{out} and lower L, the converter goes „deeper“ into the DCM and the efficiency decreases accordingly.

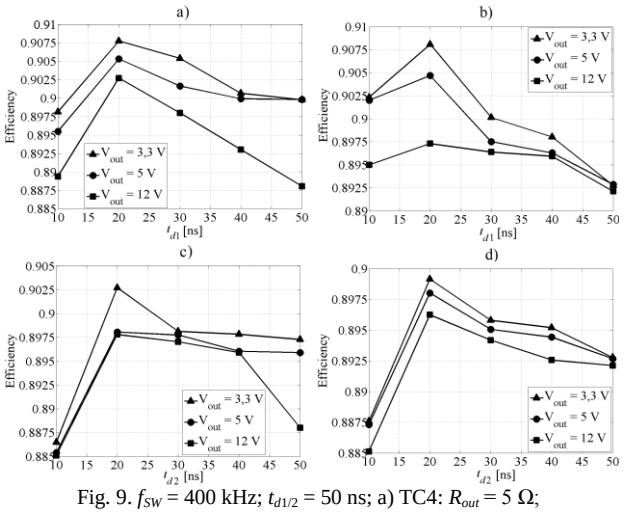


Fig. 9. f_{SW} = 400 kHz; $t_{d1/2}$ = 50 ns; a) TC4: R_{out} = 5 Ω; b) TC4: R_{out} = 2,5 Ω; c) TC5: R_{out} = 5 Ω; d) TC5: R_{out} = 2,5 Ω.

The influence of t_{d1} and t_{d2} on the power conversion efficiency is presented in Fig. 9. Due to the non-zero rise and fall times of the driving signals, short dead-times cause a shoot-through, i.e. significant decrease of the efficiency. As t_{d1} or t_{d2} increases, the power losses caused by the conduction of the low-side body diode increase accordingly. However, the time t_{d1} in TC4 (Fig. 9a and 9b) is more dominant on the efficiency than the time t_{d2} in TC5 (Fig. 10c and 10d). As t_{d1} gets longer, current through the low-side body diode delivers more charge to the parasitic output capacitance of the low-side cascode switch. This causes higher voltage spikes (Fig. 6 and 7), i.e. higher switching losses, due to the higher di/dt at the end of t_{d1}. On the other side, when t_{d2} has ended, the current through the low-side switch does not instantaneously fall to zero, but it switches its path to the Si MOSFET/SiC JFET channel.

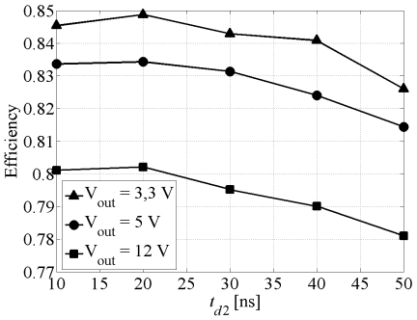


Fig. 10. TC5: f_{SW} = 1 MHz; t_{d1} = 50 ns; R_{out} = 2,5 Ω.

The f_{SW} of 1 MHz causes higher switching losses, i.e. decrease of the efficiency as shown in Fig. 10. The higher f_{SW} causes higher influence of t_{d2} on the efficiency because of the shorter period T=1/f_{SW}. Therefore, the low-side body diode conducts, i.e. causes conduction losses, for a longer part of the period T of the driving signal.

The output voltage ripple Δv_{out} is also an important figure-of-merit of the operation of the SiC JFET power switch in dc-dc converter applications. The Δv_{out} against

the R_{out} and V_{out} is presented in Fig. 11. Due to the parasitics imported by the voltage probe into the converter's output node, the Δv_{out} is calculated by multiplying the output current ripple Δi_{out}, measured by the used current probe, and the load resistance R_{out}.

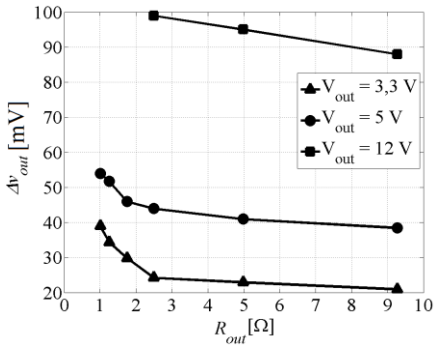


Fig. 11. The calculated output voltage ripple Δv_{out}.

In a computer power supply having a 3,3 V, 5 V or 12 V output, the output voltage ripple is normally required to be less than 1% of the output dc voltage [15]. It can be seen in Fig. 11 that the SiC JFET fulfills this condition for $R_{out} > 1{,}5\ \Omega$.

V. CONCLUSION

The system for evaluation of a normally-on SiC JFET power switch based on a buck converter is presented. The operation of the precise two-channel signal generator is described. The ability of the signal generator to adjust the pulse width and relative time-delay of the generated signals with sub-nanosecond precision is verified by the time-domain waveforms.

The transient analysis proved the proper operation of the Si MOSET/SiC JFET cascode switch as well as the performance of the complete evaluation system.

The presented system is not designed as an automated control system (using the adaptive dead-time property) like common synchronous dc-dc converters. Instead, the system enables the analysis of the influence of many signal and system parameters on the operation of the SiC JFET as a power switch in a cascode configuration.

It is shown that the near-zero dead-time, i.e. the short dead-time values, leads to the significant decrease of the power conversion efficiency because of the non-zero rise and fall times of the driving signals. However, the shorter rise and fall time increases the frequency bandwidth of the switching signals and cause significant electromagnetic emissions. Using the precise time-delay signal generator for evaluation of the operation of power switches, it is possible to estimate the necessary dead-time values for reaching the highest possible efficiency.

It was shown that the Semisouth SJDP120R085 SiC JFET in a synchronous buck converter with low output voltages (V_{out} < 12 V) operates with a power conversion efficiency higher than 0,8 for switching frequencies up to 600 kHz. Besides that, the low output voltage ripple justifies the usage of a SiC JFET in low-voltage dc-dc converter applications.

REFERENCES

[1] B.J. Baliga, „Fundamentals of Power Semiconductor Devices," Springer, pp. 23–43, pp. 98–104, pp. 465, 2008.

[2] M. Östling, R. Ghandi, C.–M. Zetterling, „SiC Power Devices – Present status, applications and future perspective," *Proc. Power Semiconductor Devices and ICs, ISPSD*, pp. 10–15, May 2011.

[3] J. Biela et al., „SiC versusSi—Evaluation of Potentials for Performance Improvement of Inverter and DC–DC Converter Systems by SiC Power Semiconductors,"*IEEE Transaction on Industrial Electronics*, vol. 58, no. 7, pp. 2872–2882, July 2011.

[4] K. Haehre et al., „Characterization and Comparison of Comercially Available SiliconCarbide (SiC) Power Switches," *Proc. Power Electronics, Machines and Drives, PEMD 2012*, pp. 1–6, March 2012.

[5] H. Qin et al., „Overview of SiC Power Devices and its Applications in Power Electronic Converters," *Proc. Industrial Electronics and Applications, ICIEA*,pp. 466–471, June 2013.

[6] P. Biljanović, „Poluvodički elektronički elementi," Školska knjiga, Zagreb, 1996.

[7] R. Pittini, Z. Zhang, M. A. E. Anderson, „Switching Performance Evaluation of Commercial SiC Power Devices (SiC JFET and SiC MOSFET) in Relation to the Gate Driver Complexity,"*Proc. ECCE Asia Downunder, ECCE Asia*, pp. 233–239, June 2013.

[8] R. Siemieniec, G. Nöbauer, D. Domes, „Stability and performance analysis of a SiC-based cascode switch and an alternative solution,"*Microelectronics Reliability*, vol. 52, no. 3, pp. 509–518, Mar. 2012.

[9] A. Rodriguez et al., „Switching Performance Comparison of the SiC JFET and the SiC JFET/Si MOSFET Cascode Configuration," *Proc. Energy Conversion Congress and Exposition, ECCE*, pp. 472–479, Sept. 2013.

[10] R. Pittini, Z. Zhang, M. A. E. Anderson, „SiC JFET Cascode Loss Dependency on the MOSFET Output Capacitance and Performance Comparison with Trench IGBTs", *Proc. Applied Power Electronics Conference and Exposition, APEC*, pp. 1287–1293, Mar. 2013.

[11] D. Aggeler et al., „D*v*/D*t*-Control Methods for the Sic JFET/Si MOSFET Cascode," *IEEE Transactions on Power Electronics*, vol. 28, no. 8, Aug. 2013.

[12] Semisouth, „SSJHB12R085-1 Half-Bridge Demo Bord User's Guide," Application note, Rev. 1.0, April 2012.

[13] Semisouth, „Normally-OFF Trench Silicon Carbide Power JFET – SJDP120R085," Datasheet, Rev. 1.4, Feb. 2011.

[14] K. Haehre et al., „Switching speed-control of an optimized capacitor-clamped normally-on Silicon Carbide JFET cascode," *Proc. Power Electronics and Motion Control Conference, EPE/PEMC*, pp. 11-1–11-5, Sept. 2012.

[15] R. W. Erickson, D. Maksimovic, „Fundamentals of Power Electronics," Springer, 2nd edition, Jan. 2001.

[16] K. Kam et al., „Analysis and Mitigation Techniques for Broadband EMI from Synchronous Buck Converter," *IEEE, Electromagnetic Compatibility Magazine*, vol. 1, no. 3, pp. 83–85, Nov. 2012.

Photocopier PID Controller with Iterative Error Compensation

K. Fukuda*, V Ceperic** and A. Baric**
* Canon INC., Office Imaging Products Device Development Center, Ibaraki, Japan
** University of Zagreb, Faculty of Electrical Engineering and Computing, Zagreb, Croatia

Kazuya.Fukuda@fer.hr

Abstract – This paper describes an implementation of a proportional-integral-derivative (PID) controller with an iterative error compensator for a DC brushless motor to improve the quality of a photocopier. The parameters of the iterative error compensator are determined by using particle swarm optimization algorithm. The proposed controller reduces the effects of repetitious unknown load torque on the DC brushless motor by compensating the motor torque based on the motor velocity from the previous sequences. The proposed controller is compared to a PID controller under various load torques.

Keywords–PID controller, derivative-free optimization, iterative learning

I. INTRODUCTION

Many sources of noise and disturbances exist in a photocopier due to its complicated mechatronics. The main source of disturbance is the load torque that is generated by feeding paper and other components during the printing. The load torque of the brushless DC (BLDC) motor is not predictable, because the load torque depends on the variable working environment of the photocopier such as humidity, temperature and its part tolerances. To achieve the high performance photocopier system under unpredictable disturbance, precise and robust motor controlling is required.

In this paper, we focus on the control of a DC brushless motor that is used in a photocopier under uncertain disturbances. The PID controller is used as a feedback controller to stabilize and to follow the target velocity.

The goal of this paper is to keep the velocity of the BLDC motor as stable as possible while using a simple controller and optimization-based controller training procedure. To minimize the effects from those repetitious disturbances and modeling uncertainty, the idea of an iterative learning feedforward control [1] is used in this paper. Theoretically, a feedforward compensator based controller can eliminate errors due to the disturbances [2], [3]. The iterative error compensator is used in this paper to reduce the effect of disturbances based on the velocity measured in the previous sequences. In the case of a photocopier, the standard feedforward control theory cannot be used because it requires precise modeling of a control structure and that is very difficult to accomplish due to the complexity of the photocopier and the wide tolerances of its parts caused by the mass production. To circumvent this problem, an algorithm based on optimization is used in this paper. Survey of the literature shows that this approach is feasible and efficient, e.g. methods that improve the behavior of the control structure in a repetitious sequence [4] [5], iterative learning control with PID feedback controller for the linear motor system in [6], genetic algorithm based iterative learning controller used to determine a manipulator's input in [7] and [8], iterative learning control based on Kalman filter for a quadrocopter follows a desired trajectory [9] and other.

The main advantages of the controller presented in this paper is its simplicity and the ability to reduce the velocity error without the requirement of the modeling of the overall control structure. The controller presented in this paper is accomplished by optimizing the iterative error compensator parameters by using the particle swarm algorithm [10], [11].

The outline of this paper is as follows: Section II explains working principle of the photocopier including the model of BLDC motor. In Section III, the proposed controller and its learning algorithm are explained. Simulation results are presented in Section IV. Finally, the conclusions are given in Section V.

II. STRUCTURE AND OPERATION PRINCIPLE OF THE PHOTOCOPIER

Fig.1 presents a simplified photocopier (laser printer) schematic that includes laser units, photosensitive drums, an intermediate transfer belt, a paper feeder, a fuser and a paper exhaust tray. The laser unit controls the exposure of the laser beam on the photosensitive drum by ON/OFF control or pulse width modulation (PWM) control, thereby forming an electrostatic latent image on the surface of the photosensitive drum respectively. The drum forms a visible image by developing the electrostatic latent image formed by the laser unit with toner, and superposes and transfers such visible images to the intermediate transfer belt. The intermediate transfer belt transfers the color visible image to the paper. The fuser fixes the color image transferred onto the paper.

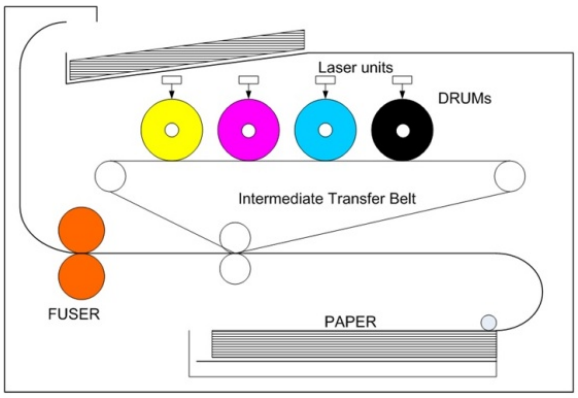


Figure 1. Schematic cross-sectional view illustrating a photocopier.

We focus on the BLDC motor that controls the velocity of the drum in the photocopier. In the following model, a three-phase BLDC motor with trapezoidal back electromotive force (EMF) is used [12] [13]. This BLDC model is based on the assumption that the stator losses are negligible. The electrical equations are:

$$V_a = Ri_a + L\frac{di_a}{dt} E_a \quad (1)$$

$$V_b = Ri_b + L\frac{di_b}{dt} E_b \quad (2)$$

$$V_c = Ri_c + L\frac{di_c}{dt} E_c \quad (3)$$

where V_a, V_b and V_c are the terminal phase voltages [V], i_a, i_b and i_c are the phase currents [A], R is the armature resistance [Ω] of the BLDC, L is the armature self-inductance [H] and E_a, E_b and E_c are the motor back EMFs [V]. In the three-phase BLDC motor, the back EMFs and the rotor angle are described by:

$$E_a = K_w f(\theta_e)\omega \quad (4)$$

$$E_b = K_w f(\theta_e - \frac{2\pi}{3})\omega \quad (5)$$

$$E_c = K_w f(\theta_e + \frac{2\pi}{3})\omega \quad (6)$$

$$\theta_e = \frac{p}{2}\theta_m \quad (7)$$

where f is a function of the rotor position whose shape is a normalized trapezoid, K_w is the back EMF constant [V/(rad.s^{-1})], θ_e is the electrical rotor angle [rad], ω is the rotor speed [rad/s], p is the number of pole pairs and θ_m is the mechanical rotor angle [rad]. The total output torque T_e [Nm], caused by each phase and the mechanical parts is represented by:

$$T_e = \frac{E_a i_a + E_b i_b + E_c i_c}{\omega} \quad (8)$$

$$T_e - T_l = J\frac{d\omega}{dt} + B\omega \quad (9)$$

where T_l is the load torque [Nm], J is the inertia of the rotor [kgm^2] and B is the damping constant [Nms/rad].

III. PID CONTROLLER WITH ITERATIVE ERROR COMPENSATION OPTIMIZED BY PARTICLE SWARM OPTIMIZATION

A. PID controller with the iterative error compensator

The block diagram of a BLDC motor control is shown in Fig. 2 while Fig. 3 shows a block diagram of the proposed PID controller with the compensator. The controller block generates the gate signals based on the rotor velocity and the hall signals. There are five main blocks in Fig. 3. The PID controller is used to follow the target velocity and to stabilize the velocity of the BLDC motor. The PID parameters are tuned by the Ziegler-Nichols (ZN) closed-loop tuning method [14]. Due to its simplicity and semi-empirical nature the ZN method is interesting in those applications where quick tuning realization (in the sense of simple step or limit-cycle test) is preferred to precise parameter adjustment. In this paper, an auxiliary iterative error compensator compensates for the ZN method's imprecision. The error memory and the torque memory store the velocity error *e(t,m-N)* and the output torque $\tau_c(t,m-N)$ of the compensator respectively which are measured in the Nth sequence, where N is the iteration number of the global best value in the particle swarm optimization (PSO) (discussed in Section III.B), t and m represent a time index and an iteration index, respectively. The proportional-derivative (PD) type iterative error compensator block (whose parameters are optimized by PSO) compensates the torque based on the velocity error from the error memory block. These iterative error compensator acts in parallel to the main PID controller enabling a partial feedback that adapts the PD action of the PID controller. The torque controller block generates the gate signals based on the hall signals from the BLDC motor and the output of the PID and the compensator. The input of the torque controller $\tau(t,m)$ is represented by:

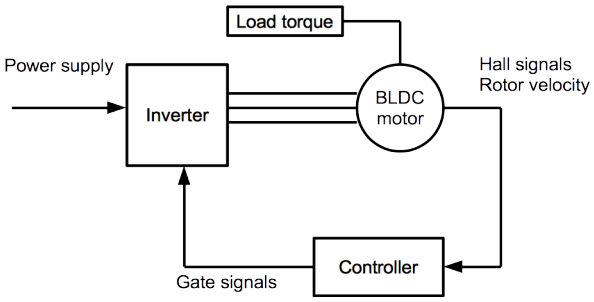


Figure 2. BLDC motor drive.

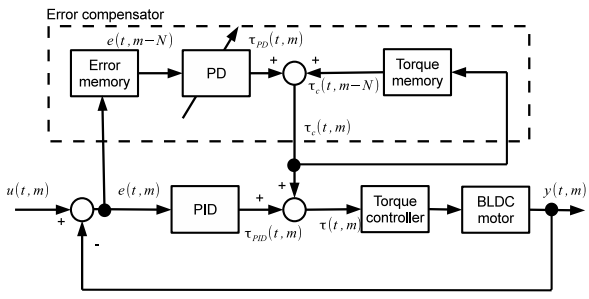


Figure 3. PID controller with proposed iterative error compensation.

$$\tau_c(t,m) = \tau_{PD}(t,m) + \tau_c(t,m-N) \quad (10)$$

$$\tau(t,m) = \tau_{PID}(t,m) + \tau_c(t,m) \quad (11)$$

where $\tau_c(t,m)$ is the output torque of the proposed compensator, $\tau_{PD}(t,m)$ is the output of the PD controller block and $\tau_{PID}(t,m)$ is the output torque of the PID controller block.

B. Particle swarm optimization for the PD-type iterative error compensator

The particle swarm optimization is used to iteratively build the PD-type iterative error compensator. The PSO is one of the evolutionary computation algorithms and is inspired by flocking birds in nature [10]. The calculation cost of the PSO is low because of the simple equations it uses. One of the benefits of using the PSO is the fact that its optimization converges fast [11].

The basic PSO algorithm consists of the velocity and position equations [10], [11] written as:

$$v_i(k+1) = \phi(k)v_i(k) + \alpha_1[\gamma_{1i}(p_i - x_i(k))] + \alpha_2[\gamma_{2i}(G - x_i(k))] \quad (12)$$

$$x_i(k+1) = x_i(k) + v_i(k+1) \quad (13)$$

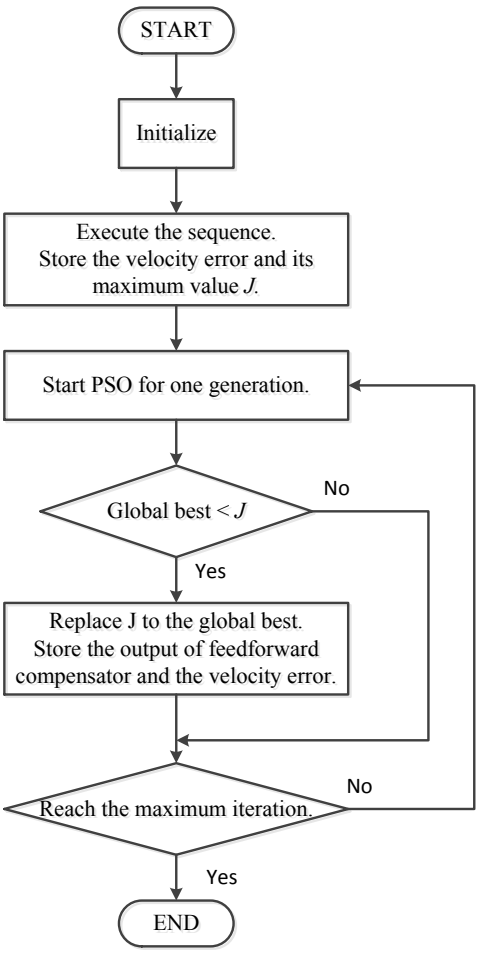


Figure 4. Flow chart of the learning iterative error compensator.

where i is the particle index, k is the discrete time index, v is the velocity of the ith particle, x is the position of the ith particle, p is the personal best position found by the ith particle, G is the global best position found by the swarm, ϕ is the inertia and α is the acceleration constant. In this paper, the PSO algorithm solves the cost function J:

$$J = \max(\|u_k - y_k\|) \quad (14)$$

where k is a natural number and $k>0$. This cost function is derived for the purpose of minimizing the maximum of the velocity error that causes an asynchronous behavior between other drums in the photocopier. Fig. 4 represents a flow chart of the learning procedure for the iterative error compensator used to reduce the maximum velocity error.

In the initialization block, the velocity error and its maximum value J are stored after each iteration. Next, the parameters of the PD controller in the iterative error compensator are set by the PSO and the cost function is evaluated. If the global best value is smaller than J, J is replaced by the global best value. The error memory block stores the velocity error and torque memory block stores the output of the iterative error compensator using the global best PD parameters. After that, the current number of iteration is checked and repeated until the maximum number of allowed iterations.

IV. RESULTS

To test the effectiveness of the proposed PID controller with the iterative error compensator, we

TABLE I. PSO PARAMETERS

Description	Value
Maximum number of iterations	100
Population size	24
Acceleration constant α_1	2
Acceleration constant α_2	2
Initial inertia weight	0.05
Final inertia weight	0.001

TABLE II. BLDC MOTOR SPECIFICATIONS

Description	Value
Nominal voltage	24 V
Stator phase resistance	0.307 Ω
Stator phase inductance	0.188 mH
Inertia	$1.21 \cdot 10^{-10}$ kgm^2
Machanical time contant	13 ms
Friction factor	0 Nms
Pole pairs	7
Torque constant	53.4 mNm/A
Stall torque	3740 mNm

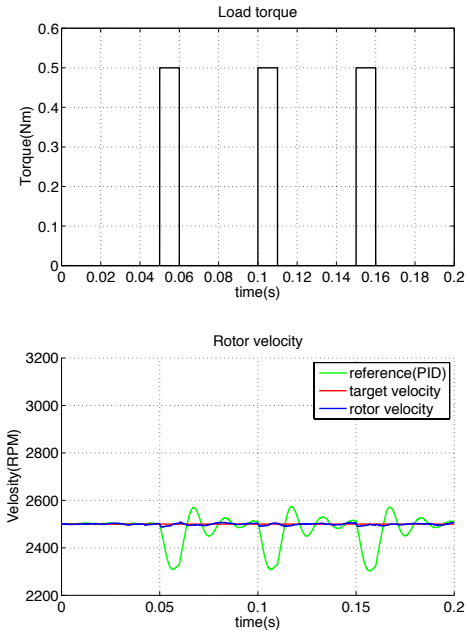


Figure 5. Velocity of the rotor under pulsed load torque.

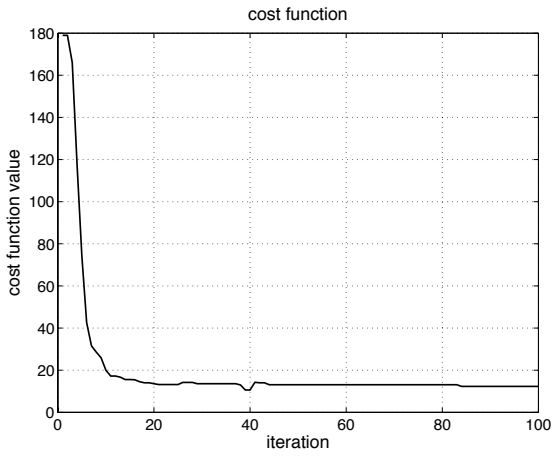


Figure 6. Convergence of the cost function J under pulsed load torque.

generate three types of disturbances: saw tooth, square pulse and uniform random disturbances. The PSO parameters for the optimization of the iterative error compensator are shown in Table I and the motor specifications used in these simulations are shown in Table II. Using these settings, the simulation results of the BLDC motor with three kinds of disturbances are shown.

Fig. 5 and Fig. 6 show the simulation results and the convergence of the cost function *J* during the optimization under the pulsed load torque. This pulsed load torque is presumed to be related to the paper feeding. In Fig. 5, the blue line shows the rotor velocity [rotation per minute (RPM)] that is optimized by the proposed controller, the green line is the rotor velocity [RPM] controlled by the PID controller without the iterative error compensator and the red line represents the target velocity [RPM] (input value to the controller). As it can be seen in Fig. 5, the proposed iterative error compensator reduces the drop of

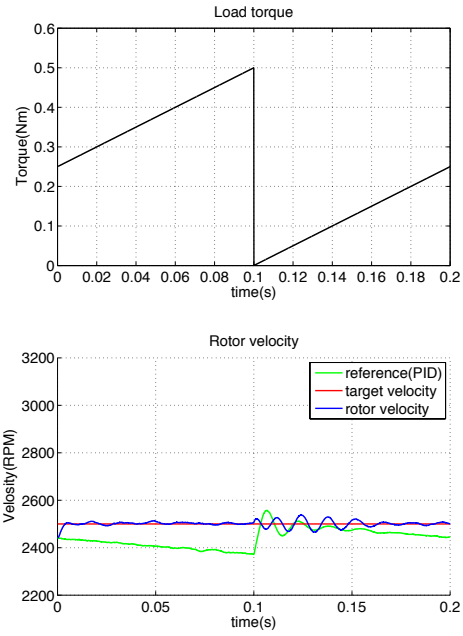


Figure 7. Velocity of the rotor under saw tooth wave load torque.

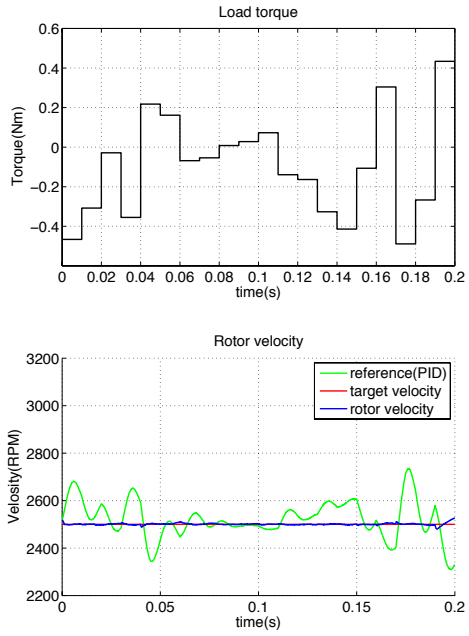


Figure 8. Velocity of the rotor under uniform random load torque (maximum amplitude: 0.5 Nm).

the velocity by 93.8%. The cost function *J* does not change significantly after 20 generations, as shown in Fig. 6. This result indicates that the iterative error compensator is able to reduce the velocity error without modeling of the control structure.

In Fig. 7, the simulation result under the saw tooth load torque is shown and the proposed compensator reduced the velocity error by 54%, where the blue line shows the rotor velocity [RPM] that is optimized by the

proposed controller, the green line is the rotor velocity controlled by the PID controller without the iterative error compensator and the red line represents the target velocity. This saw tooth disturbance is presumed to be caused by the eccentricity in the mechanical construction of the photocopier.

Fig. 8 shows the simulation results under the random signal with the maximum amplitude of 0.5 Nm. Fig. 6 shows that the result is improved 88.5%, where the blue line shows the rotor velocity that is optimized by the proposed controller, the green line is the rotor velocity controlled by the PID controller without the iterative error compensator and the red line represents the target velocity.

V. CONLUSION

In this paper, an implementation of the PID controller with the iterative error compensator for the control of the BLDC motor in a photocopier is presented. The parameters of the compensator are determined by using the particle swarm optimization. It is demonstrated that the modeling of the entire control structure is not necessary. The controller is tested under three types of disturbances commonly found in photocopiers. In each case, the simulation results show that the controller is able to reduce the velocity error of the BLDC motor caused by the unpredictable repetitious disturbances.

REFERENCES

[1] C. Q. Yang and K. L. Moore, "An optimal design of PD-type iterative learning control with monotonic convergence," Intelligent Control, 2002. Proceedings of the 2002 IEEE International Symposium, Vancouver Canada, pp. 55-60, October 2002

[2] R. Garduno-Ramirez and K.T.Lee, "Feedforward compensated multiloop control of a power plant," Power Engineering Society Summer Meeting, Seattle WA, vol. 1, pp. 206–211, July 2000.

[3] K. Ito, W. Maebashi, M. Yamamoto, M Iwasaki and N. Matsui, "Fast and Precise Positioning by Sequential Adaptive Feedforward Compensation for Disturbance," The 11th IEEE International Workshop on Advanced Motion Control, Nagaoka Japan, pp. 19-23, March 2010.

[4] S. Arimoto, S. Kawamura and F. Miyazaki, "Bettering operation of dynamic systems by learning: A new control theory for servomechanism or mechatronics systems," Decision and Control, 1984. The 23rd IEEE Conference, vol. 23, pp. 1064-1069, December 1984.

[5] F. Miyazaki, S. Kawamura, M. Matsumori and S. Arimoto, "Learning control scheme for a class of robot systems with elasticity," Decision and Control, 1986 25th IEEE Conference, vol. 25, pp. 74-79, December 1986.

[6] L. Fan and Y. Liu, "Iterative learning control for linear motor motion system," Automation and Logistics, 2007 IEEE International Conference, Jinan, pp. 2379-2383, August 2007.

[7] V. Hatzikos and D. Owens, "A genetic algorithm based optimisation method for iterative learning control systems," Robot Motion and Control, 2002. RoMoCo '02. Proceedings of the Third International Workshop, pp. 423-428, November 2002.

[8] V. Hatzikos, J. Hatonen, T. Harte and D. Owens, "Robust analysis of a genetic algorithm based optimization method for real-time iterative learning control applications," Emerging Technologies and Factory Automation, 2003. Proceedings. ETFA '03. IEEE Conference, vol. 2, pp. 396-401, September 2003.

[9] A. P. Schoellig, F. L. Mueller and R. D'Andrea, "Optimization-based iterative learning for precise quadrocopter trajectory tracking," Autonomous Robots, vol. 33, issue 1-2, pp. 103-127, August 2012.

[10] J. Kennedy and R. Eberhart, "Particle swarm optimization," Neural Networks, 1995. Proceedings., IEEE International Conference, Perth WA, vol. 4, pp. 1942-1948, November/December 1995.

[11] B. Birge, "PSOt - a particle swarm optimization toolbox for use with Matlab," Swarm Intelligence Symposium, 2003. SIS '03. Proceedings of the 2003 IEEE, pp. 182-186, April 2003

[12] M. Tsai, T.P. Quy, B.Wu and C. Tseng, "Model construction and verification of a BLDC motor using MATLAB/SIMULINK and FPGA control," Industrial Electronics and Applications (ICIEA), 2011 6th IEEE Conference, Beijing, pp. 1797-1802, June 2011.

[13] B. Tibor, V. Fedak and F. Durovsky, "Modeling and simulation of the BLDC motor in MATLAB GUI," Industrial Electronics (ISIE), 2011 IEEE International Symposium, Gdansk, pp. 1403-1407, June 2011.

[14] P. Cominos and N. Munro, "PID controllers: recent tuning methods and design to specification," Control Theory and Applications, IEE Proceedings, vol. 149, issue 1, pp. 46-53, January 2002.

Introduction of some design aspects for improved performance of the DC/AC induction motor converter

J. Podrzaj*, G. Babic*, M. Podhraski*and J. Trontelj**
* Letrika Lab d.o.o., Sempeter pri Gorici, Slovenia
** Uni. of Ljubljana, Fac. of Electrical Eng/Laboratory for Microelectronics, Ljubljana, Slovenia
jure.podrzaj@letrika.com

Abstract - The increasing trend of DC/AC induction motor converters implementation in a wide range of driving applications is becoming a fact that could not be neglected any more. This fact challenges the designers of the DC/AC induction motor converters to constantly seek performance improvements and cost reductions on an everyday basis.

The paper describes some of the design improvements of a DC/AC induction motor converter for use in light commercial electrical vehicles. The implementation of design changes is based on electrical analysis, simulation results and also on the previous prototype testing. The goal of design activity presented in the paper is a design of an improved DC/AC induction motor converter with higher performance, enhanced reliability and optimized production cost. The paper concludes with a proposal of a new DC/AC induction motor converter design and concludes with some simulation results.

I. INTRODUCTION

Nowadays various traditional driving systems (e.g. combustion driving system) are being in majority replaced with different alternative solutions. The predominant and most prevalent alternatives could be found in different variations of driving systems realised with different types of electrical motors. In applications such as light commercial vehicles (golf carts, forklifts…) or other auxiliary systems (electrical assisted steering …) the electrical driving system is commonly realised with an induction motor. The induction motor has many advantages compared to DC motor or to motor with permanent magnet. But still the motor type selection depends on several factors which are defined either by the end application or by customer. Beside the material cost and its construction robustness of an induction motor itself, it requires a more complex DC/AC induction motor converter in order to obtain a sufficient level of the driving system efficiency. Some basic information about different electrical motor types and motor control algorithms could be found in [1, 2].

For efficient driving of an induction motor, a DC/AC induction motor converter is required. A DC/AC induction motor converter is positioned between the electrical energy source/battery and the induction motor. The DC/AC induction motor converter could be splinted in to two main parts where the first part is used for high current density switching and electrical power distribution between the battery and the induction motor. The second part of the DC/AC induction motor converter represents an electrical circuit which executes the driving algorithm and services the user input commands. The driving algorithm used for DC/AC induction motor control in this paper is based on Field-Oriented-Control approach [3].

But we must be not mislead with an impression that changing to the alternative driving systems is necessarily fully justified. The decision for the transfer to electrical driving systems should consider all aspects such as performance, size, weight, mileage distances…

The DC/AC induction motor converter is the main part of such electrical driving system. Therefore this paper focuses on some DC/AC induction motor converter design aspects with the goal of optimizing performance, materials and production costs.

II. DC/AC INDUCTION MOTOR CONVERTER

The DC/AC induction motor converter discussed in this paper was designed for induction motor driving systems with battery voltage of 96V and continuously providing a phase current of $220A_{RMS}$.

The construction of the DC/AC induction motor converter is based on the patented approach [4] of the DC/AC induction motor converter assembly. For converting the battery supply to the proper power signal for the induction motor phase a power module is used. The power module is realized as two serially connected switches where the “high-side” switch is connected between the positive terminal of the battery and the induction motor phase connector while the “low-side” switch is used to apply the negative battery potential to the connected phase terminal. Fig 1. shows a block scheme of the DC/AC induction motor converter realized with three power modules, capacitor board and connection terminals to the battery and induction motor phases.

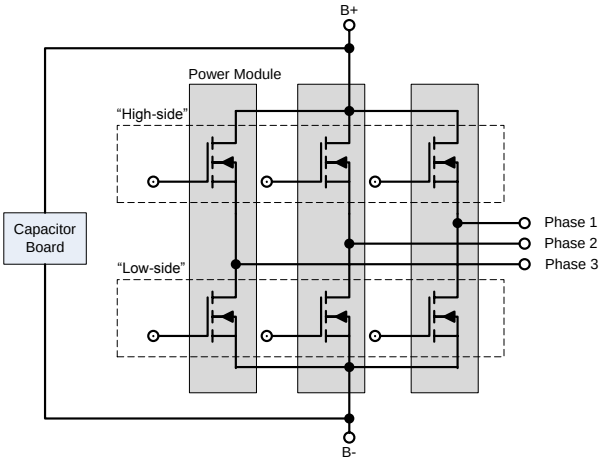


Figure 1. A block scheme of DC/AC induction motor converter – power modules with the capacitor board

Each switch, either "high-side" or "low-side", consists of a number of connected in parallel semiconductor power devices. The number of in parallel connected semiconductor power devices depends on the required switching current densities and voltage ratings. Since the battery supply voltage applied of the discussed DC/AC induction motor converter ranges up to 96V the selection of semiconductor power MOSFET devices is justified.

For more convenient production, handling and economical aspects, the DC/AC induction motor converter switching part is realized modularly as individual power modules. Fig 2. is showing a prototype sample of the power module.

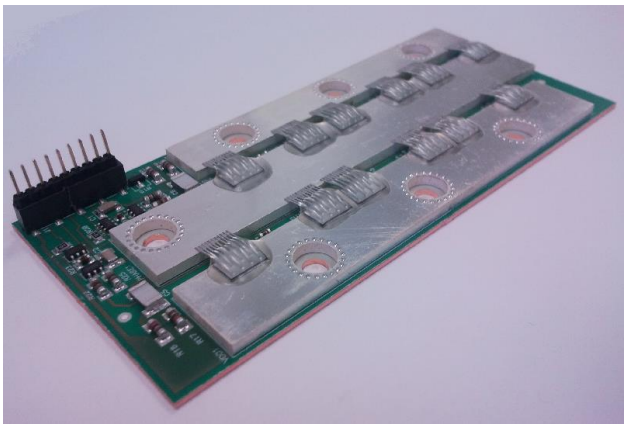


Figure 2. Power module photo

The three power modules are connected to several bus bars with different potentials. These bus bars form some sort of three dimensionally structured interconnection network which practically define the electrical circuit topology. The introduced three dimensionally structured interconnection network provides low resistance and low parasitic electrical paths between power modules and as well minimizes the influences of the parasitic inductance of all other internal electrical paths where high current densities are distributed (induction motor phases and battery terminals) [4].

An important part of the three dimensionally structured interconnection network is also formed with the capacitor board. The capacitor board is used to support the capacitor elements and additionally provides electrical connection of battery terminals B+ and B- to each of the power module. The main feature of the capacitor boards positioned near power switches is to provide adequate electrical charge necessary for the efficient switching of the power MOSFET devices. Switching of high current densities only with the connected battery source, because of the cable connection parasitic and internal battery resistance, would not be functional.

In order to improve some of the prototyped DC/AC induction motor converter performance, additional re-design actions of capacitor board and power MOSFET driver circuit have been performed. In the following paragraphs the focus of each re-design features is presented. It also must be noted that the encountered design improvements are based on the previously developed prototype and available measurement results show in Fig 3.



Figure 3. DC/AC induction motor converter prototype

A. Capacitor board

The main reason for capacitor board performance improvement was to enhance higher capacitor board ability of providing required electrical energy to the connection terminals of the power modules. Besides establishing more efficient electrical connections between capacitor board – power modules – battery terminals also higher total capacitance has been achieved.

The previous designed PCB board was realised with the standard 4-layer FR4 material with copper configuration (70 µm/ 70 µm/ 70 µm/ 70 µm) and consisted of 9 pcs of 1000 µF and 6 pcs of 1.5 µF capacitor elements. The calculated total ripple current of such capacitor block was 20,970 mA [5]. The capacitor board of the prototyped version of the DC/AC induction motor converter is shown in Fig 3.

The improved version of the capacitor board was designed with the same FR4 material configuration but the number of the capacitor elements was increased to 12 pcs of 1000 µF and to 11 pcs of 1.5 µF. The total capacitance of the capacitor board as well as the ripple current capability was therefore increased for approximately 30%.

To the design of the new capacitor board additionally also a driver circuit was added. Fig 4. shows the realized improved capacitor board with the driver circuit on a single PCB.



Figure 4. The PCB capacitor board with driver circuit of the improved prototype

Although the number of capacitors is increased from 9 pcs to 12 pcs the main feature of the improved prototype version could be found in the reduction of the PCBs. Instead of two separated PCBs, capacitor board and driver board, the new improved version only require design and manufacturing of a single PCB board. Since the PCB area of the prototyped version is increased for ≈10% the cost reduction could be identified in lower production costs of the PCB, manufacturing handling / assembly and final assembly of the PCB in to each power module and battery terminals. On the other hand, the increased number of capacitor for 30% increases the DC/AC induction motor converter performance of providing a 10% higher phase current and on the other hand also lower power losses could be expected.

In order to improve voltage distribution of the B+ and B- potential simulations of prototype version and 3D model of the improved power stage were analysed with FEM method in COMSOL® simulation environment were performed.

The distribution of the battery potential on the prototyped version of the power stage is shown in Fig 5.

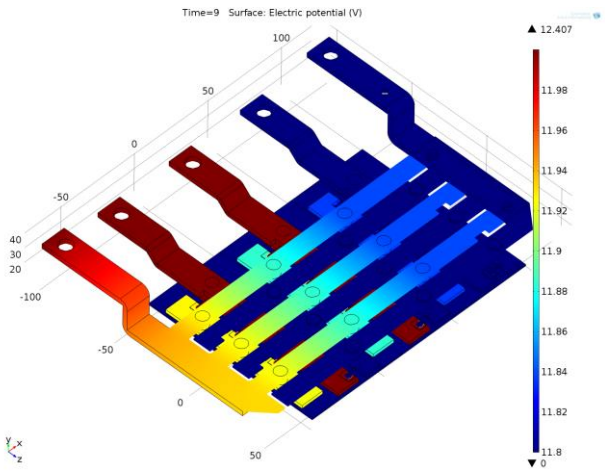


Figure 5. Distribution of the B+ battery potential of the prototyped version

The connection of B+ and B- battery terminals to the capacitor board were realised from the opposite side of positioned power modules. The simulation analysis of the potential distribution indicates that the battery terminals potentials are not equally distributed over the entire surface of the capacitor board. Therefore the B+ and B- potentials were not equally distributed to the terminals of individual power modules. Those potential differences influence the switching and also to quasi-static performance of the power MOSFET devices of the power modules.

The distribution of the battery potential on the improved prototyped version of the power stage is shown in Fig 6.

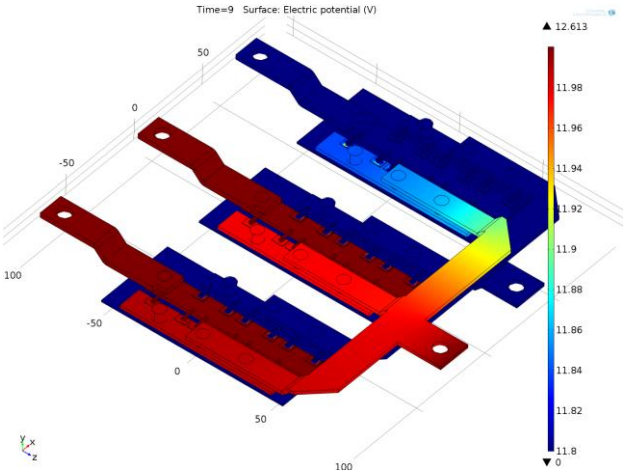


Figure 6. B+ battery voltage distribution of the improved 3D prototype version

The purpose of the improvement was to achieve more uniform distribution of the various voltage potentials to/from capacitor board and consequently also to optimize the resistance and parasitic influence of the connections. This improvement also provides better performance of the built-in capacitor elements. Fig 6. shows the B+ battery potential distribution on the power module terminals of the improved design where voltage distribution on each power MOSFET of the observed switch is more evenly distributed compared to the prototyped version shown in Fig 5.

Beside the differences of the voltage distribution the new improved version also exhibits lower absolute voltage drops from the battery connection terminals to each power module as well as to the capacitor board. The simulated voltage drop differences between both versions are estimated to be close to the factor of 2.

Both simulation results shown indicate the same situation in time domain: two switches (phase V and W) are active (conduction state) while the phase U is not active. The simulation results are indicating only the voltage distribution of battery potential distribution to each individual power MOSFET device of the connected power modules, since the influences of the capacitors placement was not included in full.

B. Driver circuit

The driver circuit of the power switches could be considered as one of the important parts of the DC/AC induction motor converter. The driver circuit represents a link between the low power circuit / converter unit and high power circuit / power modules on the other side. The functionality of the driver circuit is conditioning the logic signals from the converter unit to adequate driving signals for the power switches. In case of driving a three phase induction motor the driver circuit consists of six drivers where three drivers are used for driving the "low-side" switches and three for the "high-side" switches. Driving the "high-side" switches require floating power supply for the drivers while the "high-side" power switches are referenced to phase terminal which floats between B+ an B- battery potential during operation. Some of the gate driver circuits for power MOSFET devices together with some design consideration are presented in [6, 7]. Fig 7. shows the connection of the reference for "high-side" and "low-side" for the driver circuits.

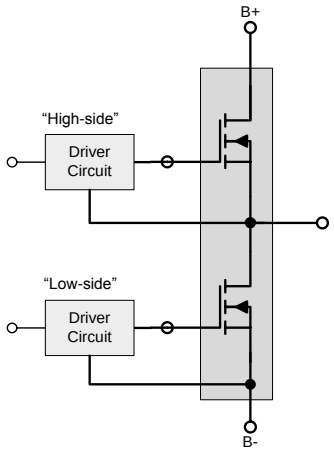


Figure 7. Driver circuit connection for "high-side" and "low-side" driver circuit

In order to achieve higher current densities switching capabilities of each power switch of power module is often realised with multiple in parallel connected power MOSFET devices. Each power MOSFET requires a driving signal with enough current capability for effective switching. The requirements for a driving signal current capability increases with either faster switching performance requirements and/or with adding additional power MOSFET switches. Although the power MOSFET device is voltage controlled the term signal current capability refers to as a function of a gate driver circuit. The signal current capability of the driver circuit directly influences on the switching performance (slew rate) of the gate voltage of the switched MOSFET. Another important issue directly correlated to the properties of controlled power MOSFET devices and as well as with the power module layout, is the ability of independent control of the switching behaviour of the power MOSFET gate terminals.

Fig 8. illustrates the electrical schematic of driving signal switching control with an embedded resistor network and diode. The control of the IC driver output current is defined with the resistance R_{ON} for switching the power MOSFET on. For controlling the IC driver sink current when the power MOSFET is switched off the total resistance of parallel connected R_{ON} and R_{OFF} resistor has to be taken into account.

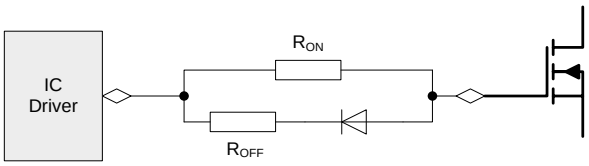


Figure 8. The driver circuit with the resistor network for controlling the switching current to the MOSFET gate terminal

A new design of the gate driver circuits for controlling parallel connected power MOSFET is realized on the same substrate on which the power devices are assembled which minimises the parasitic influences of the connections, since the gate control source/sink terminal is positioned closer to the gate terminals of the power MOSFETs. The main advantage of the proposed circuit is the option to independently control the amount of switching current provided to the MOSFET gate terminal. The proposed circuit show in Fig 9. enables the switching of the control signals of high current densities required for effective switching of the connected gate control terminals. The prototype version driver current capabilities are limited to 1A/2A source/sink current capability since the improved driver circuit is able to provide 51A of continues current in both source/sink operation.

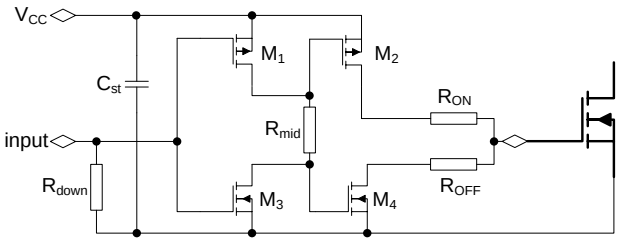


Figure 9. The driver circuit realised with non-overlapping circuit

The circuit schematic in Fig 9. shows the circuit for a single gate control. The same circuit is used for both the "high-side" and "low-side" power switches. The reference for the "high-side" power switch as well as for the driver circuit is connected to the phase potential. Since the phase potential is floating between B+ and B- battery potential an additional floating power supply source for the "high-side" driver circuit must be provided together with the capacitor element C_{st}.

The circuit consists of four MOSFET transistors (M1 – M4) connected in a full-bridge topology with an resistor R_{mid} used for preventing overlapping operation. The capacitor C_{st} is used as a bootstrap capacitor for providing the electrical charge for circuit operation. The capacitance of the bootstrap capacitor depends on the circuit power consumption and operating frequency of the application which is in this case the PWM frequency used for driving power switches within DC/AC induction motor converter. The resistors R_{ON} and R_{OFF} are used for limiting the switching on and off gate current respectively.

III. CONCLUSION

The paper introduces some design aspects for improved performance of the DC/AC induction motor converter. The improved design of the capacitor board and driver circuit performance was based on the measurements of the prototyped converter version and electrical simulations. The improved prototype also introduces implementation of a single PCB, where capacitor board is combined with driver circuits, with increased charge storage capability performance. The paper also highlights some optimization of manufacturing costs of the capacitor and driver PCB. The improved construction parts of the DC/AC induction motor converter promises the increased performance, estimation of approx. 10% high phase current capability, which will be in the future works also confirmed with additional measurements of the DC/AC induction motor converter in the laboratory as well as in the field.

ACKNOWLEDGMENT

The authors would like to thank the experts from Letrika Lab d.o.o. and Letrika d.d. for their support. The authors would also like to thank the staff of Laboratory for Microelectronic at the University of Ljubljana, Faculty of Electrical Engineering, for sharing their knowledge and experiences. The author would also like to express gratitude to the SiEVA d.o.o. which founded the research

REFERENCES

[1] D. Miljavec, P. Jereb, Električni stroji – temeljna znanja, Fakulteta za elektrotehniko, Ljubljana, 2008.

[2] A. E. Fitzgerald, C. Kingsley, S. D. Umans, "Electric machinery," McGraw-Hill, 1990.

[3] B. M. Wilamowski, J. D. Irwin, Power electronics and motor drives, 2nd edition, CRC Press, 2011, ch. 24.

[4] J. Trontelj, J. Podrzaj, "Power switching module," EPO, EP2340560 (B1), 2013.

[5] Product datasheet KMR; Nippon Chemi-Con.

[6] M. H. Rashid, Power electronics handbook, Academic Press, 2001, pp 422-426.

[7] J. Adams,"App. Note AN-1014: Using the current sensing IR212X gate drive ICs," www.irf.com.

Microcontroller implementation of dynamically adaptable control of stepper motor with continuous second derivative of speed curve

Josip Divić, Jurica Đurić and Koča Vrančić
College for Information Technologies (VSITE), Klaićeva 7, 10000 Zagreb, Croatia
josip.divic@vsite.hr, jurica.duric@vsite.hr, koca.vrancic@vsite.hr

Abstract - Stepper motor control is well known and in literature thoroughly studied subject for parameters of position, speed and acceleration.

Stepper motor speed control with dynamically adoptable parameters of target speed and maximal allowed acceleration, requiring continuous change of acceleration (continuous second derivative of speed function) presents challenge to microcontroller implementation on factors of processing power and/or available memory array.

Authors failed to find this kind of algorithm described in available literature, and consider it important for practical solutions of speed control in physical systems susceptible to oscillations.

Paper offers theoretical basics and presents methods for construction of control algorithm.

I. BASICS OF STEPPER MOTOR CONTROL

When considering control of stepper motors, three basic real-time parameters are subject to our influence: (i) windings current, (ii) step direction (CW or CCW) and (iii) time interval between two adjacent step pulses.

Winding current control and its influence is a broad subject of its own, and is not discussed here. Providing that the current is high enough to perform a valid step and constant enough to avoid excessive noise and resonance, it does not influence speed profile behavior in a substantial way.

Step direction parameter derives from angular velocity. It is generally assumed throughout this paper that all steps are in the same direction (regardless which), except in the special case when deceleration profile reaches zero speed, in which moment motor rotation can be halted or continued in another direction maintaining the same acceleration vector, thus changing direction of rotation.

Step interval is the main parameter of stepper motor movement control. Since any stepper motor, by design, travels the same angular distance α for every step performed, stepwise average angular speed is inversely proportional to the step duration, i.e. time interval δt between two adjacent step pulses:

$$\omega = \frac{\alpha}{\delta t}\ [\mathrm{rad/s}] \qquad (1)$$

Angular acceleration is achieved by differing δt for two adjacent step intervals:

$$\omega' = \frac{\delta\omega}{\delta t} = 2\alpha\frac{\delta t_1 - \delta t_2}{\delta t_1 \delta t_2 (\delta t_1 + \delta t_2)}\ [rad/s^2] \qquad (2)$$

Observe that acceleration is influenced by ratio and absolute values of δt_n and δt_{n+1}, what further complicates calculation of δt_{n+1} in attempt to control acceleration. For the constant acceleration ω', which is common request placed upon stepper motor control circuits, exact equation for calculating δt_1 to δt_n can be produced.

Total motor angular distance:

$$\theta(t) = \int_0^t \omega(t)\,dt = \frac{\omega' t^2}{2} = N\alpha\ [\mathrm{rad}] \qquad (3)$$

where N is total number of equidistant steps (integer, ≥0). Exact time for the nth step pulse is therefore:

$$t_n = \sqrt{\frac{2n\alpha}{\omega'}}\ [\mathrm{s}] \qquad (4)$$

Note that pulse at t_0 does not produce angular velocity, but merely energizes windings and brings motor shaft in the nearest discrete ("start") position. δt_n is calculated as time difference between two adjacent discrete moments [1], [2]:

$$\delta t_n = \delta t_1\left(\sqrt{n} - \sqrt{n-1}\right)[\mathrm{s}] \qquad (5)$$

Measurement of time is discrete in digital world, represented by the count of microcontroller timer in this case

$$\delta t = \frac{c}{f} \qquad (6)$$

Depending on the context, time interval can be stated in both notations further on in the paper.

Any acceleration ramp can be calculated exactly from (5), but not in the real time, at least not with computational power provided by standard microcontrollers.

Furthermore, δt_1 value has to be kept and used as an acceleration-vise constant for all subsequent δt_n calculations.

Historically, two approaches have been used to overcome this problem. First one is to use ramp table, where δt step timings are precalculated, stored in microcontroller memory array and accessed successively. Waste of memory space is obvious drawback of this solution. Benefit of this approach is that application of any sort of customized speed ramp profile is possible.

Second approach is based on pivotal work of David Austin [1]. He showed that by using Taylor series approximation, (5) can be simplified without significant loss in precision:

$$\delta t_n = \delta t_{n-1} - \frac{2\,\delta t_{n-1}}{4n+1} \tag{7}$$

Austin's approach relay on knowledge of exact angular position, i.e. counting number of steps *n* from start and awareness of number of steps *m* until stop. This approach can therefore be reliably used to conduct motor shaft rotation from known angular position θ_1 to known angular position θ_2. Both parameters need to be available for ramp calculation before movement starts.

II. RAMP TABLE LIMITATIONS

Number of steps to achieve angular speed ω starting from standstill at constant acceleration ω' can be calculated from:

$$\omega = \sqrt{2\omega'\theta}\ [\mathrm{rad/s}] \tag{8}$$

and (3). When substituted:

$$N = \frac{\omega^2}{2\omega'\alpha} \tag{9}$$

Microcontroller memory usage for ramp table is 2N bytes, since 8-bit resolution of timer counts *c* is not sufficient for smooth operation. Note that total number of steps N is inversely proportional to the acceleration, and proportionally to the second power of maximum angular speed. Standard stepper motors have 200 steps per revolution, resulting in α=0,031416 rad. At acceleration ω'=1 rad/s^2, and usual maximum applicable motor speed ω_M=6.000 rad/s (just under 1.000 revolutions per minute), ramp table memory size for 2N would yield 1,15 GB from (9).

That is absurdly large value, inapplicable for any microcontroller. It can be, however, implemented as virtual ramp table, demonstrated in chapter (VI).

III. CONTINUOUS ROTATION MODEL SPECIFICS

What if actual shaft angle (angular position) is of no relevance for control mechanism, and goal (target) angular position does not exist? What if the only available control parameters are current speed and acceleration, with target speed as dynamically changing input parameter, and with system defined maximum allowed acceleration? Furthermore, discontinuous change of acceleration can be prohibited in order to prevent inertial forces induced oscillations in the controlled mechanism.

Examples of that system are continuous process feeders of any kind, or dynamically balancing vehicles like Segway.

Authors were surprised not to find solutions for this kind of control algorithm in available literature, and have decided to design their own.

IV. ADAPTABLE SYSTEM CONTROL MODEL

Three parameters of rotating stepper motor are monitored in control model, at any given moment: (i) angular speed, represented by δt_n or ω, (ii) acceleration ω'_k and (iii) acceleration change $\omega'' = \frac{\delta\omega'}{\delta t}$. Target speed ω_d is input variable. Maximum allowed acceleration $\pm\omega'_M$ and maximum allowed change of acceleration $\pm\omega''_M$ are model constants.

In the special case when $c_n = c_d$ and acceleration ω'=0, no change of speed is required, so $c_{n+1} = c_n$. In all other cases, when $c_n \neq c_d$, motor speed must be changed until $c_n >= c_d$, without violating constraints of ω'_M and ω''_M.

V. VSITE APPROXIMATION

If continuous angular acceleration is started from standstill, equation relating acceleration and speed is

$$\omega = \omega' t\ [\mathrm{rad/s}] \tag{10}$$

From (4) and (10) yields (11)

$$\omega_n = \omega' \sqrt{\frac{2n\alpha}{\omega'}} = \sqrt{2\omega' n\alpha}\ \ [rad/s] \tag{11}$$

If we approximate that angular speed at the half angle of nth step ω_n is the average speed of that step, we get

$$\sqrt{2\omega'\alpha\left(n - \frac{1}{2}\right)} \cong \frac{\alpha}{\delta t_n} \tag{12}$$

yielding VSITE approximation for step duration

$$\delta t_n = \frac{\alpha}{\sqrt{\omega'\alpha\,(2n-1)}} \tag{13}$$

In discrete time domain it is transformed by (9) in

$$c_n = \frac{\alpha f}{\sqrt{\omega' \alpha (2n-1)}} \quad (14)$$

In order to achieve full resolution of c_n over available binary range in integer arithmetic, factor p has to be introduced

$$c_n = \frac{\alpha f p}{\sqrt{\omega' \alpha p^2 (2n-1)}} \quad (15)$$

Factor p has to fulfill both conditions

$$p \leq \frac{2^{32}-1}{\alpha f} \quad (16)$$

$$p \leq \sqrt{\frac{2^{32}-1}{\omega'_M \alpha (2n_M-1)}} \quad (17)$$

where n_M is

$$n_M = \frac{{\omega_M}^2}{2\omega' \alpha} \quad (18)$$

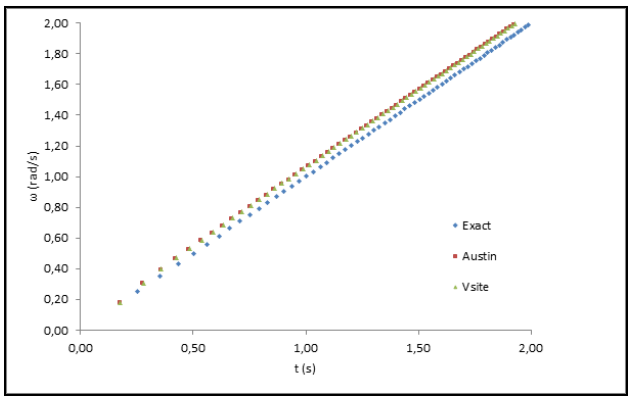


Figure 1. Comparison of constant acceleration speed ramp calculation methods

Authors argue that VSITE approximation is superior to Austin's:

- No further simplification is needed. One square root can be calculated on the fly in the integer arithmetic without exceeding the time budget.
- Each δt_n is calculated "fresh", without cumulative error of previous steps.
- δt_n is related to input parameters α, ω' and n, where n represents virtual ramp table position index. This is crucial, because it offers all the benefits of the ramp table, without consumption of the available memory.
- Formula is valid through whole domain: Compared to Austin in [1], it does not require compensation of the first step, nor special case calculation for the "zero step". Zero step is avoided altogether, movement starts with the first step as it should [3].

VI. VIRTUAL RAMP TABLE BENEFITS AND DESIGN

Ramp table normally consists of precalculated array of timer counts c for all steps from the first step after standstill c_1 until c_m, inversely proportional to ω_1 and ω_M, respectively. Each ramp table has intrinsic acceleration profile, typically constant acceleration ω'_u over the whole range.

Increase in angular acceleration, while using the same constant acceleration ramp table, can be achieved by skipping k steps (integer, $\geqslant 1$):

$${\omega_{n+1}}^2 - {\omega_n}^2 = 2\omega'_u \alpha \quad (19)$$

$${\omega_{n+k}}^2 - {\omega_n}^2 = 2\omega'_u k\alpha = 2k\omega'_u \alpha \quad (20)$$

Skipping k steps effectively reduces angular distance between two discrete motor speeds (ω_n and ω_{n+k}) from kα to α, yielding new inter-step acceleration

$$\omega'_k = k\omega'_u \quad (21)$$

Equation (21) is exact for continuous speed function, and accurate enough for discrete step speeds. Simulation shows that for typical $\alpha=\pi/100$, $\omega'=1$ and $1<k<10$, acceleration error is below 1% and diminishing from third step after stand-still onwards. For first three steps ω'_k is lower than expected, and therefore does not violate ω'_M and ω''_M constraints.

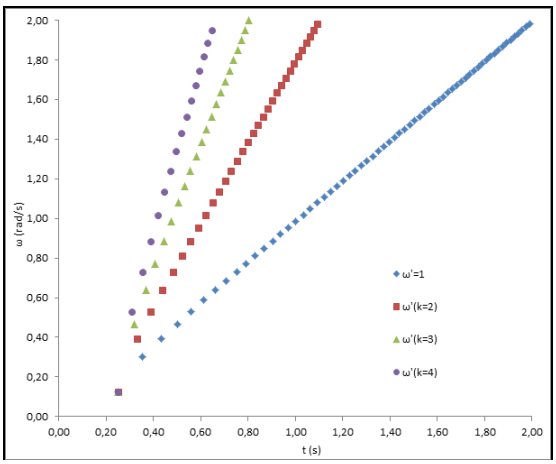


Figure 2. Dependence of acceleration profiles on factor k

Therefore, when designing virtual ramp table, ω'_u should be chosen as low as possible, and afterwards its multiples used for acceleration profiles.

VII. ACCELERATION CHANGE CONTROL

Ramping acceleration up in discrete steps k is shown in (21). Constant increase of acceleration is achieved if *k* is incremented in the constant time intervals until k_M is reached.

$$\omega''(t) = \frac{\delta \omega'_k}{\delta t} = \omega'_u \frac{\delta k}{\delta t} \quad (22)$$

In order to obey constraint of $\pm\omega''_M$, the same process has to be done at the end of speed ramp profile, ω' has to

be gradually lowered, reaching $\omega'=0$ exactly at the moment when $\omega=\omega_d$. It is possible for small changes of ω that acceleration ramp-down has to start before ramp-up is completed (before k_M is reached), so universal formula is needed for determining break point speed ω_B at which acceleration ramp-down must commence in order to land at desired speed ω_d.

We are starting from angular velocity:

$$\omega(t) = \frac{1}{2}\omega'' t^2 + \omega'_0 t + \omega_0 \qquad (23)$$

where ω'_0 and ω_0 are initial acceleration and initial speed respectively. It can be substituted with (21) to give equation for ω_b

$$\omega_b = \omega_d - \frac{1}{2}\omega'' t^2 - k\omega'_u t \qquad (24)$$

In our case, where ω_d has to be reached at the end of interval where $\omega' = \omega'_u$ (k=1), (24) gives break point speed

$$\omega_b = \omega_d - \frac{k(k+1)}{2} \times \frac{{\omega'_u}^2}{\omega''} \qquad (25)$$

Both ω'^2_u and ω'' are systemlevel constants.

If (25) is substituted with

$$\omega = \frac{\alpha f}{c} \qquad (26)$$

c_b becomes

$$c_b = \frac{(2\alpha f\omega'') \times c_d}{(2\alpha f\omega'') - k(k+1)\omega'_u{}^2 c_d} \qquad (27)$$

When condition $c_n \leq c_b$ is reached, acceleration ramp-down has to commence. Term $2\alpha f\omega''$ is separated in parentheses because it should be precalculated, it is systemlevel constant.

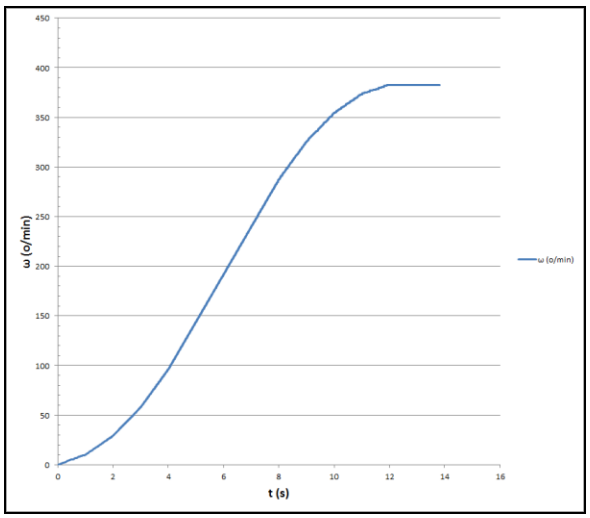


Figure 3. Continuous second derivation speed curve

VIII. CONCLUSION

Authors believe to have demonstrated method for implementation of dynamically adaptable control of stepper motor. VSITE approximation for calculating δt_n and using virtual ramp table for acceleration control allow for speed changes with continuous second derivative.

REFERENCES

[1] David Austin, December 30, 2004, Generate stepper-motor speed profiles in real time.

[2] Atmel. AVR446: linear speed control of stepper motor, revision A

[3] P.J. Siripala, Y. Ahmet Sekercioglu, A generalised solution for generating stepper motor speed profiles in real time, Mechatronics Volume 23, Issue 5, August 2013, Pages 541–547

Optimized Coupled Band-Pass Filters

Neven Mijat* and Dražen Jurišić*
*University of Zagreb/Faculty of Electrical Engineering and Computing
Unska 3, HR-10 000 Zagreb, Croatia
neven.mijat@fer.hr; drazen.jurisic@fer.hr

***Abstract*— In this paper we compare two versions of fourth-order band-pass (BP) filter section realized as coupled structure, having negative feedback around two Biquadratic sections in cascade, usually designated as Biquartic section (i.e. it is the fourth-order band-pass section). The frequency response of a filter is subject to vary from the nominal values due to the effects of aging, changing working conditions, the fabrication tolerances of passive and active elements, etc. In order to maintain the filter's characteristics inside given specifications at the time of manufacture, and as long as possible, the main requirement is to design filters with reduced sensitivity to component tolerances. The Biquartic section has significantly reduced sensitivities to the changes of passive element values, particularly within the pass band, in comparison to common cascade design. For that purpose this structure is very suitable for use as a building block of narrow band, high order BP filters. The sensitivities are further reduced by using three-amplifier biquadratic block instead of more sensitive single-amplifier block. We investigated the sensitivities of two versions of biquartic section with two extreme combinations of second-order sections q-factor values, using Schoeffler's measure, the dynamic range, and output thermal noise using analysis in PSpice. Biquartic sections used throughout the examples are realized as reduced amplifier version retaining the low sensitivities of biquartic design. They also possess simple tuning features and simple design; design equations are given.**

Keywords: Band-pass filters, Low-sensitivity filters, Low-noise filters, Biquartic section, Cascade realization, Schoeffler's sensitivity.

I. INTRODUCTION

Often in practice band–pass (BP) filters of high-order and narrow pass-band are difficult to be realized. They possess very high pole-Q factors, and therefore have very high sensitivity to component tolerances. This problem is becoming more pronounced if such band-pass filters are realized in integrated-circuit form, because on-chip component tolerances are much larger than those in discrete form. One of the efficient solutions of this problem is reducing the sensitivity using negative feedback: a coupled structure [1], [2].

In this paper we present the reduced-amplifier design of the coupled fourth-order BP filter with *low sensitivity* which is referred to as Biquart. We analyse the sensitivity of the obtained filter using Schoeffler's measure, on various examples and different designs. In the design we also wish to have the possibility of easy tuning the center frequency and the pass-band width without deteriorating the amplitude-frequency characteristics. It is also desirable that the new structure provides lower signal distortion and output thermal noise, as well.

II. HIGH-ORDER BP FILTERS DESIGN

To design high-order ($2N^{th}$–order) geometrically symmetrical BP filters we start from N^{th}-order low-pass (LP) prototype filter with the transfer function:

$$T_{LP}(S)=\left[\frac{k_{LP0}\cdot\gamma}{S+\gamma}\right]^{\eta}\prod_{i=1}^{(N-\eta)/2}\frac{k_{LPi}\cdot\omega_{LPi}^2}{S^2+\left(\omega_{LPi}/Q_{LPi}\right)\cdot S+\omega_{LPi}^2}, \quad (1)$$

where η=1 if the filter order N is odd, η=0 for N even, γ is the frequency of the negative real pole for N odd, ω_{LPi} and q_{LPi} are parameters of the $(N-\eta)/2$ complex-conjugate pole pairs, and the k_{LPi} are pass-band gains. If (1) is realized in the cascade form we take care of the sequence of realizing the individual Q-factors such that $Q_{LPi}<Q_{LPi+1}<Q_{LPi+2}<\dots$. Using common LP–BP transformation:

$$S=\frac{s^2+\omega_0^2}{Bs}, \quad (2)$$

where ω_0 is the desired center frequency, and B is the desired pass-band width of the band-pass filter, we realize the BP transfer function of the general form given by:

$$T_{CBQ}(s)=\left[T_{b0}(s)\right]^{\eta}\prod_{i=1}^{(N-\eta)/2}T_{BQi}(s),\ \eta=\begin{bmatrix}1 & \text{for } N \text{ odd}\\ 0 & \text{for } N \text{ even}\end{bmatrix}, \quad (3)$$

where we realize $(N-\eta)/2$ BP fourth-order sections in cascade, and one second-order section if the LP prototype filter was odd. The transfer function of second-order section has the form given by (4) (without index i).

Each $T_{BQi}(s)$ in (3) is fourth order transfer function and we will continue to call it Biquartic transfer function, while the network which implements such a function is called Biquartic section or "Biquart" (hereafter BQ). $T_{b0}(s)$ is second order transfer function and is referred to as Biquadratic section or "Biquad". The BP filter of higher order having the transfer function defined by (3) is referred to as "Cascade of Biquarts" (hereafter CBQ).

III. BIQUARTIC SECTION

Biquartic section, which realizes each fourth-order BP transfer function $T_{BQi}(s)$ in (3) is the main subject of this paper. It will be realized by cascading two BP Biquadratic sections (of second-order) applying a negative feedback, as shown in Fig. 1 (see [1], [2]).

Gain β represents the feedback coefficient (β>0), A is an input gain, and $T_{b1}(s)$, $T_{b2}(s)$, are the second-order BP

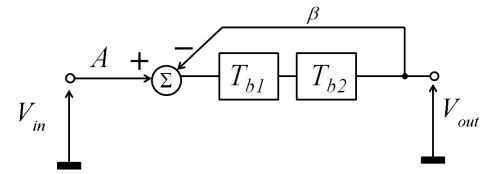


Figure 1. Biquartic BP section.

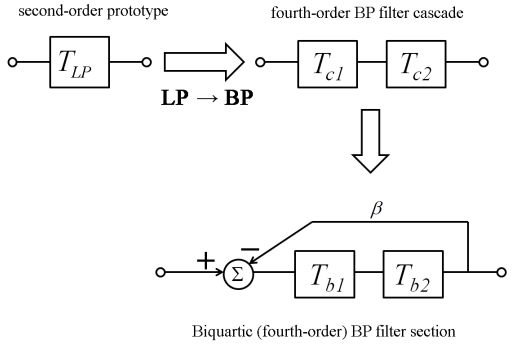


Figure 2. Complete procedure of the calculation of the fourth-order BP Biquartic section starting from second-order LP prototype via cascade.

Biquadratic sections with voltage transfer functions given by:

$$T_{bi}(s)=\frac{k_{bi}s}{s^2+(\omega_{bi}/Q_{bi})\cdot s+\omega_{bi}^2};\ i=1,2. \tag{4}$$

Voltage transfer function of the Biquartic section in Fig. 1 is given by:

$$T_{BQ}(s)=\frac{V_{out}(s)}{V_{in}(s)}=\frac{AT_{b1}(s)T_{b2}(s)}{1+\beta T_{b1}(s)T_{b2}(s)}=\frac{N_{BQ}(s)}{D_{BQ}(s)}. \tag{5}$$

The complete procedure in designing of one BP fourth-order BQ section in CBQ, from the second-order LP prototype filter over the cascaded fourth-order BP filter is shown in Fig. 2. The realization starting from the 1st-order LP producing the 2nd-order BP filters if the filter prototype order N was odd is well-known.

A. Determination of the coefficients of the cascade

In the first step in Fig. 2 it is shown the common LP–BP transformation applied on the second-order LP prototype filter and the obtained fourth-order BP filter in the cascade structure. The transfer function of a second-order LP prototype is given by (1) for N=2. Combining (2) and (1) we obtain $T_{BP}(s)=N(s)/D(s)$ in the form:

$$T_{BP}(s)=\frac{k_{LP}\cdot B^2\omega_{LP}^2\cdot s^2}{s^4+B\omega_0^2\frac{\omega_{LP}}{Q_{LP}}s^3+\left(2\omega_0^2+B^2\omega_{LP}^2\right)s^2+B\frac{\omega_{LP}}{Q_{LP}}s+\omega_0^4}. \tag{6}$$

Transfer function of the fourth-order BP cascade in Fig. 2 has the form:

$$T_{CAS}(s)=\prod_{i=1}^{2}T_{ci}(s), \tag{7}$$

where $T_{c1}(s)$ and $T_{c2}(s)$ are the second-order Biquadratic sections with BP voltage transfer functions given by:

$$T_{ci}(s)=\frac{k_{ci}s}{s^2+(\omega_{ci}/Q_{ci})\cdot s+\omega_{ci}^2};\ i=1,2. \tag{8}$$

We substitute (8) into (7) and rewrite it in the form:

$$T_{CAS}(s)=\frac{k_{c1}k_{c2}\cdot}{s^4+\left(\frac{\omega_{c1}}{Q_{c1}}+\frac{\omega_{c2}}{Q_{c2}}\right)s^3+\cdot s^2}\cdots$$
$$\cdots\frac{}{+\left(\omega_{c1}^2+\omega_{c2}^2+\frac{\omega_{c1}\omega_{c2}}{Q_{c1}Q_{c2}}\right)s^2+\left(\frac{\omega_{c1}^2\omega_{c2}}{Q_{c1}}+\frac{\omega_{c1}\omega_{c2}^2}{Q_{c2}}\right)s+\omega_{c1}^2\omega_{c2}^2}. \tag{9}$$

If we equate the coefficients multiplying the potentions of complex variable "s" in the denominator of (9) and (6) we obtain the system of equations:

$$\frac{\omega_{c1}}{Q_{c1}}+\frac{\omega_{c2}}{Q_{c2}}=B\frac{\omega_{LP}}{Q_{LP}} \tag{10}$$

$$\omega_{c1}^2+\omega_{c2}^2+\frac{\omega_{c1}\omega_{c2}}{Q_{c1}Q_{c2}}=2\omega_0^2+B^2\omega_{LP}^2 \tag{11}$$

$$\frac{\omega_{c1}^2\omega_{c2}}{Q_{c2}}+\frac{\omega_{c1}\omega_{c2}^2}{Q_{c1}}=B\frac{\omega_{LP}}{Q_{LP}}\omega_0^2 \tag{12}$$

$$\omega_{c1}^2\omega_{c2}^2=\omega_0^4 \tag{13}$$

Combining (10) and (12) we obtain the condition for determining solutions:

$$(\omega_{c2}-\omega_{c1})(Q_{c2}-Q_{c1})=0. \tag{14}$$

Two solutions are possible depending on the poles of the LP prototype:

1. $\omega_{c1}=\omega_{c2}=\omega_0$ for real poles of the LP prototype.
2. $Q_{c1}=Q_{c2}=Q_c$ for conjugate-complex poles of the LP prototype.

In what follows we choose conjugate-complex poles of the LP prototype (case 2) and the system of equations (10)–(13) is simplifying into the system:

$$\frac{\omega_{c1}+\omega_{c2}}{Q_c}=B\frac{\omega_{LP}}{Q_{LP}} \tag{15}$$

$$\omega_{c1}^2+\omega_{c2}^2+\frac{\omega_0^2}{Q_c^2}=2\omega_0^2+B^2\omega_{LP}^2 \tag{16}$$

$$\omega_{c1}\omega_{c2}=\omega_0^2 \tag{17}$$

Rearranging and grouping equations of the latest system, we arrive at expressions for Q_c, ω_{c1} and ω_{c2} of the cascade realization (Geffe equations) [3]:

$$Q_c=\sqrt{\frac{Q_{LP}\omega_0}{B\omega_{LP}}}\times\sqrt{Q_{LP}X+\sqrt{Q_{LP}^2X^2-1}};\quad X=2\omega_0/(B\omega_{LP})+B\omega_{LP}/(2\omega_0), \tag{18}$$

$$\frac{\omega_{c2}}{\omega_0}=\frac{\omega_0}{\omega_{c1}}=\frac{B\omega_{LP}}{2Q_{LP}}Q_c+\sqrt{\left(\frac{B\omega_{LP}}{2Q_{LP}}Q_c\right)^2-1}. \tag{19}$$

If we equate numerator in (9) and (6) we obtain:

$$k_{c1}\cdot k_{c2}=k_{LP}\cdot B^2\omega_{LP}^2. \tag{20}$$

B. Voltage transfer function of the Biquartic section

Consider the transfer function of the Biquartic section in Fig.1 given by (5). Since the positions of the poles of $T_{BQ}(s)$ is determined by the product $\beta k_{b1}k_{b2}$, i.e. not only by β, we will use the designation F ($F=\beta k_{b1}k_{b2}$) in order to simplify the expression. We rewrite (5) in the form:

$$T_{BQ}(s)=\frac{Ak_{b1}k_{b2}\cdot}{s^4+\left(\frac{\omega_{b1}}{Q_{b1}}+\frac{\omega_{b2}}{Q_{b2}}\right)s^3+\cdot s^2}\cdots \\ \frac{}{+\left(\omega_{b1}^2+\omega_{b2}^2+\frac{\omega_{b1}\omega_{b2}}{Q_{b1}Q_{b2}}+F\right)s^2+\left(\frac{\omega_{b1}^2}{Q_{b1}}+\frac{\omega_{b2}^2}{Q_{b2}}\right)s+\omega_{b1}^2\omega_{b2}^2} \quad (21)$$

Poles of the transfer function are determined by the denominator of the Biquartic transfer function:

$$D_{BQ}(s)=D(s)+Fs^2=D(s)+\beta k_{b1}k_{b2}s^2. \quad (22)$$

It is evident that for F=0 (β=0, k_{b1}, $k_{b2}\neq 0$) the BQ consists of the two Biquads connected in the cascade, and $D(s)$ represents the denominator of (6), i.e. of the cascade.

Consider the commonly used cascade realization and compare it with the Biquartic section (both of fourth order) shown in Fig. 2. If we equate the coefficients multiplying the potentions of complex variable "s" in (21) to (9), we obtain the system of equations:

$$\frac{\omega_{c1}}{Q_{c1}}+\frac{\omega_{c2}}{Q_{c2}}=\frac{\omega_{b1}}{Q_{b1}}+\frac{\omega_{b2}}{Q_{b2}} \quad (23)$$

$$\omega_{c1}^2+\omega_{c2}^2+\frac{\omega_{c1}\omega_{c2}}{Q_{c1}Q_{c2}}=\omega_{b1}^2+\omega_{b2}^2+\frac{\omega_{b1}\omega_{b2}}{Q_{b1}Q_{b2}}+F \quad (24)$$

$$\frac{\omega_{c1}^2\omega_{c2}}{Q_{c2}}+\frac{\omega_{c1}\omega_{c2}^2}{Q_{c1}}=\frac{\omega_{b1}^2\omega_{b2}}{Q_{b2}}+\frac{\omega_{b1}\omega_{b2}^2}{Q_{b1}} \quad (25)$$

$$\omega_{c1}^2\omega_{c2}^2=\omega_{b1}^2\omega_{b2}^2 \quad (26)$$

Because it is given above [see (18) and (19)]:

$$\omega_{c1}\omega_{c2}=\omega_0^2,\quad Q_{c1}=Q_{c2}=Q_c, \quad (27)$$

we have:

$$\omega_{b1}\omega_{b2}=\omega_0^2, \quad (28)$$

and we choose

$$\omega_{b1}=\omega_{b2}=\omega_0. \quad (29)$$

Now the system of equations (23)–(26) is simplifying into the system (note that (23) and (25) have become identical):

$$\frac{\omega_{c1}+\omega_{c2}}{Q_{c1}}=\omega_0\left(\frac{1}{Q_{b1}}+\frac{1}{Q_{b2}}\right), \quad (30)$$

$$\omega_{c1}^2+\omega_{c2}^2+\frac{\omega_0^2}{Q_c^2}=2\omega_0^2+\frac{\omega_0^2}{Q_{b1}Q_{b2}}+F. \quad (31)$$

Rearranging it we obtain the condition for determining solutions:

$$(\omega_{b2}-\omega_{b1})(Q_{b2}-Q_{b1})=0. \quad (32)$$

We have two equations (30) and (31) and three unknowns to be determined Q_{b1}, Q_{b2} and F. One of the unknowns can be freely chosen (therefore we have one degree of freedom). We choose pole Q factors of the sections T_{b1}, T_{b2} to have two limiting values of Q-factors:

1. Sections T_{b1}, and T_{b2} have identical Q-factors.
2. One of the sections T_{b1} or T_{b2} has an infinite Q-factor.

Between these two limiting cases for Q-factors there are infinite possibilities, which turn to be more sensitive, and we do not treat them separately.

C. Equal Q-factors Q_{b1}=Q_{b2} [4]

The sections T_{b1} and T_{b2} are becoming identical, and we have:

$$Q_{b1}=Q_{b2}=\frac{2\sqrt{\omega_{c1}\omega_{c2}}}{\omega_{c1}+\omega_{c2}}Q_c. \quad (33)$$

$$\omega_{b1}=\omega_{b2}=\omega_0=\sqrt{\omega_{c1}\omega_{c2}}. \quad (34)$$

Feedback factor F takes the value defined by an expression:

$$F=(\omega_{c2}-\omega_{c1})^2\left(1-\frac{1}{4Q_c}\right). \quad (35)$$

D. One Q-factor is in infinity: Q_{b2}=∞

One Q-factor is becoming infinite, while other has the value:

$$Q_{b1}=\frac{\sqrt{\omega_{c1}\omega_{c2}}}{\omega_{c1}+\omega_{c2}}Q_c. \quad (36)$$

Pole frequencies are the same as in the previous case:

$$\omega_{b1}=\omega_{b2}=\omega_0=\sqrt{\omega_{c1}\omega_{c2}}. \quad (37)$$

Feedback factor F is then given by:

$$F=(\omega_{c2}-\omega_{c1})^2+\frac{\omega_{c1}\omega_{c2}}{Q_c^2}. \quad (38)$$

IV. EXAMPLES

In this section we realize three examples of the active-RC BP filters: one for the cascade (CAS) case in the Section III.A, and two for the Biquartic (BQ) cases in the Sections III.C and III.D. Consider two approximations: *i*) Butterworth (with normalized LP prototype parameters: ω_{LP}=1, Q_{LP}=0.707107), and *ii*) 0.5dB-ripple Chebyshev (ω_{LP}=1.231342, Q_{LP}=0.863721 [a higher pole-Q case]) that are transformed into BP characteristics with normalized center frequency ω_0=1 and bandwidth B=0.1. Filters are then denormalized to the frequency 10kHz. The active realization is shown in Fig. 3 and the element values are in Table I. Corresponding amplitude-frequency characteristics are shown in Fig. 4.

Generally, the transfer function of each Biquad shown in Fig. 3 has the form given by (4), with:

$$\omega_{bi}=\sqrt{\frac{R_{Fi}}{C_{i1}C_{i2}R_{i1}R_{i3}R_{bi}}};\ Q_{bi}=C_{i1}R_{i2}\sqrt{\frac{R_{Fi}}{C_{i1}C_{i2}R_{i1}R_{i3}R_{bi}}};\quad k_{bi}=\frac{R_{Fi}}{C_{i1}R_{0i}R_{i1}};\quad i=1,2. \tag{39}$$

Note that in this realization we do not use a separate operational amplifier for negative feedback. Instead, we lead the output signal to the positive input of the first operational amplifier in the first biquad. As a consequence the feedback factor F is determined by:

$$F=\beta\cdot k_{b1}\cdot k_{b2}=\beta\cdot\left(1+\frac{R_{F1}}{R_{01}}+\frac{R_{F1}}{R_{b1}}\right)\cdot\frac{1}{R_{11}C_{11}}\cdot k_{b2}, \tag{40}$$

where

$$\beta=R_{F3}/(R_{F3}+R_{03}). \tag{41}$$

It is evident from (39) that for $R_{i2}=\infty$ the pole-Q factor is infinite (e.g., for $R_{22}=\infty$ in Fig. 3 there is $Q_{b2}=\infty$).

We choose value of R_{F3} and calculate R_{03} depending on F:

$$R_{03}=\frac{R_{F3}}{F}\left[\left(1+\frac{R_{F1}}{R_{01}}+\frac{R_{F1}}{R_{b1}}\right)\cdot\frac{1}{R_{11}C_{11}}\cdot k_{b2}-F\right]. \tag{42}$$

It is evident from (40) and (41) that for $R_{03}=\infty$ and $R_{F3}=0$ the feedback factor $F=0$ and the circuit in Fig. 3 becomes cascade of Biquads.

Using (39)−(42) we calculate normalized elements of filter examples and denormalize the filter to the center frequency f_0=10kHz and bandwidth B=1kHz. We choose denormalization resistance R_0=1.591kΩ in order to obtain denormalization capacitance C_0=10nF. The obtained values are given in Table I. In order to provide maximum dynamic range it will be needed to perform gain optimization in the next step, which is in this case not possible in the simple non-iterative way. It is needed to perform numerical optimization or the optimization in an iterative way.

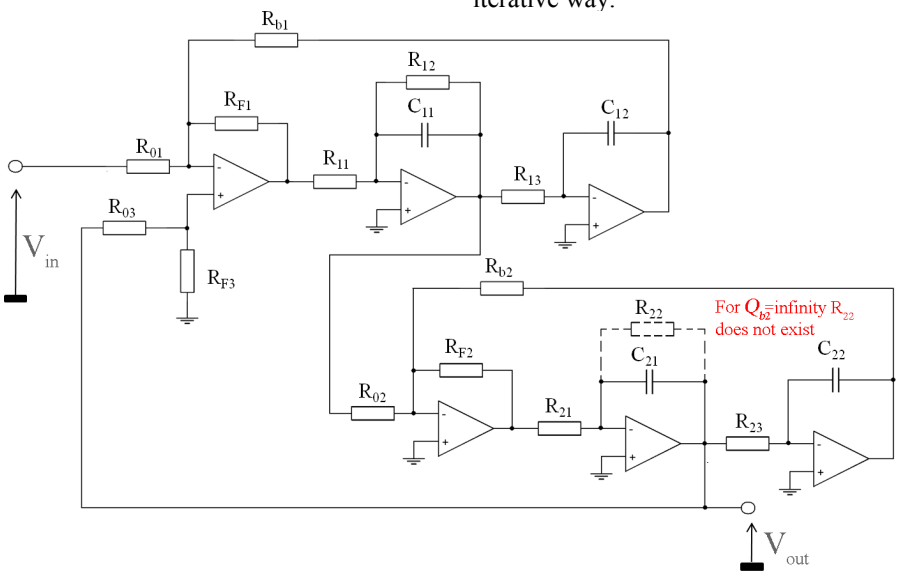


Figure 3. Realization of Biquartic section with two general purpose active-RC Biquadratic sections and feedback.

TABLE I. ELEMENT VALUES OF THE CIRCUIT IN FIG. 3 FOR THE BUTTERWORTH AND CHEBYSHEV EXAMPLES DENORMALIZED TO 10KHZ.

Elements	Butterworth			Chebyshev 0.5dB		
	CAS	BQ Q_1=Q_2	BQ Q_2=∞	CAS	BQ Q_1=Q_2	BQ Q_2=∞
R_{01}	175.923kΩ	116kΩ	95kΩ	502.16kΩ	110kΩ	100kΩ
R_{11}	1.557Ω	1.536kΩ	1.515kΩ	0.818Ω	1.475kΩ	1.412kΩ
R_{12}	22.521kΩ	22.507kΩ	11.253kΩ	31.05kΩ	22.327kΩ	11.163kΩ
R_{13}	1.895kΩ	1.633kΩ	1.617kΩ	1.681kΩ	1.746kΩ	1.612kΩ
R_{F1}	10kΩ	10kΩ	10kΩ	10kΩ	10kΩ	10kΩ
R_{b1}	9.199kΩ	10.086kΩ	10.326kΩ	19.72kΩ	9.821kΩ	11.111kΩ
C_{11}	10nF	10nF	10nF	10nF	10nF	10nF
C_{12}	10nF	10nF	10nF	10nF	10nF	10nF
R_{02}	95.784kΩ	135.714kΩ	214.364kΩ	85.749kΩ	107.843kΩ	144.144kΩ
R_{21}	1.205kΩ	1.533kΩ	1.574kΩ	1.446kΩ	1.504kΩ	1.541kΩ
R_{22}	22.521kΩ	22.507kΩ	∞	31.05kΩ	22.327kΩ	∞
R_{23}	1.972kΩ	1.608kΩ	1.591kΩ	1.75kΩ	1.629kΩ	1.591kΩ
R_{F2}	10kΩ	10kΩ	10kΩ	10kΩ	10kΩ	10kΩ
R_{b2}	9.918kΩ	10.133kΩ	10.104kΩ	9.317kΩ	10.328kΩ	10.322kΩ
C_{21}	10nF	10nF	10nF	10nF	10nF	10nF
C_{22}	10nF	10nF	10nF	10nF	10nF	10nF
R_{03}	∞	459kΩ	138.438kΩ	∞	304.758kΩ	149.566kΩ
R_{F3}	0	10kΩ	10kΩ	0	10kΩ	10kΩ

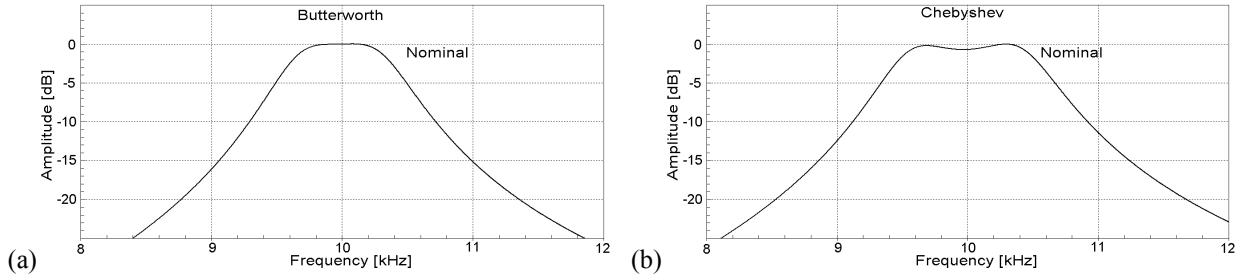


Figure 4. Simulated amplitude-frequency characteristics of filter examples in Table I using Altium Designer. (a) Butterworth. (b) Chebyshev.

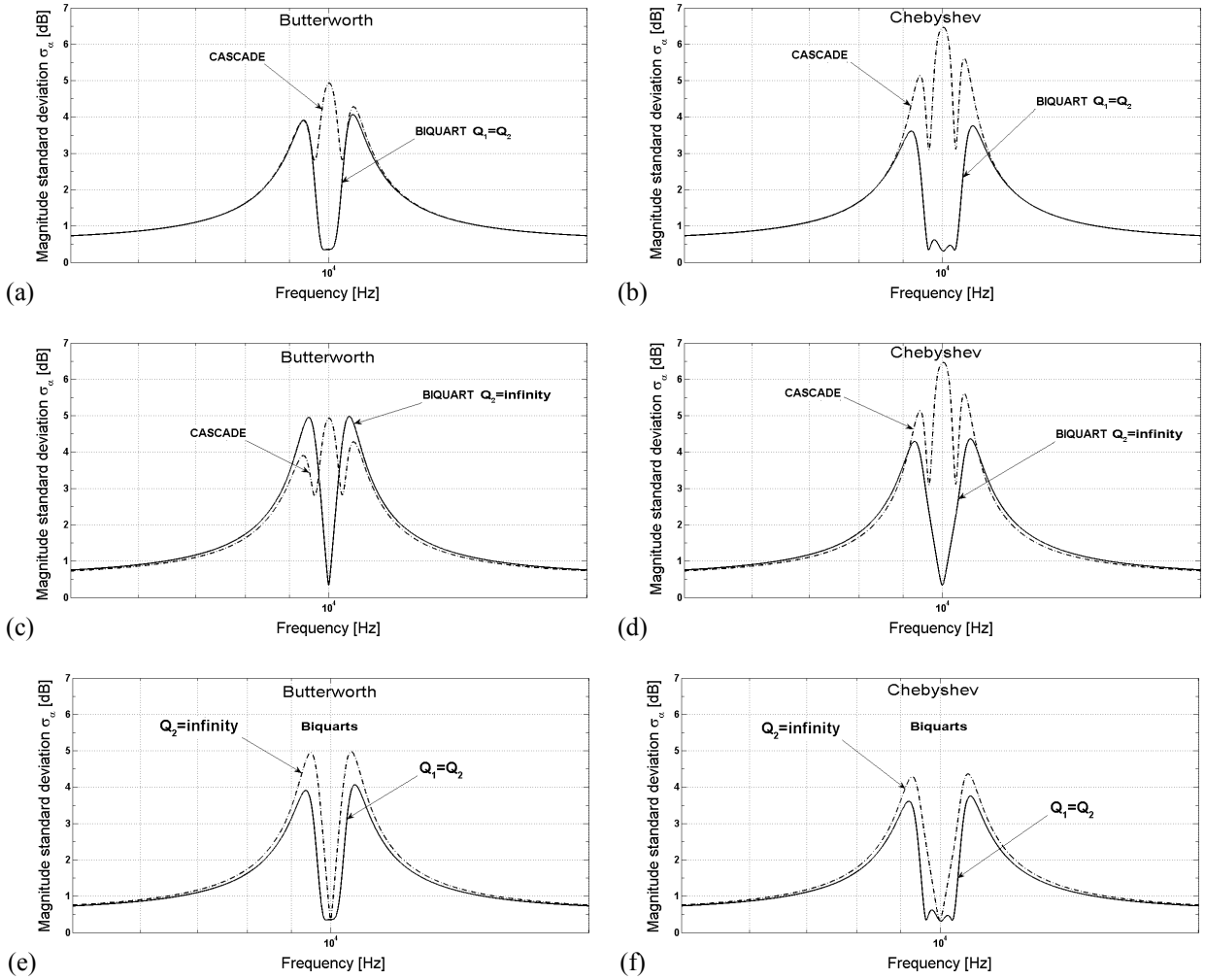


Figure 5. Simulated Schoeffler sensitivities of filter examples in Table I using Matlab. (a), (c), (e) Butterworth. (b), (d), (f) Chebyshev.

V. RESULTS OF SIMULATION AND MEASUREMENT

The sensitivity to component tolerances and output thermal noise, are compared for the cases of cascade (in Section III.A) and for the two BQ realizations: *i*) equal Q-factors: $Q_{b1}=Q_{b2}$ (in Section III.C); and *ii*) one Q-factor is in infinity: $Q_{b2}=\infty$ (in Section III.D). Sensitivity analysis is performed using MATLAB procedures given in [5], assuming relative changes of resistors and capacitors to be uncorrelated random variables, with a zero-mean Gaussian distribution and 1% standard deviation. The standard deviation (which is related to Schoeffler's sensitivities) of the variation of the logarithmic gain $\Delta\alpha$=8.68588 $\Delta|T_{BP}(\omega)|/|T_{BP}(\omega)|$[dB], with respect to passive elements, is calculated for filter examples from Table I and shown in Fig. 5. We conclude that low-sensitivity BP filters can be designed using CBQ filters with equal Q-factors: $Q_{b1}=Q_{b2}$. Using the Altium Designer program output thermal noise spectral density of filter examples in Table I was generated and shown in Fig. 6. In the simulation a model of TL081/TI (Texas Instruments) FET input opamp, having typical values, as found in the data-sheets $e_{na}(f)$=17nV/√Hz and $i_{na1}(f)\approx i_{na2}(f)$=0.01pA/√Hz is used. From Fig. 6 one can conclude that the minimum noise possesses CBQ filters with equal Q-factors: $Q_{b1}=Q_{b2}$.

Measurements were performed on the same examples realized on separate printed circuit boards using 1% passive elements and TL081/TI opamps. For each filter, the output noise spectral density was measured using a typical university lab environment. The measurement equipment consisted of a high-quality HP 4195A Network

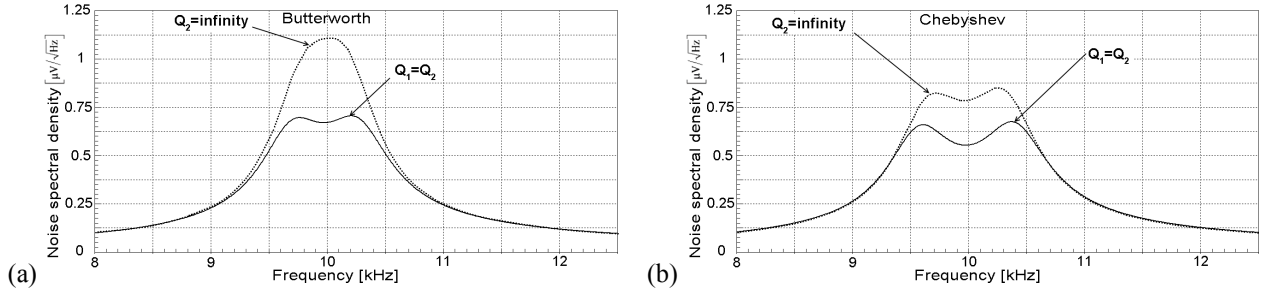


Figure 6. Simulated output thermal noise spectral densities of BQ filter examples in Table I using Altium Designer. (a) Butterworth. (b) Chebyshev.

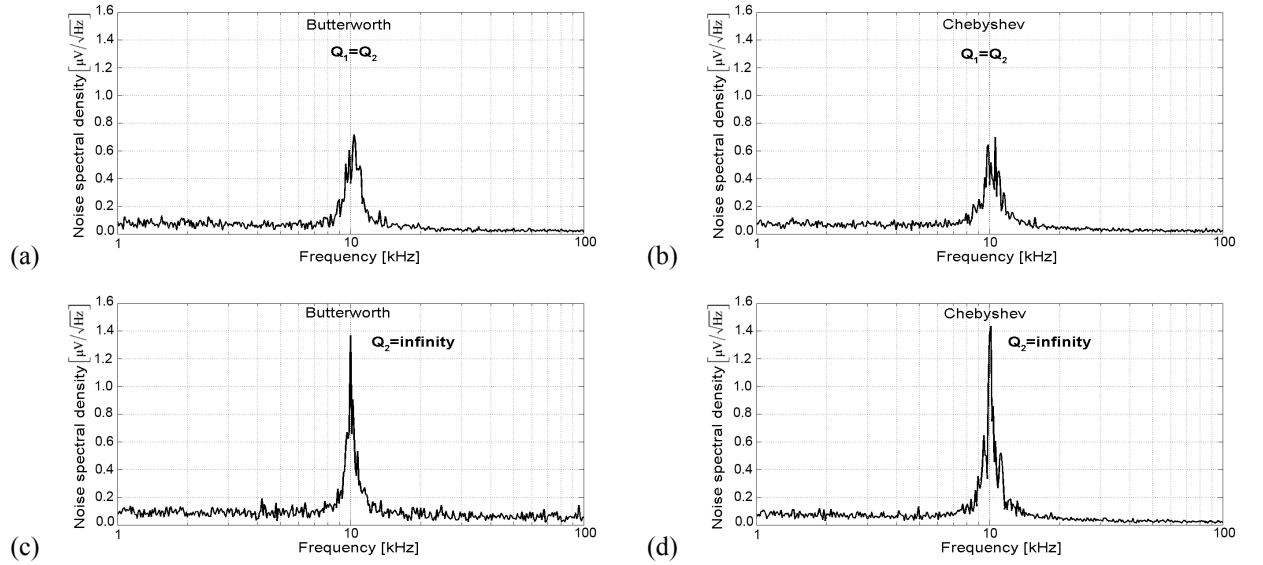


Figure 7. Measured output thermal noise spectral density of BQ filter examples in Table I. (a), (c) Butterworth. (b), (d) Chebyshev.

Analyser, which measures the spectrum of noise (Spectrum mode). Detailed description of measurement procedure and equipment is given in [6]. Measured results are shown in Fig. 7, and reconfirm the results obtained by simulation.

VI. CONLUSION

In accordance with a simulated sensitivity by Schoeffler in realizations of amplitude-frequency characteristics large discrepancies in implementation with one infinite Q factor are expected. Thus, better results are gained using Biquartic sections with identical Q factors. Deviations in amplitude-frequency characteristics are predictable in shifting of the center frequency and amplitude distortion. Observing Schoeffler sensitivities it is visible that the sections with one infinite Q factor have increased sensitivities at the band-edges.

In order to reduce the overall sensitivities we use the three-amplifier biquads as building blocks for biquartic filter because of their considerably lower sensitivities comparing to single-amplifier structures, at the expense of the increased power consumption.

The sensitivity differences between the two CBQ cases are smaller on Chebyshev filter frequency response. Since Chebyshev transfer function has higher Q-factors than Butterworth it leads us to conclusion that the differences are smaller in the highly selective BP filters. The cascade structure has considerably bigger sensitivities in all cases.

Measured spectral density of circuit output thermal noise is very close to the simulation results. We note better results in Biquartic sections with identical Q factors. Both Biquartic realizations are preferable over the common cascade structure. In all measurements and simulations wins Biquartic section with identical Q factors and imposes as an optimal realization of low- sensitive filters.

ACKNOWLEDGEMENT

The authors wish to thank M. Sc. Kristijan Bago for measurements made on the filter examples in this paper.

REFERENCES

[1] N. Mijat and G. S. Moschytz. Multiple-critical-pole coupled active filters. *Int. J. Circ. Theor. Appl.* Vol. 12, no. 3, pp. 249−268, July 1984.

[2] S. Fotopoulos and T. Deliyannis. Active RC realization of high-order bandpass filter functions by cascading biquartic sections. *Int. J. Circ. Theor. Appl.* Vol. 12, no. 3, pp. 223−238, July 1984.

[3] P. R. Geffe. Designer's Guide to Active Bandpass Filters. *EDN* 1974; 46–52.

[4] N. Mijat and G. S. Moschytz, "Sensitivity of narrowband biquartic BP active filter block," in *Proc. of IEEE International Symposium on Network Theory and Design*, (Sarajevo, Bosnia and Herzegovina), 1984, pp. 158–163.

[5] D. Jurisic. *Applications of MATLAB in Science and Engineering. Chapter 10: Low-Noise, Low-Sensitivity Active-RC Allpole Filters Using MATLAB Optimization*. Rijeka: InTech, 2011, pp. 197−224.

[6] D. Jurisic, N. Mijat and M. Ranilovic, "Power Consumption, Noise and Bode Diagram Measurement of Active Filters," in *Proc. of the 34th International Convention* MIPRO 2011, (Opatija, Croatia), May 2013, pp. 92–95.

Development and testing of complex PLL algorithm using hardware in the loop system

Tihomir Čihak*, Mladen Puškarić* and Željko Jakopović**
* KONČAR-Institut za elektrotehniku d.d., Zagreb, Croatia
** Fakultet elektrotehnike i računarstva, Zagreb, Croatia
tcihak@koncar-institut.hr

Abstract - Grid connected converters present a key element in modern systems for harvesting renewable energy sources. Each of these converters, non-dependent on the converter topology, requires a grid synchronization algorithm. This is not a new requirement for such converter types, but design and development of applicable algorithms requires a lot of time and resources. This is especially evident in high power converters on which grid codes are applied. This paper shows an accelerated development method applied on a complex PLL structure resulting with a code suitable for deployment on real-time control system.

I. INTRODUCTION

For correct operation of grid-connected converters basic prerequisite is synchronization with grid voltage accomplished through a grid synchronization algorithm. Modern synchronization algorithms have a few different approaches, and the most common ones are different versions of PLL (Phase-locked loop) structures. Just like their associated converters, they are roughly divided into PLLs for single-phase converters and PLL algorithms for three-phase converters. This paper focuses on PLL algorithms for three-phase converters only. This is done mostly, since they are regularly of higher power output, due to specific requirements of different national grid code requirements [10][12], including low voltage ride through (LVRT) capability [6][7], which are applied to these converters. Research explained further on is to be applied to an in-house developed photovoltaic grid-connected inverter which uses a form of voltage oriented vector control (VOC) [13] as shown in Figure 1. As it can be seen, quality of the extracted angle will dictate performance of the control algorithm and the converter itself.

II. DEVELOPMENT AND TESTING OVERVIEW

During the development, the idea was to create synchronization algorithm and implement the structure into the real converter without any major subsequent changes in the PLL structure which is usually necessary when carrying over simulations-based algorithms into real converters (e.g. influence of discretization). For such goals to be achieved, both PLL algorithm and the inverter had to be implemented on the system as close as possible to the target system. As an ideal platform, a hardware-in-the-loop system, shown in the Figure 3. coupled with DSP control board (same as in target converter) was used. Selection of an appropriate baseline-algorithm was done and explained in detail in authors' recent work [18]. For algorithm development, focus was set on a decoupled double synchronous reference frame PLL (DDSRF-PLL) due to its benefits over other structures [1][3] which was compared with synchronous reference frame PLL structure (SRF-PLL), and grid angle extracted from measured voltages in αβ–frame. After initial testing using simulation software (of both synchronization algorithms and grid with disturbances) and verification of the results, system was further developed using only hardware in the loop system and associated DSP.

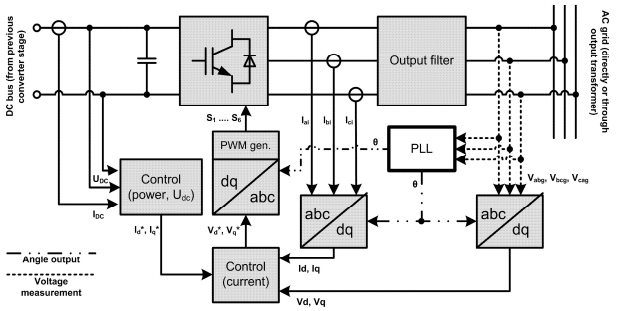


Figure 1. Block diagram of inverter VOC control algorithm

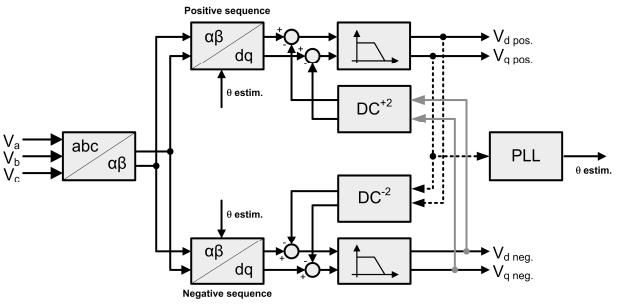


Figure 2. Block diagram of the starting PLL algorithm



Figure 3. Hardware in the loop system used for the development

Several sets of tests for various disturbances, either predefined in grid code requirements, or found as a compilation of different grid code requirements [7][10], have been applied to the PLLs. These included symmetric voltage drops, asymmetric voltage drops (i.e. short circuits), phase changes including extreme conditions equal to *phase skipping* and frequency changes. The most important test, which gave questionable results and required further research, explained in this paper, was testing of operation in grid polluted with higher harmonics. Although operating marginally satisfactory in some cases [6], in-depth testing with waveforms which could occur if the converter is changing its operation regime, i.e. switching from grid-connected operation into micro-grid (or as a worst case, into stand-alone operation with local loads) and vice versa, has shown inadequate results, as shown in Figure 4. (figure shows real-time recording from DSP system connected to the hardware-in-the-loop system). *VDN* and *VQN* are *D* and *Q* components of positive rotating transformation while *VDM* and *VQM* are inverse rotation components. *U1_JH* and *U1_JHP* are input and estimated voltage waveforms of a single phase. The found problem (i.e., bad phase and voltage estimation in harmonics-polluted grid) is known and elaborated in the literature [2]-[5]. Authors proposed pre-filtering structure based around the second-order generalized integrator (SOGI) shown in Figure 5. which suppresses higher harmonics, allowing the rest of the DDSRF structure to be optimized for required dynamics.

III. EXTENSION OF PREFILTERING STRUCTURE

Although results showed that large amounts of higher harmonics have been successfully attenuated [18], some deficiencies have been encountered. These are based around inherent downsides of the SOGI structure [14], most easily seen and described as phase estimation error found in system response during start-up and in responses in similar step-changes in the input. Although it seems that system can be further optimized, this is not the case as the system is to be used in phase estimation of grid voltages with high amount of harmonics, and any further tuning of the system results with the distorted output which is then transferred through the succeeding DDSRF structure and into the estimated phase. A possible solution, as described in the [8] was therefore tested.

A. Extended system description

Second order generalized integrator is described with transfer function according to (1). Transfer functions across the direct and quadrature components (outputs of the system) are given with (2) and (3) respectively. These equations show the idea behind the presented structure, i.e., the filter bandwidth is not a function of the central frequency but filter gain and the amplitudes of both outputs are the same as in the input signal. A simple modification to second order generalized integrator structure shown in Figure 5. can be made as shown in Figure 7. First difference between previous SOGI structure and filtering structure presented here is in redirecting both integrator outputs as error signals back to the integrator inputs of the phase-shifted signal, while this was true only for the quadrature integrator output in the case of generalized integrator filter. Second difference is that input error signal is now split into two branches with separate gains (outer loop with single gain is not shown in Figure 5.). This filtering structure has, again, two separate outputs with 90° phase offset. For such structure transfer functions across direct and quadrature component are given with (4) and (5) respectively.

$$G_{filt}(s) = \frac{s \cdot \omega_S}{s^2 + \omega_s^2} \tag{1}$$

$$G_D(s) = \frac{k \cdot s \cdot \omega_S}{s^2 + s \cdot k \cdot \omega_S + \omega_s^2} \tag{2}$$

$$G_Q(s) = \frac{k \cdot \omega_s^2}{s^2 + s \cdot k \cdot \omega_S + \omega_s^2} \tag{3}$$

$$G_{DK}(s) = \frac{k_{DK} \cdot s - k_{QK} \cdot \omega_S}{s^2 + s \cdot k_{DK} - k_{QK} \cdot \omega_S + \omega_s^2} \tag{4}$$

$$G_{QK}(s) = \frac{k_{QK} \cdot s + k_{QK} \cdot \omega_S}{s^2 + s \cdot k_{DK} - k_{QK} \cdot \omega_S + \omega_s^2} \tag{5}$$

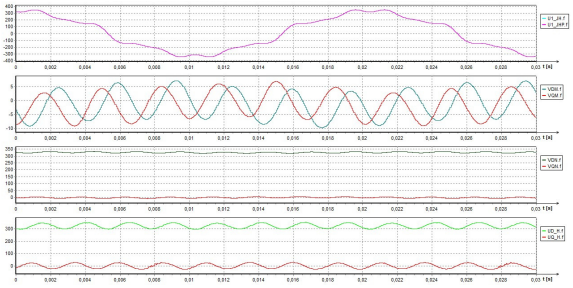


Figure 4. Baseline PLL algorithm behaviour in harmonics-polluted grid

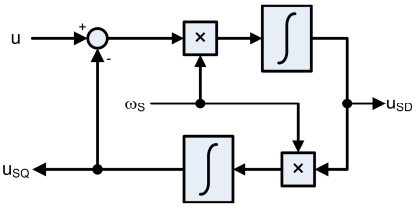


Figure 5. Schematic of second-order generalized integrator (SOGI)

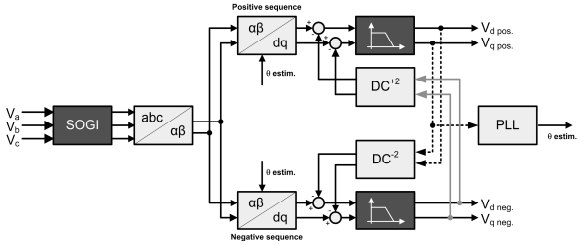


Figure 6. Modified DDSRF-PLL structure with added pre-filtering structure

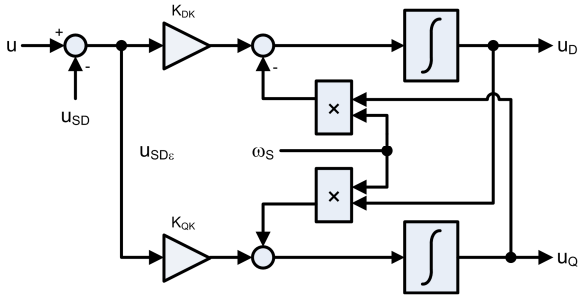


Figure 7. Transformation of SOGI structure

B. Output components' estimation

Two different approaches were tested for designing the presented filtering structure. First is an iterative, heuristic process which starts with generalized integrator settings. This is accomplished by setting the K_{QK} factor to 0 which basically converts the extended structure back to generalized integrator filter. K_{DK} factor is now tuned. Time constant (convergence) of the SOGI structure is given with (6), while the time constant of the extended general integrator structure is given with (7).

$$\tau = \frac{2}{k \cdot \omega_S} \tag{6}$$

$$\tau = \frac{2}{k_{DK}} \tag{7}$$

Since literature suggests that settling time for second order system is $t_S=4.6\times\tau$, [17], it is implied that, in practice, optimal gain for SOGI structure is *k=sqrt(2)*. For frequency of 50Hz this equals to settling time of roughly 20ms. For clear understanding, Figure 8. and Figure 9. give (initial) step response of second order generalized integrator structure for several different settling times. Shown tuning range is from 0.8 to √2 in 0.2 steps. Further increase of K_{DK} factor does make improvements, but, as it can be seen from Figure 8. , improvements are limited and a phase offset error (i.e. delay if viewed as sinusoidal values) after startup always exists, as shown in Figure 9. After a base value for K_{DK} was set, K_{QK} factor was changed in steps (equivalent outputs for K_{QK} settings are not shown). Acquired results show that gain increase creates perceivable change in the filter behavior. More precisely, maximum error at the output can be reduced by a large amount (more than 40% in tested cases), but in the same time, as a downside, overshoot increases and with further gain increase, inadmissible oscillations occur. Such behavior will be seen as *hunting* of the input signal which will be transferred to DDSRF structure resulting false values of phase angles. This deficiency will be passed further on into the rotating transformations and current controllers which destabilize the control system. It should be emphasized that it can be seen that while extending SOGI structure, position of the multiplication with ω has been changed prior to summation point. Due to this fact, a scaling factor, approximately same as the angular frequency, had to be incorporated into the gain blocks so that the results are comparable with the starting general integrator structure. Also, it should be noted that throughout the paper, graphs dedicated to extended structure are designated as *Kalman* due to reasons explained further on in the text.

After the system was tested using clean input sine signal and iterative procedure gave a satisfactory response, a setup was extended using a higher harmonic generator. Basic setup in simulation models uses percentage harmonic shares as described in [7]. For previously described implementations in real-time system this had to be changed due to limitations of the hardware in the loop system (only first four higher harmonic components were used on hardware in the loop system). For testing in simulation software as described further on, all harmonic components where used. Also, a set of output gain blocks named *weight* was added by which all harmonic components can be changed by same amount, i.e. this allowed us to change the output THD. Structure was tested with THD between 2.5% and 20% but test results are given for 7.5% total harmonic distortion. System was tested in the same way as in the case without harmonics. Results show the same qualitative change in filter behavior, but the optimal setting found using iterative procedure differs from the one for the clean signal. Obviously, analytical calculation must include influence of signal distortion to be valid for such environment. Therefore, another iterative procedure was done with distorted input signals, and this resulted with different settings as shown in Figure 11. Results show that the system, although the gains were found satisfactory, does not behave optimally. Iterative procedure with noisy input signals also proved very time consuming, as it does not select correct pairs of K_{DK} and K_{QK} gains at the same time.

C. Alternative approach

Due to problems described previously, a different method for finding filtering structure gains was selected. This procedure is the optimal gain calculation using the Kalman state estimator. Due to this method, *Kalman filter* designation is used further on in the text (*Kalman* designation in all figures). This method uses a state-space model of the structure, i.e. the part of the filtering structure in Figure 7. from gain outputs to integrator outputs (i.e. oscillators), process and measurement noise data. For this approach, procedure described in [15] and [16] was used. Output of the estimated direct component is measured output *y*, while the gain outputs are noise inputs *w*. Measured inputs *u* are non-existent. According to mentioned literature, for (continuous) plant model described as (8) and (9), and which satisfies requirements given with (10) and (11), a Kalman state estimator can be constructed.

$$\dot{x} = \mathbf{A}x + \mathbf{B}u + \mathbf{G}w \tag{8}$$

$$y = \mathbf{C}x + \mathbf{D}u + \mathbf{H}w + v \tag{9}$$

$$\mathbf{E}\left(w\right) = \mathbf{E}\left(v\right) = 0, \mathbf{E}\left(ww^T\right) = \mathbf{Q} \tag{10}$$

$$\mathbf{E}\left(vv^T\right) = \mathbf{R}, \mathbf{E}\left(wv^T\right) = \mathbf{N} \tag{11}$$

$$\dot{\hat{x}} = \mathbf{A}\hat{x} + \mathbf{B}u + \mathbf{L}\left(y - \mathbf{C}\hat{x} - \mathbf{D}u\right) \quad (12)$$

$$\begin{bmatrix} \hat{y} \\ \hat{x} \end{bmatrix} = \begin{bmatrix} \mathbf{C} \\ \mathbf{I} \end{bmatrix} \hat{x} + \begin{bmatrix} \mathbf{D} \\ \mathbf{0} \end{bmatrix} u \quad (13)$$

$$L = \left(\mathbf{P}\mathbf{C}^T + \bar{\mathbf{N}}\right)\bar{\mathbf{R}}^{-1} \quad (14)$$

$$sys : \mathbf{A}, \left[\mathbf{B}\mathbf{G}\right], \mathbf{C}, \left[\mathbf{D}\mathbf{H}\right] \quad (15)$$

Kalman filter is described with (12) and (13), while the filter optimal gain L is found solving algebraic equation (14). This equation was solved using numerical computing tools which require the system to be expressed using matrices given with (15). Given the continuous structure model in form (8) and (9), according to [8], gains shown in Figure 7. can be found analytically by solving (14) which gives gains expressed as in (17) and (16). q are elements of process noise covariance data matrix (diagonal matrix $\boldsymbol{Q}$) and r is scalar representing measurement noise covariance data matrix $\boldsymbol{R}$. Such representation gives the both filter gains with values that are dependent on the two parameters, q and r as previously defined. First parameter selection was done on clean input signal, as in previous cases. Parameter q was set at $q=1$, while the parameter r was variated between 1×10^{-4} and 1×10^{-8}. For each of the settings, gains were found and the simulation of the filtering structure was performed. Some of the results are given in Figure 12. Best solution found in this case yielded very high gains and the tracking of the input signal was extremely good, as it can be seen from selected (reference) output, compared to SOGI structure (see Figure 14. Similarly to results in previous procedure, selected settings for clean input signal did not work correctly when the input signal was highly distorted. Therefore, values in process noise covariance data matrix (i.e. parameter q) had to be modified. A range of values has been tested but satisfactory results were found for $q=0.05$. Parameter r was changed in the same range of values as in the previous case. Behavior of the system during the tuning is presented in the Figure 13. Alteration of parameter values in matrices resulted in generally lower gains than those found in the previous case, and, as presented in Figure 15. gave a system which did not track the input signal as good as in the case of clean input signal, but still much better than any of settings found by iterative process (see Figure 10. and Figure 11. for reference).

$$k_{QK} = \omega_S \cdot \left(1 - \sqrt{1 + \frac{q}{r \cdot \omega_s^2}}\right) \quad (16)$$

$$k_{DK} = \sqrt{\frac{2 \cdot q}{r} - k_{QK}^2} \quad (17)$$

More precisely, there is no initial delay with the used procedure and the settling times are lower with same or better output waveform. Higher gains, as it can be seen in Figure 13. , will not give better results as the output signal becomes increasingly distorted in that case. Also, further tuning can be considered obsolete as presented structure is a pre-filter for DDSRF which can handle small distortions in the output of the pre-filter. To further examine filter behavior, linear analysis has been performed on developed filtering structure and compared to the SOGI. Results are given as Bode plots. Figure 16. shows Bode plot for filter settings found with iterative procedure. Results show that extended filter will have generally similar frequency response to SOGI filter if the input signal is not very distorted, but the phase angle characteristic is much flatter in frequency range under the resonant frequency. This feature is the main reason for correct phase-tracking of the input signal. Same analysis done on the filter designed using Kalman estimator method shows that found gain values give a different system behavior. Namely, high-gain model has completely flat amplitude characteristics extending much above the filter resonant frequency, which, when phase characteristics is taken into account, converts the filter structure into low-pass filter rather than band-pass filter of the starting SOGI structure. This is the reason why such filter does not work as expected with highly distorted input signals. Tuning the system for satisfactory operation with large amounts of higher harmonics leads, again, to a different filtering behavior, as shown in Figure 17. Gain values are here much lower and give the system much like the one shown in Figure 16.

D. Real-time system implementation

Designed Kalman filter was tested on real-time system based on the Texas Instruments DSP coupled to the hardware in the loop system shown in Figure 3. DSP was programmed using proprietary graphical programming tool GRAP, which uses highly optimized assembler code. Results shown in the following pictures are taken through the same proprietary software. System response with the settings as those in Figure 14. (high gains) are given in Figure 18. Bottom picture shows direct (estimated) component output overlaid over first harmonic of the input signal. Figure 19. shows the high-gain filter behavior with distorted input signal. The results show, as expected, bad filter performance. Figure 20. shows the case with filter settings optimized for high input signal distortions (for distorted signal only). It is clear that the system behavior is as expected through simulation software, although the simulated structure is created (and parameters found for) continuous system, while the target system is necessarily discrete. The slight difference does exist and can be seen comparing Figure 20. to Figure 15. , but this is also due to difference in input signal harmonic content. In conclusion, it is not necessary to include discretization effects in filter design for a control system with such a high operating frequency.

IV. CONCLUSION AND FURTHER WORK

Implemented extended PLL structure proved to be adequate for the converters on which the grid code requirements apply (high-power grid-connected ones). The tested algorithm gave excellent results for all disturbances, both theoretically and practically. Addition of filtering structure extended into Kalman filter solved the observed problems with initial PLL, initial delay on input step changes and optimization of the structure with

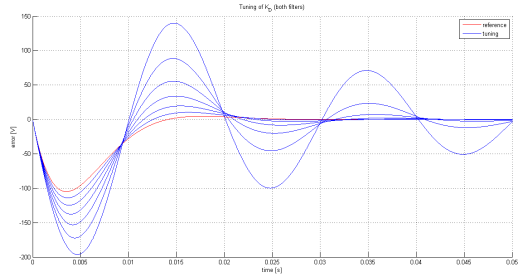


Figure 8. Influence of K_D value on filter output (error output)

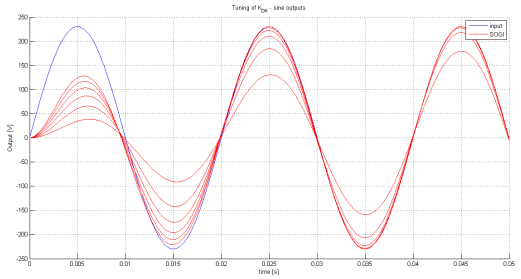


Figure 9. K_D value influence on filter output (sine output)

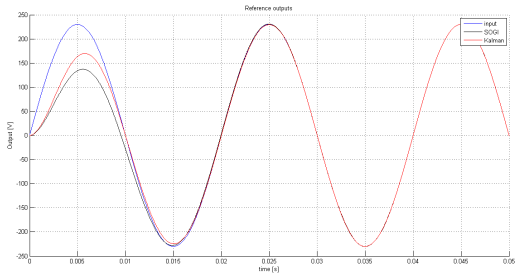


Figure 10. Reference settings for non-distorted input signal (sine output)

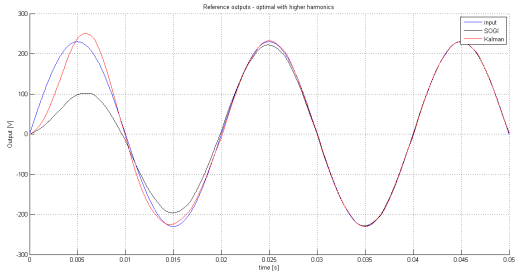


Figure 11. Reference settings for distorted input (sine)

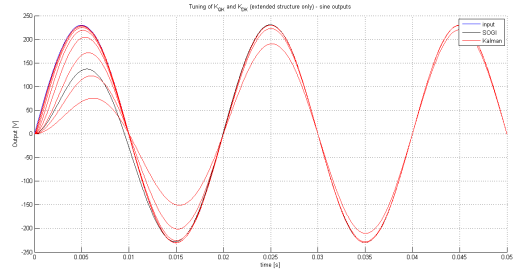


Figure 12. Tuning by Kalman-estimator procedure with clean input signal (sine output)

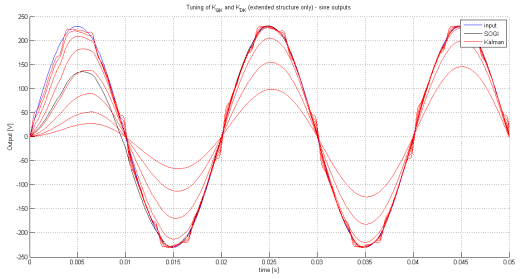


Figure 13. Tuning by Kalman-estimator procedure with distorted input signal (sine output)

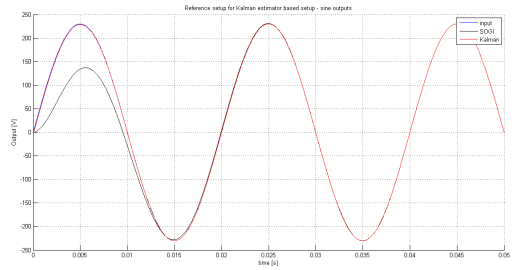


Figure 14. Reference settings for clean input (sine)

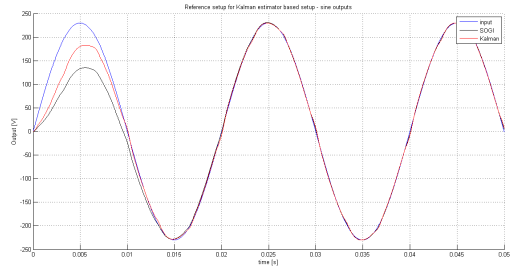


Figure 15. Reference settings for noisy input (sine)

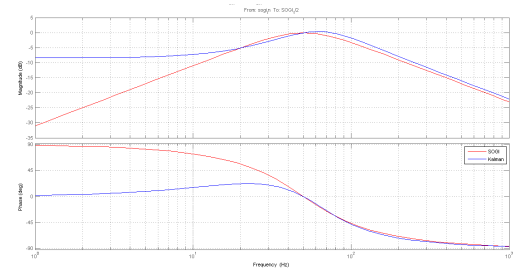


Figure 16. Filter setting derived from SOGI technical optimum (K_{DK}=√2, K_{QK}=0.5); iterative setting for clean signal

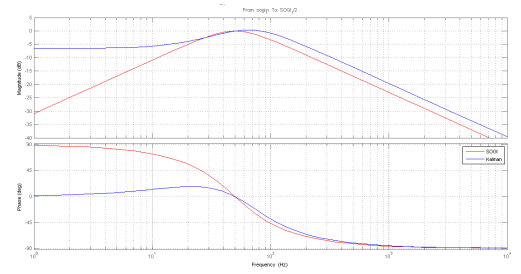


Figure 17. Setting found as optimal with Kalman-estimator procedure for noisy signals (K_{DK}=657, K_{QK}=281)

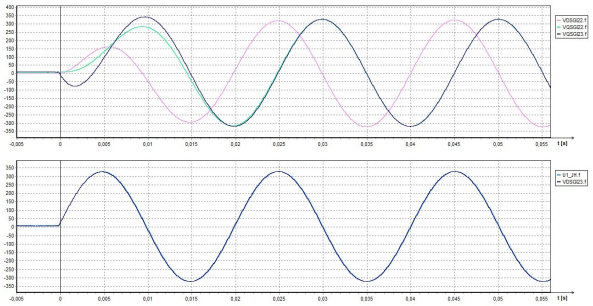


Figure 18. Filtering structure behavior on real-time system (clean input signal, high gains)

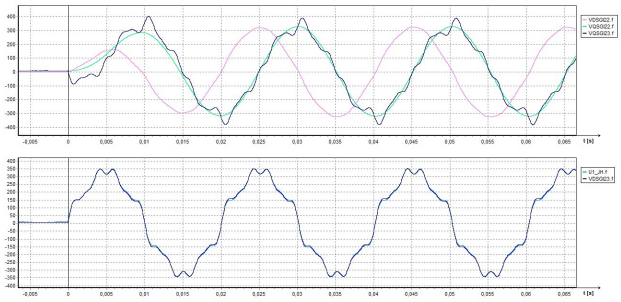


Figure 19. Filtering structure behavior on real-time system (distorted input signal, high gains)

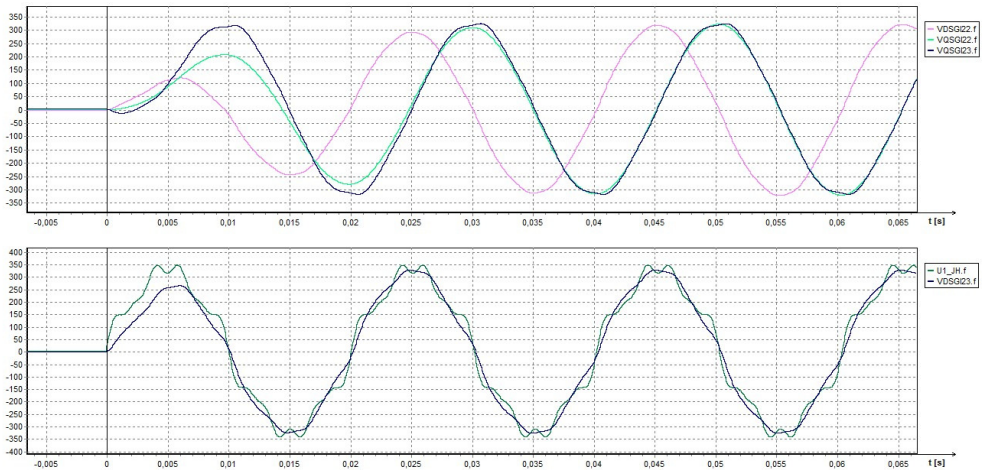


Figure 20. Filtering structure behavior on real-time system (distorted input signal, low gains)

high input signal distortions (large quantities of higher harmonics). Kalman estimator design procedure also proved to be a valuable tool as the system can be designed with exact, optimal set of parameters, given that the input variables are found correctly. Hardware in the loop system proved itself to be a valuable research tool as the implementation could be done directly in the real-time system and influences of the real-time system on algorithm and system stability, such as discretization effects, could be observed. In future, stability of the proposed PLL algorithm will be further analyzed and results from testing on a real converter will be shown.

REFERENCES

[1] R. Teodorescu, M. Liserre, P. Rodríguez, "Grid Converters for Photovoltaic and Wind Power Systems", John Wiley & Sons, 2011

[2] P. Rodríguez, A. Luna, I. Candela, R. Teodorescu, F. Blaabjerg, "Grid Synchronization of Power Converters using Multiple Second Order Generalized Integrators", IECON 2008, 34th Annual Conference of IEEE, pp. 755 – 760

[3] P. Rodríguez, J. Pou, J. Bergas, J.I. Candela, R.P. Burgos, D. Boroyevich, "Decoupled Double Synchronous Reference Frame PLL for Power Converters Control", Power Electronics, IEEE Transactions on, Vol.22, No.2, March 2007

[4] S. J. Lee, J. K. Kang, S. K. Sul, "A New Phase Detection Method for Power Conversion Systems Considering Distorted Conditions in Power System", IEEE industry applications conference thirty-fourth IAS annual meeting, 1999, vol.4, pp. 2167 - 2172

[5] J. A. Suul, "Control of Grid Integrated Voltage Source Converters under Unbalanced Conditions", Ph.D. Thesis, Norwegian University of Science and Technology, Trondheim, March 2012

[6] Ministarstvo gospodarstva, rada i poduzetništva, "Mrežna pravila elektroenergetskog sustava", Narodne novine, No.177/04, Croatia, 2004

[7] Bundesverband der Energie- und Wasserwirtschaft e.V., "Generating Plants Connected to the Medium-Voltage Network", Germany, June 2008

[8] K. De Brabandere, et al, "Design and Operation of a Phase-Locked Loop with Kalman Estimator-Based Filter for Single-Phase Applications", IECON 2006, 2006

[9] T. Gehlhaar, "Grid Code Compliance beyond simple LVRT", Germanischer Lloyd, Germany, 2012

[10] F. Iov, A.D. Hansen, P. Sorensen, N.A. Cutululis, Mapping of grid faults and grid codes", Riso National Laboratory, Denmark, 2007

[11] Luna, A., Citro, C., Gavriluta, C., Candela, I., Rodriguez, P., Hermoso, J., Advanced PLL structures for grid synchronization in distributed generation, ICREPQ'12, 2012

[12] Sourkounis, C., Tourou, P., Grid Code Requirements for Wind Power Integration in Europe, Conference Papers in Energy journal, Volume 2013 (2013)

[13] T. Čihak, M. Puškarić, Ž., Jakopović, "Modified vector control appropriate for synthesis of all-purpose controller for grid-connected converters", EDPE2013 Conference, Croatia, 2013

[14] S. Golestan, M. Monfared, F. D. Freijedo, J. M. Guerrero, "Dynamics Assessment of Advanced Single-Phase PLL Structures", IEEE Transactions on Industrial Electronics, Vol.60, No.6, June 2013

[15] G.F. Franklin, J.D. Powell, M.L. Workman, "Digital Control of Dynamic Systems, Second Edition", Addison-Wesley, 1990.

[16] F. Lewis, "Optimal Estimation", John Wiley & Sons, Inc, 1986.

[17] Z.Vukić, Lj. Kuljača, "Automatsko upravljanje", Kigen, Zagreb, 2005.

[18] T. Čihak, M. Puškarić, N. Težak, "Testing of PLL structures on on real time systems", Ee2013, Republic of Serbia, October 2013

Software simulation of LDPC codes and performance analysis

C. Sandu, I. Florescu, C. Rotaru
University POLITEHNICA Bucharest,
Iuliu Maniu 1-3 Blvd , Bucharest, ROMANIA

Abstract – In this paper we will analyze the influence of using Low Density Parity Check LDPC codes in different channels environment. The performances of LDPC codes are estimated in several scenarios. The simulation software gives us the possibilities to choose between different communications channels according to the encoding matrix used. Also we can adjust a number of parameters, thing that help to find the optimal solution depending on the decoding error rate or Bit Error Ratio (BER). Optimal combinations between parameters, the encoding matrix and structures design are obtained, related to required performance or Signal Noise Ratio (SNR) ranges. The sets of parameters and criteria introduced are discussed as the simulation results, also.

I. INTRODUCTION

Robert Gallager introduced Low Density Parity-Check, just a decade after Shannon's work was published. Gallager made a crucial innovation by introducing iterative decoding algorithms; this showed that significant fraction of channel capacity can be achieved at low complexity [1]. In the early 1980's, Tanner has provided a graphic representation of LDPC and other coding scheme, which are now referred by Tanner's diagram.. He proposed that the problem be solved by decomposing decoding codes in simpler codes from their structure and he introduced the concept that is now known as min-sum decoding algorithm. Tanner also has foreseen the huge potential of parallelism, allowing simple codes to be processed simultaneously due to their low complexity. [2]

A low density parity check code is a code with a linear matrix having a parity check matrix sparsely populated. Considering a regular LDPC code with parity matrix given by **H**:

$$\mathbf{H} = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 & 0 & 1 \end{bmatrix}$$

Figure.1

We assign to each bit of a code word a V_S: $s \in \{1...m\}$, where each variable is also corresponding to a column in **H** matrix. Therefore, each row of **H** represents a restriction on parity check with code, c_r: $r \in \{1...m\}$. The participation of variable v_s in restriction control is suggested by an element different from zero at the position (r, s) in **H**. Variable nodes and check nodes can also have values associated with them. Symbols v_s and c_r are used to designate both the nodes and to represent the values. We call the number of one elements in the row or columns as *weight.* Each line in **H** is a parity check constraint. If the values assigned to the variable stack are a valid code word, then $c_r = 0$ for all $r \in \{1...m\}$.

$v_1+v_3+v_5+v_7=c_1$

$v_1+v_4+v_6+v_8=c_2$

$v_2+v_3+v_6+v_7=c_3$

$v_2+v_4+v_5+v_8=c_4$

Gallager defines a regular structure for LDPC codes as follows: ***a regular LDPC code with (n, j, i) is a code block of length n that has a parity check matrix of weights for row and column, exactly j on column and exactly i on column.*** [1]

By contrast, irregular LDPC codes allow variables to participate in various numbers of controls and enable possibility to apply restrictions to control numbers different from variables. An irregular LDPC code can be described by the weight vector distribution line and column of the parity check matrix of his.

LDPC codes are a class of linear block codes. Thus, they can be encoded using the generator matrix but in absence of structure within the code, as the data block increases this method requires to the decoder large storage and related processing requirements. Although H is built to be sparsely populated, the code generator matrix, G, is generally charged. Therefore encoders can become considerably slower and lager when block n is increased. [3] macaky

Iterative decoding of LDPC codes: the code word transmitted, x, is exposed to noise from the channel. The decoder provides an estimate of the transmitted information, based on code restrictions and received vector y. Message values are calculated at each node based on local code restriction. **H** matrix is easily calculated due to the low number of components. Decoding algorithms with multiple levels of decision shall use the information received. Considering the channel used a priori vector of values is generated by y and performs the function of transmission for decoder. An a posteriori vector of values is returned when

decoding process ends. *x* is produced by imposing a tougher admissions on this vector.

Sum-product algorithm can be used to perform decoding light admittance coefficient diagram, when each node is considered a local processor. Bidirectional edges are used to carry messages between processors during each iteration step.

The algorithm of two level of decision decoding can be applied for additive white Gaussian noise AWGN, binary erasure and symmetric binary channels. If the information is carried by the receiving vector, then it is downloaded.

LDPC codes can be considered a serious competitor to turbo codes. They admit a wide range of tradeoffs between performance and decoding complexity and they can do an asymptotically better performance than turbo codes. High encoding complexity is one of the major criticism.

The contribution of this paper is to provide a BER analysis in magnetic channels and binary channel. Also, this paper presents two versions of sum-product algorithm with their BER analysis. [4]

The remainder of this paper is organized as follows. Section II discusses about the communication channels that we approach. In Section III, we present the software simulator, improvements and functional terms. Section IV presents simulation results. Conclusions are drawn in Section V.

II. COMMUNICATION CHANNELS AND ALGORITHMS USED

The rapid growth of storage capacities and transfer rates imply increased density information recording disk, which requires the development of efficient coding methods and efficient methods for information detection on *magnetic channels (PR IV, EPR IV, E2PR IV)* which is necessary to preserve the integrity of the recorded data. Coding information is essential in magnetic recording channel structure, thus preventing major problems than can affect the timing of the read-write head with magnetic disk and data integrity. For studying magnetic channel we will use the channel with AWGN), this type of channel has a good practical relevance.

The sum-product algorithm SPA is used in binary channel. A binary channel with a stationary ergodic noise is a binary-input, binary-output channel whose output in response to a transmitted binary vector ***t*** is given by

r = ***t*** + ***n*** *mod* ***2,*** where ***n***, the noise vector, has a probability distribution that is a) independent of ***t*** and b) stationary and ergodic. [5].

Algorithm I: SPA decoder: (probability domain):

1. Initialization:

Sending messages: $\mu_{vs\to cr}(0)=0$, $\mu_{vs\to cr}(1)=0.5$ and $\mu_{cr\to vs}(0)=\mu_{cr\to vs}(1)=0.5$ (1)

2. Priority setting:

Using each symbol received is calculated: $u=2y_s/\sigma^2$ (2)

Calculate: $p_s(0)=e^u/(1+e^u)$ and $p_s(1)=e^{-u}/(1+e^{-u})$ (3)

3. Control variable:

We suppose $\delta_{VS}=\mu_{VS\to cr}(0)-\mu_{VS\to cr}(1)$ and calculate: $\delta_{Cr\to v}=\prod_{vs\in\Gamma(cr)\setminus v}\delta_{vs}$ (4)

From each control c_r to each variable $u\in\Gamma(c_r)$ is transmitted: $\mu_{cr\to u}(0)=1/2(1+\delta_{cr\to u})$ and $\mu_{cr\to u}(1)=1/2(1-\delta_{cr\to u})$ (5)

4. Variable check:

From each variable v_s for each $c\in\Gamma(u_S)$ is transmitted:

$$\mu_{v_S\to c}(0)=\gamma p_S(0)\prod_{c_r\in\Gamma(v_S)\setminus c}\mu_{c_r\to v_s}(0) \quad (6)$$

and

$$\mu_{v_S\to c}(1)=\gamma p_S(1)\prod_{c_r\in\Gamma(v_S)\setminus c}\mu_{c_r\to v_s}(1) \quad (7)$$

The normalization factor γ is selected so that $\mu_{vs\to c}(0)+\mu_{vs\to c}(1)=1$. (8)

5. A posteriori calculations:

For each symbol is calculated:

$$q_s(0)=\gamma p_s(0)\prod_{c_r\in\Gamma(v_s)}\mu_{c_r\to v_s}(0) \quad (9)$$

and

$$q_s(1)=\gamma p_s(1)\prod_{c_r\in\Gamma(v_s)}\mu_{c_r\to v_s}(1) \quad (10)$$

The normalization factor γ is selected so that $q_s(0)+q_s(1)=1$.

6. Stop/ continues:

Is calculated $\hat{x}$ from: $\hat{x}_s=q_s(1)>1\setminus 2$ (11)

If H $\hat{x}=0$ then operation is completed successfully, so if the limit of iterative operation is reached, operation

should be terminated and declared a failure, otherwise return to step 3 is done.

Algorithm II: SPA decoder: (probability registration domain):

1. Initialization:

Initialize messages: $\mu_{vs\to cr} = \mu_{cr\to vs} = 0$ (12)

2. **Priority setting**:

Using each symbol received is calculated: $u = 2y_s / \sigma^2$ (13)

3. Control variable:

From each control c_r to each variable $v \in \Gamma(c_r)$ is transmitted:

$$\mu_{c_r \to v} = 2\tanh^{-1}\Big(\prod_{v_s \in \Gamma(c_r)\setminus v} \tanh(1 \setminus 2\mu_{v_s \to c_r})\Big) \quad (14)$$

4. Variable check:

From each variable v_s for each $c \in \Gamma(u_S)$ is transmitted:

$$\mu_{v_s \to c} = p_s + \sum_{c_r \in \Gamma(v_s)\setminus c} \mu_{c_r} \to v_s \quad (15)$$

5. A posteriori calculation:

For each symbol is calculated:

$$q_s = p_s + \sum_{c_r \in \Gamma(v_s)} \mu_{c_r \to v_s} \quad (16)$$

6. Stop/ continues:

If all check al passed $\prod_{v_s \in \Gamma(c_r)} q_s > 0$ (17)

then the operation ends being declared a success, else if the iteration limit is reached then the operation ends being declared a faiure, otherwise we must return to step 3.

Algorithm III is based on Gallager work [1]:

III. SOFTWARE SIMULATOR

In this paper we used CAMAG software to analyze the influence of LDPC codes in magnetic channels [6]. This in house build software is dedicated to read/write processes for multiple encoding/decoding processes, R/W parameters and interference noise. In our case we used CAMAG packages for LDPC codes with the above algorithms.

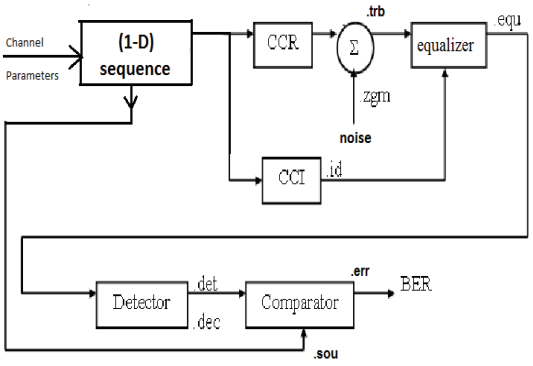


Figure 2. Block diagram of CAMAG[7]

The first block performs a precoding of data to reduce propagation of errors that could occur in detection. The formula for precoding is: (1+D2) mod 2 for PR4 channel; (1-D+D2+D3) mod2 for EPR4 channel and (1-2D+2D3+D4) mod2 for E2PR4 channel. For read/write head we used 1-D operator to simulate the characteristic differential [7].

We used for binary channels a implementation of algorithms for binary channels that are based on the sum-product algorithms and low-density pseudo-random irregular matrices based on M. Neal Radfors's programs [8].

Creating a matrix can be done by in ways: first one is based on randomly distribution of the same amount of 1s between columns, the second one similar, but also try to make rows having the same number of 1s. The software can create only R=1/2 LDPC matrix. Parity bits generating is done using spare matrix properties of **H = [A|B].**

For decoding LDPC code is used either iterative belief propagation, either sum-product algorithm (SPA). We tested in this paper Hard-Decision (bit-flip) decoder.

For decoding 0/ 1 message, if the majority is 1 is choose 1, else is choose 0. This is a simple implementation can be an introduction to message passing algorithms.

Others algorithms for decoding that derivate from SPA algorithm are Log-Domain SPA decoder and Simplified log-domain SPA decoder.

IV. RESULTS

We used for our test a matrix with ½ rate which has *M*=500 and *N*=1000, where *M* is number of rows and *N* is number of columns, recording density is 3. The LDPC matrix was created with the second method where we have approximately the same number of ones on each row. We considered *Eb/No* ration from 0 to 3. Input message is 1000 bits long and BER calculation includes parity bits, for brevity.

In first chart we made a comparison between PR IV magnetic channel and binary channel.

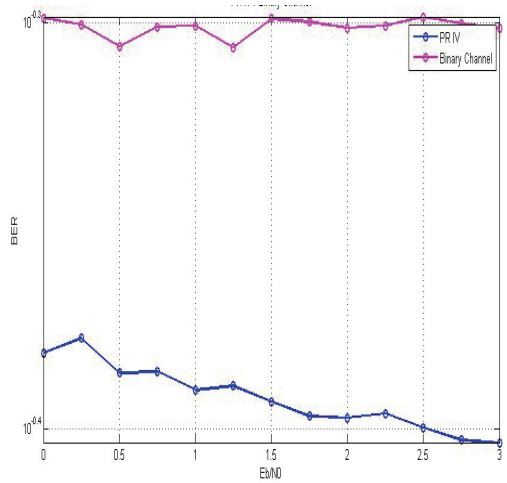


Figure 3. PR IV magnetic channel and binary channel comparison

In the second chart is the comparison between EPR IV magnetic channel and binary channel.

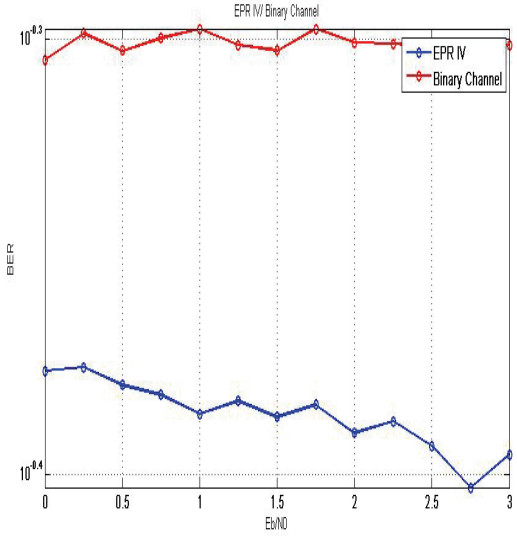


Figure 4. EPR IV magnetic channel and binary channel comparison

The third chart compare E2PR IV magnetic channel with the binary channel.

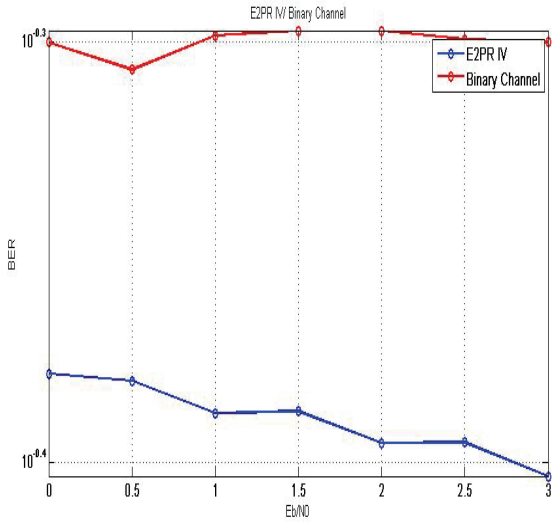


Figure 5. E2PR IV magnetic channel and binary channel comparison

V. CONCLUSION

We can observe that over scenario and parameters that we have tested, the difference between magnetic channels PR IV, EPR IV and E2PR IV and binary channel is not significant. BER obtained on the binary channel was in every scenario better than the BER obtained on magnetic channels.

We can see that the BER drops with the complexity of the channel but the cost for implementation represents an important criteria.

REFERENCES

[1] R. G. Gallager, Low-Density Parity-Check Codes. Cambridge, MA: M. I. T. Press, 1963.

[2] R. M. Tanner, "A recursive approach to low complexity codes," IEEE Trans, In- form. Theory, vol IT-27, pp. 533-547, Sept. 1981.

[3] D. J. C. MacKay, “Near Shannon limit performance of low density parity check codes,” Electron. Lett., vol. 33, pp. 457–458, Mar. 1997

[4] S. Howard**,** Two new LDPC decoding algorithms: Soft bit decoding and sub-min-sum decoding, in 3rd Analog Decoder Workshop, Banff, Canada, 2004.

[5] David J. C. MacKay, Good Error-Correcting Codes Based on Very Sparse Matrices, IEEE Transactions on Information Theory, Vol. 45, No. 2, pp. 339 – 431 March 1999

[6] Victor Soare, Ştefan Stăncescu, “Modular Partial Response Channel Software” MeleCon 2000, May 2000, pp. 786-789.

[7] St. Stancescu, C. Sandu, Random Interleaver Design in Turbo Coding for Digital Magnetic Recording Channel,Information & Communication Technology Electronics & Microelectronics (MIPRO), 2013 36th International Convention, pp. 137-139 May 2013

[8] Radford M. Neal, Software for Low Density Parity Check Codes Dept. of Statistics and Dept. of Computer Science, University of Toronto

Smart Home Automation System for Energy Efficient Housing

Nikša Skeledžija, Josip Ćesić, Edin Kočo, Vladimir Bachler, Hrvoje Nikola Vučemilo, Hrvoje Džapo
University of Zagreb, Croatia
Faculty of Electrical Engineering and Computing
Unska 3, Zagreb
Email: { niksa.skeledzija, josip.cesic, edin.koco, vladimir.bachler, hrvoje-nikola.vucemilo, hrvoje.dzapo } @fer.hr

***Abstract*—This paper presents a concept and implementation of modern smart monitoring and control system for building automatization. The system is designed to enable significant reduction of energy consumption and carbon footprint by increasing the energy efficiency of the building under control. The system consists of a *Linux*-based remotely accessible main embedded control unit, a custom designed programmable logic controller named *littlePLC*, and a propriatery low-power Wireless Sensor Network (WSN). The energy flow is optimized by using a Model Predictive Control (MPC) algorithm that runs on the main control unit. The main control unit communicates with *littlePLC*, which serves as an interface that controls the parameters and state of HVAC systems in the building. The feedback information for MPC is gathered by means of the WSN, which consists of various sensor node types, such as temperature, air pressure, humidity, VOC and CO_2. The WSN nodes are connected in a star type network topology, with a communication HUB connected to the main control unit. The information gathered by WSN are used in the MPC algorithm in order to calculate and estimate the requirements for heat corrections, with respect to ventilation and weather predictions.**

I. Introduction

One of the main concerns today is energy generation, consumption and preservation. Since the reserves of fossil fuels are limited, the use of alternative energy sources need to be encouraged. Large part of the engineering community considers solar energy being one of the most perspective alternative sources, whether directly by using solar heat and light or indirectly by using the wind, tidal changes and similar effects generated by the Sun's emitted energy. Apart from considering sustainable sources, it is also important to improve the efficiency at which the energy is consumed.

The focus in the research community is to reduce energy wasted in residences in order to improve global energy efficiency. Residential buildings consume a major part of all produced electrical energy, in particular, the residential and commercial sectors use almost 40% of overall energy used in the world, while in European countries 76% of energy goes towards comfort control in buildings and houses – heating, ventilation and air conditioning (HVAC) [1]. According to some other studies, at least 20% of total energy produced is consumed in residences [2].

On one hand, today's architecture often puts an emphasis on passive methods for improving energy efficiency. On the other hand, an approach that is not as widespread in residential buildings is active control of humidity, ventilation or air conditioning systems. By using such active monitoring and control approaches, significant improvements in energy efficiency can be achieved. Such smart buildings can produce energy, consume it, store it for later use and even return it to the electrical grid [3]. Having efficient distribution of the produced and stored energy between HVAC systems, a building can achieve a positive influence on the network, meaning that it requires zero energy from the electrical grid [4].

Generally, there is a large and increasing demand for integrated system solutions that would facilitate a design of energy efficient buildings by employing smart and adaptive control techniques. These systems must provide means for advanced HVAC monitoring and control, sensory data gathering, system interoperability, flexiblity, blending into ecosystem of mobile and web tehnologies etc. Large amount of work has been undertaken in the field of smart home monitoring and control and various approaches for building systems for energy efficient buildings has been proposed [5]–[7]. There are also numerous of proprietary commercial closed system solutions. In this research we propose a modular, open system architecture for smart automation of energy efficient home, optimized for low-cost, flexibility and interoperability. The proposed implementation was built from the state-of-the-art open hardware and software components, as well as custom designed electronic modules, that provide an optimal price-performance ratio and that can be easily adapted for various usage scenarios. The system enables easy customization, both in terms of adding new software functionality or additional hardware peripheral extensions. An important part of the system solution is Model Predictive Control (MPC) framework, which is the key to energy efficient smart home usage scenario.

The paper is organized as follows. The Sec. II brings the general idea of the home automation system through a short overview. The Sec. III briefly describes a commercial embedded processor as an important element of the smart house. The advanced control approach used for controlling such house, namely the MPC, is given in Sec. IV. WSN developed by the team is shown in Sec. V. A custom made PLC named *littlePLC* is presented in Sec. VI. Finally, the conclusion and future work are given in the last section.

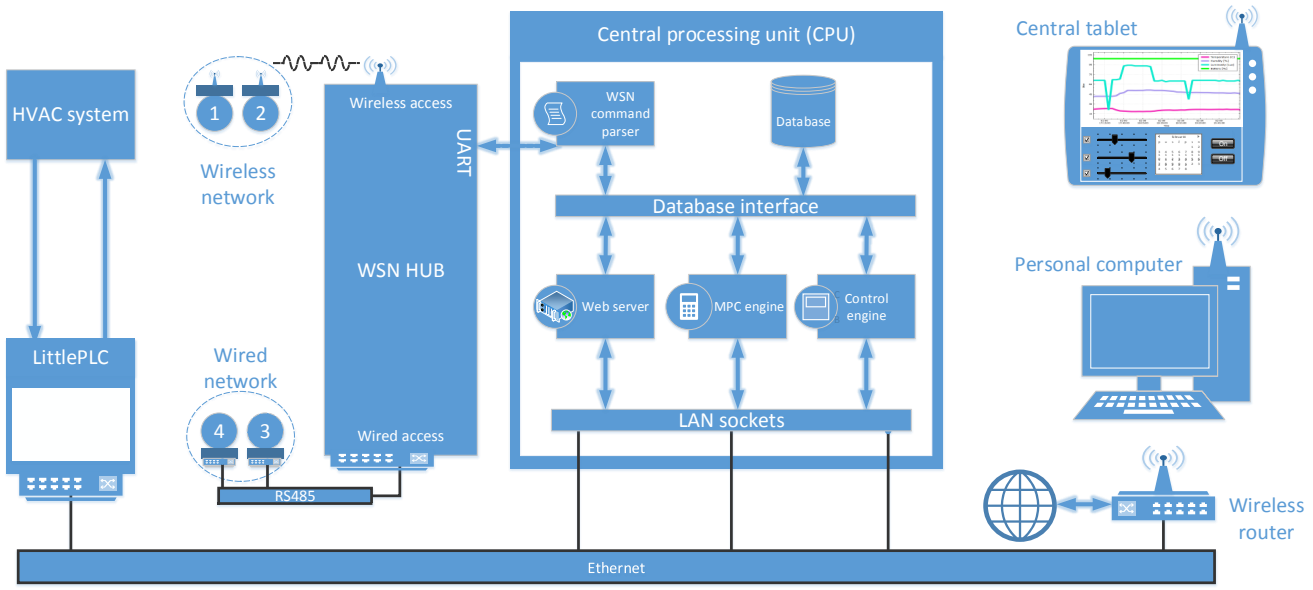


Fig. 1: Block diagram of the system

II. System overview

The overall system structure shown in Fig. 1 can be logically divided into four subsystems:

- Central Processor Unit (embedded processor)
- Model Predictive Control algorithm (*MPC*)
- Programmable Logic Controller (*littlePLC*)
- Wireless Sensor Network (*WSN*)

The CPU provides the necessary computing resources that need to run the control algorithm. The MPC algorithm uses a mathematical model of the building, input from the WSN and predictions of future conditions from the weather forecast service to efficiently maintain good living conditions in the home. The WSN has a star topology. It consists of a HUB connected to the main processor and nodes scattered around the building. Nodes provide various measurements such as temperature, air pressure, humidity, etc. The (*littlePLC*) controls the actuators of the system, namely a heat pump for heating and cooling. It was designed as an alternative to commercially available PLCs while paying heed to the industries standards.

III. Central Processor Unit

The CPU chosen for the project is *BeagleBone Black*. It runs the MPC algorithm and interfaces towards each of the subsystems. An embedded processor was chosen for its small size and power consumption [8]. A *Linux* distribution is installed on the board, allowing for a number of advantages, such as using script based programming languages (*i.e. Python*) or database programming languages (*i.e SQL*). Thread management provided by an operating system is also necessary, since the system will have multiple tasks implemented in separate processes. The processor gathers data from the WSN via a custom hardware plug-in (HUB). The HUB is designed to fit as a cape plug-in for the *BeagleBone Black*. Communication is achieved such that the HUB sends standardised data packages over serial communication which are then written into a database on the *Beaglebone*. Data gathered from the WSN serves as input for the MPC algorithm.

IV. Model predictive control

In the vein of previously presented general idea of sustainable, affordable and energy efficient concept of the house, energy management stands in the focus of the concept. Thus, an optimal control system approach which respects constraints resulting from desired occupant comfort as well as actuators' constraints was to be used. Such requirements on optimal energy consumption were satisfied using MPC approach as suggested in recent works [9]–[11].

The general idea of MPC is to use a model of the process as well as the weather forecast (i.e. outdoor temperature, insolation, humidity, etc.) in order to predict the future evolution of the system and to compute a control actions which ensure the fulfilment of demands. In particular, a custom made heat pump serves as a controlled actuator, while insolation and outdoor temperature are considered as disturbances within such model. Ventilation system available in the house also contributes notable heating flows and therefore is included into computation of demands on the actuator.

The basic concept of MPC is to use a dynamic model, constraints of the process and current measurements to forecast system behaviour and to produce the optimal decision according to performance criteria – the control move at the current time. Therefore, the control action is obtained by solving a finite horizon optimal control problem at each sampling instant in which the initial state is the current state of the house. The result of this optimization

is a finite control sequence and only the first action in this sequence is applied to the plant. A block diagram for this approach is given on Fig. 2.

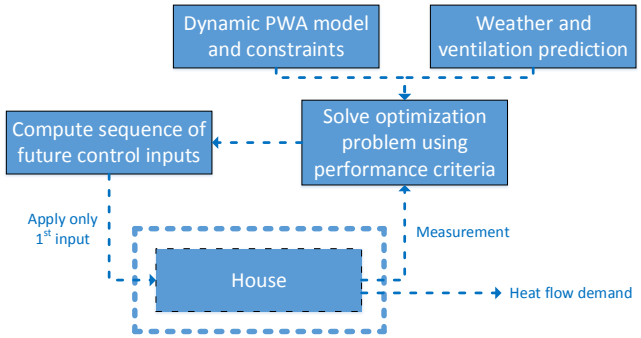


Fig. 2: Block diagram of the MPC control approach

A. MPC framework

Apart from typical materials that are often used in buildings and houses, recent studies have proved the efficiency of Phase Change Material (PCM) which has a dominant non-linear behaviour [12]. Consequently, the model of the house which arises from this fact can not be well described as pure linear one. Still, the characteristics of PCM can be well described using a piecewise affine (PWA) model, which brings the same formulation of the model of the house.

Aforementioned PWA time invariant dynamic system is given with the following continuous system:

$$\dot{x}(t) = \begin{cases} A_1\, x(t) + B_1\, u(t) & \text{if } \delta_1 = 1, \\ \quad \vdots & \\ A_s\, x(t) + B_s\, u(t) & \text{if } \delta_s = 1, \end{cases} \tag{1}$$

where $\delta_i \in \{0, 1\}$, $\forall i = 1, .., s$ are variables satisfying the condition

$$\sum_{i=1}^{s} \delta_i = 1, \tag{2}$$

at the particular moment in time. The previous representation of the model arises in gainscheduling, where the linear model is switched among a finite set of models, according to changes of the operating conditions. δ_i can be considered being a function of state of the system as well as input variable. In principle,

$$[\delta_i = 1] \leftrightarrow [x; u] \in \mathcal{C}_i \tag{3}$$

where $\mathcal{C}_i$ is a subspace of the space of feasible x and u.

The model given in (1) represents a continuous time model, which is discretized using zero-order hold discretization approach with 1 h sampling time in order to be used within the discrete MPC, with states x and inputs $u = [u^a\ u^d]$. Input u^a corresponds to heat demands for controlled actuator, namely the heat pump and u^d corresponds to non-controlled disturbances. The MPC takes into account the predictions $U^d = (u^d_{t|t}, u^d_{t+1|t}, ..., u^d_{t+N-1|t})^T$ obtained from the weather forecast service and determines the controlled input $U^a = (u^a_{t|t}, u^a_{t+1|t}, ..., u^a_{t+N-1|t})^T$ on the horizon N minimizing a cost function (subscript t represents a discrete moment). This approach also respects constraints of the actuator and can be formalized as follows:

$$\begin{aligned} &\min_{U^a} \textstyle\sum_{k=0}^{N-1} \|u^a_{t+k|t}\| \\ &\text{s.t. } u^a_{t+k|t} \in \mathcal{U}^a \end{aligned} \tag{4}$$

where $\mathcal{U}^a$ represents the feasible space for the actuator. As stated previously, the MPC problem also takes into account the limitations on states of the system. Namely, $x^a_{t+k|t} \in \mathcal{X}$, where $\mathcal{X}$ represents the limitations of the system states (both spaces $\mathcal{U}^a$ and $\mathcal{X}$ are constrained using linear inequalities). In principle, space $\mathcal{X}$ inherently possesses limitations on the temperature inside the house. The remark regarding previous discussion is given in the following subsection. Detailed solution to such formulated problem is given in [13].

B. Implementation remarks

The mathematical model of the house used within this project was gained in the vein of [14], where the model of the house is generated through suitable RC network. The state vector x contains the voltage levels on corresponding capacitances, which corresponds to temperatures of particular wall or room. Such gained mathematical model was consisted of 40 states.

The first step towards implementation of the MPC algorithm was to identify the mathematical model of the system (1). For this reason, an accurate model of the house was built using a commercial software for dynamic multi-zone simulation application to study thermal indoor climate conditions. The response of such model was thus used for validation of mathematical representation of the house developed in Matlab. This control approach assumes the accuracy of the mathematical representation of the house. Therefore, some identification technique still needs to be developed to be used on-sight, since the house changes its characteristics in time.

In order to have well estimated states of the house a wireless sensor network was built. More details on it follow in the next section.

V. Wireless Sensor Network

As discussed in previous section, the CPU requires various measurements as feedback for the control algorithm. Wireless approach was chosen since it requires no additional wiring. The wireless nodes are self contained and can be placed around the house without impairing its visual appeal. The sensor nodes send measurements to the WSN HUB, which then relays the data to the CPU. The advantage of such a topology is that the idle radio times of the battery powered nodes are significantly reduced. The HUB is connected to a fixed power supply and its radio frequency (RF) connection is online perpetually. Each node has a sleep interval in which the device powers down and consumes as little electrical energy as possible. The node wakes up to send its

measurements and, for a short period it listens for incoming traffic from the HUB. This allows for the HUB to send commands to the node, i.e. command for setting the sleep interval duration. The node spends an short amount of time in receive mode if it is not called out by the HUB. This allows for quick transition back to the power down mode, saving energy and extending battery life.

Power consumption was reduced by using low power microcontroller units (MCU). The chosen processor for the wireless nodes is the *PIC18F26K22* microcontroller. It interfaces the measurements from the sensors placed on the node and communicates with the RF transceiver. Generally, the *PIC18F* series of microcontrollers offers an extremely low-power management technology suitable for reducing the energy consumption of sensor nodes.

Even though most basic sensor nodes are designed to be wireless, some sensors require a wired connection. For example, most CO_2 sensors require a warm up period during which the heater on the sensor sets the adequate temperature for gas percentage readings, whence the power consumption duty cycling is not possible. Another example of such a sensor is a smoke detector. The appearance of smoke requires an instant detection for safety reasons. In such situations, active time cannot be duty cycled with sleep modes and the battery drain is increased drastically. A wired solution (RS485 communication with a 12 V power supply) is proposed as an alternative in this kind of situations.

A single sensor node consists of an MCU board, containing either a regulated battery power supply or a 12 V supply line, a processing unit and a communication interface. The sensor unit is a plug-in board. The block diagram of a sensor node is shown on Fig. 3. Sensor node duty cycling and estimated

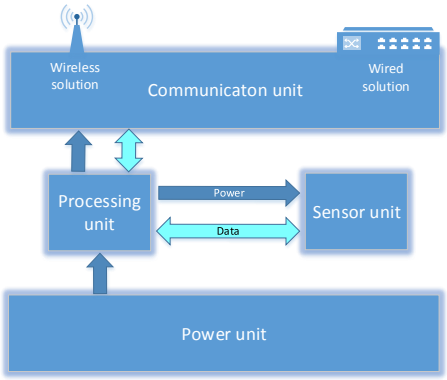


Fig. 3: Sensor node diagram

current drain are given on Fig. 4, where T_{mc} is a measurment collection time, T_{mr} is a measurement report time, T_{nc} is node configuration time and T_s is sleep time. Active time is sum of T_{mc}, T_{mr} and T_{nc}.

Voltage from the battery is regulated using a low drop-out linear regulator. The batteries selected are Li-Po cells as they have a high energy density and provide enough energy to achieve long up times. Battery lifetime can be roughly estimated with omitted self discharge current as:

$$t_{up} = \frac{E_b}{I_{on}\,\delta + I_{sleep}\,(1-\delta)}, \tag{5}$$

where E_b represents battery capacity, δ stands for duty cycle coefficient (share of active time in overall time), while I_{on} and I_{sleep} represent average current in active and sleep mode, respectively.

Example: Using a 1000 mAh Li-Po battery with a duty cycle of 0.01% while assuming active current of 20 mA and sleep current of 1 μA yields $t_{UP} = 38.05$ years.

The transceiver integrated circuit chosen for the wireless sensor nodes is *nRF24L01+* [15]. It is often used in ultra low power wireless node designs [16], [17]. The chosen integrated circuit is a RF transceiver operating in the 2.4 GHz frequency band. It was chosen since it provides multiple useful features, such as automatic packet handling, 6 data pipes, power down modes, etc.

The sensing units are different for every node type. Sensor boards can be designed independently since the plug-in header predicts the usage of standard serial interfaces (e.g. SPI, I2C, UART). Proposed sensor nodes include a temperature measurement node, an air pressure and humidity measurement node, an ambient light sensor node, a gas and CO_2 detector.

VI. LITTLEPLC

The PLC is a standard component of industrial automatization. It is a device dedicated to controlling electrical machines (i.e. motors, compressors, pumps, heaters, etc.), monitoring system operation, and providing a control interface to the supervising personae. There are many commercially available PLCs, most of them developed for highly demanding tasks of industrial control. Using such a device for control of a house is not financially viable. A small PLC was designed to significantly reduce the cost of the whole system of automatization.

LittlePLC was designed for use in milder environments. It does not have an industrial grade temperature operating range, nor is resistant to an aggressive atmosphere, but it provides a large advantage over a standard PLC's processing power. It was designed to provide a platform for easier implementation of complex control algorithms with a lower price than industrial equipment.

An ARM based MCU is the main processing unit of *LittlePLC* (in particular the *STM32F407* microcontroller [18]). The device was designed as a modular system that allows the variation of the number of inputs and outputs using expansion modules. Similarly to standard PLC's, *LittlePLC* provides analog and digital inputs and outputs (Fig. 6). All inputs, outputs and communications with the outside world are galvanically separated from the microcontroller to ensure reliable operation and to protect the internal circuitry. There are several ways of communicating with external devices or users (RS485, Ethernet, GSM, SD card, as shown on Fig. 6). The *littlePLC* is shown on Fig. 5.

LittlePLC can be divided into following subsystems:

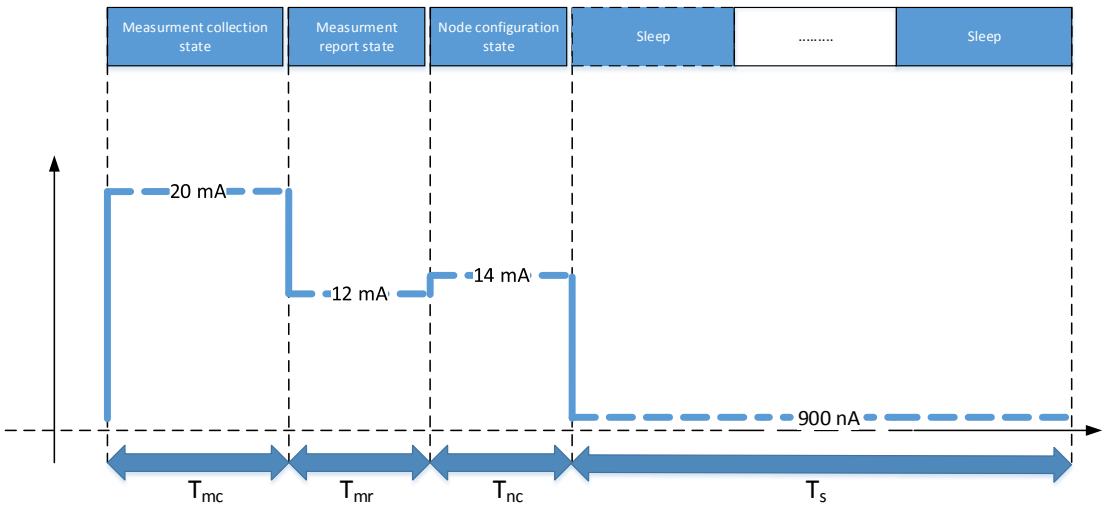


Fig. 4: The operation cycles of the sensor node and provides estimated currents for each state

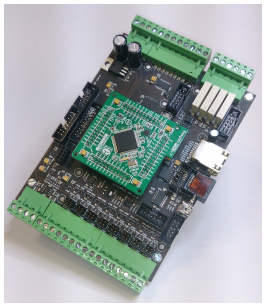


Fig. 5: littlePLC

- Motherboard (*MCU, real time clock, SD card interface, etc.*)
- Communication layer (*GSM, ethernet and USB communication modules*)
- Control layer (*analog and digital outputs and inputs, RS485 standard bus*)

In the system described in this article, *littlePLC* communicates with the CPU via Ethernet. The CPU issues HVAC control commands to the *LittlePLC*, based on the MPC algorithm results. *LittlePLC* then issues controls to the HVAC systems. (i.e heat pump, condensing heater, ventilation, air conditioning, etc.) Any errors that may appear during operation are reported to the CPU. The main advantages of using *LittlePLC* for HVAC systems control are lower price, and increased processing power.

VII. Conclusion

In this paper we presented an architecture and implementation of a modular smart automation system for energy efficient housing that was optimized for low-cost, flexibility of system configuration, and interoperability with other technologies. The core of the proposed system consists of the Central Processor Unit (CPU), Wireless Sensor Network (WSN), Programmable Logic Controller (*littlePLC*), and Model Predictive Control algorithm of the building HVAC system (MPC). The loosely coupled system framework enables easy adaptation for various specific usage scenarios, both in adding the new hardware system components, and the additional software services. *Linux*-based CPU promotes an open system approach design which enables adding a new functionality as new system process that integrates with rest of the system through the dedicated IPC messaging protocol. The WSN was optimized for energy efficient smart home monitoring usage scenario, and it provides low cost multifunctional sensor nodes, with ultra low power consumption and simple messaging protocol, without an unnecessary overhead often met in similar existing solutions. Furthermore improvements in lowering the overall system costs and in providing the extra features usually met in high-end industrial PLCs have been achieved by means of the custom developed *littlePLC*. Finally, the MPC control algorithm plays a key role in a system that enables significant energy efficiency improvements.

All described system components were tested in a prototype implementation of the smart home automation system. The initial development phase encompassed the design and implementation of hardware and software components of CPU, WSN and *littlePLC*, which were put together to form a basic core of the proposed system. The experimental tests showed that the proposed architecture and components were well chosen both in terms of achieving the low cost of the overall solution and satisfactory system performance.

The future work will encompass the installation of the described system into the actual smart building, assessment of system components performance under realistic working conditions, adjustment of the MPC algorithm to the realistic house model, and experimental tests under the conditions of real-world environment. The most interesting part of the future research will be the answer how much energy savings can be achieved by using the proposed system, what is

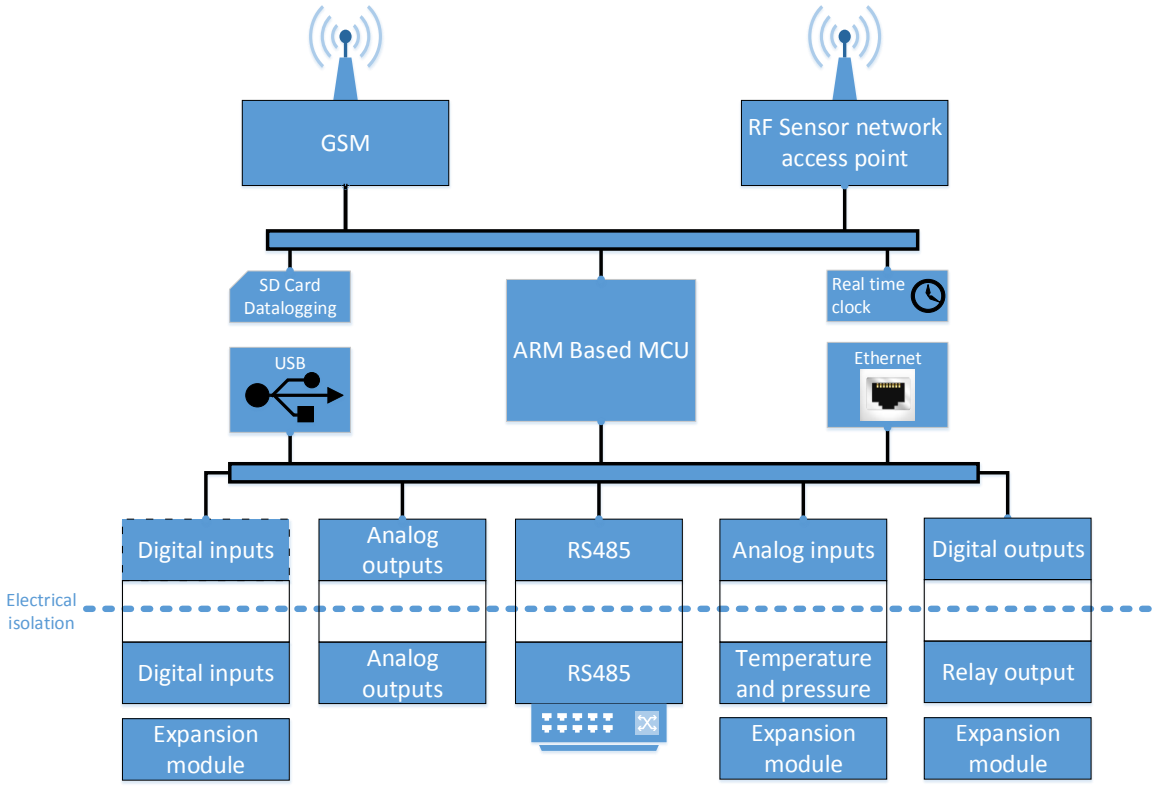


Fig. 6: Block diagram of the littlePLC

impossible to answer beforehand, but also difficult to assess without the custom developed system for this purpose. Beside the basic system core described in this paper, we plan to extend the basic system functionality by providing additional modules and services, such as smart entry control subsystem, human behaviour patterns learning and monitoring, remote access system control, configuring and controlling system configuration over smartphone or tablet etc.

ACKNOWLEDGMENT

We would like to thank to the members of the UNIZG team of the University of Zagreb for their collaboration on *mem**brain*** project (*http://www.membrain.com/*).

REFERENCES

[1] "Energy efficiency policies for new buildings," *International Energy Agency, Energy Efficiency Requirements in Building Codes*, 2008.

[2] S. Basuroy, J. W. Chuah, and N. Jha, "Making buildings energy-efficient through retrofits: A survey of available technologies," in *Power and Energy Society General Meeting (PES), 2013 IEEE*, July 2013, pp. 1–5.

[3] B. Morvaj, L. Lugaric, and S. Krajcar, "Demonstrating smart buildings and smart grid features in a smart energy city," in *Energetics (IYCE), Proceedings of the 2011 3rd International Youth Conference on*, July 2011, pp. 1–8.

[4] B. Todorovic, "Towards zero energy buildings: New and retrofitted existing buildings," in *Exploitation of Renewable Energy Sources (EXPRES), 2011 IEEE 3rd International Symposium on*, March 2011, pp. 7–14.

[5] D.-M. Han and J.-H. Lim, "Smart home energy management system using ieee 802.15.4 and zigbee," *Consumer Electronics, IEEE Transactions on*, vol. 56, no. 3, pp. 1403–1410, Aug 2010.

[6] K. Gill, S.-H. Yang, F. Yao, and X. Lu, "A zigbee-based home automation system," *Consumer Electronics, IEEE Transactions on*, vol. 55, no. 2, pp. 422–430, May 2009.

[7] H. ElKamchouchi and A. ElShafee, "Design and prototype implementation of sms based home automation system," in *Electronics Design, Systems and Applications (ICEDSA), 2012 IEEE International Conference on*, Nov 2012, pp. 162–167.

[8] "Beaglebone black system reference manual," 2013, revision A4.

[9] A. E.-D. Mady, G. Provan, C. Ryan, and K. N. Brown, "Stochastic model predictive controller for the integration of building use and temperature regulation." in *25th AAAI Conference on Artificial Intelligence*, 2011.

[10] F. Oldewurtel, A. Parisio, C. Jones, M. Morari, D. Gyalistras, M. Gwerder, V. Stauch, B. Lehmann, and K. Wirth, "Energy efficient building climate control using stochastic model predictive control and weather predictions," in *American Control Conference (ACC), 2010*, June 2010, pp. 5100–5105.

[11] F. Oldewurtel, A. Parisio, C. N. Jones, D. Gyalistras, M. Gwerder, V. Stauch, B. Lehmann, and M. Morari, "Use of model predictive control and weather forecasts for energy efficient building climate control," *Energy and Buildings*, vol. 45, no. 0, pp. 15 – 27, 2012.

[12] P. A. Mirzaei and F. Haghighat, "Modeling of phase change materials for applications in whole building simulation," *Renewable and Sustainable Energy Reviews*, vol. 16, no. 7, pp. 5355 – 5362, 2012.

[13] A. Bemporad and M. Morari, "Control of systems integrating logic, dynamics, and constraints," *Automatica*, vol. 35, no. 3, pp. 407 – 427, 1999.

[14] M. Vasak, A. Starcic, and A. Martincevic, "Model predictive control of heating and cooling in a family house," in *MIPRO, 2011 Proceedings of the 34th International Convention*, May 2011, pp. 739–743.

[15] "nrf24l01+ single chip 2.4ghz transceiver, product specifications," 2008, version 1.0.

[16] S. Sonavane, V. Kumar, and B. Patil, "Designing wireless sensor network with low cost and low power," in *Networks, 2008. ICON 2008. 16th IEEE International Conference on*, Dec 2008, pp. 1–5.

[17] D. Zhao and C. Peng, "A small low-power reliable communication module in a wireless monitoring system," in *Bioinformatics and Biomedical Engineering, 2007. ICBBE 2007. The 1st International Conference on*, July 2007, pp. 1194–1197.

[18] "STM32 32bit arm cortex mcu - STM32F407 datasheet," 2013, rev 4.

Remote Laboratory for Testing Processor Cores in FPGA Device

Kristjan Saksida and Andrej Trost
University of Ljubljana, Faculty of Electrical Engineering, Ljubljana, Slovenia
Andrej.Trost@fe.uni-lj.si

Abstract – In the paper we present generic hardware verification structures for efficient testing of the custom processor cores in FPGA devices. Hardware verification environment consists of hardware debug structures: control machine, memory access port, input and output data buffers and interface logic. The verification structures are integrated and synthesized with the processor core under test and implemented on the FPGA device. The structures can be customized and easily integrated in the hardware development flow. A software support for the hardware verification consists of a data acquisition driver running on the server with HTML5 graphical interface. The software enables either local testing or setup of a remote laboratory for testing of the processor cores. The application of the remote laboratory in the educational process is presented. The presented hardware verification structures are optimized for testing the soft programmable processor cores and are vendor independent. The software support is based on open languages and protocols and the scripting tools enable quick customization. We present advantages of our solution compared to commercial general purpose on-chip logical analyzers. The benefit of our approach is that it can be used with the standard programmable design tools on the low cost platforms and provides two abstraction levels of debugging.

I. INTRODUCTION

Field Programmable Gate Array (FPGA) devices are used for building Systems-on-Chip (SoC) composed of processor cores and custom logic [1]. There are a lot of hard and soft general purpose processor cores available for the system designers and the capacity of modern FPGA devices enables implementation of many cores in a single device. Application specific circuits and systems can be optimized using custom designed processor cores. The custom processor core development is also part of the digital systems educational process [2].

Software simulation of a complex processing core is time consuming process and typically only some basic tests are performed. Comprehensive tests are executed in hardware using implemented processor core and hardware verification structures. Hardware verification on the FPGA device is much faster and more accurate. A drawback of hardware verification is limited visibility and control of the internal circuit operation. The FPGA vendors provide embedded logical analyzers to enhance visibility of the internal signals and provide some basic control [3]. A full visibility can be provided using scan chains [4] and better control by integrating test bench in the hardware [5].

We are focusing on testing processor cores used in FPGA devices for performance non critical tasks as a replacement for complex state machine logic. The processor core can be tailored to specific application requirements and the task of testing is to eliminate bugs in the design process. The design error can be either in the core logic, the processor software or in the interface logic. For the efficient debugging process it is important to provide tools covering various levels of abstraction [6]. The logical analyzer tools for hardware debugging cover only the lowest hardware abstraction level. In the paper we focus on the specific hardware verification structures for testing embedded processors used in the programmable devices.

Programmable devices are used in digital electronics education process and several remote laboratories were developed for supporting distance learning [7-10]. Programmable development boards are connected to a computer used for design implementation, board configuration and hardware testing. The remote laboratories provide server environments for distant control of the design and hardware verification process.

In the paper we present general requirements for testing embedded processor cores, setup of a remote digital design laboratory and our remote processor testing environment. The processor testing remote laboratory is used in the educational process.

II. EMBEDDED PROCESSOR TESTING REQUIREMENTS

In order to provide generic hardware verification structures for testing different processing cores, we identified common features of soft processor cores in FPGA devices.

A hardware verification environment should cover at least two abstraction levels. In the low abstraction level, the important signals should be sampled in order to detect processor logic and bus protocol related bugs. In the higher level, data transactions on the processor data bus should be sampled and initiated in order to emulate processor environment. Specifically, peripheral bus transactions should be saved in internal buffers for control and analysis of the processor running a test program.

A. Embedded Processing System

An embedded processing system in FPGA device consists of (a) a processing core, (b) program memory and (c) interfaces. Small processors used for control tasks often

execute a bare-metal application saved in memory blocks inside the FPGA device. A peripheral bus is used for communication with the interfaces and the rest of the system. We can summarize testing requirements for the various parts of the system:

a) Hardware verification interface should provide control of the processor core by driving reset and clock signals. The important processor control signals should be sampled with a logical analysis module.

b) The verification interface should provide access to the program memory in order to load test programs.

c) The verification interface should be connected to the peripheral bus emulating transactions to the external interfaces.

B. Educational Embedded Processor Core

In order to demonstrate the embedded processor testing system, we use a small generic processor architecture presented in [11]. The processor core and a program memory are described in a hardware description language VHDL. The core architecture supports 4 to 32-bit data bus and is optimized for FPGA devices. The core design is used as a soft CPU template in the educational process. Only a basic set of instructions are given in the initial design and the students are encouraged to extend the instruction set with other general and application specific instructions.

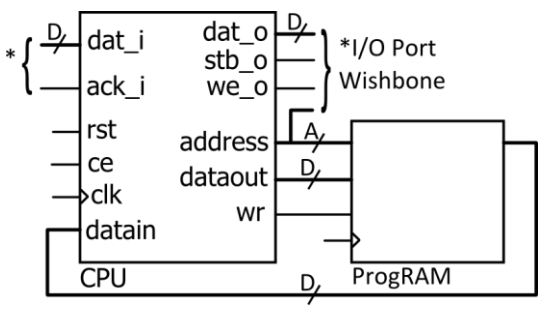


Figure 1. Educational embedded processing system

Block diagram of the educational embedded processing system is presented in Figure 1. The system is composed of a CPU core and a program memory. Processor control signals are: clock (clk), reset (rst) and clock enable (ce) used for stopping the processor during test in order to enable long time verification with limited hardware buffer resources [12]. This signal can be replaced by dedicated clock resources and clock multiplexers in the hardware test structure.

The generic processor architecture provides D-bit data bus and A-bit address bus. The processor peripheral interface is a Wishbone compatible master input and output bus. Wishbone is an open-source synchronous bus architecture, which is commonly used in digital systems on a chip [13].

III. REMOTE LABORATORY SETUP

The remote laboratory software is based on open languages and protocols PHP, MySQL, HTML, JavaScript and JSON running on Apache server, as shown in Figure 2. Remote laboratory client side requires only a web browser which supports HTML 5 canvas object. The web page requires HTML version 5 due to the use of a canvas feature used for displaying waveforms of the sampled data.

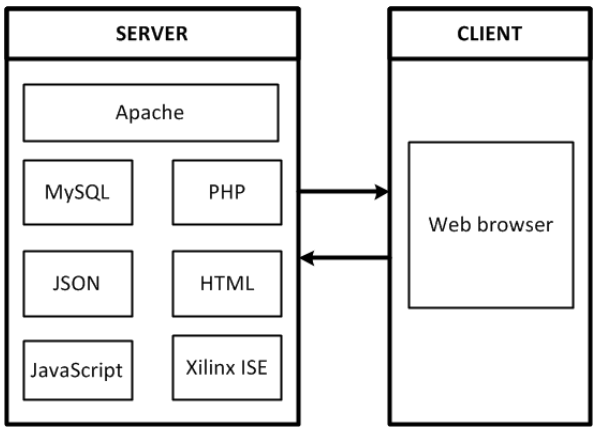


Figure 2. Software components of a remote laboratory

The Xilinx ISE tools for design synthesis, implementation and FPGA configuration file formatting are installed on the server. The PHP script calls individual tools of the Xilinx ISE enabling remote design implementation flow and programming of the FPGA device on the board.

The user on the client side needs an account, which enables him to log in to the server and get 30 minutes of a reserved laboratory session. The MySQL provides a base of users and user login feature.

Server side PHP script sends instructions and data from the web page through the serial port to the board and collects verification output data. The received data is decoded and sent to JavaScript in JSON format. On the client browser the JavaScript routine draws signal waveform into the canvas area.

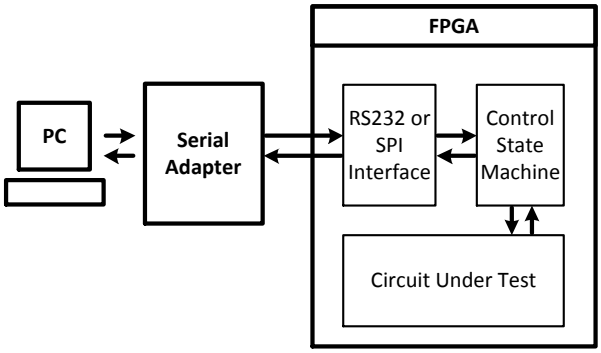


Figure 3. Hardware verification structures

The hardware of our remote laboratory is composed of a Xilinx Spartan-3E FPGA development board Basys2. The development board is powered through USB device configuration connector. A serial adapter on expansion connector is used for serial data connection to the host computer. Figure 3 shows hardware verification structures used for testing circuits on the development board.

The circuit under test is connected to a control state machine (CSM) providing input and output registers and control signals. The input and output register length is adjustable using generic VHDL parameters. When testing embedded processors, the control state machine enables writing and reading to the processor memory. The width and the depth of the memory can also be adjusted in 4-bit slices covering most of the processor bus width configurations. The CSM is connected to serial interface, asynchronous RS232 at 115200 baud or synchronous SPI at 3 Mbit/s. The selected serial interfaces provide enough bandwidth for our testing purpose and occupy a small area in the FPGA device. Commands on the serial interface are used for reading and writing the registers and the processor memory.

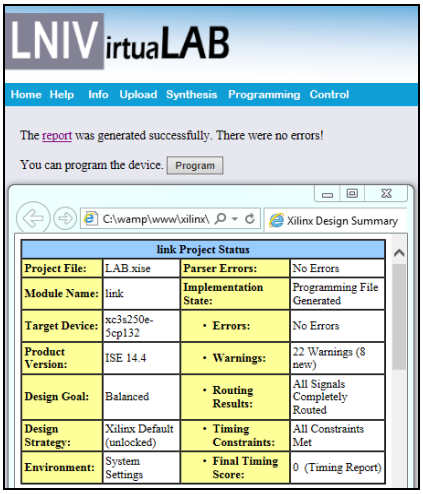


Figure 4. Client view of the virtual laboratory

A generic hardware test structure is available to the user which connects his circuit under test, uploads VHDL files to the server and runs remote implementation process. When the process is finished, a report provided from the tools is available as presented in Figure 4 and the user can program the board and start hardware verification.

IV. Embedded Processor Testing System

During development of the educational remote laboratory, we realized the need for more specialized processor verification structures. The specialized processor testing system components are presented in Figure 5. The design in programmable device contains three top level components: a serial verification interface (IF), a control state machine (CA) and a test wrapper (TEST).

The test wrapper includes the CPU core and processor specific hardware verification structures. The processing system program memory component is replaced with memory access port (MAP). The memory access port contains control logic and additional memory bus, used for accessing memory during test. A selected set of processor control signals is connected to trace data buffer (TDBUF). The TDBUF is similar to the logical analyzer component used for low abstraction level signal sampling and debugging. The processor peripheral input (pin) and output (pout) Wishbone bus are connected to data buffers IDBUF and ODBUF. The data buffers are designed as FIFO memory blocks holding the data for the higher abstraction level debugging. The buffers occupy only a small portion of FPGA resources. Table 1 summarizes data buffers area utilization for different data width and buffer depth of 31 locations. Figure 6 shows almost linear Look-Up Table (LUT) utilization depending on a data buffer depth for a fixed data width of 12 bits.

TABLE I. FPGA UTILISATION OF 31 LOCATION DATA BUFFERS

Data Width	Slices	FF	LUT	LUT %
8	19	13	70	1
12	34	14	90	1
16	40	14	110	2
32	72	16	192	3
64	134	16	350	7

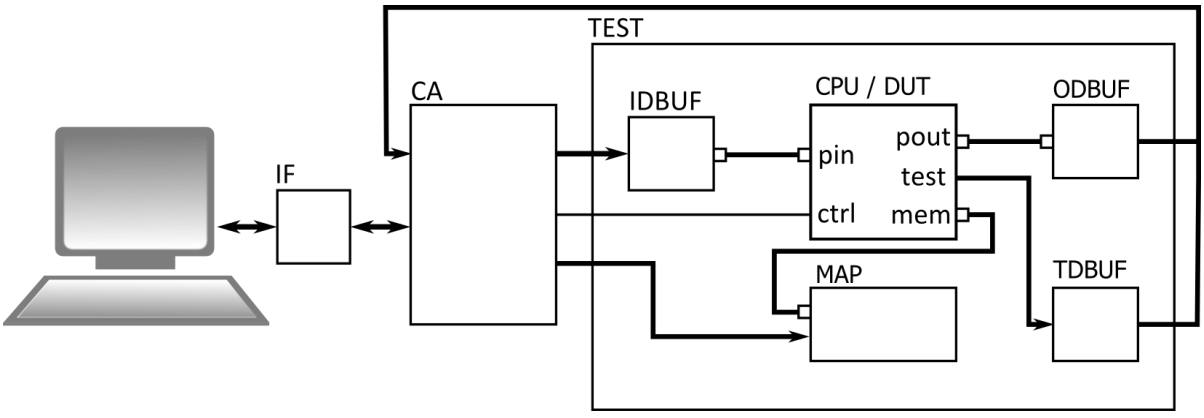


Figure 5. Embedded processor testing system components

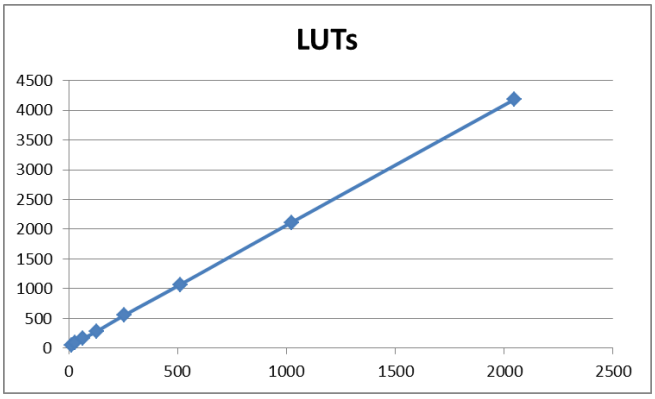


Figure 6. Data buffers LUT utilisation depending on the buffer depth

The control state machine requires 7% and serial UART interface 5% of the available LUT in 100k gate Spartan-3E device. A simplified state diagram of the control state machine is presented in Figure 7. The system starts in the state Idle and the serial commands change the state to:

- Prog – writing program data to memory access port,
- Set – setting the input data buffer,
- Read – reading output data buffer,
- ReadT – reading trace buffer,
- Reset – resetting the CPU core,
- runN – running the CPU for specified number of cycles or until breakpoint and
- runT – running the CPU until trace buffer is full.

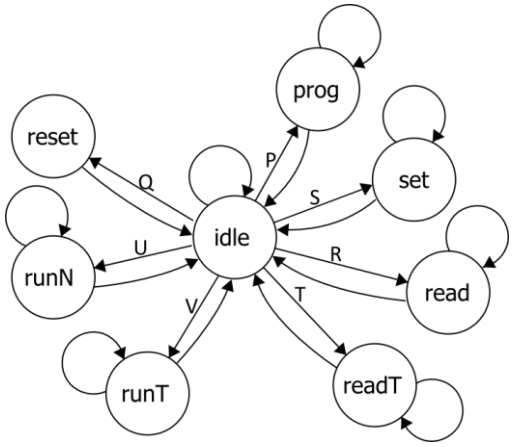


Figure 7. Simplified state diagram of the control state machine

Peripheral data bus can be used for debugging with a breakpoint. One of the peripheral addresses is used to trigger break in the run cycle. This setup requires almost no additional hardware resources and provides convenient means of high abstraction level processor debugging.

A. Educational Remote Laboratory

For testing the educational processing system, we set an educational remote laboratory OpenLAB. The goal of the OpenLAB is remote testing of the processing core designed at the digital integrated circuits course. The processor testing flow in OpenLAB is presented in Figure 8.

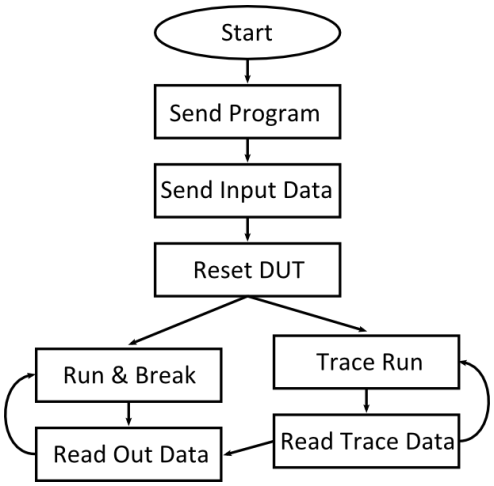


Figure 8. OpenLAB processor testing flow

The user is first required to send a test program to the processor already implemented on the board. The code should be in raw hexadecimal format. Then some input data is set on the web page to be stored to the input FIFO buffer (IDBUF) and the processor is reset.

The low abstraction level test is initiated with trace button. The processor is run until trace buffer is full and the output data is presented on the signal waveform. The higher level of test is performed by running the program for desired amount of cycles or to the breakpoint and reading out the data. The system allows consecutive execution of both levels of testing as many times as required by the test program.

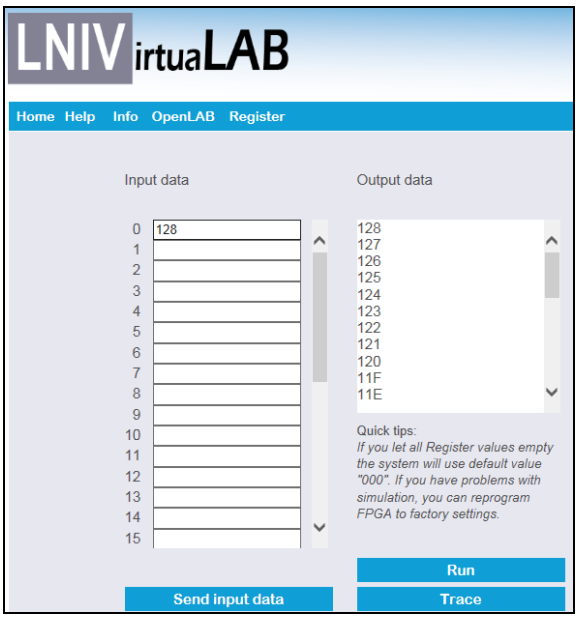


Figure 9. OpenLAB part of the web page

Figure 9 shows part of the OpenLAB web page for testing educational 12-bit processing system. There are 31 data cells for the input data buffer and listing of the output data buffer in 3 digit hexadecimal format. There is also an input window to enter the program code, which is transferred to RAM after pressing the Program button. A Factory settings button is used for complete board reset and loading the initial test code.Upon pressing Run button, the processor is enabled for the preset number of cycles and the processed output data is displayed in the output data area. Figure 9 shows an example output of a small program reading one input value (128), counting down and sending the results to output bus.

The waveform on the Figure 10 presents the sampled processor signals:

- data input (data_in) and output (data_out) bus,
- address bus (adr),
- execution state (st), memory control (wr) and
- Wishbone port control signals (stb, we).

We can observe behavior of listed signals through 32 clock cycles. The processor is stopped after the trace and by pressing the trace button again we are able to sample next 32 cycles, so the trace length is not limited to the hardware buffer size.

V. CONCLUSIONS

In the paper we presented hardware verification structures and remote laboratory used for testing embedded processor cores. Testing a complex hardware on the programmable development board requires implementation of the hardware verification interfaces. The FPGA vendors provide logical analyzer modules for low level in-circuit testing. These modules are designed for specific FPGA families and are not included in the free versions of the tools.

We initially designed a generic hardware verification remote laboratory, where device under test is connected to the generic input and output registers connected to the test environment. Testing the embedded processors with such environment proved to be tedious and error prone work and a more specialized environment was required for the successful implementation in the educational process. We identified common features of the soft embedded processors and designed generic processor testing components. The Wishbone bus used for peripheral connections is a common SoC bus structure and bridges are available for other processor busses.

The hardware components enable testing on two levels of abstraction which is more effective for understanding the core operation and eliminating the design bugs. By using the presented technologies we are able to implement an educational remote laboratory focused on the debug process. Our further work will focus on the test system automation and testing on even higher abstraction levels.

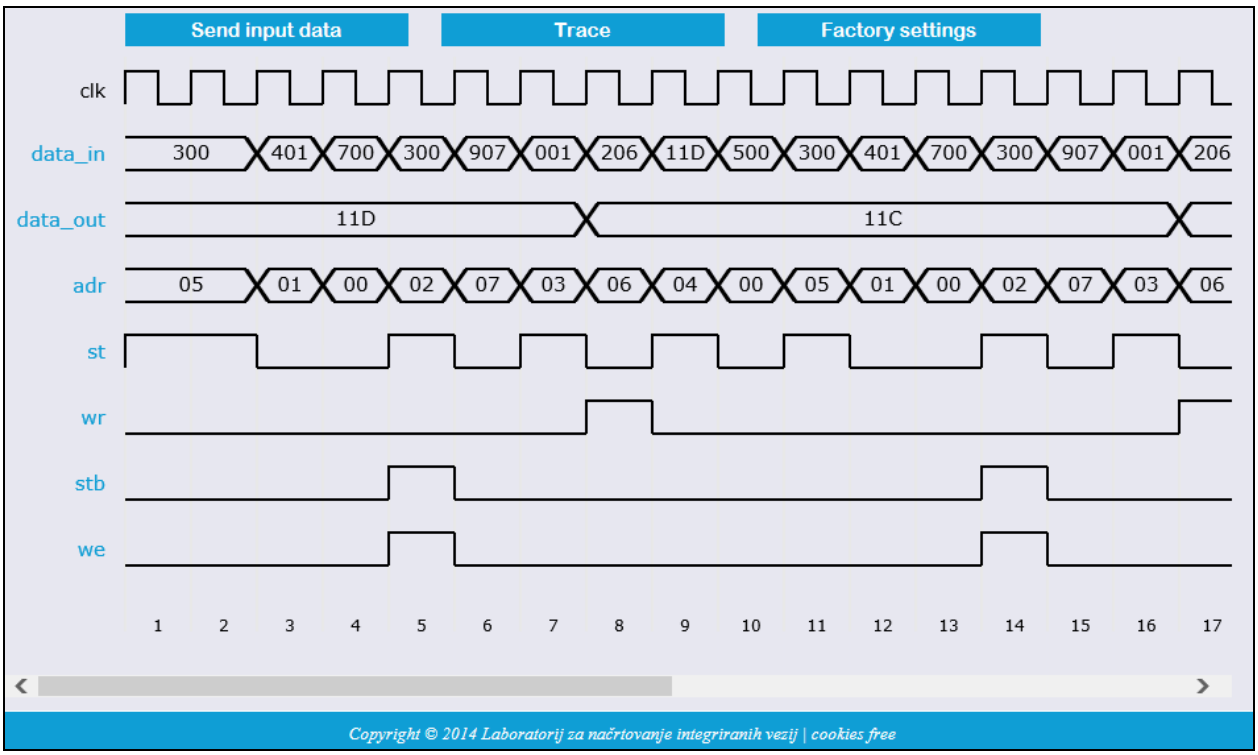


Figure 10. Signal trace waveform

REFERENCES

[1] W. Wolf, FPGA-Based System Design, Prentice Hall, New Jersey, 2004

[2] D. Hanna, R. E. Haskell, "Learning Digital Systems Design in VHDL by Example in a Junior Course", Proceedings of the ASEE North Central Section Conference, Charleston, West Virginia, March 2007.

[3] Xilinx, ChipScope Pro 14.1 Software and Cores User Guide, Xilinx, Inc., 2012

[4] I. Mavroidis, I. Mavroidis, I. Papaefstathiou, "Accelerating Emulation and Providing Full Chip Observability and Controllability at Run-Time", IEEE Design & Test of Computers, vol. 26 (6), pp. 84-94, 2009

[5] S. Banerjee, T. Gupta, S. Gupta, "Logic emulation with forced assertions: A methodology for rapid functional verification and debug", 5th Asia Symposium on Quality Electronic Design, ASQED 2013, pp.312-320, August 2013

[6] Roesler, E.; Nelson, B., "Debug methods for hybrid CPU/FPGA systems," *Field-Programmable Technology, 2002. (FPT). Proceedings. 2002 IEEE International Conference on* , vol., no., pp.243,250, 16-18 Dec. 2002

[7] C. Quintas et all, "Digital electronics learning system based on FPGA applications", Frontiers in Education Conference, Indianapolis, USA, 2005

[8] M. Drutarovský, J. Šaliga, I. Hroncová, "Hardware Infrastructure of Remote Laboratory for Experimental Testing of Fpga Based Complex Reconfigurable Systems", Acta Electrotechnica et Informatica, vol. 9 (1), pp. 44-50, 2009

[9] D. Pop et all, "An Online Lab to Support a Master Program in Remote Engineering", Frontiers in Education Conference, Rapid City, USA, 2011

[10] S. Sitnikov et all, "Remote Access Laboratory for Analog and Digital Electronic Course", Eurocon 2011 Proceedings, Lisbon, Portugal, April 2011

[11] A. Trost, A. Žemva, "Design of Custom Processors for the FPGA Devices", Elektrotehniški vestnik, vol. 79 (1-2), pp. 55-60, 2012

[12] J. Viejo, J. I. Villar, J. Juan, A. Millan, E. Ostua, J. Quiros, "Long-term on-chip verification of systems with logical events scattered in time", Microprocessors and Microsystems, vol. 36 (5), pp. 402-408, 2012

[13] WISHBONE, Revision B.4 Specification, 2010, http://opencores.org/opencores,wishbone

Towards an energy efficient SoC computing cluster

Z. Krpić*, G. Horvat**, D. Žagar ** and G. Martinović*
*Department of Software and Computer Engineering, Faculty of Electrical Engineering Osijek, Croatia
**Department of Communications, Faculty of Electrical Engineering Osijek, Croatia
e-mail: krpic@etfos.hr

Abstract – High performance computing (HPC) systems are omnipresent in processing huge amounts of data. However, power consumption of these systems is becoming far from negligible, forcing software developers and hardware providers to deviate from usual performance plans, and to address power consumption and costs. As a result, small-scale applications are becoming infeasible to run on those systems, making room for the power-efficient HW newcomers, such as SoC (System on Chip), to take that role. The idea of composing a large number of small SoCs in a computing cluster can result in a powerful, green, and a simple platform capable of delivering enough performance for the small applications while maintaining low energy consumption and costs. In this paper we present a system composed of four LAN-interconnected dual-core ARM based SoC devices, powered by a single power supply. We investigate applicability of such a system by monitoring its performance and energy efficiency and stability while using common HPC paradigms: message passing and thread programming.

Keywords: energy efficiency, SoC, computer cluster, HPC, performance, stability

I. INTRODUCTION

In today's technologically advanced society energy is becoming a very expensive and scarce resource, making energy efficiency as one of the most dominant and omnipresent issues in virtually all existing computing systems. In some systems the total energy costs of running can exceed the initial investment in only few years, signifying the need to drastically reduce power consumption of the existing systems [1, 2].

On the other hand, High Performance Computing (HPC) systems are characterized by the ability to perform very complex tasks very effectively performance-wise. The energy efficiency of these systems is not a primary concern [1, 3], although "performance no matter what" approach is slightly fading away, a trend which was started by the "Green Destiny" project 2001 [12]. New metrics appeared to re-evaluate HPC computing, such as number of floating point operations per watt consumed starting a completely new era of power efficient HPC. The green aspect of the HPC systems is popularized the Green500 list [13], currently topped by the machine capable of achieving roughly 4500 MFLOPS/W. Still, the scale of current HPC systems is still too large, making their use for the small-scale applications infeasible. Smaller and more end-user oriented parallel applications along with the importance of the energy awareness in the future computing systems [4] imply that new steps have to be undertaken in the design of energy efficient HPC systems.

The emergence of the new architectures (e.g. see [7]) opens up new possibilities regarding small-scale HPC computing, expanding domain of HPC use, and starting a stream of new approaches. One of the approaches in the increase of HPC energy efficiency by using small, low cost, low power ARM based devices as the compute cluster nodes [5, 8, 10, 11]. Following the trend of using ARM processors in HPC clusters in this paper we propose an ARM based cluster composed of 4 dual core A20 SoC (System on Chip) using a Cubieboard2 embedded system (Fig. 1).

The Cubieboard2 is a SoC based on ARM Cortex A-7 dual core processor running on 1GHz, containing 1 GB of DDR3 memory and 4GB of internal NAND flash. From communication point of a view the Cubieboard2 includes 10/100 Ethernet and USB 2.0 ports. Maximum power consumption of the board is set at 2A at 5VDC input (10W). On the operating system side is supports various UNIX based operating system. We have tested performance, energy efficiency and stability of our proposed platform in order to evaluate its applicability as a small-scale HPC machine [6, 9].

The rest of the paper is organized as follows: Section II describes the test platform and metrics used within the paper. In Section III we show performance results of our platform, alongside with platforms power and stability

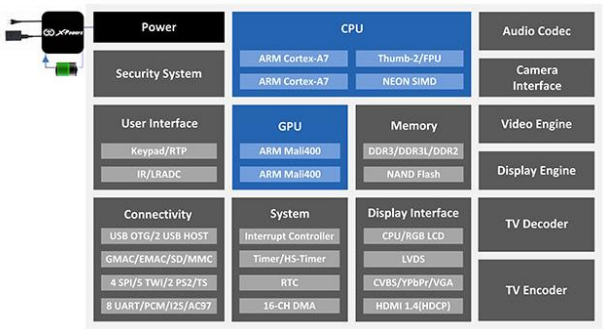


Figure 1. Block diagram of A20 processor [6]

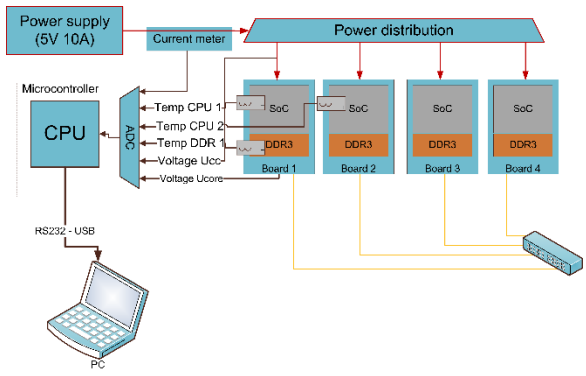


Figure 2. Proposed measurement testbed

aspects. Finally, Section IV concludes the paper along with the plans for our future work.

II. MEASUREMENT TESTBED

Testbed platform used for measurements consists of a small-scale computing cluster composed of four SoC based boards connected to a power supply through a power distribution hub. Constant voltage is controlled using a digital power supply set at 5.0V, maximum available electrical current is 10A. The boards are interconnected via 10/100BASE-T Ethernet Switch operating at full-duplex speed, used for remote access and command exchange between boards. The platform setup (measurement testbed) is shown in Fig. 2.

As various parameters affects the performance, stability and power consumption of SoC based systems the proposed testbed also incorporated three temperature sensors, two voltage sensors and a Hall-effect based electrical current sensor to monitor the effect of these parameters on the aforementioned parameters. In order to digitally represent data Atmel's AVR microcontroller was used to perform analogue to digital conversion and sample the data in regular intervals. The functionality of the measurement testbed is divided into performance, power and temperature components.

A. *Performance measurement*

The test algorithm is a distributed matrix multiplication

$$\mathbf{X} = \mathbf{M} \times \mathbf{N}, \quad \mathbf{M}, \mathbf{N} \in \Re^{n \times n}, \tag{1}$$

which is a large scale SIMD operation consisting of add and multiply operations. Many of HPC benchmarks, such as NPB (NAS Parallel Benchmarks) [14] rely on highly optimized routines for matrix-vector and matrix-matrix product from parallel libraries such as PBLAS and SCALAPACK making the aforementioned operations to be considered as a widely-accepted HPC benchmark. The complexity of (1) is $O(n^3)$, although there are more efficient methods, such as generalized Coppersmith–Winograd algorithm, having a complexity of $O(n^{2,3727})$, [15]. Nevertheless, the goal is to put a high computational load to the processing units and their interconnection, making (1) as a practical option. The total number of operations is considered to be $2n^3$, with each iteration having one multiplication and one addition. We ignore other operations, such as loop iterators, since their number is not significant to the total number of operations.

The algorithm has been parallelized using message passing interface (MPI) using C++ compiled with a g++ compiler

B. *Electrical current measurement*

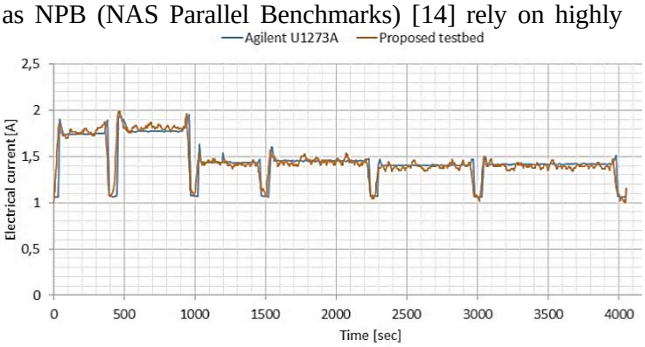
The process of electrical current measurement is a process of outmost importance in analyzing power efficiency of SoC systems. In order to create self-sustaining system capable of real-time measuring and data delivery towards PC an ACS712 Hall-based electrical current sensor was used to convert electrical current value to voltage value. The transfer function of a sensor can be approximated as:

$$I[A] = (U - U_0) \cdot k, \tag{2}$$

where U is the output voltage of a sensor, U_0 is the bias voltage at I = 0A (can be represented as $U_{CC}/2$) and k is the slope of a transfer function.

The voltage value of a Hall sensor is fed towards a 10bit ADC integrated within the microcontroller. After ADC conversion the voltage value is converted to electrical current value and sent towards PC at regular intervals using a UART communication.

As the error in electrical current measurement must be lowered to a minimum the proposed testbed is compared with Agilent U1273A True RSM Multimeter to estimate the amount of the introduced error from the proposed measurement testbed. The process of measurement validation was performed using a logging Multimeter (Agilent U1273A) and the proposed testbed. The result of a measurement validation are shown on Fig. 3 and Fig. 4.



Figue 3. Measurement validation of the proposed testbed (electrial current)

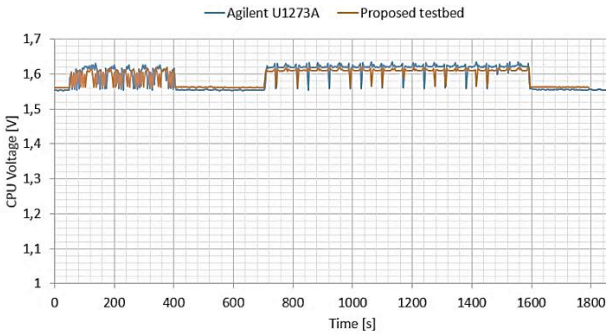


Figure 4. Measurement validation of the proposed testbed (voltage)

As seen from Fig. 3-4 the values obtained from the proposed testbed are in correlation with the data from the Agilent multimeter. The oscillations of the testbed data are due to the higher sampling frequency and the fact that the Agilent Multimeter uses a low-pass filter to filter the high frequency data. Also, the overall error of the measurement (if Agilent multimeter is used as a reference meter) is below 5%. Main reason of using the proposed testbed for electrical current measurement is the ability to distribute current sensing among multiple SoC boards, enabling the measurement of individual power consumption for SoC systems within a cluster.

C. *Voltage measurement*

Second part of the system consist of two voltage probes used to measure voltage values within the system. The most important voltage is the power supply voltage that tends to vary upon different types of load. The voltage measurements have been conducted using a voltage divider connected to an ADC input. Depending on the maximum voltage measured the voltage divider is constructed accordingly. As the voltage of a power supply is important in calculating the consumed power the voltage measurement are validated using a logging Multimeter (Agilent U1273A). From the results it can be concluded that the proposed testbed error is less than 3% compared to reference meter.

D. *Temperature measurement*

Last of the testbed analysis incorporates temperature measurements of SoC cores and memory. We have chosen to measure three temperature values: the temperature of CPU core for SoC Board1 and Board2, respectively, and the temperature of DDR3 memory for SoC Board1. The temperature measurement is performed using analog LM35 temperature probe that outputs linear voltage proportional to temperature. The transfer function of a temperature sensor can be approximated as:

$$T[°C] = U \cdot k_t, \tag{3}$$

where U is the output voltage of a sensor, k_t is the slope of a transfer function $(k_t = 100°C/V)$. Thus temperature values can be digitally represented for each temperature sensor. To enhance the precision of the measurements each sensor was calibrated using Agilent U1273A and K-thermocouple. Accuracy of the temperature sensors can be estimated within the range of +/- 0.2 ^{0}C.

III. EXPERIMENTAL RESULTS

Similar to the classification of the testbed setup in Section II we classify the measurement results of our platform. The first measurement shows performance aspects of our cluster, making the initial insights of its usability, and setting base for other measurements. Secondly, we analyze the power consumption and its energy efficiency, and finally we consider potential stability issues using temperature measurements during specific load profiles.

TABLE I. PERFORMANCE TEST PROFILES

Profile no.	Platform configuration		Software configuration
	Number of nodes (boards)	Total number of cores	Type of operations
1.	4	8	FLOP
2.	4	8	IOP
3.	4	4	FLOP
4.	4	4	IOP
5.	2	4	FLOP
6.	2	4	IOP

A. *Performance results*

Since the platform consists of low power nodes (SoC Boards), the performance is not the main objective aimed by this platform. However, considering that other measurements rely on performance (e.g. power efficiency) we first present and discuss the platform computational power. The tests are divided into 6 different profiles, shown by Table I. We used 2 types of software configurations to stress out the possible differences regarding different types of the load, one using integer operations (IOPs) and one using floating point operations (FLOPs). Both software configurations were conducted on three different hardware configurations: 4 nodes full load (8 cores), 4 nodes half load (4 cores total, load balanced) and 2 nodes full load (4 cores) to analyze different computation and communication setups.

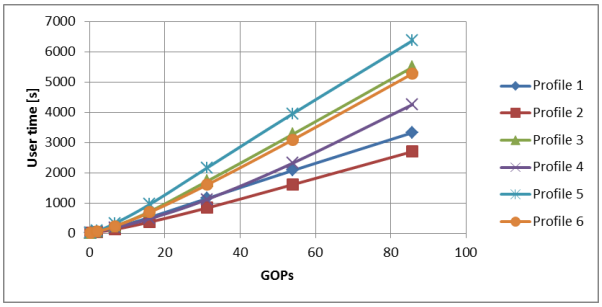
We observed user time (total execution time)

$$t_{user} = t_{comp} + t_{comm} \tag{4}$$

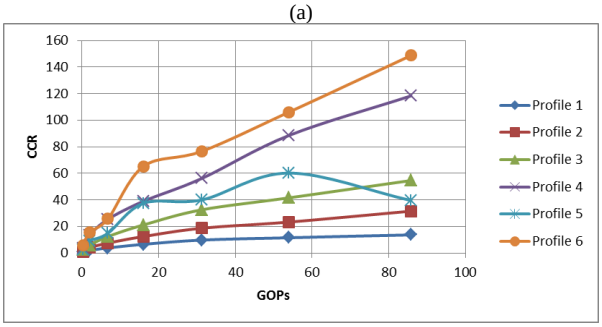
and computation to communication ratio (CCR)

$$CCR = \frac{t_{comp}}{t_{comm}}, \tag{5}$$

since those metrics illustrate the most important performance output of our platform. Measurements of both metrics are given in Fig. 5. User time shows good scalability of the proposed platform with the problem size increase (Fig 5a), although load balanced profiles (3rd and 4th) tend to deviate from this rule because of the bigger

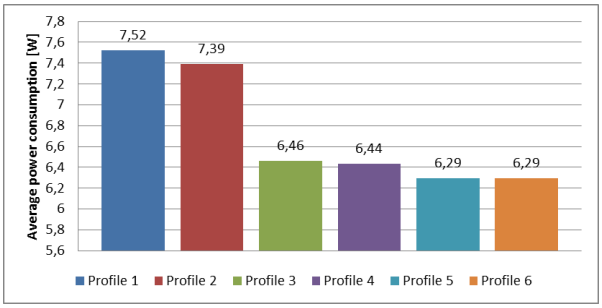


(a)

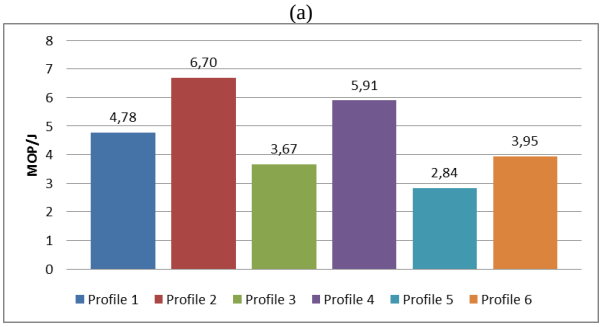


(b)

Figure 5. Performance results: (a) user time, (b) computation to communication ratio



(a)



(b)

Figure 6. Power consumption: (a) average power consumption, (b) energy efficiency

overheads. However, load balanced profiles show better user time results than full node load due to the smaller number of cache misses and light shared memory access of the former. The best configuration performance-wise is shown to be the 8-core profiles, achieving up to 90 MIPS (mega integer operations per second) and 75 MFLOPS (mega floating point operations per second) which is almost three times the performance of a 60W 2.4GHz Pentium 4 desktop CPU tested in [16] (achieving roughly 33 MIPS). On the other hand, the CCR metric (Fig. 5b) is in favor of the profiles 5 and 6, regardless of their lowest user time performance, while 8-core profiles are showing the worst results (because of the network congestion increase). The latter implies that by putting tasks in close configurations could be used for software with the smaller computational load with significant communication demands. Profile 5 results shows the network bottleneck for more complex and intensive loads, making sudden drop of CCR beyond 50 GOPs implying the network congestion beyond this point, where the increased computational performance cannot compensate for the communication overheads.

B. *Power efficiency*

We calculated the cluster instantaneous power consumption using current and voltage measurements explained in Section II.B and Section II.C respectively. The power was calculated using:

$$P_{avg} = U \cdot I, \tag{6}$$

which enables the calculation of the profile power efficiency by:

$$\mu = \frac{n_{OPs}}{P_{avg}}. \tag{7}$$

The platform power efficiency is expressed in number operations per second per watt (OP/Ws, or more commonly millions of OP/Ws, MOP/Ws), which is equivalent to the operations per Joule (OP/J, MOP/J), as a common energy efficiency metric of computing systems used in [12, 13, 16]. The true value of our tested cluster lies in its small physical size, and its low power consumption which rarely peaked 8 W. Fig. 6a shows the average power consumption of the test profiles from Table I. The highest power consumption occurred during test profile 2, indicating that floating point operations apply more stress to the platform than the integer ones. The smallest power stress comes from the profiles 5 and 6 due to their low communication traffic, and the fact that 2 nodes are not used at all (their memory and peripherals are idle). Additionally, profiles 5 and 6 show the smallest power consumption, but since their low performance cannot make up for it, that makes them the least energy efficient of all tested profiles. The overall energy efficiency of the proposed platform ranges from 2.8 to 6.7 MOP/J, making our platform up to 7 times more efficient than a 3.14GHz dual-core AMD PC platform (PP2) used in [16]. If we consider performance comparable to the power-hungry desktop processors, such a small-scale cluster makes an affordable alternative for the small-scale distributed processing.

C. *Stability*

Stability was tested by measuring temperatures during different types of varying loads, shown in Fig. 7. The Linux shell scripts were made to quickly alternate from load to idle, making short load bursts in order to stress the platform. The core time increase during stress is shown in Fig. 8. Additionally, we used 2 different burst types, the first with small run times of approximately 5 s (light load, longer period of time) and second with half a minute load (heavy load, shorter period of time), both interrupted with 2 s pauses. Bigger idles (5 min) were made between short-burst and long-burst tests in order to cool the platform, and to simulate some idle times in real applications. We recorded temperatures of the frontend node, T(node1), one

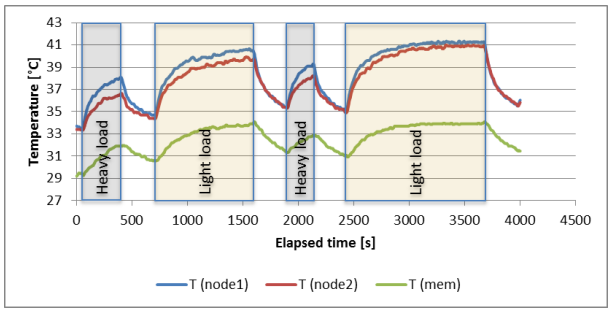


Figure 7. Stability, temperatures

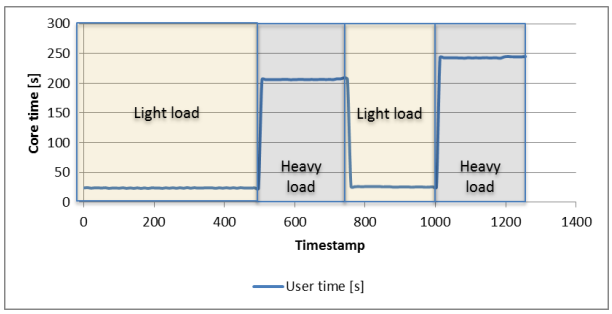


Figure 8. Stability, core time

worker node T(node2) and main memory T(mem). The temperature measurements were made according to the procedure from Section II.D under the ambient temperature of 24.6 °C.

Fig. 7 shows the temperature trace, making an obvious difference from idle time and load time on one hand, and light and heavy load on the other. The temperatures rise up to 41°C for the CPU and 34°C for the memory which is only 16.4°C and 9.4°C increase over the ambient temperature. Fig. 8 shows the behavior of the total user time during light and heavy load in the second experiment which shows no significant increase of the execution time during time (and the temperature stress). Only a small increase of the user time is visible towards the end of the execution, which is negligible. Therefore, the platform can be considered stable during varying load and temperature increase.

IV. Conclusion

This paper presents a low-power, energy-efficient HPC cluster based on four ARM SoC based boards. The usability of the proposed computational cluster is analyzed by testing the three most important metrics in the HPC domain: performance, power consumption and stability via 6 different test profiles.

In order to measure power consumption and other related parameters the testbed incorporates a microcontroller dedicated in analogue to digital conversion and sampling. The parameters monitored consist of three temperatures, two voltages and total electrical current of the boards. The proposed measurement testbed was compared with existing referential easement device and from the results it can be concluded that the error of the proposed measurement testbed is below 3% compared with reference meter.

Performance results show that our simple test cluster, although limited by its hardware, can deliver enough computational power for the small-sized distributed applications, with averages up to 90 MOP/s. Slight network bottlenecks occurred, but only during heavy loads due to the lack of communication optimization of the benchmark software.

Additionally, tested platform's power consumption rarely exceeded 8 W, with the highest averages of ~7.5 W and it energy efficiency reached up to 6.7 MOP/J (4.8 MFLOP/J) which makes it a perfect candidate for the energy-efficient small-scale computing.

The stability of the proposed platform was analyzed using temperature measurements under different load stress tests, showing no sign of problems, and keeping user times constant at all temperature levels. The slight increase of the user time occurred under high temperatures, but they were negligible.

Acknowledgment

This work was sponsored by the Ministry of Science, Education and Sports of the Republic of Croatia under project grant No. 165-0362027-1479, 165-0362980-2002 and Proof of Concept (BICRO) grand No. PoC5_12_2.

References

[1] N. Rajovic, A. Rico, N. Puzovic, C. Adeniyi-Jones, A. Ramirez, "Tibidabo: Making the case for an ARM-based HPC system", Future Generation Computer Systems, Available online 7 August 2013.

[2] D. Göddeke, D. Komatitsch, M. Geveler, D. Ribbrock, N. Rajovic, N. Puzovic, A. Ramirez, "Energy efficiency vs. performance of the numerical solution of PDEs: An application study on a low-power ARM-based cluster" Journal of Computational Physics, Volume 237, 15 March 2013, Pages 132-150.

[3] Z. Ou; B. Pang; Y. Deng; N., J.K.; Ylä-Jääski, A.; Pan Hui, "Energy- and Cost-Efficiency Analysis of ARM-Based Clusters," Cluster, Cloud and Grid Computing (CCGrid), 2012 12th IEEE/ACM International Symposium on , vol., no., pp.115,123, 13-16 May 2012

[4] R. V. Aroca and L. M. Garcia GonçAlves "Towards green data centers: A comparison of x86 and ARM architectures power efficiency" J. Parallel Distrib. Comput. 72, 12 (December 2012), 1770-1780

[5] Principi, E.; Colagiacomo, V.; Squartini, S.; Piazza, F., "Low power high-performance computingon the Beagleboard platform," Education and Research Conference (EDERC), 2012 5th European DSP , vol., no., pp.35,39, 13-14 Sept. 2012

[6] Cubieboard2 Technical specifications, available at: http://cubieboard.org

[7] K.R. Hoffman, P. Hedge, ARM Cortex-A8 vs. Intel Atom: architectural and benchmark comparisons. Technical Report. University of Texas at Dallas. 2009. On-Line. Access date: March 2012.

[8] G. Horvat, Z. Balkic, D. Zagar, "Real-time WSN communication for access control applications," Telecommunications and Signal Processing (TSP), 2013 36th International Conference on , vol., no., pp.321,325, 2-4 July 2013

[9] N. Rajovic, L. Vilanova, C. Villavieja, N. Puzovic, A. Ramirez, "The low power architecture approach towards exascale computing", Journal of Computational Science, Volume 4, Issue 6, November 2013, Pages 439-443, ISSN 1877-7503, http://dx.doi.org/10.1016/j.jocs.2013.01.002.

[10] Z. Lin, P. Chow, "ZCluster: A Zynq-based Hadoop cluster," Field-Programmable Technology (FPT), 2013 International Conference on , vol., no., pp.450,453, 9-11 Dec. 2013

[11] Yu Huang; Reddy, S.M.; Wu-Tung Cheng, "Core-clustering based SoC test scheduling optimization," Test Symposium, 2002. (ATS '02). Proceedings of the 11th Asian , vol., no., pp.405,410, 18-20 Nov. 2002

[12] M. Warren, E. Weigle and W-C Feng, "High-Density Computing: A 240-Node Beowulf in One Cubic Meter", Proc. ACM/IEEE High-Performance Networking and Computing Conference, Baltimore, MD, USA, 2002.

[13] The Green500 List, available at http://www.green500.org/, accessed February 2014.

[14] D. Bailey, "NAS Parallel Benchmarks," in Encyclopedia of Parallel Computing, D. Padua, Ed. Springer US, 2011, pp. 1254–1259.

[15] V. Williams, "Breaking the Coppersmith-Winograd barrier", unpublished manuscript, 2011.

[16] Z. Krpic; G. Martinovic; I. Crnkovic, "Green HPC: MPI vs. OpenMP on a shared memory system", Proceedings of the 35th International Convention MIPRO, 2012, pp. 246-250

Electronic interpretation of selected lessons in fuzzy mathematics

Lovre Perković*, Darko Kovačević *, Toma Karađole*, Mate Jović**
*Faculty of Maritime Studies, University of Split, Split, Croatia
**Technical School Imotski, Imotski, Croatia
dkovac@pfst.hr

Abstract - Dedicated electronic hardware has been used to interpret selected lessons in fuzzy mathematics. DeMorgans Laws were simulated in an electronic environment and that simulation was then used as a base to illustrate how the law of excluded middle is violated by fuzzy sets. In that context the problem of fuzzy grade of inclusion is briefly analysed and simulated electronically.

I. INTRODUCTION

The theory of fuzzy logic can be developed using the concepts of fuzzy sets similar to how the theory of crisp bivalent logic can be developed using the concepts of crisp sets. Specially there exist an isomorphism between sets and logic [2]. This means that a good foundation in fuzzy sets is necessary to develop the theory of fuzzy logic. On the other hand, in real practice, a number of fuzzy logic gates and circuits were developed to build more complex circuits. In this paper we are going to show how some elements of fuzzy mathematics can be interpreted in hardware environment.

In the second and third section of this paper we will deal with DeMorgans law(s) and use the obtained results to show some difference between crisp and fuzzy sets (law of excluded middle in fourth section).

Later (sections five and six) , we will show how a dedicated hardware detector can be used to indicate the grade of inclusion of two triangular fuzzy sets.

The detector of grade of inclusion is designed using *min-max-Op-Amp* technique [3] [4].The second idea of this paper is to design dedicated fuzzy hardware that can be used for different purposes as process modelling, control, etc.

II. UNITY SQUARE

A set (Cartesian product) $S_x \times S_y$, where

$S_x = \{x : x \in [0.1],\ \ x \in R\}$

and

$S_y = \{y : y \in [0.1],\ \ y \in R\}$,

is called the *unity square* and is denoted as S_1^2. Its (first) subset Δ consisting of ordered pairs, i.e. elements $(x, y = x)$ is called a diagonal of S_1^2 [1]. Its (second) subset $\overline{\Delta}$ consisting of ordered pairs, i.e. elements $(x, y = 1 - x)$ is called a transverse of S_1^2.

The diagonal can be seen (perceived) as a fuzzy set (1):

$$A = \{(x, y = \mu_A(x)) : x \in [0,1], \mu_A(x) \in [0,1]\}, \qquad (1)$$

where the function $\mu_A(x)$ is the corresponding membership function. In the very same way, the *transverse* can be seen (perceived) as a fuzzy set (2) :

$$\overline{A} = \{(x, y = \mu_{\overline{A}}(1-x)) : x \in [0,1], \mu_{\overline{A}}(1-x) \in [0,1]\} \qquad (2)$$

where $\mu_{\overline{A}}(1-x)$ is the corresponding membership function. It is obvious that these two membership functions are correlated over the purposed *unity square* (3):

$$\mu_{\overline{A}}(1-x) = 1 - \mu_A(x) \ \text{ for all } \ x \in [0,1] \qquad (3)$$

In the theory and practice of fuzzy sets the membership function $\mu_{\overline{A}}(1-x)$ is usually denoted as $\mu_{\overline{A}}$ and respresents the complement of the membership function $\mu_A(x)$.

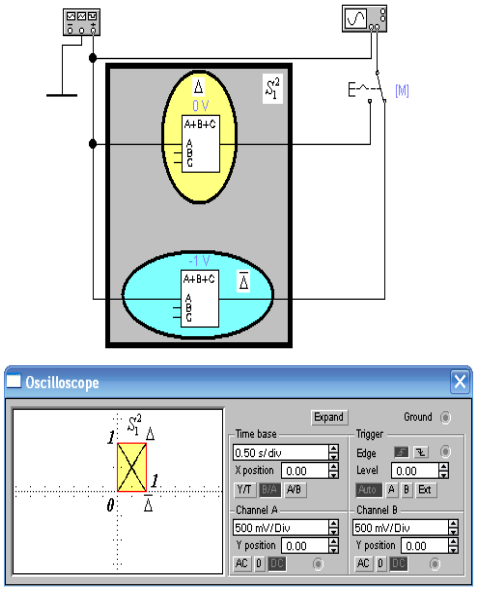


Figure 1. *Unity square*

III. SOME PROPERTIES OF FUZZY SETS; DEMORGAN'S LAWS

Consider fuzzy sets *A* and *B*, in the universe *X* , whose membership functions are given by terms (4 and 5):

$$\mu_A(x) = \min(1, \max(0, x)) \text{ for } 0 \le x \le 1 \quad (4)$$

$$\mu_B(x) = \max(0, \min(1, 1-x)) \text{ for } 0 \le x \le 1 \quad (5)$$

and shown in Fig. 2. Oscillographs a) and b) shown in Fig. 2 represent unlimited membership functions defined above, while oscilograph c) represents limited membership parts according to given relations. Their intersection is a fuzzy set (D) containing all the elements that are common to both sets, in a „fuzzy sense“. The membership function of the resulting set $A \cap B$ is given by (6):

$$\mu_D \equiv \mu_{A\cap B}(x) = \min(\mu_A(x), \mu_B(x)) \quad \forall(x) \in X \quad (6)$$

A graphic i.e., oscillograph (membership function $\mu_D(x)$) representation of the intersection of the two given fuzzy sets is shown in Fig. 3. Now, consider a fuzzy set *D* in the same *X*. Its complement is a fuzzy set whose membership function is given by (7):

$$\mu_{\bar{D}} \equiv \mu_{\overline{A\cap B}}(x) = 1 - \min(\mu_A(x), \mu_B(x)) \quad \forall(x) \in X \quad (7)$$

A graphic representation of the complement of the fuzzy set *D* is shown in Fig. 4. DeMorgan's laws are particularly useful in processing (simplifying) expressions of crisp sets (and logic), but they can be used in the „fuzzy domain“, as well. This means that the first DeMorgan's Law holds for our case as it is shown Fig. 4. One must know that, for the sake of simplicity, only straight line parts of given membership functions, i.e. lines $y = x$ and $y = 1 - x$, respectively were used to proof the following relation (8) (see Fig. 2 and 5):

$$\overline{A \cap B} = \overline{A} \cup \overline{B} \quad (8)$$

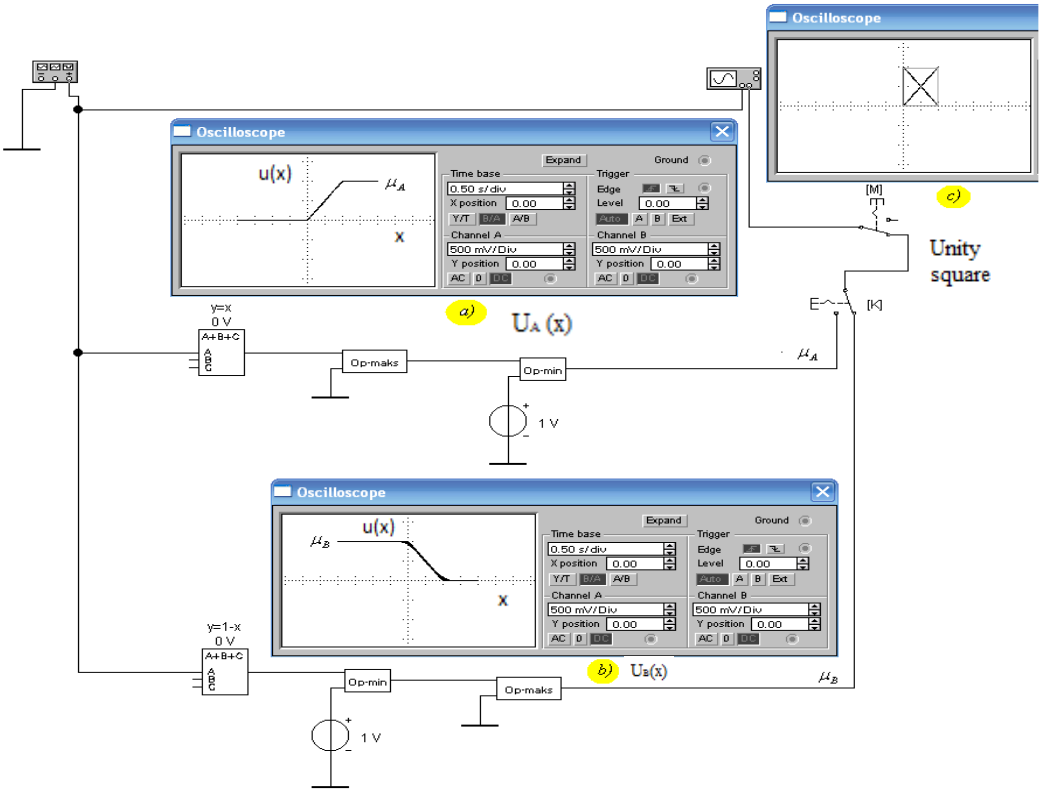


Figure 2. Complementary character of two circuits generating two membership functions is obvious

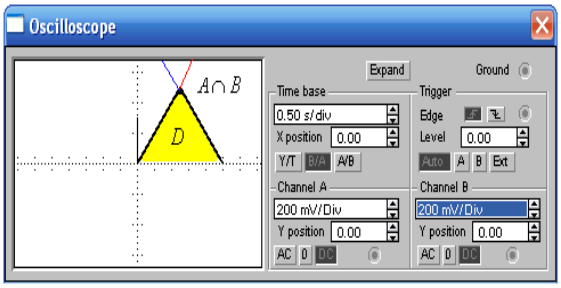


Figure 3. Two fuzzy sets in the same universe *X* are intersected using *Op-Amp* [2]

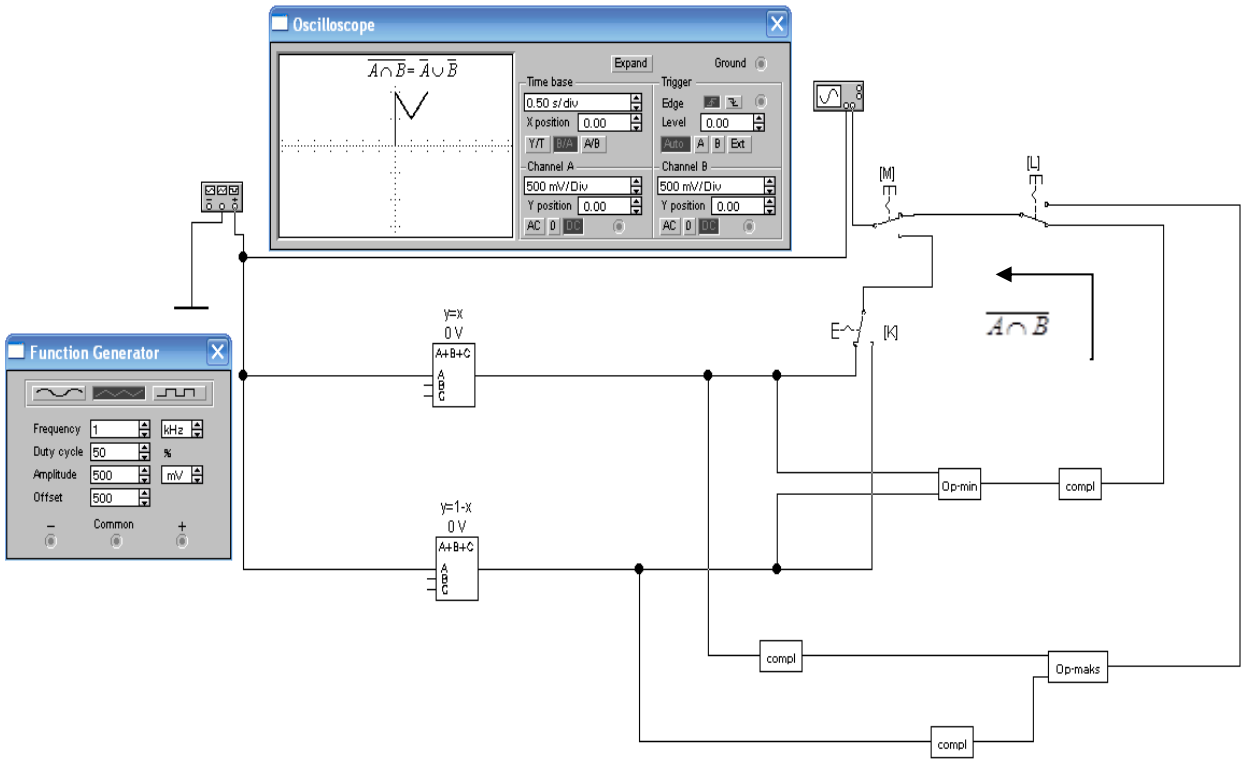


Figure 4. De Morgan's law $\overline{A \cap B} = \overline{A} \cup \overline{B}$

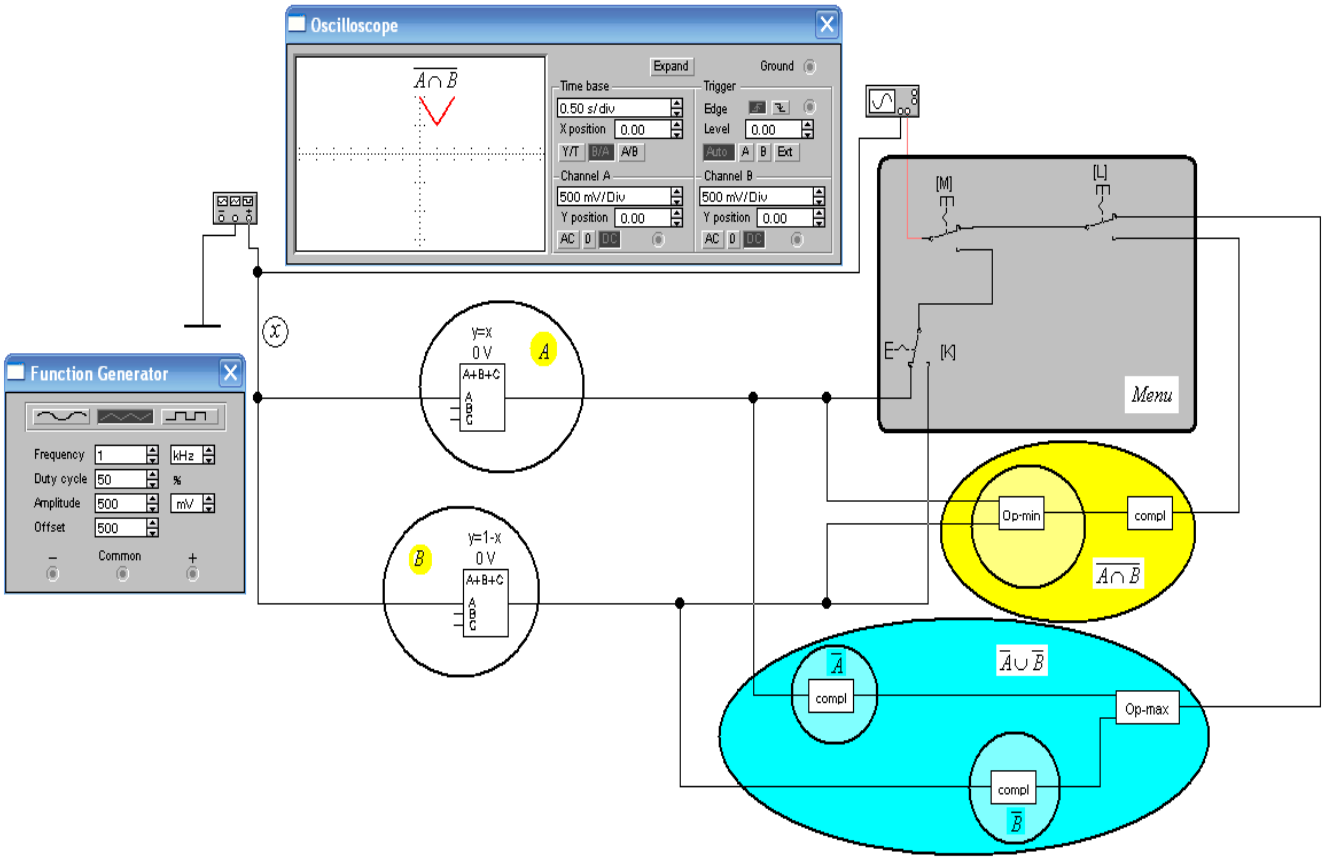


Figure 5. De Morgan's law $\overline{A} \cup \overline{B} = \overline{A \cap B}$

Note that the second DeMorgans Law expressed as a term $\overline{A \cup B} = \overline{A} \cap \overline{B}$ can be interpreted in hardware environment on a very similar way. In this context, Fig. 6 shows that the laws of exclusion, which are satisfied by crisp sets, are not satisfied by fuzzy sets. In particular the union of a fuzzy set $A \cap B$ and its complement $\overline{A \cap B}$ is not equal to the universal set, but is a subset of the universal set. Furthermore, the intersection of a fuzzy set $A \cap B$ and its complement $\overline{A \cap B}$ is not a null set and may have non-zero membership value(s) (see Fig. 5). These exclusion properties of a fuzzy set are not a shortcoming. In fact, they enable fuzzy logic to be more practical and versatile than crisp logic.

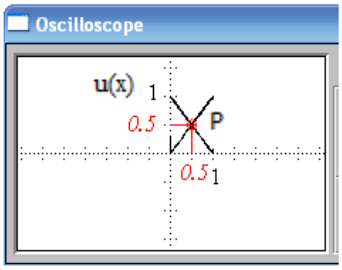


Figure 6. Fuzzy set $A \cap B$ and its complement $\overline{A \cap B}$

IV. SOME PROPERTIES OF FUZZY SETS; THE LAW OF EXCLUDED MIDDLE

We are going to use another example to illustrate how the law of excluded middle is violated in fuzzy set environment. A triangular membership function (see Fig. 7) is used to represent a fuzzy set *A*. Its complement $\overline{A}$ is shown on the same oscillograph which is obtained by subtracting the original membership function from 1 (see Fig. 8 and Fig. 9). The original circuit shown in Fig. 8 is converted to sub circuit called compl. (see Fig. 9) The union of *A* and $\overline{A}$ is performed using *max* gate as it is shown in Fig. 9.

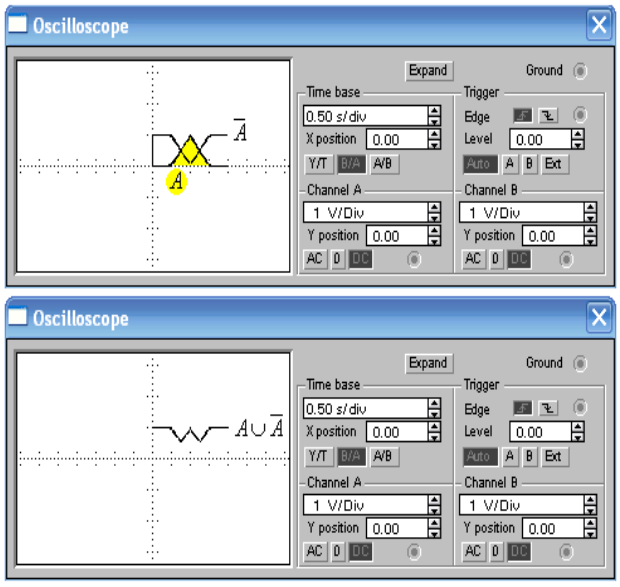


Figure 7. An example of exclude middle in fuzzy sets

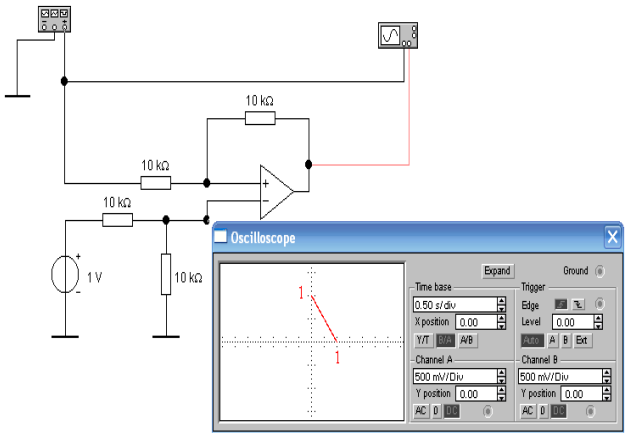


Figure 8. Simple *Op-Amp* complement circuit (compl)

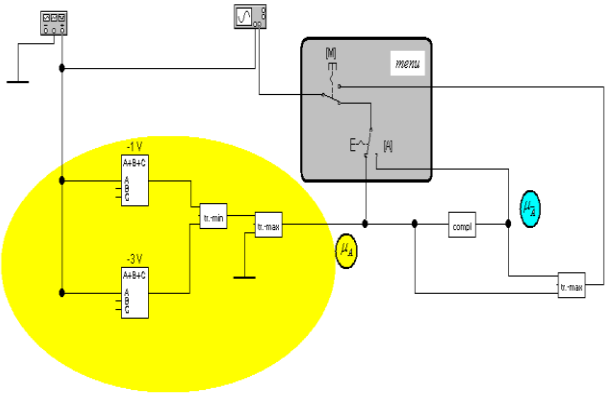


Figure 9. Electronic interpretation of an example of excluded middle in fuzzy sets

Now it is convenient to introduce a notion of, so called, *relative complement* [6].

V. RELATIVE COMPLEMENT

Operation of *relative complement* in the domain of fuzzy set can be modeled in hardware as relative complement gate accorcing the following definition [6].
Definition: Assume *A* and *B* are two fuzzy subset of *X*. The *relative complement* of *B* with respect to *A*, denoted $E = A - B$, is defined as the fuzzy subset *E* of *X* (9), where for each $x \in X$:

$$E(x) = \max[0, A(x) - B(x)] \qquad (9)$$

For example let's asume:

$$A = \{0/a,\ 0.2/b,\ 0.5/c,\ 0.6/d,\ 0.9/e,\ 1/f\}$$

and

$$B = \{1/a,\ 0.8/b,\ 0.5/c,\ 0.4/d,\ 0.1/e,\ 0/f\}$$

then

$$E = A - B = \{0/a,\ 0/b,\ 0/c,\ 0.2/d,\ 0.8/e,\ 1/f\}.$$

If $A = y = x$ and $B = y = -x+1$ are fuzzy sets from the „*unity square*“ then *E* can be expressed in an electronic environment as it is shown in Fig. 10.

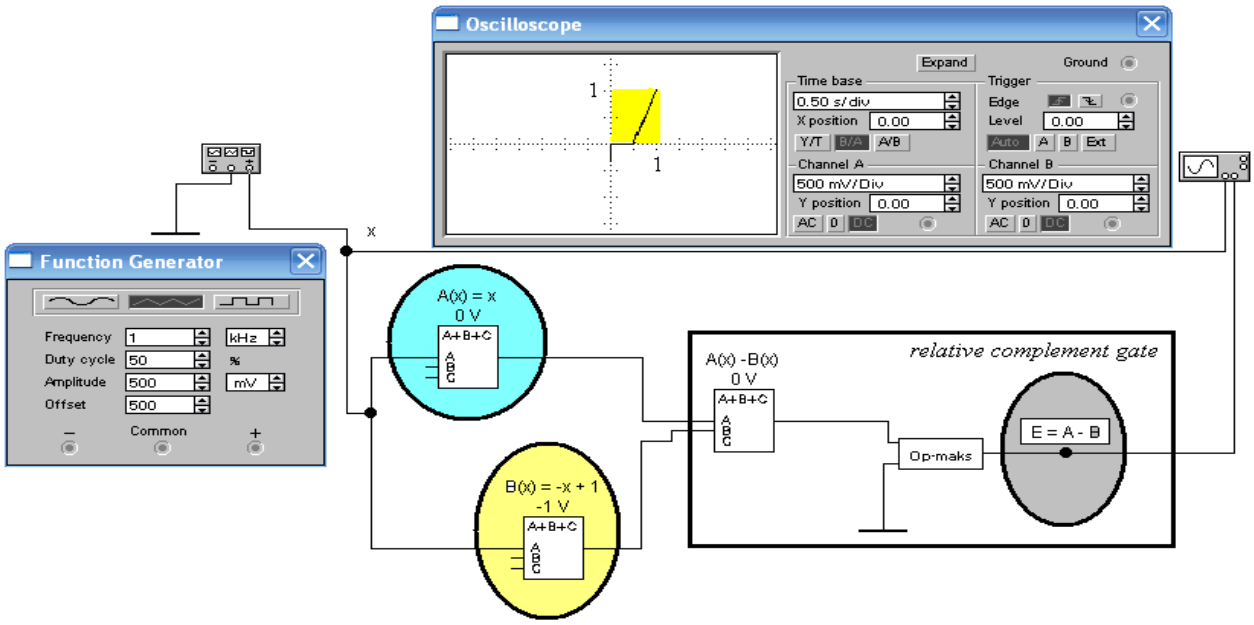


Figure 10. Operation of the relative complement in the domain of unity square; membership functions $y = x$ and $y = - x + 1$ represent fuzzy sets *A* and *B* respectively . One can easily find out that the previous example is related to this electronic model.

VI. SOME PROPERTIES OF FUZZY SETS; GRADE OF INCLUSION

In the context of fuzzy logic, a set, say *A*, may be partially included in another set, say *B*. Then it is convenient to define a grade of inclusion of a fuzzy set *A* in another fuzzy set *B*. This may be interpreted as the membership function of the fuzzy relation $A \subset B$. In a specific application one can use min (instead more general *T-norm*) to define a grade of inclusion (10) [2]:

$$\mu_{A\subset B}(x) = \begin{cases} 1 & \text{if } \mu_A(x) \le \mu_B(x) \\ \mu_B(x) & \text{otherwise} \end{cases} \qquad (10)$$

Consider an electronic fuzzy membership generator (see Fig. 10) that can deliver two fuzzy sets shown on the top of Fig. 12 [4]. According to relation (10), the grade of inclusion of *A* in *B* is given by solid line on the oscillograph (see bottom of Fig. 12). That line is generated by so called detector of grade of inclusion showed in Fig. 12. Besides standard transistor min and *max* gate, the detector includes comparator (*m-M-comp*) [5]. The design of that comparator is based on *min-max-Op-Amp* technique [5] as it is shown in Fig. 11. *Max* gate is used along with very steep straight line (*Op-Amp*) and ground level line to establish the raising edge of the comparator. In the second design step a min gate is used to limit the raising edge to a certain level (battery voltage).

Characteristic shown in Fig. 11 can be used to express (AI – model) wholesaler agent behaviour in the process of newspaper distribution and control [5]; in the very beginning of a tourist season wholesaler agent ignores small and sporadic orders coming from local operators. Then, as the season "grows" local orders are accepted up to a certain level. After, according to publisher's wish the control is dismantled by setting order(s) to a certain maximum level(s). Such and similar characteristics (models of human acting) can be regarded as a kind of *event filters* [5].

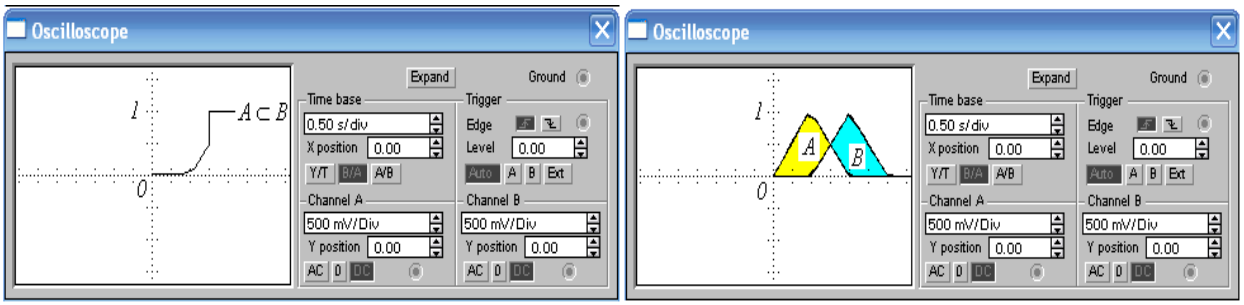


Figure 11. An example of *grade of inclusion*

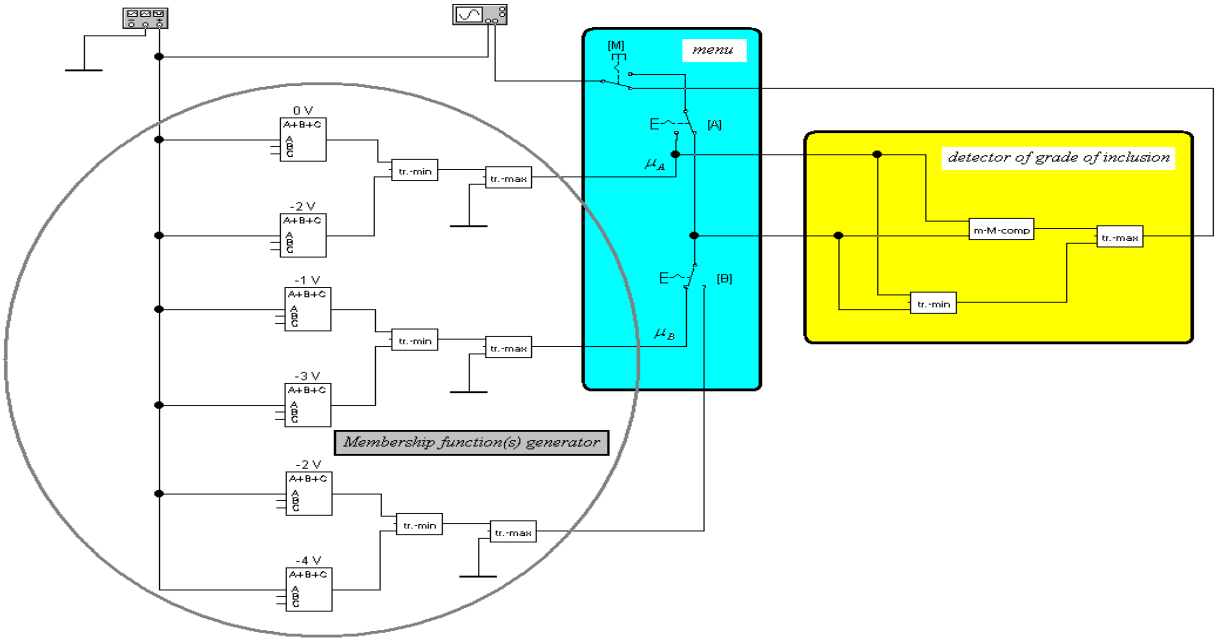


Figure 12. Electronic simulation of a problem; *grade of fuzzy set inclusion*

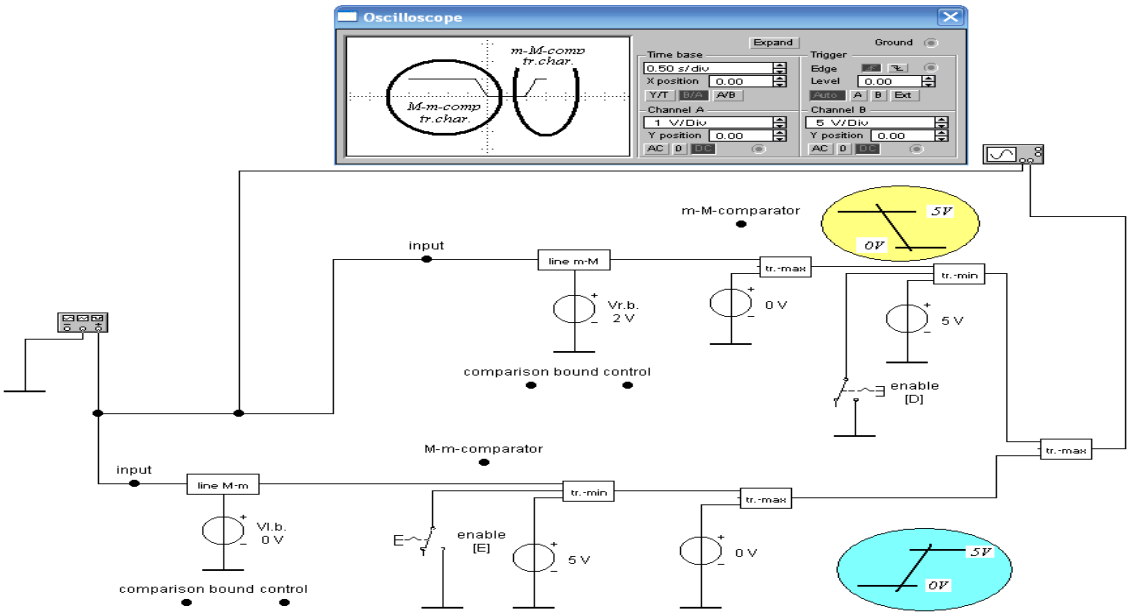


Figure 13. Electronic design of *m-M-comp* and *M-m comp*

VII. CONCLUSION

Electronic hardware environment was used to present some mathematics of fuzzy sets. DeMorgans Law electronic presentation was extended to the problem of exclude middle in fuzzy sets. That problem was modelled in two different ways. In the context of fuzzy logic mathematics original detector of grade of inclusion has been designed. That detector has been used to interpret the grade of inclusion for two triangular fuzzy sets. The design of the presented detector is based on so called *min-max-Op-Amp* technique. Future investigations of hardware implementations of functions used in fuzzy mathematics can be based on *min-max-mid op-Amp* technique.

REFERENCES

[1] S.Kurepa, Uvod u matematiku, Tehnička knjiga Zagreb, Zagreb, Hrvatska, 1975, pp. 50

[2] F.O.Karray, C.Silva, Soft Computing and Intelligent System Design, Pearson Education Limited, Harlow, Essex, England, 2004, pp. 65-82.

[3] D.Kovačević, A.Kovačević, Fuzzy Logic Controller Based on Standard Operational Amplifiers, Chapter 13, Edited by J.F.Baldwin, John Wiley & Sons,Chichester, England, 1996, pp. 231-244.

[4] D.Kovačević, Master Thesis; Hardware realisation of fuzzy logic operations

[5] D.Kovačević, Phd Thesis; Hardware Modelling of Newspaper Distribution and Sell Process Control based on Intelligent Agent Network Realisation

[6] R.Yager, P. Filev, Essentials of Fuzzy Modelling and control, John Wiley & Sons New York, 1994, pp. 8-10.

Fuzzy gates interchangeability

Darko Kovačević, Toma Karađole and Jure Vrsalović
Falcuty of Maritime Studies, University of Split, 21000 Split, Croatia
dkovac@pfst.hr

Abstract - The idea of this paper is related to possible conversions of min, max and mid fuzzy gates functions. If we want to convert basic min and max gate function to another logical function, for some reason, that can be done on several ways. For the beginning one can convert only gate inputs. In the second case only gate outputs can be inverted. The third kind of the gate interchangeability is done by the combination of inverting inputs and outputs. At the end, it is possible to design min, max and mid gates to be controllable through gate inputs information (inputs voltages).

I. INTRODUCTION

In the second section of this paper we are going to show that the function(s) of fuzzy gate(s) can be interchanged in the way similar to standard logic gate function interchangeability. The simplest way of "gate change" is achieved by adding a complement gate on *min* or *max* fuzzy gate inputs. The complement gate can be added on fuzzy gates outputs, if one wants to change a gate function on the other way. Finally, the complement gates can be simultaneously added to inputs and outputs to change agate function. On the other hand (Section III.), fuzzy *min* or fuzzy *max* can be used as a design tool for other kinds of fuzzy gates in a manner similar to design of complex gates in digital electronics [1].
Also, we are going to show that standard fuzzy gates can be used to support logical operations in the domain of crisp sets. Note that standard digital gates, in principle, cannot support fuzzy logic operations.

A dedicated fuzzy logic gates ($\max/^{c}\min$) with inherent (designed) interchangeability mechanisms will be briefly explained in the fourth section.
At the end of this paper, a new kind of a complex fuzzy gate called *mid* gate will be shown. That complex gate can perform one of three functions*: min*, *max* and *mid* what depends of straight gate input information.

II. USING COMPLEMENT GATES TO CONVERT GATES

Consider a fuzzy set A in the universe X. Its complement $\overline{A}$ is a fuzzy set whose membership function is defined by:

$$\mu_{\overline{A}}(x) = 1 - \mu_A(x) \quad \text{for all } x \in X$$

as illustrated in figure 1.

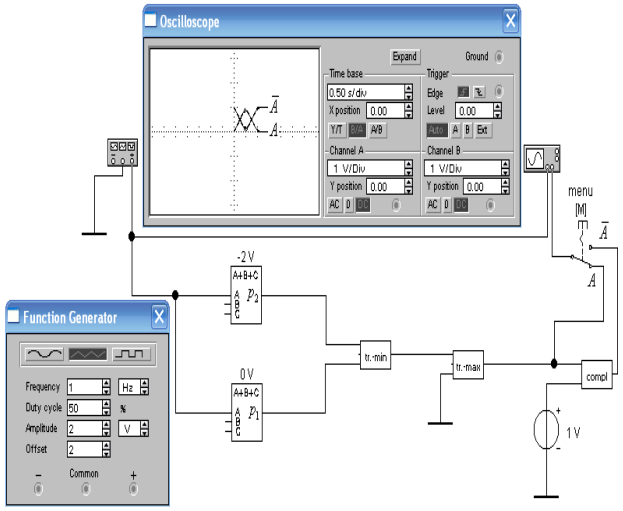


Fig. 1. Fuzzy-set complement of fuzzy logic NOT

The complement in fuzzy sets corresponds to the negation (NOT) operation in fuzzy logic, just as in crisp (bivalent) logic.
In the „unity square" (see Fig. 2) we can define two sets B and C by two membership functions

$$\mu_B(x) = x$$
$$\mu_C(x) = 1 - \mu_B(x)$$

The "unity square" can be used as an explanation tool for gate interchangeability, DeMorgan's laws etc.

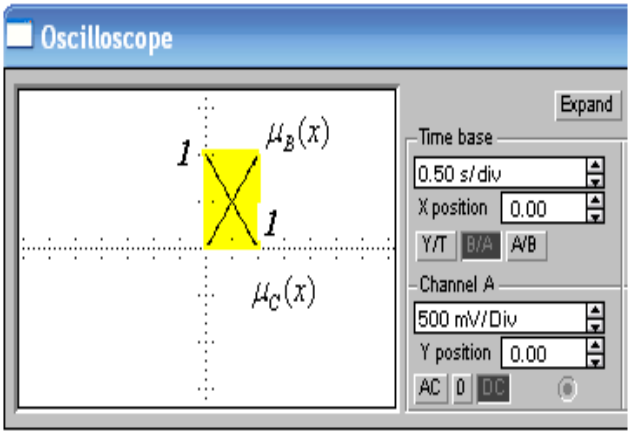


Fig. 2. „Unity square" along with its characteristic membership functions $\mu_B(x)$ and $\mu_C(x)$ for $x \in [0, 1]$

A. Inverting outputs

Sometimes there is necessity to convert basic *min* and *max* gates to another logic function. [1] That can be done easily with the use of complement gates.
Ones, only output is inverted to get desired logic function (see Fig. 3.). Inverting the outputs leads us to easy predictable results; *max gate* is converted to $\overline{\max}$ *gate*. On the same way $\overline{\max}$ *gate* can be converted to *max gate*.
In a very similar fashion it is possible to design conversion of *min gates* to $\overline{\min}$ *gates* (and vice versa) as it is shown in Fig. 4. Note that $\overline{\min}$ *gates* can be designed in quite different way (see Fig. 5).

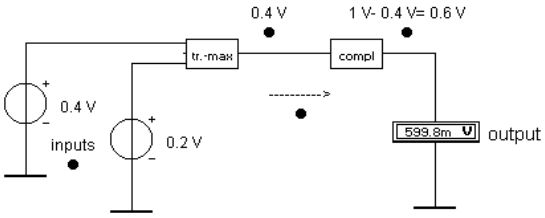


Fig 3. Conversion by complementing output; $\overline{\max}(x, y) = \overline{\max}(0.2,0.4) = \overline{(0.4)} = 0.6$. In the case that inputs are 5Vand 0V respectively, then the output will be 0V

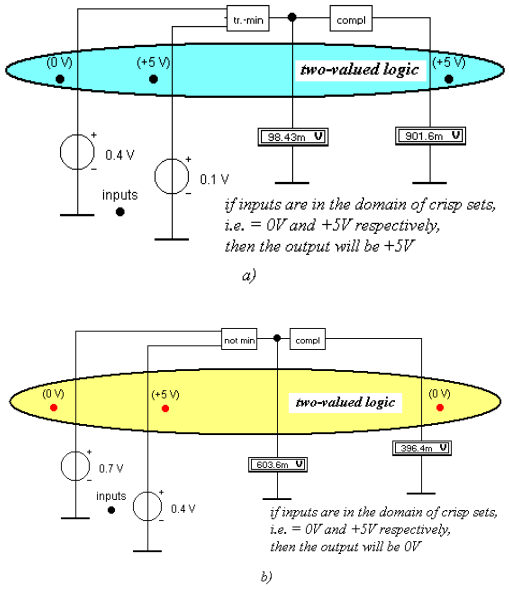


Fig 4. Conversion: a) *min gates* to $\overline{\min}$ gates and b) $\overline{\min}$ gates to *min gates*

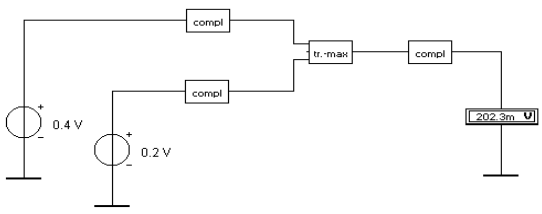


Fig 5. *Max gate* to $\overline{\min}$ *gate* conversion

B. Inverting inputs

In this section, gate inputs will be inverted to achieve conversion. For instance, if both inputs of *max* gate are inverted, the gate becomes a $\overline{\min}$ *gate* (see Fig. 5.).

C. Inverting inputs and outputs

It is possible to invert inputs and outputs of *min (max)* gates to get different logic gates. For example, if inputs to *min gate* are inverted along with inversion of the gate output, then *max gate* will be designed (see Fig. 6.).
In the very same way, max gate be converted to min as it is shown in Fig. 7.
In a similar way, a $\overline{\max}$ *gate* can be converted to $\overline{\min}$ *gate* (see Fig. 8.).

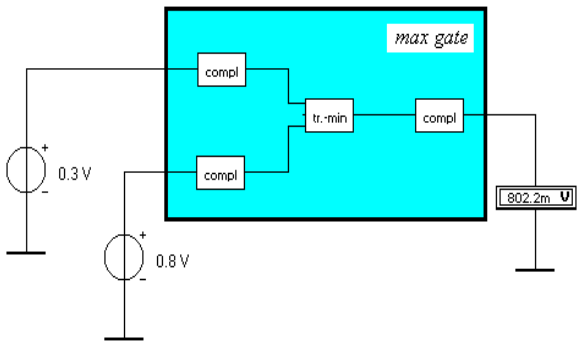


Fig 6. Converting *min gate* to *max gate*; for inputs 0.3V and 0.8V the output will be 0.8V (see Fig.2. for reference)

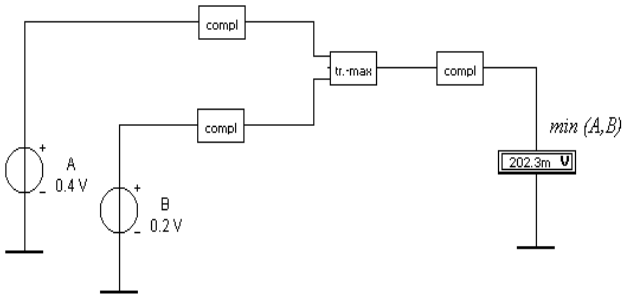


Fig. 7.Interchanging gate function; min (0.4V, 0.2V) = 0.2V (refered to «unity squre»)

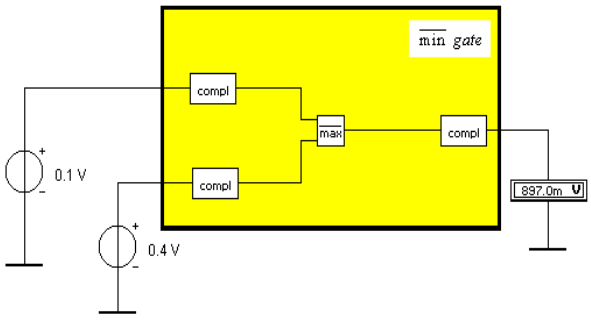


Fig 8. $\overline{\min}(0.1V,0.4V) = 0.9V$ (refered to «unity squre»)

III. $\overline{\min}$ GATE AS A UNIVERSAL GATE

Like in classic digital circuit environment, $\overline{\min}$ *gate* can be used as a sort of "universal gate"; when wired they can be used to create any of other logic functions (see figure 9 and 10). According to Fig. 9 $\overline{\min}$ *gates* are used to design (realize) *min gate*. In similar manner, universal $\overline{\min}$ gates can be used for the realization of max gates (see Fig. 10.).

Using "universal" $\overline{\min}$ *gates* it is possible to design classic *ex-or* gates that support crisp set operation (bivalent logic) as it is shown in Fig. 11 Note that such kind of *ex-or* gates cannot support logic operations in the domain of fuzzy sets [2]. Fuzzy ex-or gates have been designed on an entirely different way [2]. Fuzzy ex-or gates have been designed on an entirely different way [2]. In general, multiple use of one kind of inverting fuzzy gates (e.g. $\overline{\min}$) is good enough to make any combinational function in crisp logic domain.

A picture gallery (see Fig.11.) shows functionality of *ex-or* gate designed in *min-max-Op-Amp* technique to handle crisp set operation.

So far, we can conclude that dedicated fuzzy gates, depending on current inputs can be used to handle logic operations in domains of fuzzy and crisp sets.

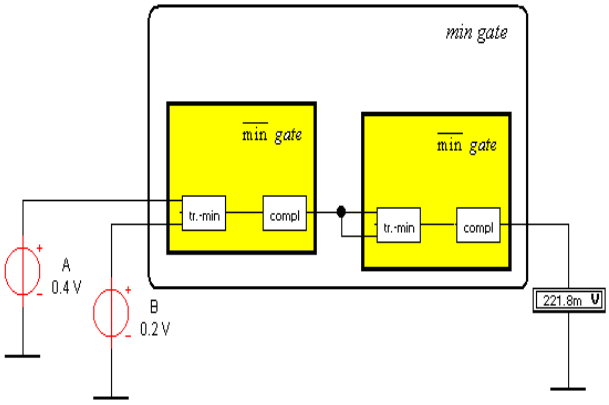


Fig. 9. Wired "universal" $\overline{\min}$ *gates* perform *min function*

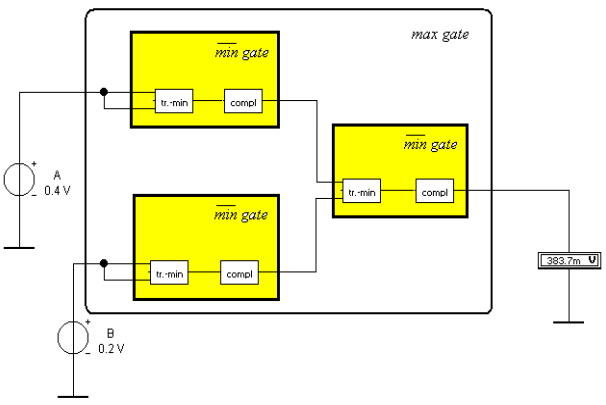
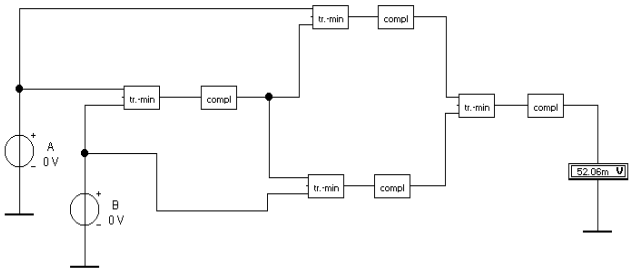


Fig. 10. "Universal" $\overline{\min}$ *gates* design of *max function*

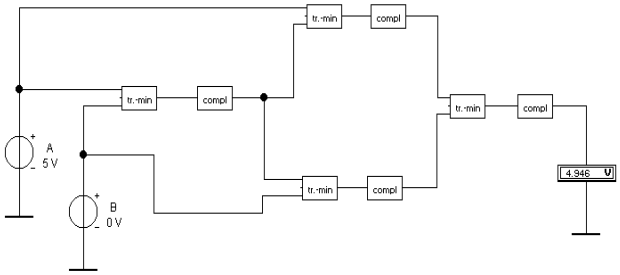
IV. INFORMATION CONTROL OVER MIN (MAX) GATES TO ACHIEVE MAX (MIN) GATES INTERCHANGEABILITY

A. Changable min and max gates (min/max gate)

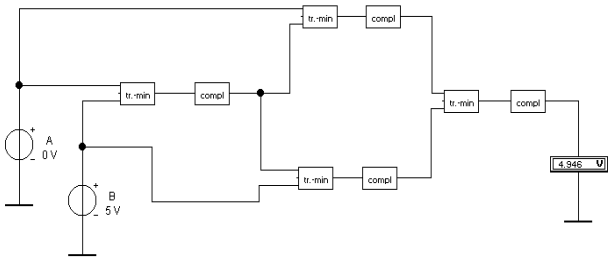
Min or *max* gate designed in *Op-Amp* technique [3] can be easily, „in hardware" , transformed to perform *max* or *min* function, respectively (see Fig. 12. and 13.). We can talk about information controled interchangeability, ie. *min/c max* gates.



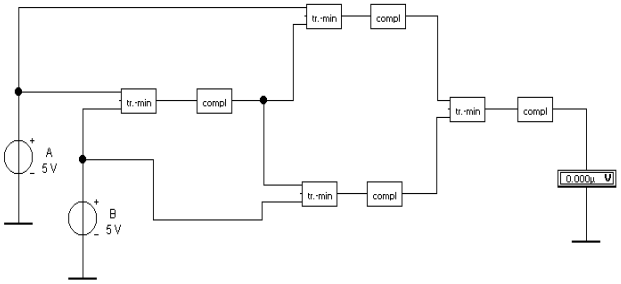
a) Ex-or (A, B) = "0" (0 V)



b) Ex-or (A, B) = "1" (5 V)



c) Ex-or (A, B) = "1" (5 V)



d) Ex-or (A, B) = "0" (0 V)

Fig 11. Classic ex-or designed in $\overline{\min}$ *gates* technique
We have used ex-or gate to notify ability of using fuzzy gates in crisp logic operations.

The mentioned interchangeability is based on mathematical definition of min (max) gate operation [2]:

$$\min(x,y)=\frac{1}{2}(x+y-|x-y|)$$

$$\max(x,y)=\frac{1}{2}(x+y+|x-y|)$$

Changing the sign of the third term the logic operation of the *min (max) gate* can be exchanged (see Fig.12. and Fig. 13.).

The interchangeability of the original three inputs *min (max) gate* in the *Op-Amp* technique can be achieved in the very same way :

$$\min z,\min(x,y))=\frac{1}{2}(z+\min(x,y)-|z-\min(x,y)|)$$

$$\max(z,\max(x,y))=\frac{1}{2}(z+\max(x,y)+|z-\max(x,y)|)$$

The sign of the third term must be interchanged to perform min or max operation of three inputs *min/ᶜmax* gate.

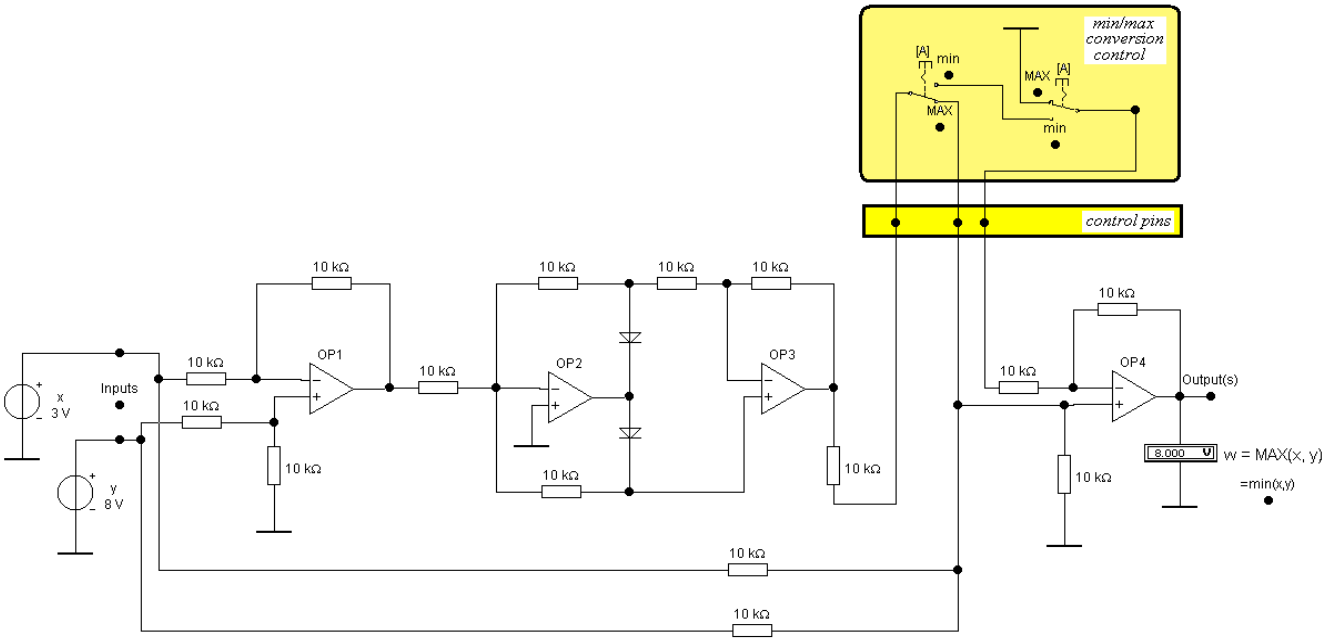


Fig. 12. The state of two inputs *min/ᶜmax* gate is set to perform *max* function. Switch(es) A ($\overline{A}$) flips from one position to another interchanging the circuit logic function.

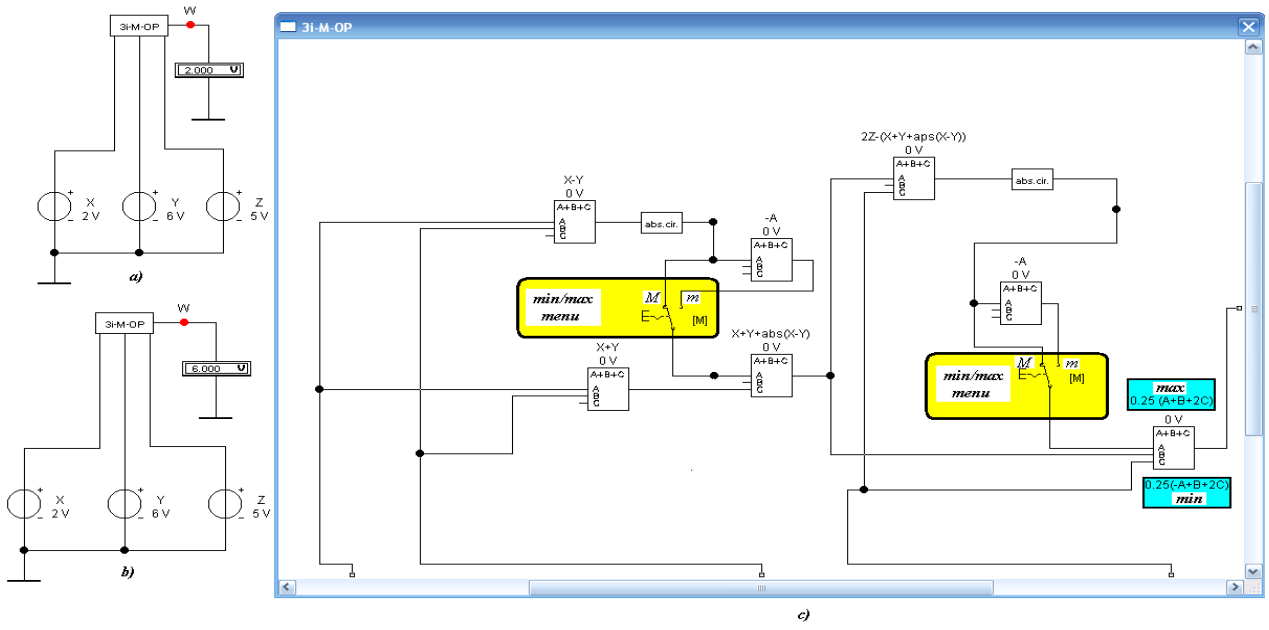


Fig. 13. The state of three inputs $\max/^{c}\min$ gate;
a) min gate performing, b) *max* gate performing, c) circuit scheme

Interchangable *min/c max* gates can be used as an element of an anology *electronic sorting device* (*sorting network*) capable to exchange the way of sorting analog data (input voltages).

Sorting network are comparison networks that always sort their inputs (see Fig.14.) [4]. A comparison network is composed solely of wires and comparators (see Fig. 14.). A comparator shown in Figure 14., is an electronic device with two inputs, *x* and *y* and two outputs *x'* and *y'*, that performs the following function:

$$x' = \min(x, y)$$

$$y' = \max(x, y)$$

If the gate control changes, the $\max/^{c}\min$ gate state will be interchanged and therefore outputs of the electronic comparator will be exchanged; *x'* switches from 5.000 V to $x' = \max(x, y) = 9.000$ V and y' switches from 9.000 V to $x' = \min(x, y) = 5.000$ *V* . In that way it is possible to design controlled sorting; monotonically increasing sorting network (see Fig. 15.) can be transformed in monotonically decreasing sorting network (see Fig. 16.) due to input information. Doing sort in an analog hardware can be important in different applications. In an AI model scenario of the newspaper distribution and sell process, an intelligent agent (process operator) wants to establish, so called, *pilot based coupling control* over a certain newspaper [2]. Prior to any decision he (she) must do sorting of current orders and remnants to get correlation coefficients among the newspaper and newsstands in the area of distribution. For that purpose orders (remnants) can be marked adding a small denoting voltage to orders (remnants) in a row. Numerical codes of newsstands are introduced in sorting as it is shown in figure 15.

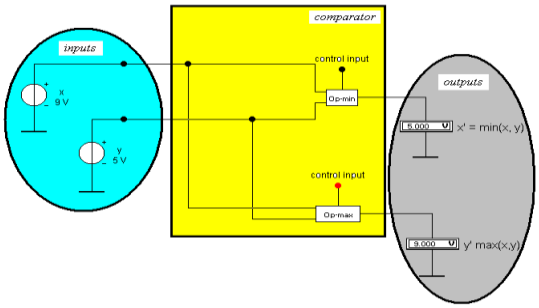


Fig. 14. A electronic comparator with applied inputs *x* and *y* generates outputs *x'* and *y'*.

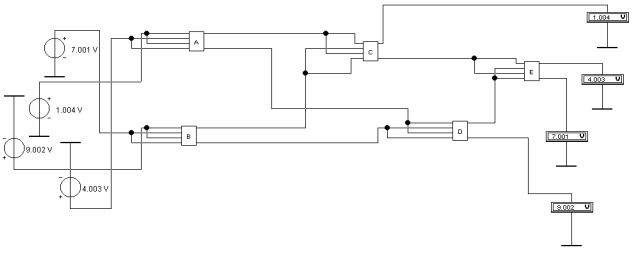


Fig. 15. Monotonically increasing sorting network.Input voltages remnants are deliberatively marked (1,2,3,4) to correlate newstands and corresponding remnants

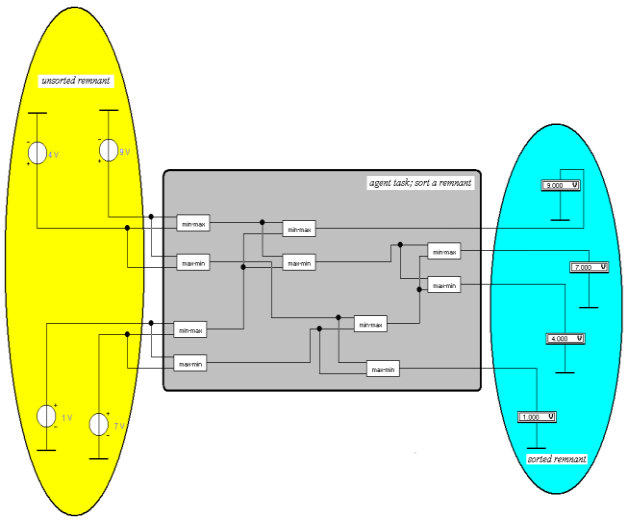


Fig 16. Agent task; sort the remnant prior to any control decision; monotomically decreasing sorting network

B. Mid gate interchangeability

The operation of *mid* which plays an important role in modelling a part of certain kind of intelligent (hardware) agent can be expressed as [2]:

$$\text{mid}(x,y,z) = y \quad \text{iff} \quad x < y < z$$

and is independent of input's possitioning, i.e. gate action will be correctly performed pying no regard to arragment of applied input voltages.

In the text to follow we will (slide)show how *mid* gate can perform *min* and *max* operations, as well (see Fig. 17,18 and 19.).

One must take charge of *control menu* (control switch M) to decide upon using *mid* or *min*(*max*) gates. Once, when you want to choose between *min* and *max* gate, you need to apply (control input C) zero voltage (information) for *min* gate (see Fig. 18.) and +5V (information) for *max* gate performing (see Fig. 19.), respectively.

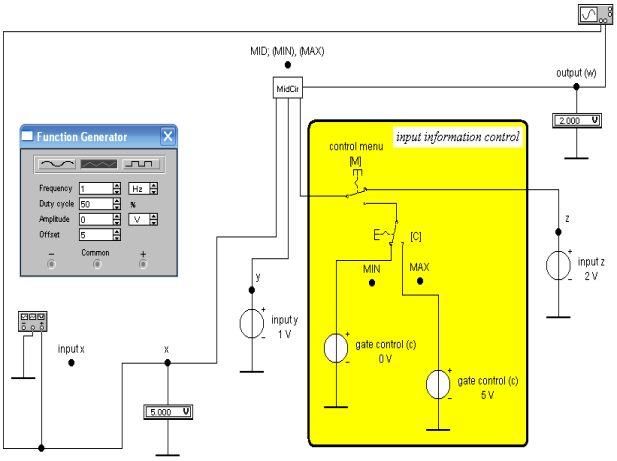


Fig. 17. The response of *mid* gate when control menu switch is set to applied input z;
mid(x,y,z) = *mid*(1,2,5) = *mid*(5,1,2) = *mid*(2,1,5) = ...2

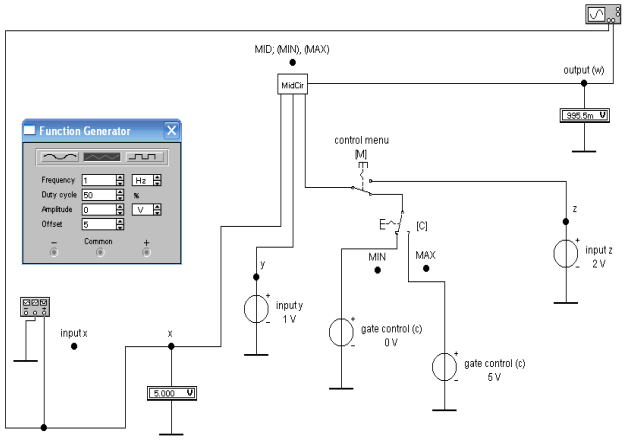


Fig. 18. $mid \xrightarrow{M-C_{\min}} \min(5,1) = 1$

When control menu M is set to input gate control switch C, one can select „0“ to establish *min* operation of *mid* gate.

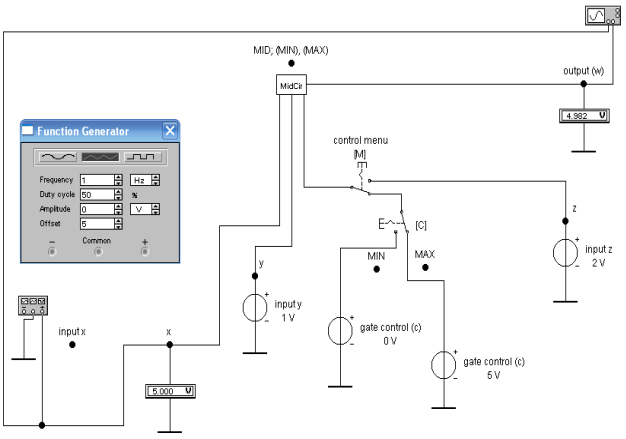


Fig. 19. $mid \xrightarrow{M-C_{MAX}} \max(5,1) = 5$

When gate control is set to +5V in C-menu, the *mid* gate will perform *max* function.(bivalent or fuzzy)

Performing *min* or *max* functions accordingly, means that the complex *mid* gate can be included in fuzzy gates family. Aditional promising features of *mid* gate will be the subject of the future work.

V. CONCLUSION

Different ways of fuzzy *min*, *max* and *mid* gates interchangeability have been investigated in this paper. We have started interchangeability experiments using complement gates in the manner that is known in digital electronics; complementing inputs and outputs (or both). Then we have designed voltage control over original min and max gates to achieve inherent interchangeability. At the end, an information control over inputs has been used to pick out among logic operations performed by original, so called, mid gates.

Finally, we can conclude that interchangeability in the domain of fuzzy gates can be done through:

- gate combinations
- control incorporated in dedicated hardware design
- controlled information applied to gate inputs

Additional promising features of mid gate, in the area of control, will be the subject of the future work.

ACKNOWLEDGMENT

The result presented in the paper obtained in the scientific research project No. 250-2502209-2364 and the international research project “ The possibilities of reducing pollutant emissions from ships in the Montenegrin and Croatian Adriatic implementing Anex of MARPOL convention” supported by the Ministry of Science Education and sport of the Republic of Croatia.

REFERENCES

[1] D.Kovačević, Hardware realizations of elements and functions in fuzzy sets theory, Master thesis, FESB-Split, Croatia, 1993.

[2] D.Kovačević, Distribution and Sell Process Control Modelling Based on Intelligent Agent Network Realization, PhD thesis, FESB-Split, Croatia, 2012.

[3] D.Kovačević, A.Kovačević, Fuzzy Logic controller Based on Standard Operational Amplifiers, Edited by J.F.Baldwin, John Wiley & Sons, Chichester, England, 1996, pp. 231-244

[4] T.H. Cormen at al, Introduction to Algorithms, The MIT Press, Cambridge Massachusetts, USA, 2001, pp. 704-709.

Solving mathematic problems in electronic environment, finding global and local minimum and maximum function values

Ivan Madunić, Darko Kovačević. Lovre Perković and Toma Karađole.
Falcuty of Maritime Studies, University of Split, Split, Croatia
dkovac@pfst.hr

Abstract - A dedicated hardware simulation tool is used to solve specific mathematical as well as electronic problem: finding global minimum and maximum function values (coordinates in Cartesian coordinate system). A new class of peak detectors based on fuzzy *min* and *max* gates has been proposed. Hardware methods, having pure mathematical background, in detecting x-coordinates of function extremes are explained, too.

I. INTRODUCTION

There are numerous applications in which it is necessary to determine the peak value of some input waveform.

The simplest, but very problematic, method is to use diode and condenser. The highest point of the input waveform charges up condenser, which holds that voltage value while diode is back-biased.

A better circuit, using *Op-Amp* and feedback from capacitor solves some simple peak detector problems but new ones are introduced; finite slew rate of *Op-Amp*, input bias current etc. [1].

The most of these problems can be solved by more complex circuits [1]. One of such circuits is shown (see figure 1).

In this paper we are going to introduce a new class of simple peak detectors that are based on *min-max-Op-Amp* technique [2]. These circuits will be used in a very special applications; finding extremes of different functions in a time domain or in a dedicated phase plain domain [3].

Obtained information about extremes can be used in different purposes (for example in a process control).

In the second section we will refer to the mathematical presentation of a function extremes very briefly, *min* and *max* detectors definitions will be presented in the third and fourth section.

Finding global maximum and global minimum of a function in an electronic environment (EWB - analog circuit domain) will be presented in the fifth section of this paper.

A "hard" problem for "math-community", i.e. finding maximum and minimum of discontinuous functions can be elegantly solved using *min* and *max* detectors as it is described in two last sections of this paper.

The concept of this paper is based on picture preferred presentation for the sake of clarity [2].

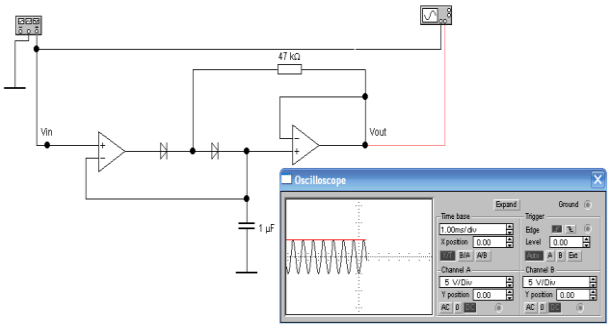


Figure 1. Improved peak detector

II. PROPERTIES OF CONTINUOUS FUNCTIONS; MINIMUM AND MAXIMUM

If f (v_{in}) is continuous in the interval $[V_a, V_b]$ then it has the following properties (among others properties not mentioned here):

- f (v_{in}) is a bounded function: there are numbers V_m and V_M such that $V_m < f(v_{in}) < V_M$ for all $v_{in} \in [V_a, V_b]$
- f (v_{in}) has a largest and a least value in the interval $[V_a, V_b]$.

-The least value of f(v_{in}) on $[V_a, V_b]$ is called the minimum off(v_{in}) on $[V_a, V_b]$, the largest value is the maximum off(v_{in}) on $[V_a, V_b]$ and the difference between the two is called the oscillation off(v_{in}) on $[V_a, V_b]$ as it is shown (see figure 2) [4].

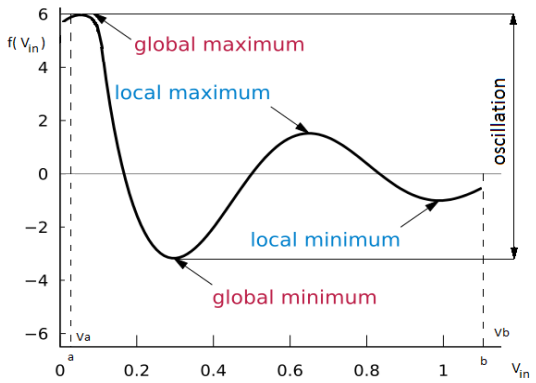


Figure 2. Oscillation of a function

III. MIN PEAK DETECTOR

A transistor *min* circuit along with a capacitor is used to design a peak detector (see figure. 3.) capable to detect a voltage minimum (V_{min}) of a function $y \equiv f(V_{in})$ (1) :

$$y_{min} \equiv V_{min} = V_{out_{min}} = \min(V_{in}, V_{out}) \qquad (1)$$

A buffer is used for the sake of isolation. Note that antoher max circuit can be used for isolation, too.

IV. MAX PEAK DETECTOR

A transistor max circuit having a feedback from a capacitor is used to design a peak detector (see Figure. 4.) capable to detect a voltage maximum (V_{MAX}) of a function $y \equiv f(V_{in})$ (2):

$$y_{MAX} \equiv V_{MAX} = V_{out_{MAX}} = \max(V_{in}, V_{out}) \qquad (2)$$

Again, an *Op-Amp* is used for the sake of isolation. An option is to use high input differential amplifier to compensate peak error [5].

V. FINDING GLOBAL MAXIMUM AND GLOBAL MINIMUM ELECTRONICALLY

A. *Example No.1.*

In this section we are going to show how(global) maximum and (global) minimum of a certain function can be found using *min* and *max* detectors respectively
Plot the graph of the function

$$f(x) = \frac{2x}{1+x^2} \text{ for } x \in \Re .$$

and show that $f(x)$ is continuous on its whole domain. Find its maximum and minimum values and show that it attains every value between these extremes.
Find the difference (oscillation) between maximum and minimum. Use EWB or similar simulation programme to achieve these tasks.

B. *Graph*

The electronic graph of the function is shown (see figure 5). and enlarged in figure 6. From those figures we can see that the shown part of the function is a continuous curve, indeed. The graph is displayed in the phase space (oscillograph) for the function that is designed gradually (see figure 5).

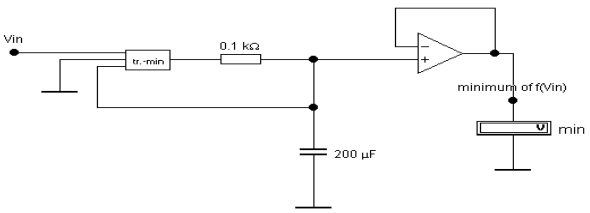


Fig. 3. Simple *min* peak detector with *Op-Amp* follower.

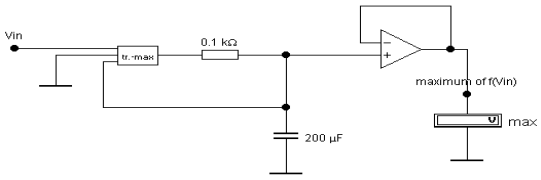


Figure. 4. Simple max peak detector with Op-Amp follower

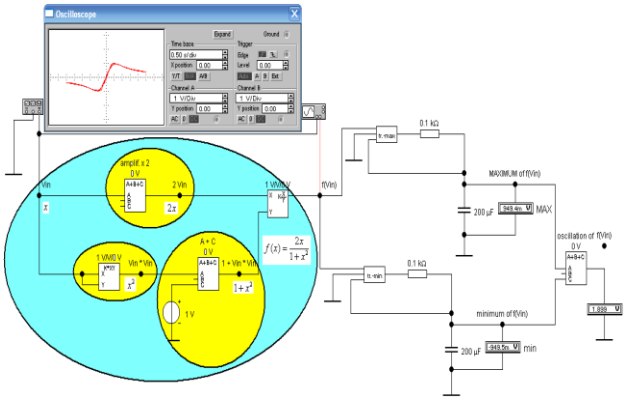


Figure. 5. Voltmeters are showing minimum and MAXIMUM values of the function

C. *Continuity*

One can use one of oscilloscope pointer lines (see Figure 6.) and move it from left-hand end to right-hand end without noting any disruption along the shown length of the curve (function). In mathematics we can prove this more formally as follows. Select any point x_0 of the domain of the function. Then we have to show that the term $|f(x) - f(x_0)|$ can be made as small as we please by taking x_0 sufficiently close to x. The term (3):

$$\frac{2x}{1+x^2} - \frac{2x_0}{1+x_0^2} = \frac{2(1-xx_0)(x-x_0)}{(1+x^2)(1+x_0^2)} \to 0 \qquad (3)$$

converges to zero as $x \to x_0$

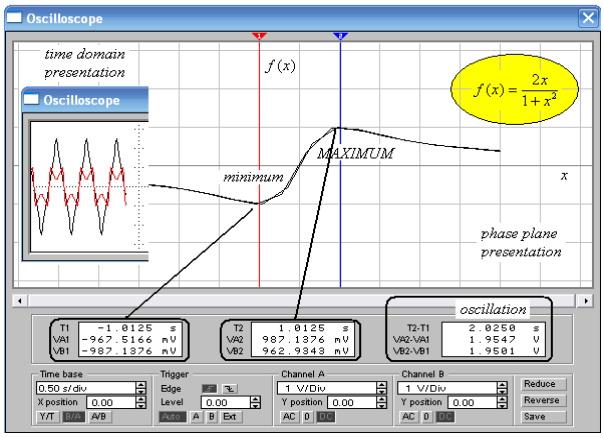


Figure 6. Oscillograph of the function and measurements extremes and oscillation on the function $f(x) = \frac{2x}{1+x^2}$

In an electronic environment an oscilloscope having two moveable pointer lines can be used to simulate described mathematical continuity model (3) as it is illustrated (see figure 7).

D. *Maximum and minimum:*

The proof that the function takes every value between its least and largest values actually determines those values in this simple example.
Let us find values of *x* that corresponds to the given *y* in the range of $f(x)$.

From the term $y = \frac{2x}{1+x^2}$ we can easily come to the term $yx^2 - 2x + y = 0$. This gives two values for x for each $y \in (-1, 1)$ (terms (4) and (5)):

$$x_1 = \frac{1-\sqrt{1-y^2}}{y}, \quad y \neq 0 \qquad (4)$$

and

$$x_2 = \frac{1+\sqrt{1-y^2}}{y}, \quad y \neq 0 \qquad (5)$$

The maximum and minimum values for *y* are -1 and 1 respectively, and corresponding values for *x* are -1 and 1. In electronic environment getting corresponding values of *x* is a little bit tricky.
First, the electronic model of the relation $yx^2 - 2x + y = 0$ must be realized as it is shown (see figure 8. and figure 9.). Then one must vary Y_{MAX} input to get corresponding *x* coordinates of the function extremes.

E. *Example No.2.*

Design a complete electronic model (*an extreme detector*) for the function $f(x) = \max(0; -(x-1)^2 + 3)$ to get its extreme (take x and y coordinates) in a compact circuit. The function f(x) is a shifted inverted parabola limited to positive values f(x)>=0 using the max circuit (see figure 11.). *Max peak detector* is used to get Y_{MAX} and the corresponding model for finding X_{MAX} completes the required electronic model.

Figure 11. illustrates every step in the model design. Note that a standard squarer is used to obtain square-rooter circuit (see figure. 10.).

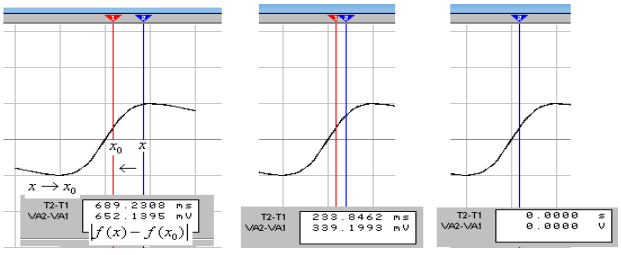


Figure. 7. Tracking continuity by an oscilloscope;

$$f(x) \to f(x_0) \text{ as } x \to x_0$$

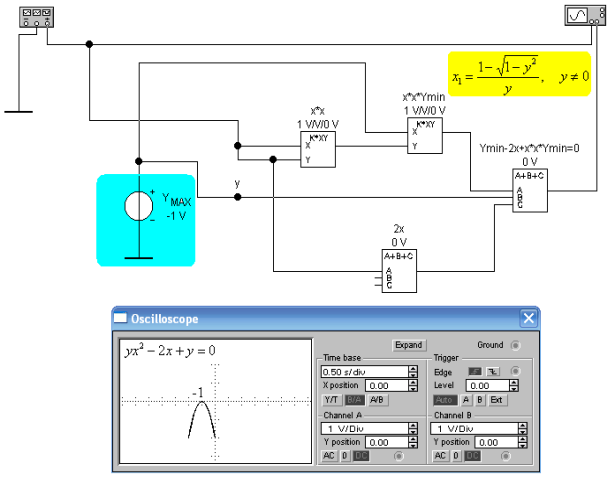


Figure. 8. Electronic solution (x-coordinate detector) for

$$x_{min} = \frac{1-\sqrt{1-y^2}}{y}, \quad y \neq 0$$

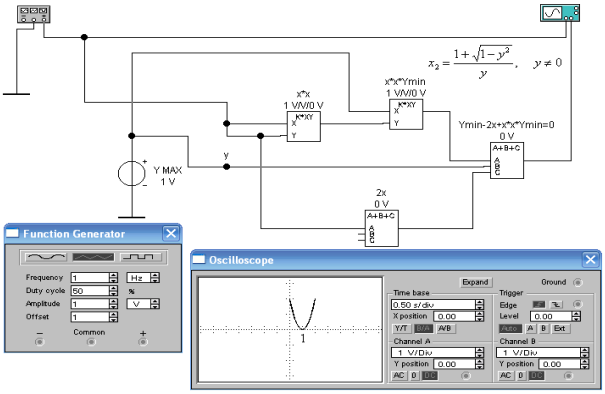


Figure. 9. Electronic solution (x-coordinate detector) for

$$x_{max} = \frac{1+\sqrt{1-y^2}}{y}, \quad y \neq 0$$

The output of the square root circuit can be defined as (see figure 10.):

$$y = A(x - y^2)$$

$$\frac{y}{A} = x - y^2$$

$$y^2 = x - \frac{y}{A}$$

For $A \to +\infty$ the output is $y^2 = x$ and finally

$$y \neq \sqrt{x}$$

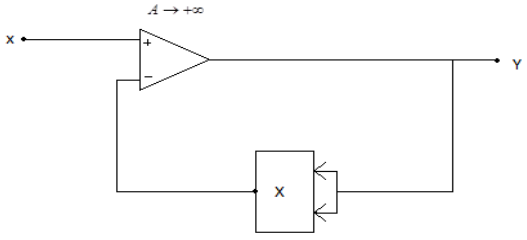


Figure 10. Model of square root as an inverse function

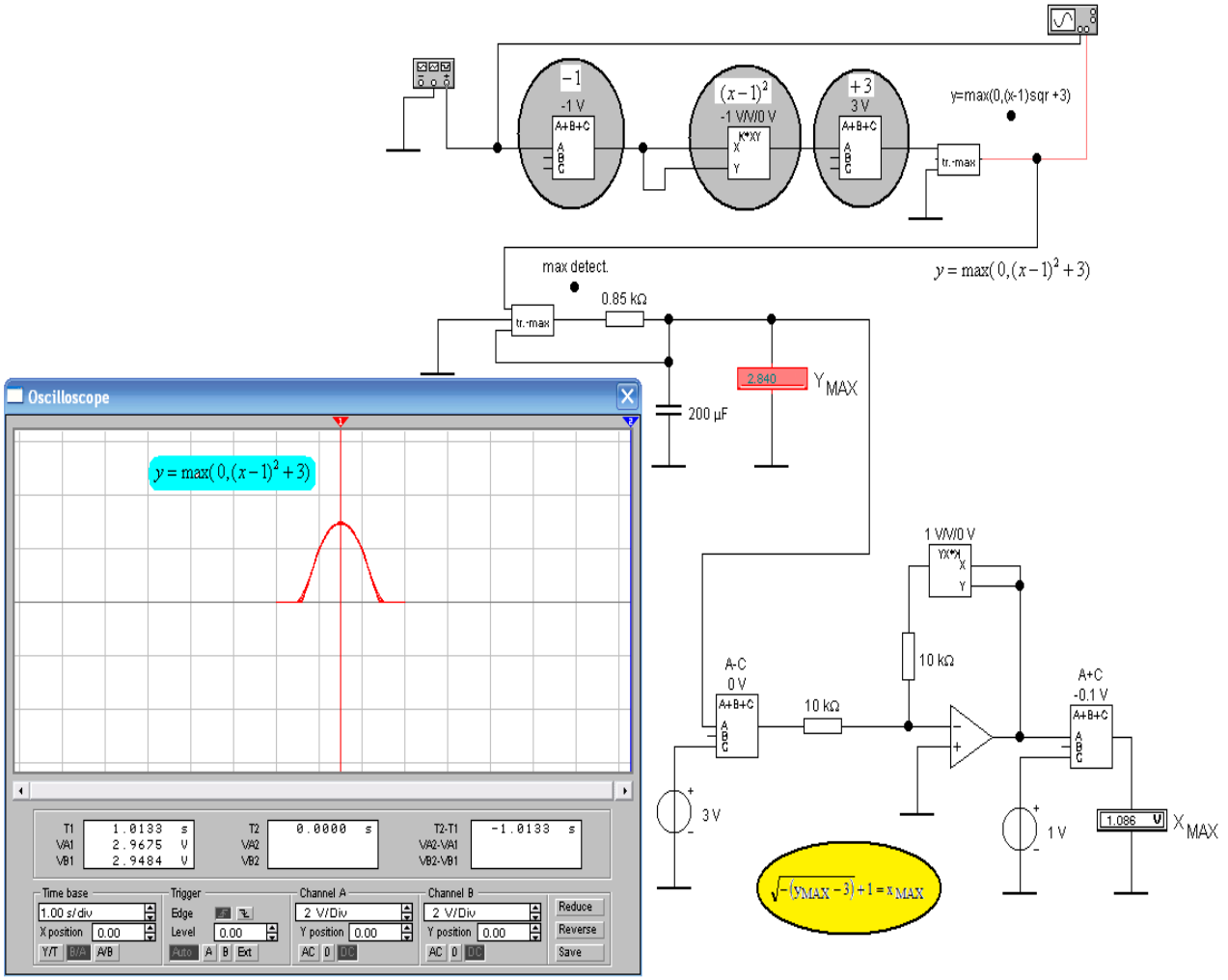


Figure 11. Electronic solution for the function $y = \max(0;-(x-1)^2+3)$ and the function extreme (*x*,*y* coordinates):

$$y_{MAX} \equiv V_{out_{MAX}} = \max(V_{in}, V_{out}),\ V'_{out} = \sqrt{-(y_{MAX}-3)}+1 = x_{MAX}$$

VI. FINDING MAXIMUM OF DISCONTINUOUS FUNCTION

When locating minimum and maximum of discontinuous functions mathematicians sometimes do some tricks, if possible [4]. For an electronic engineer finding the maximum of the function (6)

$$f(x) = \min((x)^2, (x-2)^2) \text{ for } x \equiv u_{in} \in [0,2] \quad (6)$$

is not „a big deal“ (see figure 12.). Input variable x corresponds to input voltage u_{in}. The function *f(x)* in electrical environment can be expressed as $u_{out} = \min((u'_{in})^2, (u'_{in}-2)^2)$ and the function domain $x \in [0,2]$ can be realized as $u'_{in} = \min(U_b, \max(0, u_{in}))$ where $U_b = 2$ V. Then a *min* detector is used to find out the extreme of the function (see figure 13.).

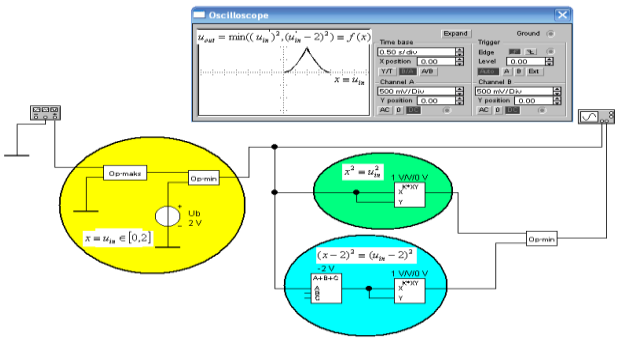


Figure. 12. Electronic realisation of the function $f(x) = \min((x)^2,(x-2)^2)$ for $x \equiv u_{in} \in [0,2]$

Realisation of the circuit (function) begins with so called *interval generator* $u'_{in} = \min(2; \max(u_{in}, 0))$, and then two common multipliers are used to generate the parabola u^2_{in} and the shifted parabola $(u_{in}-2)^2$.

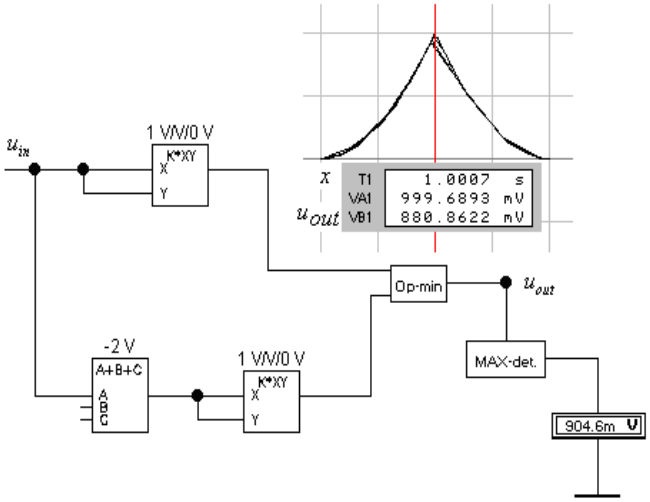


Figure 13. Electronic maximum of the function

$$f(x) \equiv u_{out} = \min((u_{in}')^2, (u_{in}' - 2)^2)$$

for $u_{in}' = \min(2, \max(0, u_{in}))$

un buffered max detector is used to "find out" the extreme of the function having discontinuity at $u_{in}' = 1$.

VII. FUZZY HEIGHT

Height of a fuzzy set

The height (hgt) of a fuzzy set is the maximum value of its membership function. Specifically, for a fuzzy set *A* in a universe *X*, with membership function μ, the height is given by term (7):

$$\mathrm{hgt}(A) = \sup_{x \in X} \mu_A(x) \qquad (7)$$

In a electronic hardware environment (see figure 14. and figure 15.) height of a fuzzy set can be interpreted as (8):

$$\mathrm{hgt}(A) = \max(f(V_{in})) \qquad (8)$$

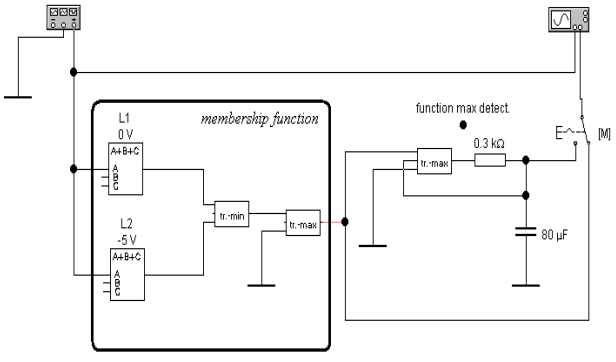


Figure 14. Fuzzy membership function *hgt peak detector. For detector output isolation and their max gate can be used efficiently used.*

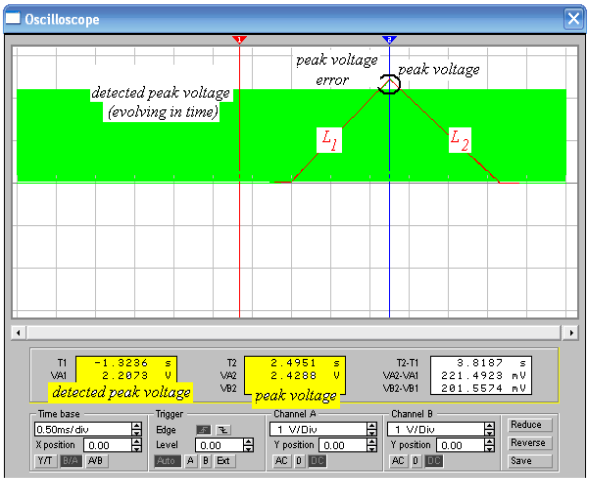


Figure 15. For the membership function $\mu_A(x) = \mathrm{maks}(0, \min(L_1, L_2))$ max peak detector error using EWB is about 8%. When using Multisim ver.13 the error is about 1 %

The height may be termed the modal grade of the fuzzy set, and the corresponding element value (i.e., the value of *x* where the membership grade peaks) may be termed the modal element value, or simply modal point [5].

VIII. CONCLUSION

A simple, new class of hardware peak detectors based on *min-max (Op-Amp)* technique is introduced and some applications in mathematical environment were briefly explained. Virtual electronic work-bench was used to find extremes of some functions, even in situations where classic mathematical methods failed. These circuits can have a additional (extended) characteristic; they are able to compare peak values with a pre-set voltage levels. Future work related to both kind of *min-max Op-Amp* peak detectors must go in to directions:

- hardware improvement of detectors
- finding the way to process (detect) local minimum and maximum

REFERENCES

[1] P.Horowitz, W.Hill, The Art of Electronics, Cambridge University Press, Cambridge, USA, 1989, pp. 217-220.

[2] D.Kovačević, Distribution and Sell Process Control Modelling Based on Intelligent Agent Network Realization, PhD thesis, FESB-Split, Croatia, 2012.

[3] D.Kovačević, N. Pribačić, Idea About a Different Lesson in a Different Classroom, 10th International Educational technology Conference, Proceedings Book Volume 1, Instanbul, Turkey, May 2011, pp. 496-501.

[4] G.James, Mothern Engineering Mathematics, Addison-Wesley Publishing Company, Harlow, England, 1996, pp. 462-463.

[5] B.Clagton, Linear integrated circuit Applications, The macmillan Press LTD, London, 1975, pp.5

[6] F.O.Karray, C.Silva, Soft Computing and Intelligent System Design, Addison-Wesley Publishing Company, Harlow, England, 2004, pp. 89.

DC VIS

International Conference on

DISTRIBUTED COMPUTING, VISUALIZATION AND BIOMEDICAL ENGINEERING

Steering Committee

Chairs: Karolj Skala, Ruđer Bošković Institute, Zagreb, Croatia
Roman Trobec, Jožef Stefan Institute, Ljubljana, Slovenia
Uroš Stanič, Biomedical Research and Innovative Society, Ljubljana, Slovenia

Members: Enis Afgan, Ruđer Bošković Institute, Zagreb, Croatia
Piotr Bala, N. Copernicus University, Toruń, Poland
Leo Budin, University of Zagreb, Croatia
Borut Geršak, University Medical Center Centre Ljubljana, Slovenia
Simeon Grazio, "Sisters of Mercy" University Hospital Centre, Zagreb, Croatia
Yike Guo, Imperial College London, UK
Gordan Gulan, Orthopaedic Clinic Lovran, Croatia
Ladislav Hluchý, Slovak Academy of Sciences, Bratislava, Slovakia
Peter Kacsuk, MTA SZTAKI, Budapest, Hungary
Aneta Karaivanova, Bulgarian Academy of Sciences, Sofia, Bulgaria
Charles Loomis, Paris-Sud University, Orsay Cedex, France
Luděk Matyska, Masaryk University, Brno, Czech Republic
Željka Mihajlović, University of Zagreb, Croatia
Damjan Miklavčič, Univeristy of Ljubljana, Slovenia
Laszlo Szirmay-Kalos, Technical University of Budapest, Hungary
Tibor Vámos, Hungarian Academy of Sciences, Budapest, Hungary
Matjaž Veselko, University Medical Center Centre Ljubljana, Slovenia

PAPERS

BIOMEDICAL ENGINEERING

Twenty-nine-year stimulation of the common peroneal nerve in a hemiplegic patient

P. Pečlin*, J. Krajnik**, R. Aćimović-Janežič**, U. Stanič*** and J. Rozman*
* ITIS d. o. o. Ljubljana, Centre for Implantable Technology and Sensors, Republic of Slovenia
** University Rehabilitation Institute (Employed during investigation), Ljubljana, Republic of Slovenia
*** Jožef Stefan Institute (Employed during investigation), Ljubljana, Republic of Slovenia
polonapeclin@gmail.com, uros.j.stanic@gmail.com, jnzrzmn6@gmail.com

Abstract - Functional electrical stimulation (FES) of the common peroneal nerve (CPN), to correct gait, has been shown to be a useful means for the restoration of functional movement in the lower extremities of hemiplegic individuals. Motor function over a twenty-nine year period of selective stimulation of the CPN with an implantable stimulator to correct drop-foot in a 56 year old patient with left-sided hemiplegia, was evaluated. Partial restoration of dorsal flexion and well balanced eversion of the affected foot was achieved by selective stimulation of superficial region of the CPN innervating mostly the tibialis anterior and partly peroneus longus and peroneus brevis muscles. It was shown that the system is suitable for very long-term FES providing improved muscle function. Namely, even after 29 years of daily FES of the CPN, a functional and reliable recruitment of fibres within the CPN was still achieved, thus providing an appropriate dorsal flexion and eversion of the affected foot. It was found that the stimulation-associated-mechanical trauma to the CPN elicited some minor non-age-related loss of the larger myelinated nerve fibres. Electrophysiological and biomechanical measurements however, have not revealed any functional signs which could be attributed to the damage of the CPN.

I. Introduction

Functional neuromuscular stimulation (FNS) of the common peroneal nerve (CPN) to correct gait has been shown to be a useful means for the restoration of functional movement in hemiplegia [1, 2, 3]. FNS is applied during the swing phase of the affected leg dorsiflexing the foot so that patients walk faster and more steadily [1, 2, 3].

Raising the affected foot, the patient releases the heel switch that turns the external unit on, thus transmitting the radio-frequency signal via the antenna to the implant and the nerve. At the beginning of the stance phase, FNS is turned off by closing the heel switch [1, 2, 3, 4, 5]. In FNS of a CPN, the recruitment of slow motor units was desired to optimize postural holding and function against gravitational forces [6, 7].

The present study was aimed at demonstrating the results on functionality of the implant implanted in a hemiplegic patient and comparing these results with those obtained with the implant implanted ten years before. The main goal, however, was to preserve an appropriate dorsal flexion and eversion of the affected foot with selective FNS of the particular superficial region of the CPN innervating mostly the tibialis anterior (TA) and partly peroneus longus (PL) and peroneus brevis (PB) muscles.

II. Materials and Methods

It was demonstrated in peripheral nerves that electrodes placed proximal to the branching point provide the ability to activate selectively an individual muscle [8, 9]. The CPN lies between the tendon of the biceps femoris and lateral head of the gastrocnemius muscle, winds around the neck of the fibula, between the PL and the bone, and divides beneath the muscle into the superficial peroneal nerve, deep peroneal nerve and recurrent tibial nerve. They arise just distal to the head of the fibula, and are directed forwards, diverging in their course, beneath the PL muscle.

It was presumed that at the level above bifurcation of the CPN, fibres belonging to TA muscle lie mostly within the particular region of the deep peroneal nerve and fibres belonging to PB and PL lie mostly within the particular region of the superficial peroneal nerve [9]. In order to achieve strong dorsal flexion and moderate eversion of the foot, three muscles TA, PB and PL (Fig. 1a), must be activated via selective stimulation of a superficial region of the CPN (Fig. 1b), comprising both aforementioned particular regions.

Informed consent was obtained from a 56-year-old patient with left-sided spastic hemiparesis, who recovered from an ischemic stroke occurred in January 1983. The first generation implantable stimulator (not shown in this paper) was implanted in 1984. However, wire console shaped loop anode broke out after ten years of daily use.

In the year 1995 the implant was replaced with the second generation monopolar one (Fig. 1c). The figure shows the monopolar half-cuff and the bulk body of the implant (dimensions of 22X22 mm). A platinum common electrode (geometric surface ~100 mm^2) was integrated on one side of the body. The leads exited the body on one of four sides through the feed-through. The half-cuff with a platinum electrode (Fig. 1c) was constructed considering both, the aforementioned requirements and the results of modelling monopolar selective FNS of superficial regions and fibres with different diameters in the CPN [3, 10, 11, 12].

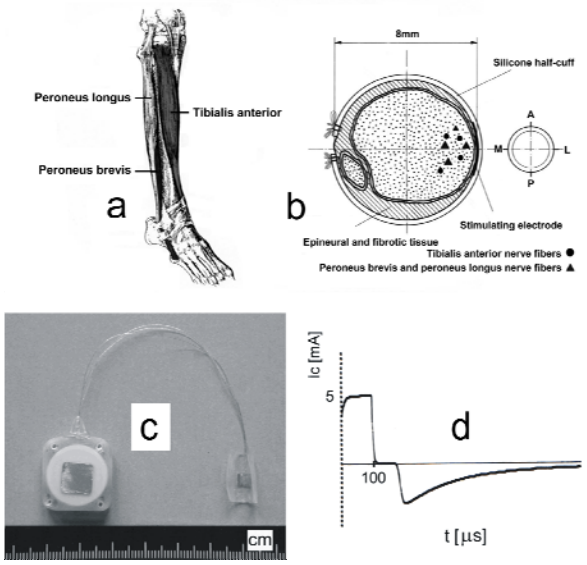


Fig. 1 (a) Muscles to be stimulated, (b) Fixation of half-cuff on the CPN, (c) Implantable stimulator, (d) Stimulating pulse.

The stimulus (Fig. 1d) was a current biphasic pair with rectangular cathodic component i_c which could be preset within the external unit between 1 and 10 mA and exponential decaying anodic component i_a. The width of the cathodic component t_c could be set by the patient from 0 to 500 µs, while the frequency f and the time delay t_d between the cathodic and the anodic part of the stimulus was preset at 22 Hz and at 50 µs, respectively [13].

This implant provided satisfactorily functional results for next ten years. However, due to forces generated by muscle activities and forces imposed from outside of the body the lead wires of the stimulating electrode broke. Consequently, the implant was replaced with the same monopolar one in March of 2008.

In local anaesthesia the CPN behind the lateral head of the fibula was exposed. To activate mainly the TA, PL, and PB, the stimuli had to be delivered to the cathode situated above the specific site on the CPN. To find this site, the half-cuff was moved above different sites until position eliciting optimum movement of the foot was detected [2]. Afterwards, the site was tested using different values of i_c and t_c while the f and the t_d remained preset at 20 Hz and 50 µs, respectively. Finally, the half-cuff was fixated at four corners as shown in Fig. 2a and 2b.

The external unit (Fig. 3) consists of: (a) transmitter that generates RF signals, (b) antenna that transmits RF signals through the skin to the implant and (c) heel switch which triggers the transmitter synchronously with the swing phase.

FNS started 14 days after surgery while gait analyses were performed 6 months after surgery. In the gait kinetics measurement performed in the year 2000, movements of the left leg segments were tracked with a 6-camera Vicon Motion System (model; Vicon 370 Datastation). For recording surface electromyograms (EMG) of 10 muscles a Noraxon (Myosystem 1200, Finland) was connected to the 64-channel analog data input.

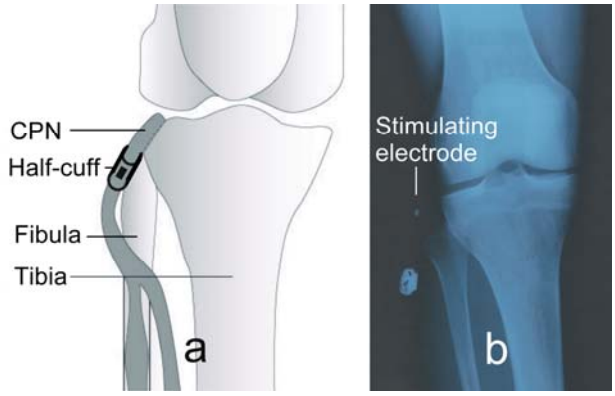


Fig. 2 (a) Site of half-cuff fixation, (b) X-Ray picture of the implant.

An off-line analyses were performed by a PC running the Vicon Polygon software.

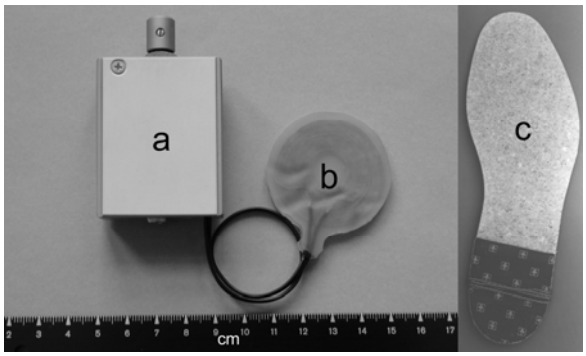


Fig. 3 (a) External transmitting unit, (b) Transmitter antenna and (c) Synchronizing heel switch.

In the year 2009 however, the measurements and analyses were performed using a 6-camera Vicon motion analysis system (model Vicon MX; Oxford Metrics Ltd, Oxford, United Kingdom) collecting ground reaction force data at 100 Hz. Illumination, video data collection, and analog-to-digital conversion of transducer input (force plate and EMG data) were synchronized and controlled by the Vicon MX, which was interfaced and controlled by a Pentium-based PC running the MS Windows operating system. For recording surface electromyograms (SEMG) of 12 muscles a Noraxon (Telemyo 2400T, Finland) was connected to the 64-channel analog data input. EMG signals were collected at 1080 Hz and were synchronized with the Vicon system for simultaneous collection. The off-line analyses were performed using the Vicon Polygon and Vicon Nexus software.

In both trials the patient landed on the two force platforms (model OR 6-7; Advanced Medical Technology, Inc, Watertown, MA, USA), which were used to collect ground reaction force data were collected at 1080 Hz and synchronized with the Vicon system.

Standard analyses included basic statistical analysis of a set of temporal and spatial parameters of gait, calculation and averaging of three dimensional

goniograms in the hip, knee, and ankle, rectifying and averaging of EMG signals, and finally calculation and averaging kinetic variables (forces, torques, powers) in the main joints of lower extremities. Two separate sessions of trials were performed with each measurement: one without stimulation and one with stimulation. All other conditions remained unchanged. The measured data were analysed using the method of Grieve [14] and compared to the normative data taken from Winter [15] if available, if not, the data from our own data base was used.

III. Results and Discussion

The results showed that after fifteen years the speed of stimulated gait significantly increased. After twenty-five years, however, the difference in the speed of stimulated and even non-stimulated gait increased even more.

It was observed that EMG activity of the left muscle triceps surae (TS) was weak during the swing phase while a slight increase in EMG activity during the weight acceptance phase was observed. Within the remaining part of the gait cycle a constant lowered tonic activity was present. This pattern was described as a typical sign of spasticity of the TS. Furthermore, ankle goniograms and length of the muscle soleus (SOL) and muscle gastrocnemius (GAST) demonstrate contractures in TS. Besides missing push off, one consequence of impairment in GAST was also weak knee flexion, especially at the beginning of the swing phase. Consequently, a patient has performed the swing phase with explicit circumduction, which was also confirmed by significantly increased external pelvic rotation during the left swing phase. The patient compensated for his inability to perform the left swing phase by generating additional power in the knee and hip, and with the right side. As additional compensation, the missing left push off also caused a shortening of the right step. It could be concluded from EMG of the left TA that this muscle was actually not affected and the pattern seemed to be almost normal. In the ankle goniogram recorded in the last measurement, a significant move towards dorsal flexion (~10°) could be observed while the shape of the first one was almost identical. Perhaps this difference could be attributed to different footwear used in the two 10-year separated measurements. Slightly increased EMG activity and the rather stable shape in all muscular groups in the second measurement could perhaps be attributed to a higher walking speed.

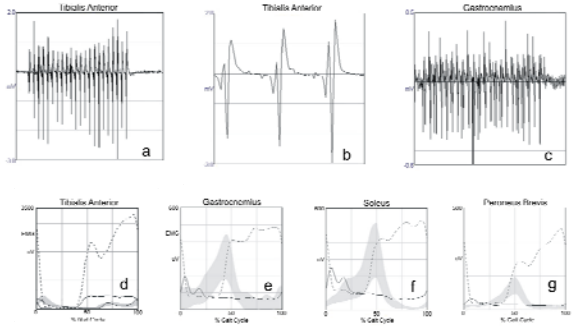


Fig. 4 (Upper) SEMG in the TA, plot (a), and plot (b), and GAST, plot (c). Plots (a) and (c) show entire FNS time interval while plot (b) shows a shape of activation in motor units, elicited by an individual stimulating pulse (M-wave). (Lower) Rectified and averaged SEMG in muscles TA, GAST, SOL and PB during FNS. Solid lines - measurement without FNS, dashed lines - measurement with FNS and grey thick line - normative data.

It could be seen in Fig. 4 (lower part), representing EMG recorded in TA, PB, SOL and GAST, that all these muscles were stimulated. Perhaps, some other muscles like, for instance, TS innervated by fibres of the CPN whose stimulation was not considered in the model, were also stimulated. Accordingly, a rectified EMG activity in the TA, PB, and TS (Fig. 4, lower) had the following maximum values: 3 mV, 1 mV and 0.5 mV, respectively. It was observed that without stimulation there was no normal EMG activity of the flexor TA while EMG activity of the PL could be considered as normal. However, with selective stimulation of the CPN both EMG activity of the TA and PL could be observed.

Plots (a), (b) and (c) in Fig. 5 clearly demonstrate the effectiveness of stimulation by showing a position of center of pressure (COP) along the sole (X direction) (COPX) and, especially plot (d) by showing the ankle goniogram. Furthermore, correction of ankle inversion, which should be elicited by stimulation of the PL and PB, was favorable in the measurement performed in the year 2000 as shown in Fig. 5.

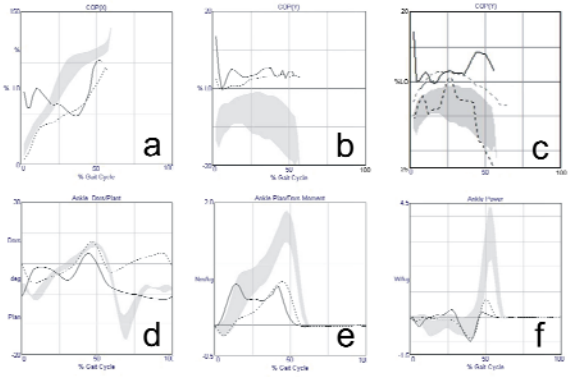


Fig. 5 (Upper) A trajectory of COP along the sole. Plot (a) and plot (b) measured at 2009, plot (c) measured at 2000. (Lower) An influence of FNS on increase in dorsal flexion plot (d), on torque in the ankle plot (e) and, on push off power plot (f). Solid lines-measurement without FNS, dashed lines-measurement with FNS and grey thick line-normative.

Dorsal flexion in the swing phase was significantly larger, thus enabling heel contact. Moreover, a co-activation of TS, elicited by unexpected stimulation, actually stabilized the ankle at heel contact and additionally contributed to heel contact becoming more steady. FNS significantly improved the shape of the ankle torque plot in the sagittal plane. However, as a consequence of severely affected TS, amplitude in the mid stance was heavily decreased. In fact, during this interval, TS was only able to produce passive tonic activity. Immediately after heel lift off, stimulation was turned on, thus causing activity to appear in these muscles and, consequently, contributing to the push off. This could be seen in plot (e) of Fig. 5, showing the torque in the ankle.

Results also revealed that left support lasted longer. Furthermore, more steadily heel contact decreased compensational pelvic movement and thus stabilized flexion/extension within the hip.

IV. CONLUSION

In comparison to similar systems developed elsewhere [2], the presented system requires a relatively simple procedure of implantation and is suitable for a wide population of patients.

Electrophysiological measurement results obtained using the method of needle electromyography showed a minimum sign of functional damage induced by the cuff or by the stimuli to the CPN. Namely, within the twenty-nine year period of FNS a slightly reduction (about 5 m/s) of conduction velocity of the CPN and weak pathological spontaneous activity, in the TA and in the PL was observed [6]. A value of the conduction velocity measured in 1993 was 55 m/s, seven years later, in 2000 was 52 m/s and nine years later in 2009 was 50 m/s.

It was also observed that cuff implantation resulted in thickened epineurium and endoneural connective tissue. A connective tissue growth, in greater extent in the CPN segment located closer to the implantation site. On the other hand, vascularization was appeared almost normal.

In this particular case, an obvious cause of damage to the CPN could include the use of a tight elastic band over the implanted region. Examination of the legs however, did not show a more severe loss of muscle control over the legs and feet and muscles atrophy.

ACKNOWLEDGMENT

This work was funded by the following research grants: J2-3415 and P3-0171 from the Ministry of Higher Education, Science and Technology, Ljubljana, Republic of Slovenia, and HPRN-CT-2000-00030 from the European Commission.

REFERENCES

[1] Liberson WT, Holmquest HJ, Scot D, Dow M. Functional electrotherapy: stimulation of the peroneal nerve synchronized with the swing phase of the gait of hemiplegic patients. Arch Phys Med 1961; 42:101-105.

[2] Waters RL, McNeal DR, Faloon W, Clifford B. Functional electrical stimulation of the peroneal nerve for hemiplegia. J Bone and Joint Surg 1985;67:792-793.

[3] Rozman J, Krajnik J, Gregorič M. Selective Stimulation of the Common Peroneal Nerve for Hemiplegia: Long-Term Clinical Follow-up. Basic Appl Myol 2004;14:223-229.

[4] Rozman J, Pihlar B, Strojnik P. Surface examination of electrodes of removed implants. Scand J Rehab Med Suppl. 1988;17:99-103.

[5] Garozzo D, Ferraresi S, Buffatti P. Surgical treatment of common peroneal nerve injuries: indications and results. A series of 62 cases. J Neurosurg Sci. 2004;48:105-12.

[6] Johnson MA, Polgar J, Weightman D, Appleton D. Data on the distribution of fibre types in thirty-six human muscles-an autopsy study. J Neurol Sci 1973;18:111-129.

[7] Comerford MJ, Mottram SL. Functional stability re-training: principles and strategies for managing mechanical dysfunction. Manual Therapy 2001;6:3-14.

[8] Sunderland S. Nerve and Nerve Injuries. 2nd ed. Baltimore: WB Saunders; 1978, 925-991.

[9] Sweeney JD, Ksiensky DA, Mortimer JT. A nerve cuff technique for selective excitation of peripheral nerve trunk regions. IEEE Trans 1990;BME-37:706-715.

[10] McNeal DR, Bowman BR. Selective activation of muscles using peripheral nerve electrodes. Med & Biol Eng & Comput 1985;23:249-253.

[11] Warman EN, Grill WM, Durand D. Modeling the effects of electric fields on nerve fibers: Determination of excitation thresholds. IEEE Trans Biomed Eng 1992;BME-39:1244-1254.

[12] Rattay F. Ways to approximate current-distance relations for electrically stimulated fibres. J Theor Biol 1987;125:339-349.

[13] Roblee LS, Rose TL: The electrochemistry of electrical stimulation. Proceedings, 12th Ann Conf IEEE Eng in Med & Biol Soc, Philadelphia, 1-4 Nov, 1990,1479-1480.

[14] Grieve DW. Gait patterns and the speed of walking. Biomed. Eng.1986; 3:119-122.

[15] Winter DA. Biomechanics and motor control of human gait, University of Waterloo Presss, Waterloo, Ontario, Canada N2L 3G1, 1991.

A method for time-domain analysis of respiratory sinus arrhythmia

R. Trobec and V. Avbelj
Department of Communication Systems, Jožef Stefan Institute, Ljubljana, Slovenia
{roman.trobec; viktor.avbelj}@ijs.si

Abstract - Respiration-related heart rate variability – respiratory sinus arrhythmia (RSA) – is clearly related to vagal activity. Usually it is analysed in frequency domain. An improved method for the assessment of the RSA in time-domain is proposed and evaluated on a control group of 12 healthy volunteers (9 male, 3 female; age 31-70). The method is based on a measurement of respiration rate, combined with a reliable beat-to-beat heart rate determination. RR intervals of each respiration cycle are superimposed to obtain average RSA. The study of RSA in the time-domain can provide more information about the adaptation process and possible parasympathetic response of the heart, e.g. after heart surgery, because the shape of a typical RSA curve and its phase (relative to the measured breathing) are obtained. The time domain analysis of RSA is simple and does not need forced timing of respiration. We confirmed that the proposed methodology could be used in ambulatory and remote monitoring environments to follow the long-term changes in the RSA.

I. INTRODUCTION

Respiratory sinus arrhythmia (RSA) is a respiratory-related heart rate variability (HRV) observed universally in vertebrates [1]. The largest part of this variability is mediated through the vagus, so RSA has been used as an index of cardiac vagal control [2, 3]. It was reported [4] that different methods of heart surgery could influence the heart rate variability [5, 6]. With transplantation, the donor heart is completely disconnected from the central nervous system, but a small level of RSA still remains. This residual RSA is of non-neural origin and can be related to perfusion pressure [7] and to mechanical stretch of the donor sinus node in the atrium wall caused by changes in venous return [8]. Small values of RSA can arise from imperfections in measurement due to slight time variations of the fiducial point in the QRS complex caused by changes in the cardiac electrical axis during breathing [9, 10].

The study of RSA after heart transplantation can provide a unique model to assess the process of adaptation and possible re-innervation of this complex transplanted organ [11]. Most studies have used the frequency domain approach, together with programmed breathing, periodic neck suction and other manoeuvres [12, 13, 14]. Only rarely has the time-domain approach been used [15, 16].

The standard techniques for assessment of the parasympathetic re-innervation such as scintigraphy, positron emission tomography, or baroreflex stimulation, are rather complex or can influence the conscious of patients biasing the results. This study was designed to characterize RSA in cardiac transplanted recipients while allowing spontaneous breathing. We hypothesise that the phase between the spontaneous breathing signal and the measured RSA could be an indicator of the vagal re-innervation. The shape of the RSA near the maximal intratoracial pressure can give further information about the HRV control loop. The method for assessing RSA was first developed for an unpublished preliminary investigation of sympathetic re-innervation of transplanted hearts in our previous study [4]. We have further refined the method to become a more sensitive tool for investigating RSA. We confirmed that the proposed methodology could be used in ambulatory and remote monitoring environments [17, 18] to follow the long-term changes in the RSA.

II. METHODS

The study [4] was approved by the National Medical Ethics Committee of the Republic of Slovenia. Written informed consent was obtained from all the subjects before the study. A control group of 12 healthy volunteers (9 male, 3 female; age 31-70) was enrolled. None of the control subjects was receiving any medication or had a history of cardiovascular disease. 15 patients (age 30-71) 0.5-8.1 years after heart transplantation were also included in the study for an ongoing detailed analysis, not published yet.

The surface electrocardiogram, using bipolar lead CM5, was recorded with high sample rate of 1800 Hz and resolution of 2 μV (DEKG recording device, Intekom, Slovenia). For each subject, an 8.5-minute ECG recording was obtained in the supine position after 5 minutes rest. Breathing was recorded simultaneously (16-times lower sample rate) with a small mass temperature sensor placed in front of the nose (thermistor 10K3MCD1, Betatherm, Ireland).

The recordings were analyzed offline with a custom-made computer program. The moments of the maximal amplitude of the R-waves (R) in the ECG were determined by an automated computer-based peak detection algorithm [19]. The time intervals between neighbouring R peaks (RRI), together with the absolute times of the R peaks, were used in further analysis. The fiducial point in the respiration cycle (E) was defined as the maximal rate of change in the exhaled air temperature. The intervals between consecutive exhalations are denoted by EEI. The E event was determined again by an

The study was supported in part by the Slovenian Research Agency under grant P2-0095..

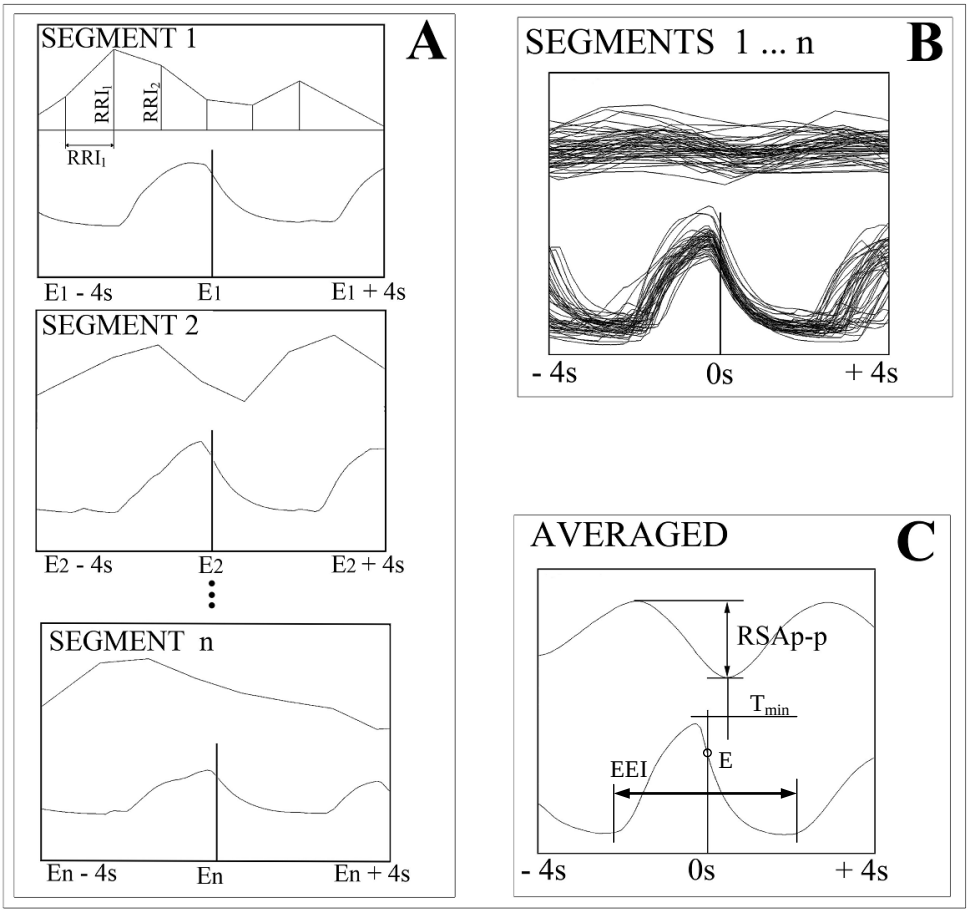


Figure 1. The method for RSA shape extraction. (A) The time (R_i) at the peak point of R-wave was determined for all R-waves in the ECG recording. HRV was presented as a series of R events at R-wave peak times with amplitudes $RRI_i = R_i-R_{i-1}$. All RRI points were connected by lines and re-sampled at 100 Hz. The E events were obtained from the respiration recording, calculating maximal negative slope in the respiration signal. (B) 8-second segments of HRV and respiration signal with E events placed in the middle were superimposed and (C) averaged, giving the RSA pattern and respiration curve. RSAp-p is the peak-to-peak RR-interval deviation during respiration cycle, EEI is average respiration time, and T_{min} is interval between E and the nearest minimum in RSA.

automated computer-based detector; nevertheless, a final visual inspection was applied to ensure correct time determination of the R and E events. HRV patterns of 8 seconds segments (± 4 s around E) were superimposed and averaged with each respiration cycle. The main reason for the HRV is the respiration and therefore we termed the heart variability as RSA. The time difference between the fiducial point E in the respiration cycle and the nearest minimum in the RSA is denoted by T_{min}. The time-domain analysis of RSA and respiration was performed on the resulting two averaged signals. The method for RSA shape extraction is shown in Figure 1 for a person from the control group.

III. RESULTS

Typical results of the described method, obtained from a healthy subject, are given in Figure 2. On the left plot (a), aligned HRV segments in the range of ± 4 seconds around E event, which are supposed to happen at time 0 s, are shown with green curves. Average RSA segments over 2 minutes windows are shown with colour curves. We can conclude that just small time invariance is present in the 8.5 minutes recording because all averaged signals are quite similar. The RSA is coherent with the respiration with RSAp-p = 60 ms. Aligned respiration signals are shown in the same way on the central panel (b). The sinusoidal RSA signal shape follows from logical deduction that RRI increases if the inner thorax pressure decreases and vice versa. Note that RRI is increasing soon after the maximum intensity of exhalation (E event).

On the right panel (c) the coherence and the phase between the RSA and the respiration signal are given for comparison with the frequency-domain analysis. As expected, there is a significant coherence in the respiration frequency band with a stable phase of $\varphi = -110^{\circ}$, which is in relation with the delay between the maximal values of RSA and E events in the respiration signals. However, no further information about the RSA shape is obtained from the frequency-domain analysis.

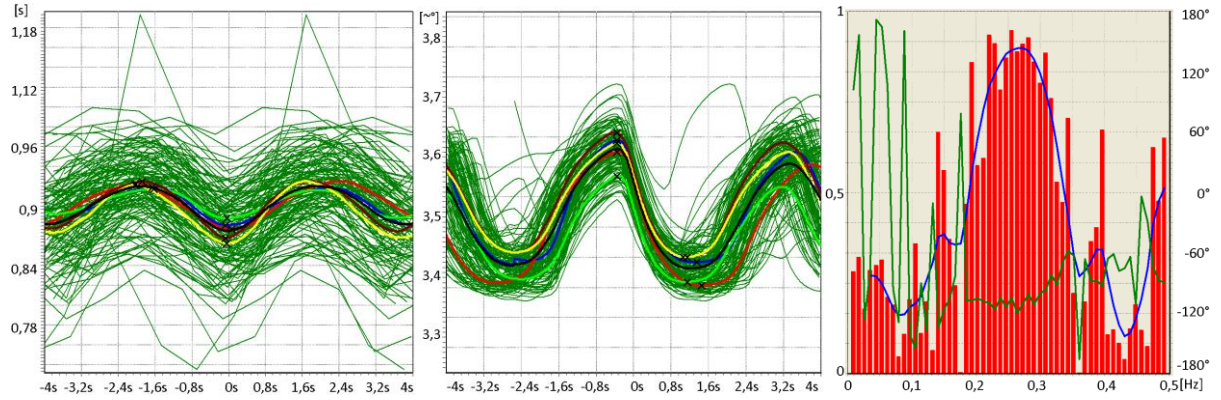
Measurements of the control group are quite uniform and all similar to the representative case shown in Figure 2, which is confirmed by the statistical data in Table 1. RSA signals increase after a fixed delay T_{min} from the fiducial point of the maximum intensity in the exhalation cycles (E event), indicating decreased heart rates as a consequence of increased intrathoracic pressures around E event. The phases of both series possibly reflect the cause-effect relation between exhalation and decreased heart rate.

TABLE I. STATISTICAL VALUES OF HEART BEATS, RESPIRATION, RSA AND T_{MIN} FOR THE CONTROL GROUP.

	Controls (n=12)
RRI [ms]	822 ± 134
EEI [s]	3.97 ± 0.59
RSAp-p [ms]	23 ± 15
T_{min} [s]	0.19 ± 0.17

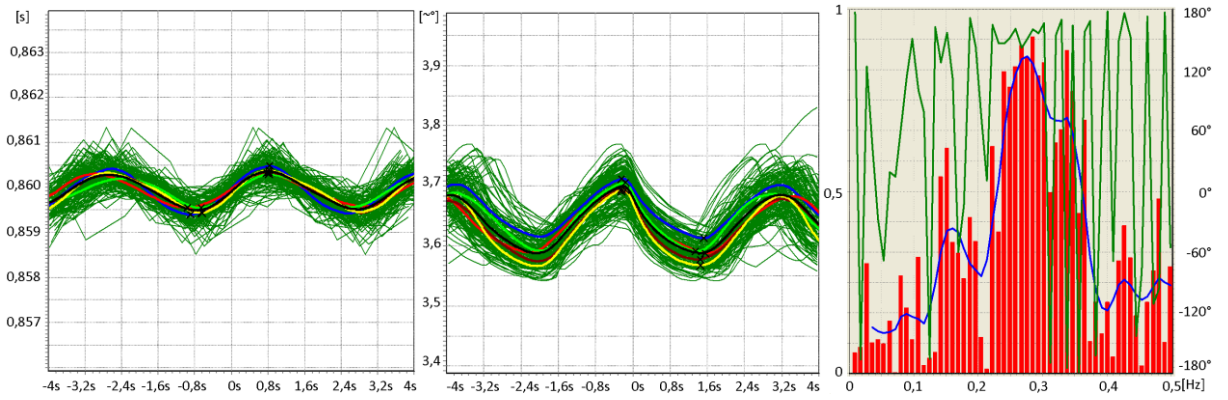
between HRV and respiration peak, and it seems that this value is also unreliable.

Preliminary results from measured RSA and respiration curves from the remaining patients are quite divergent in amplitudes and shapes and therefore more in-depth analysis are needed. However, they are evidently statistically different from the control subjects.



(a) (b) (c)

Figure 2. (a) Aligned HRV segments (green), around E event at time 0 s, for a typical control subject. Average RSA from 2 minutes windows are shown with different colours. (b) Same as in (a) but for the respiration signal from thermistor. (c) Coherence between RSA and the respiration signal with corresponding phase.



(a) (b) (c)

Figure 3. (a) Aligned HRV segments (green), around E event at time 0 s, for a patient with an active pacemaker. Note that the y-scale is smaller as in Figure 2. Average RSA from 2 minutes windows are shown with different colours. (b) Same as in (a) but for the respiration signal from a thermistor. (c) Coherence between RSA and the respiration signal with corresponding phase.

An additional control measurement of a patient with a pacemaker is shown in Figure 3. The pacemaker was set to a constant frequency in order to evaluate the performance of the proposed method. The RSAp-p should be absent in this case, however we obtained 0.9 ms. This result is probably a consequence of variations in the R-wave shapes because of movements of the thorax and the ECG electrodes during breathing. We can conclude that only RSAs with RSAp-p > 1 ms can be identified with our method and considered as valid.

We still have sinusoidal RSA signal shape, but it seems that "effect" (RRI increase) is here app. 800 ms before "cause" (thorax pressure increase at E event). The phase $\varphi = 150°$ (see Figure 3(c)), expresses just the delay

IV. CONCLUSION

The time domain analysis of RSA is simple and does not need the metronomic breathing and other manoeuvres, like stimulation of carotid baroreceptors by vacuum pulses, that can influence the parameters of conscious breathing. The method presented can be applied retrospectively to ECG measurements in cases where breathing pattern is measured simultaneously or can be reconstructed from the ECG data [20]. The proposed methodology can be used in ambulatory environment to measure the vagal activity or to follow the level of functional re-innervation of transplanted hearts.

The proposed time-domain analysis could be a complementary method of RSA investigation because of

several reasons. HRV segments in aligned sections were averaged and therefore inherently filtered to produce a diagram of RSA with the evident response to the respiration cycle, regarding the phase and shape of the control mechanisms. The frequency components of the RSA corresponding to the breathing stay unaffected, while other components, either short-term or long-term, are reduced due to the averaging process. The method does not require periodic signals and therefore allows spontaneous breathing. It allows sub-interval analysis in order to explore time invariability of each measurement. The method is appropriate for analyses of signals with extremely low variability.

The proposed time-domain analysis of RSA can elucidate the vagal activity, level of re-innervation and other processes after heart surgery. We demonstrated, on a small investigated group, that RSA shape, amplitude and phase are stable and consistently support the cause-effect relation in control subjects. However, preliminary analysis of measurements from patient after heart transplantation indicates that the amplitude and phase are not stable and do not consistently support the cause-effect relation. The preliminary results also indicate that the shapes of RSA curves are significantly different between patients, which remains to be clarified with further research.

ACKNOWLEDGMENT

The authors gratefully acknowledge the help of Andrej Šmigoc in preparing the presentation of the method for RSA shape extraction in Figure 1.

REFERENCES

[1] E.W. Taylor, D. Jordan, and J.H. Coote, "Central control of the cardiovascular and respiratory systems and their interactions in vertebrates," Physiol Rev., vol. 79(3), pp. 855-916, July 1999.

[2] D.L. Eckberg, "Human sinus arrhythmia as an index of vagal cardiac outflow," J Appl Physiol., vol. 54(4), pp. 961-966, April 1983.

[3] J.A. Taylor, C.W. Myers, J.R Halliwill, H. Seidel, and D.L. Eckberg, "Sympathetic restraint of respiratory sinus arrhythmia: implications for vagal-cardiac tone assessment in humans," Am J Physiol Heart Circ Physiol., vol. 280(6), H2804-2814, June 2001.

[4] S.S Lovric, V. Avbelj, R. Trobec, D. Zorman, P. Rakovec, S. Hojker, B. Gersak, and M. Milcinski, "Sympathetic reinnervation after heart transplantation, assessed by iodine-123 metaiodobenzylguanidine imaging, and heart rate variability," Eur J Cardiothorac Surg., vol. 26(4), pp. 736-741, October 2004.

[5] S. Frljak, V. Avbelj, R. Trobec, and B. Gersak, "Beat-to-beat heart rate and QT interval variability in patients undergoing coronary artery bypass graft surgery," Computers in Cardiology, vol. 29, pp. 661-664, 2002.

[6] J.M. Kalisnik, V. Avbelj, R. Trobec, D. Ivaskovic, G. Vidmar, G. Troise, and B. Gersak, "Assesment of cardiac autonomic regulation and ventricular repolarization after off-pump coronary artery bypass grafting," Heart Surgery Forum, vol. 9 (3), pp. 169-175, 2006.

[7] D.P. Slovut, J.C. Wenstrom, R.B. Moeckel, C.T. Salerno, S.J. Park, and J.W. Osborn, "Beat-to-beat modulation of heart rate is coupled to coronary perfusion pressure in the isolated heart," J Appl Physiol., vol. 86(2), pp. 694-700, February 1999.

[8] L. Bernardi, F. Keller, M. Sanders, P.S. Reddy, B. Griffith, F. Meno, and M.R. Pinsky, "Respiratory sinus arrhythmia in the denervated human heart," J Appl Physiol., vol. 67(4), pp. 1447-1455, October 1989.

[9] A. Rashkovska, and V. Avbelj, "Signal processing methods for ST variability assessment in ECG," Proc. of 36th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2013, pp. 331-334, May 2013.

[10] A. Rashkovska, R. Trobec, M. Depolli, and V. Avbelj, "Time and frequency analysis of non-uniform sampling," Proc. of 33rd International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2010, pp. 236-240, May 2010.

[11] K. Nytrøen, J. Myers, K.N. Chan, O.R. Geiran, and L. Gullestad, "Chronotropic responses to exercise in heart transplant recipients: 1-yr follow-up," Am J Phys Med Rehabil, vol. 90, pp. 579-588, 2011.

[12] L. Bernardi, B. Bianchini, G. Spadacini, S. Leuzzi, F. Valle, E. Marchesi, C. Passino, A. Calciati, M. Vigano, M. Rinaldi, et al., "Demonstrable cardiac reinnervation after human heart transplantation by carotid baroreflex modulation of RR interval," Circulation, vol. 92(10), pp. 2895-903, November 1995.

[13] P. Uberfuhr, A.W. Frey, and B. Reichart, "Vagal reinnervation in the long term after orthotopic heart transplantation," J Heart Lung Transplant., vol. 19(10), pp. 946-950, October 2000.

[14] I. Pinhas, E. Toledo, D. Aravot, and S. Akselrod, "Bicoherence analysis of new cardiovascular spectral components observed in heart-transplant patients: statistical approach for bicoherence thresholding," IEEE Trans Biomed Eng., vol. 51(10), pp. 1774-1783, October 2004.

[15] O. Gilad, C.A. Swenne, L.R. Davrath, and S. Akselrod, "Phase-averaged characterization of respiratory sinus arrhythmia pattern," Am J Physiol Heart Circ Physiol., vol. 288(2), pp. H504-510. September 2004.

[16] P. Loula, V. Jantti, A. Yli-Hankala, "Respiratory sinus arrhythmia during anaesthesia: assessment of respiration related beat-to-beat heart rate variability analysis methods," Int J Clin Monit Comput., vol. 14(4), pp. 241-249, November 1997.

[17] A. Rashkovska, I. Tomašić, K. Bregar, and R. Trobec, "Remote monitoring of vital functions - Proof-of-concept system," Proc. of 35th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2012, pp. 446-450, May 2012.

[18] A. Rashkovska, I. Tomašić, and R. Trobec, "A Telemedicine application: ECG data from wireless body sensors on a Smartphone," Proc. of 34th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2011, pp. 262-265, May 2011.

[19] I. Tomasic, S. Frljak, and R. Trobec, "Estimating the universal positions of wireless body electrodes for measuring cardiac electrical activity," IEEE Transactions on Biomedical Engineering, vol. 60(12), pp. 3368-3374, 2013.

[20] R. Trobec, A. Rashkovska, and V. Avbelj, "Two proximal skin electrodes - A respiration rate body sensor," Sensors, vol. 12(10), pp. 13813-13828, 2012.

Review of Regularization Techniques in Electrocardiographic Imaging

Matija Milanič*, Vojko Jazbinšek** and Rok Hren**
*J. Stefan Institute, Ljubljana, Slovenia
**Institute of Mathematics, Physics and Mechanics, Ljubljana, Slovenia
Email: vojko.jazbinsek@imfm.si

***Abstract*—Regularization methodologies are an integral part in dealing with ill-posedness of inverse problem in electrocardiograhy, expressed in terms of potential distribution on the epicardium. In order to systematically evaluate various regularization techniques under controlled conditions, we employed progressively more complex idealized source models (from single dipole to triplet of dipoles) to calculate body surface potentials, which served as an input data to the inverse problem. In total, we examined, 13 different regularization techniques and found that non-quadratic methods (total variation algorithms) and first-order and second-order Tikhonov regularizations outperformed other methodologies.**

I. Introduction

It has been shown that there is an infinite number of internal electrical heart sources that can produce the same potential distribution on the body surface [1], [2]. Therefore, to solve such ill-posed inverse problem, an equivalent source model has to be presupposed. Some of the models, e.g., single dipole or single quadrupole, are rather crude approximations of the entire heart's electrical activity, while other models, e.g., epicardial potentials, can be actually measured. The inverse solution that employs the potential distribution on the epicardial surface as an equivalent source model or electrocardiographic imaging (ECGI) [3] has been widely studied in electrocardiography due to its inherent ill-posedness [4]–[7]. A plethora of regularization techniques have been applied to gauge wildly oscillating inverse solutions, however, there is a growing need to compare, structure, and unify those diversified regularization methods. In particular, such a comparison requires the use of the same volume conductor and the same cardiac source models in order to eliminate other sources of variation in results. In this study, we employed the same volume conductor model as in a related study [8] to systematically evaluate the performance of 13 different regularization techniques using progressively more complex idealized source models (from single dipole to triplet of dipoles).

II. Methods

A. Problem formulation

The equivalent potential distribution on the epicardial surface can be found from a known potential distribution on the torso surface - which can be in principle obtained using multichannel ECG devices [9] - by solving generalized Laplace's equation subjected to Cauchy boundary conditions [4]–[7]. Such a boundary-value problem must be in an arbitrarily shaped volume conductor approximated on a discretized solution domain as the system of linear equations. For the homogeneous and isotropic model of the human torso, this can be achieved by means of the boundary element method (BEM), which relates the potentials at the torso nodes (expressed as an m-dimensional vector Φ_B) to the potentials at the epicardial nodes (expressed as an n-dimensional vector Φ_E)

$$\Phi_B = \mathbf{A}\Phi_E \tag{1}$$

where $\mathbf{A}$ is the transfer coefficient matrix ($m \times n$) and $n < m$. The transfer coefficient matrix depends entirely on the geometric integrands which can be calculated analytically [10], [11]. In principle, the epicardial potential distribution could be simply found in the form of a pseudo inverse; however, the matrix $\mathbf{A}$ is ill-conditioned, i.e., its singular values are limiting to zero with particular gap of separation in the singular value spectrum, which yields an unstable solution. A number of approaches have been developed to control the wild oscillations of the solution and the ones we have used and described in more details in the related study [8] are divided into three groups

1) Tikhonov regularizations: zero order (ZOT) [12], [13], first order (FOT) [7], [14] and second order (SOT) [13],
2) iterative techniques: truncated singular value decomposition [15] (zero order (ZTSVD), first order (FTSVD), and second order (STSVD)), conjugate gradient [16] (zero order (ZCG), first order (FCG), and second order (SCG)), ν-method (NU) [15], and MINRES method [16],
3) non-quadratic techniques [7], [17]: total variation (FTV), and total variation with Laplacian (STV).

We used geometry based on the torso shaped outer boundary, defined with m=771 nodes, and an internal, barrel-shaped cage with n=602 electrodes, which surrounded all sources (surrogate for the epicardial surface) [8].

B. Selection of sources

We created a cylindrical source space, consisted of 744 dipoles in 248 positions, which were arranged in 8 axial planes 10 mm apart along the polar z-axis, see Fig. 1. There were three perpendicular dipoles in each position: normal (radial, $\vec{p}_\rho$) and tangential (along the polar angle, $\vec{p}_\varphi$) to the cage side surface, and alongside the polar axis ($\vec{p}_z$). In the first and second source (axial) planes there were 7 source positions (Fig. 1c), one in the center and 6 arranged along a concentric circle with a radius of 10 mm (nodes of hexagon with side 10 mm). The center of the bottom plane was at (-16, 45, 230)

mm in the Cartesian coordinate system of the torso-cage model. In the next two source planes (Fig. 1d), there were 19 positions on each plane, 7 as in the first and second planes and 12 arranged on an additional concentric circle with a radius of 20 mm (nodes of a dodecagon with side 10 mm). In the next two source planes, there were 37 positions on each plane (Fig. 1e), 19 as in the previous two planes, and 18 on an additional concentric circle with a radius of 30 mm (equivalent to the nodes of an octadecagon with a side length of 10 mm). In the last two source planes, there were 61 positions on each plane (Fig. 1f), 37 as in the previous two planes and 24 on an additional concentric circle with a radius of 40 mm (nodes of tetracosagon with sides of 10 mm).

From the above source space, we selected three different idealized source models:

1) single dipoles at 16 different locations and 3 different orientations (in total 48 combinations),
2) pairs of dipoles at 324 combinations, and
3) triplets of dipoles at 24 different combinations.

For single dipole source, we selected sources in four planes at levels (z_i in mm): bottom plane ($z_0 = 230$, Fig. 1c), two planes in the middle ($z_1 = 260$, Fig. 1d and $z_2 = 270$, Fig. 1e), and upper plane ($z_3 = 300$, Fig. 1f). On each plane, we selected four positions (in cylindrical coordinates ρ – radial distance, φ - polar angle, and z – height), one in the center (polar axis) and three on the outermost circle of a given plane (ρ_0=10, ρ_1=20, ρ_2=30, ρ_3=40 mm) at polar angles φ=-180°, -120° and -60°, see Figs. 2a–c.

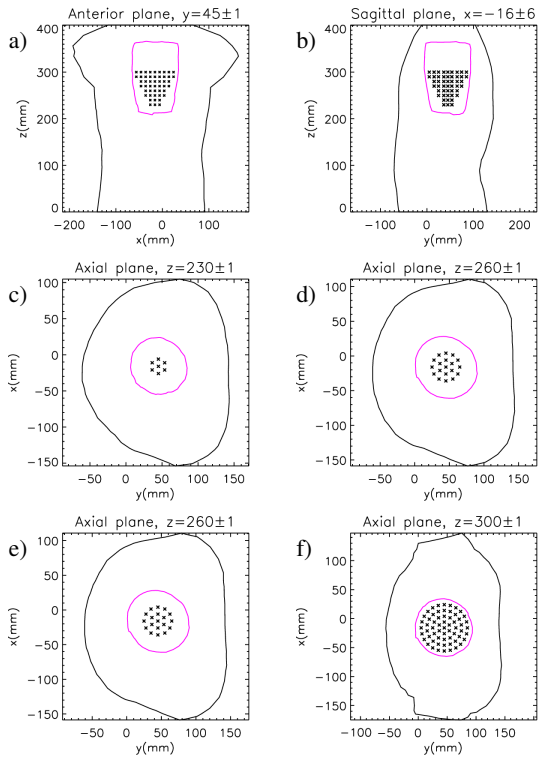


Fig. 1. Cross-sectional views of dipole positions denoted by a) Anterior, b) Sagittal and (c through f), Axial planes at different z levels. Torso and cage borders are displayed with black and magenta colors, respectively. On each plane, sources within x or y or z levels ± tolerance are displayed. The polar axis of the cylindrical cage is at x=-16 mm and y=45 mm.

For dual dipole sources, we selected single dipole sources in three planes at levels (z_i): two planes in the middle ($z_1 = 260$ and $z_2 = 270$), and in the upper plane ($z_3 = 300$), see Figs. 2d–f. Note, that we kept the same notation of planes as in the case of single dipole sources. On each plane, we selected only dipoles in the outermost circle with radial distances ρ_1, ρ_2 and ρ_3, respectively. We put the first single dipole on the right side (polar angle φ=-180°). The other dipoles are then positioned in angular steps of $\Delta\varphi_i$ in the anterior part

$$\varphi = -180^\circ + \Delta\varphi_i,\ -180^\circ + 2\Delta\varphi_i,\ \ldots, 0^\circ,$$

where $\Delta\varphi_1$=30°, $\Delta\varphi_2$=20° and $\Delta\varphi_3$=15°, with 6, 9 and 12 additional dipole positions in the three selected planes, respectively. We combined dipoles along the same direction (3 parallel $(\vec{p}_\rho, \vec{p}_\rho)$, $(\vec{p}_\varphi, \vec{p}_\varphi)$, $(\vec{p}_z, \vec{p}_z)$ and 3 anti-parallel $(\vec{p}_\rho, -\vec{p}_\rho)$, $(\vec{p}_\varphi, -\vec{p}_\varphi)$, $(\vec{p}_z, -\vec{p}_z)$ combinations), as well as dipoles 6 orthogonal combinations ($(\vec{p}_\rho, \vec{p}_\varphi)$, $(\vec{p}_\rho, \vec{p}_z)$, $(\vec{p}_\varphi, \vec{p}_\rho)$, $(\vec{p}_\varphi, \vec{p}_z)$, $(\vec{p}_z, \vec{p}_\rho)$, $(\vec{p}_z, \vec{p}_\varphi)$).

For 3-dipoles sources, we selected single dipole source positions that approximately form nodes of equilateral triangle. We formed six different triangles, which are named according to either an approximate cross-sectional plane where triangle nodes are positioned (Anterior and Sagittal) or a part of the cage (torso) where most nodes are positioned (Right) with added approximated distance between dipoles (side of a given triangle) in mm. We selected six triangles. The first node of Anterior–80 (Fig. 3a) and Sagittal–80 triangles is positioned in the center of the bottom plane ($z_0 = 230$).

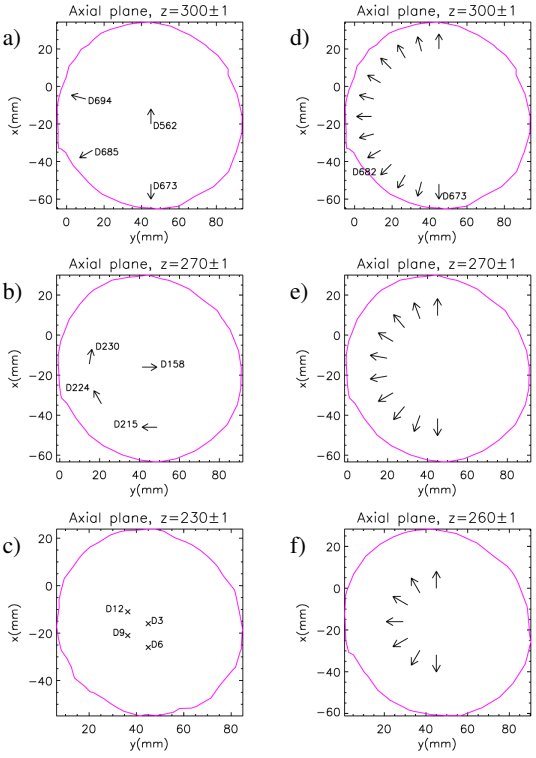


Fig. 2. Selected positions for single dipoles in different planes: a) z_3=300 mm (radial dipoles $\vec{p}_\rho$ are displayed), b) z_2=270 mm (dipoles along the polar angle $\vec{p}_\varphi$ are displayed) and c) z_2=260 mm (axial $\vec{p}_z$ dipoles are displayed). Labels denote dipole numbers in 744–dipoles source space (Fig. 1). Selected positions of dual dipole sources in planes: d) z_3=300 mm, e) z_2=270 mm and f) z_2=260 mm. Dipoles normal to the cage side ($\vec{p}_\rho$) surface are displayed.

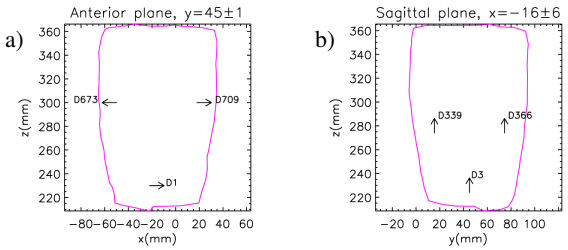


Fig. 3. Positions of triple dipole sources in a) Anterior–80 (radial $\vec{p}_\rho$ are displayed) and b) Sagittal–60 (axial $\vec{p}_z$ are displayed) triangles.

The other two nodes are positioned in the outermost circle (ρ_3=40 mm) of the upper source plane ($z_3 = 300$), with polar angles φ=(-180°, 0°) and (-90°,90°), respectively, see Fig. 3. Similar scheme was used also for constructing Anterior–60 and Sagittal–60 (Fig. 3b) triangles, where the other two nodes are positioned in the outermost circle (ρ=30 mm) of the source plane at z=280 mm. The first node of the Right–57 triangle is positioned in the center of the source plane at z=260 mm and the other two nodes are positioned in the outermost circle of the upper source plane at z_3, with polar angles φ=(-180°, -90°). The first node of the Right–42 triangle is positioned in the center of plane at z=250 mm and the other two nodes are positioned in the outermost circle of the source plane at z=280 mm, with polar angles φ=(-180°, -100°). For each triangle, we formed four 3-dipoles sources

1) normal with dipoles $(-\vec{p}_z, \vec{p}_\rho, \vec{p}_\rho)$
2) tangential with dipoles $(\vec{p}_\rho, -\vec{p}_z, \vec{p}_z)$
3) along the polar angle $(\vec{p}_\varphi, \vec{p}_\varphi, \vec{p}_\varphi)$, which are also all tangential to the cage side surface
4) axial with all dipoles along polar axis $(\vec{p}_z, \vec{p}_z, \vec{p}_z)$.

C. Data simulation and evaluation

For all selected single, dual and triple dipole sources, we simulated data at 602 leads of the cylindrical cage and at 771 nodes on the torso surface using boundary element method (BEM) [18]. In order to test the inverse solution, we added to the BEM calculated potential maps on a body (torso) surface 10 different random noise distributions at level (S/N=40 dB), where $S/N = 20\log_{10}\frac{\mathrm{RMS(signal)}}{\mathrm{RMS(noise)}}$. For each simulated data with added noise on the torso surface, we applied all 13 regularization techniques from section II-A to reconstruct the potential distribution on the cage surface. As a measure of reconstruction accuracy, we used the relative error (normalized RMS error) $\mathrm{RE} = ||\Phi_\mathrm{E}^\mathrm{r} \Phi_\mathrm{E}^\mathrm{m}||_2 / ||\Phi_\mathrm{E}^\mathrm{m}||_2$, and the correlation coefficient, $\mathrm{CC} = \Phi_\mathrm{E}^\mathrm{r} \cdot \Phi_\mathrm{E}^\mathrm{m} / ||\Phi_\mathrm{E}^\mathrm{r}||_2 ||\Phi_\mathrm{E}^\mathrm{m}||_2$, where $\Phi_\mathrm{E}^\mathrm{m}$ are the directly-computed cylindrical-cage potentials and $\Phi_\mathrm{E}^\mathrm{r}$ are the reconstructed potentials. Since there are no specific clinical thresholds for RR and CC, we also compared qualitative features of both simulated and inversely computed potential maps, (e.g., areas of negative potentials and positions of extrema).

III. Results

Table I summarizes average reconstruction results for single, dual and triple dipole models with 40-dB input noise. Results show that NU and MINRES are the worst for all source configurations. Zero order methods (ZOT, ZCG, ZSTSVD) are also much worse than their first and second order equivalents (FOT, SOT, FTSVD, STSVD, FCG, SCG, FTV, STV), so we display results for those methods only in other tables. Best reconstruction results for all source configurations are on average obtained with FOT, SOT, FTV and STV methods. FTSTVD and STSVD methods have good performance only for single dipoles. FCG and SCG methods under-perform in reconstruction of dual dipoles.

TABLE I. Average $\overline{\mathrm{RE}}$±SD and $\overline{\mathrm{CC}}$*

Method	Single dipoles		Dual dipoles		Triple dipoles	
	$\overline{\mathrm{RE}}$±SD	$\overline{\mathrm{CC}}$	$\overline{\mathrm{RE}}$±SD	$\overline{\mathrm{CC}}$	$\overline{\mathrm{RE}}$±SD	$\overline{\mathrm{CC}}$
ZOT	0.42±0.11	0.90	0.50±0.13	0.85	0.49±0.12	0.86
FOT	0.28±0.16	0.95	0.43±0.17	0.88	0.40±0.20	0.89
SOT	0.36±0.12	0.94	0.43±0.14	0.89	0.42±0.13	0.90
ZTSVD	0.46±0.13	0.89	0.53±0.14	0.83	0.51±0.13	0.85
FTSVD	0.29±0.17	0.94	0.46±0.19	0.86	0.48±0.22	0.87
STSVD	0.31±0.20	0.92	0.48±0.21	0.84	0.45±0.23	0.85
ZCG	0.42±0.12	0.90	0.52±0.14	0.84	0.50±0.12	0.86
FCG	0.31±0.18	0.93	0.48±0.19	0.85	0.42±0.20	0.88
SCG	0.30±0.15	0.94	0.44±0.18	0.87	0.39±0.17	0.90
NU	0.59±0.15	0.78	0.65±0.15	0.73	0.65±0.17	0.72
MINRES	0.44±0.14	0.88	0.56±0.17	0.80	0.53±0.12	0.84
FTV	0.29±0.16	0.94	0.42±0.16	0.89	0.40±0.18	0.90
STV	0.28±0.15	0.94	0.42±0.17	0.89	0.39±0.18	0.90

*Averaged over all selected groups of dipole sources

TABLE II. Single dipoles in different directions

Method	radial $\vec{p}_\rho$		polar angle $\vec{p}_\varphi$		polar axis $\vec{p}_z$	
	$\overline{\mathrm{RE}}$±SD	$\overline{\mathrm{CC}}$	$\overline{\mathrm{RE}}$±SD	$\overline{\mathrm{CC}}$	$\overline{\mathrm{RE}}$±SD	$\overline{\mathrm{CC}}$
FOT	0.27±0.16	0.95	0.30±0.18	0.93	0.26±0.15	0.95
SOT	0.32±0.10	0.94	0.35±0.13	0.92	0.30±0.11	0.95
FTSVD	0.29±0.17	0.94	0.31±0.19	0.93	0.27±0.16	0.95
STSVD	0.31±0.20	0.93	0.33±0.22	0.91	0.29±0.20	0.93
FCG	0.31±0.18	0.93	0.33±0.19	0.92	0.29±0.17	0.94
SCG	0.29±0.16	0.94	0.33±0.16	0.93	0.27±0.15	0.95
FTV	0.29±0.11	0.97	0.33±0.19	0.92	0.30±0.16	0.94
STV	0.28±0.13	0.96	0.31±0.18	0.93	0.26±0.15	0.95

TABLE III. Single dipoles in different planes

Method	z_0=230 mm		z_1=260 mm		z_2=270 mm		z_3=300 mm	
	$\overline{\mathrm{RE}}$	$\overline{\mathrm{CC}}$	$\overline{\mathrm{RE}}$	$\overline{\mathrm{CC}}$	$\overline{\mathrm{RE}}$	$\overline{\mathrm{CC}}$	$\overline{\mathrm{RE}}$	$\overline{\mathrm{CC}}$
FOT	0.25	0.97	0.17	0.98	0.27	0.95	0.41	0.88
SOT	0.31	0.95	0.24	0.98	0.29	0.95	0.44	0.86
FTSVD	0.26	0.97	0.17	0.98	0.29	0.95	0.44	0.86
STSVD	0.27	0.96	0.17	0.98	0.30	0.94	0.51	0.81
FCG	0.28	0.96	0.18	0.98	0.31	0.94	0.47	0.84
SCG	0.26	0.96	0.18	0.98	0.28	0.95	0.49	0.87
FTV	0.30	0.95	0.20	0.97	0.28	0.95	0.38	0.89
STV	0.29	0.96	0.17	0.98	0.26	0.96	0.38	0.90

Table II displays results for single dipoles in different directions. Results show that dipoles oriented along the polar angle (tangential to the cage surface) are slightly worse than results for dipoles oriented along the polar axis and for radial dipoles (normal to the cage surface). Table III displays results for single dipoles in different planes. Results show that the quality of reconstruction depends on a distance of sources from the cage surface. The best results are obtained for the plane z_1, where the selected dipoles on the outer circle are ∼27 mm from the side surface. The worst results are obtained for the plane z_3 where the selected dipoles are ∼10 mm from the side surface. The selected dipoles on the plane z_2 are ∼18 mm from the side surface. The selected dipoles on the bottom

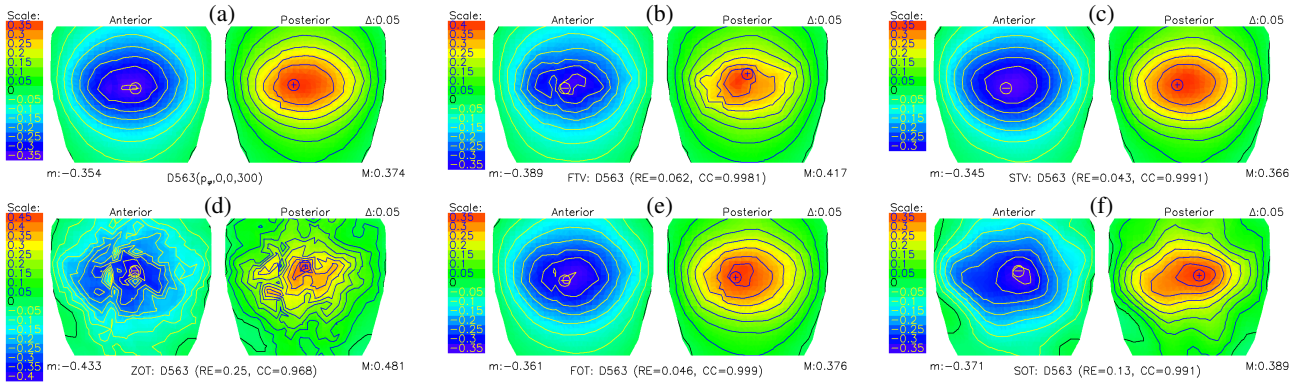


Fig. 4. a) Cage potential distribution due to single dipole positioned in the center axial plane z_3=300 mm, oriented towards the anterior side of the torso. D563(p_φ,0,0,300) bellow the map denotes dipole number in the 744-dipole source space from Fig. 1, dipole orientation and coordinates in cylindrical coordinate system, respectively. Note, that this dipole is located in the same position as the dipole D562 in Fig. 2a. "M" denotes maximum potential, "m" denotes minimum and Δ step between iso-potential lines potential for a given map, respectively. Inversely computed cage potentials using b) FTV and c) STV (first and second order total variation algorithm), d) ZOT, e) FOT and f) SOT (zero, first and second order Tikhonov regularization methods. Below each reconstruction map, RE and CC values are displayed. Each map is displayed on the anterior and posterior cage side, where vertical direction corresponds to cage height in the range $z \in [218, 360]$ mm, and horizontal direction corresponds to polar angle multiplied with the cage radius at a given height (note, that in Figs. 1-3 torso and cage cross-sections in the zx, zy and xy planes are shown).

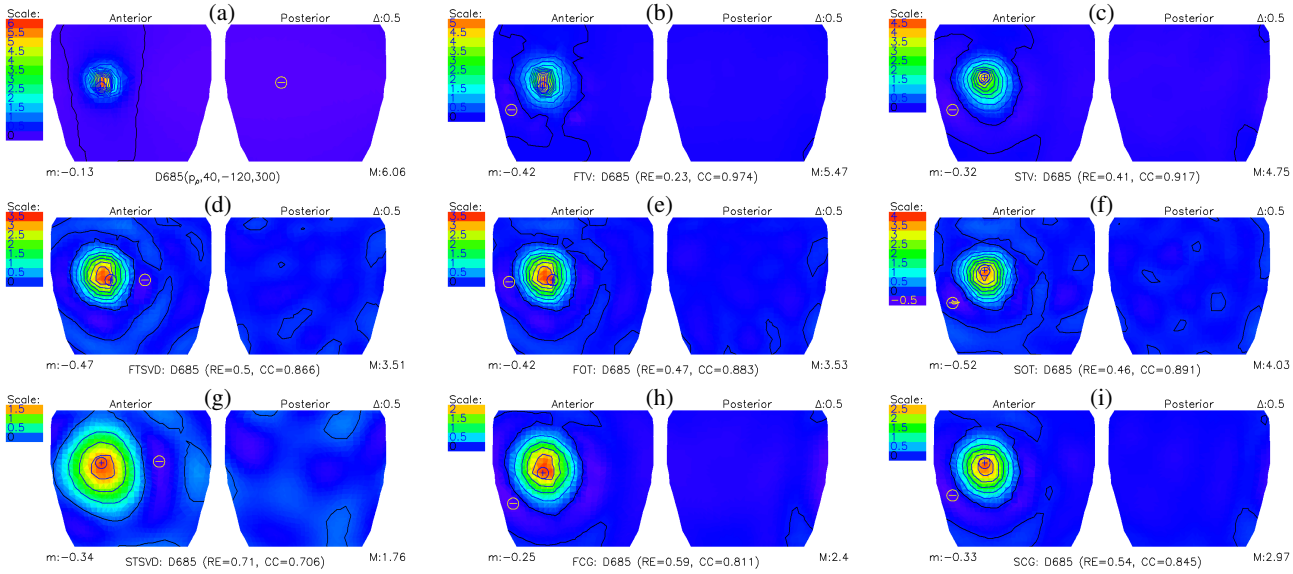


Fig. 5. a) Cage potential distribution due to single dipole positioned closed to the cage surface (ρ=400 mm, φ=-120°, dipole D685 in Fig. 2a) in the axial plane z_3=300 mm, oriented towards the cage surface (p_ρ). Inversely computed cage potentials using b) FTV, c) STV, d) FTSVD, e) FOT, f) SOT, g) STSVD h) FCG and i) SCG.

plane z_0 are ∼33 mm from the side surface, but they are only ∼18 mm from the cage bottom surface.

Fig. 4 compares directly simulated cage potentials with inversely computed potentials using ZOT, FOT, SOT, FTV, and STV, for a single dipole source near the cage center. Isopotential map forms a nice symmetrical dipolar pattern distributed over the whole cage surface in this case. Results clearly shows that ZOT performs worse than FOT and SOT, similar conclusion can be reached for other zero-order methods (ZTSVD and ZCG) as well as for MINRES and NU methods. The lowest/highest values of RE/CC were obtained with STV and FOT methods.

Fig. 5 compares directly simulated cage potentials with inversely computed potentials using FTV, STV, FOT, SOT, FTSVD, STSVD, FCG, and ZCG, for a single dipole source near and oriented normal to the cage edge. In this case, isopotential map has non-zero values in a rather small region, while FTV method clearly outperforms other methods.

Table IV displays results for dual dipoles in different planes. Like in the case of single dipoles (Table III), the quality of reconstruction depends on a distance of sources from the cage surface. Table V displays results for different orientations of dual dipoles. It seems that dual dipole orientations have no effect on the quality of reconstructed results, the anti-parallel orientations gives

TABLE IV. DUAL DIPOLES IN DIFFERENT PLANES

Method	z_1=260 mm		z_2=270 mm		z_3=300 mm	
	$\overline{\text{RE}}$±SD	$\overline{\text{CC}}$	$\overline{\text{RE}}$±SD	$\overline{\text{CC}}$	$\overline{\text{RE}}$±SD	$\overline{\text{CC}}$
FOT	0.21±0.081	0.97	0.36±0.09	0.93	0.59±0.08	0.80
SOT	0.29±0.055	0.96	0.37±0.08	0.93	0.56±0.09	0.82
FTSVD	0.23±0.090	0.97	0.39±0.12	0.91	0.62±0.11	0.77
STSVD	0.27±0.093	0.97	0.40±0.12	0.90	0.67±0.10	0.73
FCG	0.23±0.080	0.97	0.41±0.09	0.91	0.66±0.08	0.75
SCG	0.22±0.095	0.97	0.37±0.10	0.92	0.61±0.08	0.79
FTV	0.24±0.092	0.96	0.37±0.11	0.92	0.54±0.13	0.83
STV	0.22±0.083	0.97	0.36±0.10	0.93	0.56±0.10	0.82

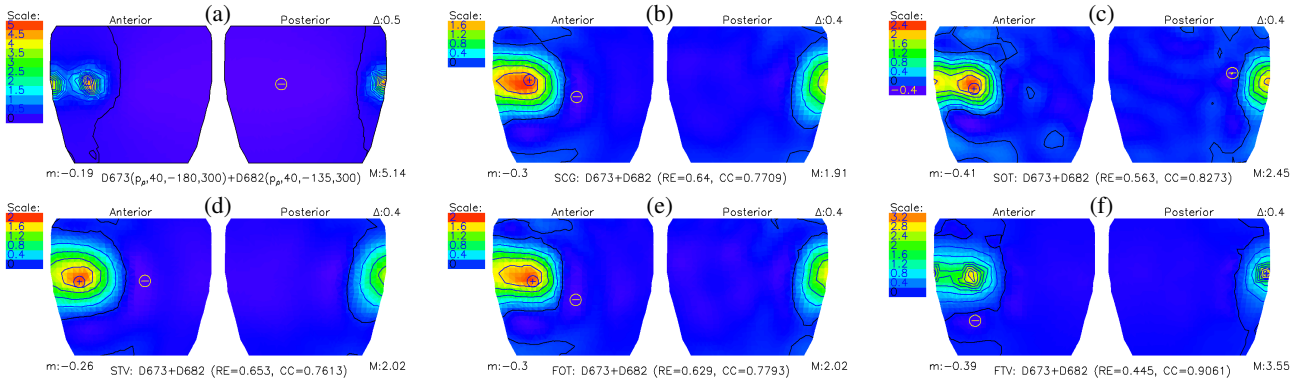


Fig. 6. a) Cage potential distribution due to a pair of parallel dipoles, 30.6 mm apart, positioned close to the surface of the cage in the axial plaen $z_3 = 300$ mm (dipoles D673 and D682 in Fig. 2d). Inversely computed cage potentials using b) SCG, c) SOT, d) STV, e) FOT, and f) FTV.

TABLE V. DIFFERENT ORIENTATIONS OF DUAL DIPOLES

Method	parallel		orthogonal		anti-parallel	
	$\overline{\text{RE}}$±SD	$\overline{\text{CC}}$	$\overline{\text{RE}}$±SD	$\overline{\text{CC}}$	$\overline{\text{RE}}$±SD	$\overline{\text{CC}}$
FOT	0.40±0.16	0.90	0.42±0.17	0.89	0.47±0.18	0.85
SOT	0.41±0.12	0.90	0.43±0.12	0.89	0.47±0.18	0.86
FTSVD	0.43±0.18	0.88	0.44±0.18	0.87	0.52±0.22	0.81
STSVD	0.46±0.21	0.85	0.46±0.21	0.85	0.53±0.22	0.80
FCG	0.46±0.18	0.87	0.47±0.18	0.86	0.52±0.21	0.82
SCG	0.42±0.16	0.89	0.43±0.17	0.88	0.49±0.21	0.84
FTV	0.39±0.14	0.91	0.41±0.14	0.90	0.46±0.21	0.85
STV	0.40±0.15	0.90	0.41±0.16	0.90	0.46±0.20	0.86

on average only slightly worse results than parallel and orthogonal orientations. The same conclusion is also valid for different combinations of parallel, anti-parallel and orthogonal dual dipoles.

Fig. 6 compares directly simulated cylindrical-cage potentials with inversely computed potentials using SCG, SOT, STV, FOT, and FTV, when the two dipoles were 30.6 mm apart. It is evident that two distinct extrema were reconstructed only when using SOT (relative error of 0.56) or FTV (0.44); it is interesting that for this specific example, the Laplacian was a more suitable operator than the gradient when applying Tikhonov regularization, while the opposite was true for the total variation technique. When taking averages, however, the difference among regularization techniques in the reconstruction of two-dipole potential distributions on the cylindrical-cage surface was small and results were nearly identical when using, FOT (0.43±0.17), SOT (0.43±0.14), FTV (0.42±0.17), or STV (0.42±0.17). These observations point toward the often-neglected notion that qualitative features of reconstructed maps may show different comparative assessment than do quantitative summary statistics (e.g., relative error), and that in some instances, smaller relative errors may not mean more potent qualitative discrimination of localized events.

Table VI displays results for different groups of triple dipoles. Like in the case of single dipoles (Table III) and dual (Table IV), the quality of reconstruction depends on a distance of sources from the cage surface. The worst results are for Anterior–80 and Sagittal–80 triangles, where all nodes are close to the cage surface. On the other hand, we got almost prefect results for triple dipoles on nodes of Right-42 triangle, which are all quite deep inside the cage. Dipoles in this configuration are also relatively

TABLE VI. DIFFERENT GROUPS OF TRIPLE DIPOLES

Method	Anterior-60		Right-42		Sagittal-60	
	$\overline{\text{RE}}$±SD	$\overline{\text{CC}}$	$\overline{\text{RE}}$±SD	$\overline{\text{CC}}$	$\overline{\text{RE}}$±SD	$\overline{\text{CC}}$
FOT	0.43±0.23	0.86	0.08±0.02	0.99	0.30±0.05	0.95
SOT	0.38±0.06	0.92	0.25±0.08	0.96	0.35±0.04	0.94
FTSVD	0.42±0.22	0.87	0.11±0.02	0.99	0.35±0.11	0.93
STSVD	0.41±0.19	0.90	0.09±0.02	0.99	0.34±0.07	0.94
FCG	0.37±0.08	0.93	0.09±0.03	0.99	0.32±0.06	0.95
SCG	0.34±0.08	0.94	0.13±0.05	0.99	0.29±0.04	0.95
FTV	0.38±0.14	0.92	0.10±0.02	0.99	0.34±0.11	0.93
STV	0.37±0.07	0.93	0.08±0.03	0.99	0.32±0.05	0.95
Method	Anterior-80		Right-57		Sagittal-80	
	$\overline{\text{RE}}$±SD	$\overline{\text{CC}}$	$\overline{\text{RE}}$±SD	$\overline{\text{CC}}$	$\overline{\text{RE}}$±SD	$\overline{\text{CC}}$
FOT	0.60±0.09	0.80	0.54±0.06	0.84	0.46±0.05	0.89
SOT	0.57±0.09	0.81	0.53±0.04	0.85	0.45±0.05	0.89
FTSVD	0.62±0.14	0.77	0.60±0.13	0.78	0.52±0.13	0.84
STSVD	0.63±0.11	0.76	0.64±0.13	0.75	0.60±0.13	0.79
FCG	0.62±0.08	0.77	0.58±0.07	0.81	0.52±0.07	0.85
SCG	0.58±0.08	0.81	0.54±0.07	0.84	0.48±0.06	0.88
FTV	0.57±0.13	0.81	0.54±0.10	0.83	0.44±0.08	0.90
STV	0.59±0.10	0.80	0.52±0.05	0.85	0.44±0.06	0.89

TABLE VII. DIFFERENT ORIENTATIONS OF TRIPLE DIPOLES

Method	normal $-\vec{p}_z, \vec{p}_\rho, \vec{p}_\rho$		axial $\vec{p}_z, \vec{p}_z, \vec{p}_z$		tangential $\vec{p}_\rho, -\vec{p}_z, \vec{p}_z$		tangential $\vec{p}_\varphi, \vec{p}_\varphi, \vec{p}_\varphi$	
	$\overline{\text{RE}}$	$\overline{\text{CC}}$	$\overline{\text{RE}}$	$\overline{\text{CC}}$	$\overline{\text{RE}}$	$\overline{\text{CC}}$	$\overline{\text{RE}}$	$\overline{\text{CC}}$
FOT	0.42	0.89	0.33	0.93	0.50	0.82	0.35	0.92
SOT	0.43	0.89	0.42	0.90	0.45	0.88	0.39	0.92
FTSVD	0.49	0.83	0.34	0.92	0.55	0.79	0.37	0.92
STSVD	0.50	0.82	0.37	0.91	0.54	0.80	0.40	0.89
FCG	0.46	0.86	0.35	0.92	0.48	0.85	0.39	0.90
SCG	0.42	0.88	0.34	0.93	0.45	0.87	0.36	0.92
FTV	0.31	0.95	0.35	0.92	0.51	0.83	0.42	0.89
STV	0.39	0.90	0.35	0.92	0.46	0.87	0.35	0.92

close to each other (∼42 mm), so they are not likely to generate a complex potential distribution on the cage surface.

Table VII displays results for different orientations of triple dipoles. The worst results are obtained for tangential $(\vec{p}_\rho, -\vec{p}_z, \vec{p}_z)$ orientation. Fig. 7 shows results for a such triple dipole source in Anterior–80 (Fig. 3a). It is interesting to note, that all reconstructed maps have similar pattern for all presented methods, but none can focus 3 dipole sources: one on the bottom source plane (z_0), which is not clearly seen in this view of BEM calculated map (Fig. 7a), and the other two on the left

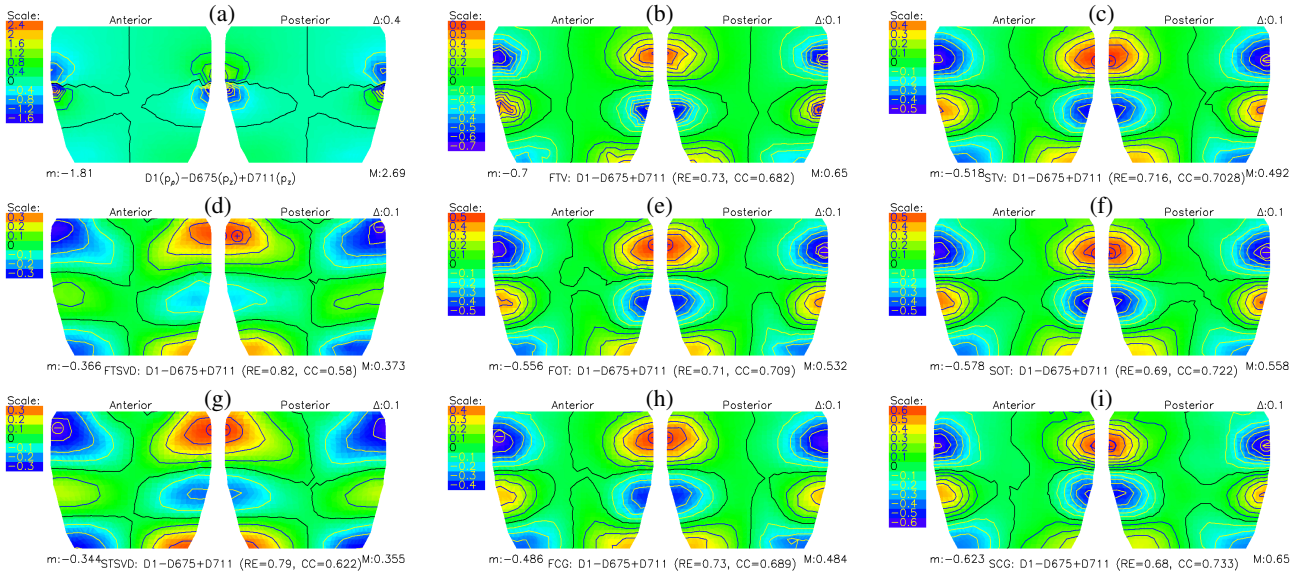


Fig. 7. a) Cage potential distribution due to triple dipoles in Anterior–80 triangle positioned closed and oriented tangentially $(\vec{p}_\rho, -\vec{p}_z, \vec{p}_z)$ to the cage surface. Inversely computed cage potentials using b) FTV, c) STV, d) FTSVD, e) FOT, f) SOT, g) STSVD h) FCG and i) SCG.

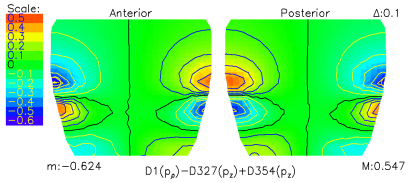


Fig. 8. Cage potential distribution due to triple dipoles in Anterior–60 triangle oriented tangentially $(\vec{p}_\rho, -\vec{p}_z, \vec{p}_z)$ to the cage surface.

and right side of the upper source plane(z_3). However, one can clearly deduce from reconstructed maps the presence of 3 dipoles, but a bit deeper because the dipolar patterns are more spread. In fact, these reconstructed maps look more related to the corresponding source in Anterior–60 triangle, where dipoles are closer to each other (60 instead of 80 mm) and are not so close to the cage surface, see Fig. 8.

IV. Conclusions

The main finding of our study is that - for idealized source models - non-quadratic methods (total variation algorithms) and first-order and second-order Tikhonov regularizations outperformed other regularization methodologies. This study is a necessary step in comprehensive validation of various regularization methodologies by comparing them systematically and under controlled conditions.

References

[1] E. Frank, "Spread of current involume conductors of finite extent," *Ann. N. Y. Acad. Sci.*, vol. 65, pp. 980–1002, 1957.

[2] D. B. Geselowitz, "Dipole theory in electrocardiography," *Am. J. Cardiol.*, vol. 14, pp. 301–306, 1964.

[3] R. L. Lux, "Electrocardiographic mapping. noninvasive electrophysiological cardiac imaging," *Circulation*, vol. 87, pp. 1040–1042, 1993.

[4] H. Oster, B. Taccardi, R. L. Lux, P. R. Ershler, and Y. Rudy, "Noninvasive electrocardiographic imaging: reconstruction of epicardial potentials, electrograms, and isochrones and localization of single and multiple cardiac events," *Circulation*, vol. 96, pp. 1012–24, 1997.

[5] Y. Rudy and B. J. Messinger-Rapport, "The inverse problem in electrocardiography: Solutions in terms of epicardial," *Crit. Rev. Biomed. Eng.*, vol. 16, pp. 215–268, 1988.

[6] R. Hren, G. Stroink, and B. M. Horáček, "Accuracy of single-dipole inverse solution when localising ventricular pre-excitation sites: simulation study," *Med. Biol. Eng. Comp.*, vol. 36, pp. 323–329, 1998.

[7] P. Colli-Franzone, E. Macchi, and C. Viganotti, "Oblique dipole layer potentials applied to electrocardiology," *Ann. Biomed. Eng.*, vol. 37, pp. 902–912, 2009.

[8] M. Milanič, V. Jazbinšek, R. S. MacLeod, D. H. Brooks, and R. Hren, "Assessment of regularization techniques for electrocardiographic imaging problem," *J. Electrocardiol.*, vol. 47, pp. 20–28, 2014.

[9] R. Trobec, "Computer analysis of multichannel ecg," *Comput. Biol. Med.*, vol. 33, pp. 215–226, 2003.

[10] J. de Munck, "A linear discretization of the volume conductor boundary integral equation using analytically integrated elements," *IEEE Trans. Biomed. Eng.*, vol. BME–39, pp. 986–990, 1992.

[11] A. Ferguson, X. Zhang, and G. Stroink, "A complete linear discretization for calculating the magnetic field using the boundary element method," *IEEE Trans. Biomed. Eng.*, vol. 41, pp. 455–60, 1994.

[12] A. Tikhonov and V. Arsenin, *Solutions of Ill-Posed Problems*. Washington, DC: Winston, 1997.

[13] D. H. Brooks, G. F. Ahmad, R. D. MacLeod, and G. M. Maratos, "Inverse electrocardiography by simultaneous imposition of multiple constraints," *IEEE Trans. Biomed. Eng.*, vol. BME–46, pp. 3–18, 1999.

[14] P. Colli-Franzone, B. Taccardi, and C. Viganotti, "An approach to inverse calculation of epicardial potentials from body surface maps," *Adv. Cardiol.*, vol. 21, pp. 50–54, 1978.

[15] P. C. Hansen, *Rank-Deficient and Discrete Ill-Posed Problems*. Philadelphia, PA: SIAM, 1998.

[16] M. Hanke, *Conjugate Gradient Type Methods for Ill-Posed Problems*. Harlow: Longman Scientific & Technical, 1995.

[17] R. S. M. D. H. Brooks, K. G. Srinidhi and D. R. Kaeli, "Multiply constrained cardiac electrical imaging methods," in *Subsurface Sensors and Applications*, C. Nguyen, Ed., vol. 3752. Piscataway, NJ: Proc SPIE, 1999, pp. 1770–1771.

[18] A. Barnard, I. Duck, M. Lynn, and W. Timlake, "The application of electromagnetic theory to electrocardiography," *Biophys.J.*, vol. 7, pp. 443–491, 1967.

Single-Dipole and Dual-Dipole Inverse Solutions in Electrocardiography

Vojko Jazbinšek and Rok Hren
Institute of Mathematics, Physics and Mechanics, Ljubljana, Slovenia
Email: vojko.jazbinsek@imfm.si

***Abstract*—In this study, we closely examined the performance of inverse solution in terms of equivalent dipole source model. To simulate potential distribution on the body surface, we employed an analytical model of a single current dipole (or a pair of current dipoles) placed within the homogeneous isotropic volume conductor consisting of two non-concentric spheres. Using these data, we evaluated the accuracy of recovering both location and orientation of the single or dual dipole sources. In total, we examined 24 different dipole locations and found that the location of fitted single and dual current dipoles virtually coincided with the original source for high S/N ratios.**

We extended investigation to more complex geometry, where a computer model of the human ventricular myocardium was used to simulate activation sequences initiated at eight sites positioned on the epicardial surface of the atrio-ventricular ring. From these sequences, 117-lead body surface potentials were simulated on a realistic torso surface and were then used to localize single and dual accessory pathways employing single or dual equivalent dipoles. Average localization errors were 5 and 12 mm for the single and the dual accessory pathways, respectively, what could be useful additional information prior to electrophysiological study.

I. Introduction

Mathematical modeling of both source and volume conductor remains one of the prerequisites for any quantitative interpretation of electrocardiographic data. One of the most often used approaches to solving the inverse problem in electrocardiography involves calculating the position and moment of an equivalent single dipole source in the model of the human torso [1]–[3]. It has been widely recognized that the equivalent single dipole approximation is adequate only when cardiac activity is confined to a single, relatively small volume, e.g. arising owing to ectopic foci and accessory pathways between the atria and ventricles. It thus appears intuitive that, in the presence of multiple ventricular events, more complex equivalent generators consisting of two [4] or more equivalent dipoles would achieve better accuracy and therefore a better understanding of the underlying cardiac process. This is particularly important when localizing those accessory atrioventricular connections presented in Wolff-Parkinson-White (WPW) syndrome that consist of dual parallel fibers [5]. As radiofrequency ablations of such multiple aberrant pathways are considered difficult [6], it is desirable to obtain as much information as possible regarding the accessory pathways location and anatomy (a single pathway as against multiple ones) prior to the invasive procedure.

In this simulation study, we explored the possibility that approximation by two dipoles could be used to localize regional activation of dual accessory pathways. In the first part, we employed simple analytical conducting sphere model that may serve as an efficient tool in examining performance of electrocardiographic inverse solution in terms of models, which use either single current dipoles or dual current dipoles. In the second part, we employed an anatomical computer model of the human ventricular myocardium to simulate activation sequences initiated on the epicardial surface of the atrio-ventricular ring, and from these sequences we simulated body surface potentials on a realistic torso surface. The results may be useful in providing additional means for assisting electrophysiology study during radiofrequency ablations of dual accessory pathways.

II. Methods

A. Conducting sphere model

In the first part of this study, we used a pair of homogeneous and isotropic non-concentric spheres and placed either a single current dipole or a pair of two dipoles inside of the smaller sphere. The potentials at an arbitrary point $\vec{r}$ inside a sphere with radius R and conductivity σ can be calculated analytically [7]

$$V(\vec{r}) = \frac{\vec{p}}{4\pi\sigma R^3}\cdot\left\{\frac{R^3(\vec{r}-\vec{r}_p)}{|\vec{r}-\vec{r}_p|^3}+\frac{1}{r_{pi}^3}\left(\vec{r}-\frac{r^2}{R^2}\vec{r}_p\right)+\frac{1}{r_{pi}}\left[\vec{r}+\frac{(\vec{r}\cdot\vec{r}_p)\vec{r}-r^2\vec{r}_p}{R^2(r_{pi}+1)-(\vec{r}\cdot\vec{r}_p)}\right]\right\}, \quad (1)$$

where $\vec{p}$ and $\vec{r}_p$ are dipole moment and position, and

$$r_{pi}=\left[1+\left(\frac{r_p r}{R^2}\right)^2-2\frac{(\vec{r}\cdot\vec{r}_p)}{R^2}\right]^{1/2}. \quad (2)$$

Note, that in the original formula (13) in [7], the $(\vec{r}\cdot\vec{r}_p)$ is expressed as $(r_p r\cos\varphi)$, where φ is the angle between $\vec{r}$ and $\vec{r}_p$. For the surface potential, $r\equiv R$, $Rr_{pi}=|\vec{R}-\vec{r}_p|$, (1) is simplified to formula derived by Brody et al. [8]

$$V(\vec{R}) = \frac{\vec{p}}{4\pi\sigma R^3}\cdot\left\{2\frac{R^3(\vec{R}-\vec{r}_p)}{|\vec{R}-\vec{r}_p|^3}+\frac{R}{|\vec{R}-\vec{r}_p|}\left[\vec{R}+\frac{(\vec{R}\cdot\vec{r}_p)\vec{R}-R\vec{r}_p}{|\vec{R}-\vec{r}_p|+R-r_p\cos\varphi}\right]\right\}. \quad (3)$$

We approximated the thoracic surface by the homogeneous conducting sphere with the unity radius ($R=1$), and the epicardial surface by a smaller sphere with a radius of $R_E=0.5$, positioned eccentrically

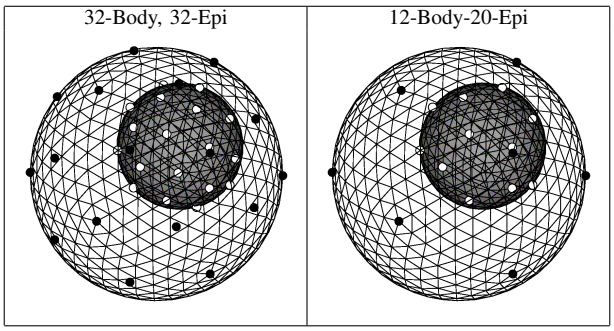


Fig. 1. Lead systems used in the fitting procedure: i) 32-leads on the body, ii) 32-leads on the epicardium and iii) combination of 12-leads on the body and 20-leads on the epicardium. Leads on the body (outer sphere) are denoted with •, leads on the epicardium(inner sphere) are denoted with ◦.

$\vec{r}_{\mathrm{E}} = (0.1, -0.2, 0.3)$. The body and the epicardial surfaces were tessellated using 1280 and 720 triangles (642 and 362 nodes), respectively. The tessellation of the body surfaces was generated by refinement of icosahedron in four steps and the epicardial surface was generated by refinement of truncated icosahedron in two steps. In this paper, we tested three electrode systems, each having 32 leads, see Fig. 1.

We put the single dipole source model in the total of 24 locations, 12 in nodes of icosahedron positioned near the center of the body surface and 12 in nodes of icosahedron positioned near the center of the epicardial surface, see Table I. In each of the positions, we put 3 single current dipoles pointing along axes of the Cartesian coordinate system, which gives total of 72 single dipole sources.

For the dual dipole model, we combined node positions of both source icosahedrons from Table I to construct 12 most close positioned pairs (mean distance 0.178 ± 0.069), 12 most distant positioned pairs (mean distance 0.696 ± 0.026), and 12 pairs of corresponding nodes shifted by $\vec{r}_{\mathrm{E}}$ (median, $||\vec{r}_{\mathrm{E}}|| = 0.374$), see Table II. In each of the pairs, we put either 3 parallel or anti-parallel or perpendicular dipoles pointing along axes of the Cartesian coordinate system. For each type of distance (close, median, distant) and each type of direction (parallel, anti-parallel, perpendicular), we therefore obtain 36 dual dipole sources.

For all single and dual dipole sources, we calculated potential maps on three lead systems from Fig. 1 using (1). In order to test the inverse solution, we added to the analytically calculated maps 7 different noise levels (*S/N*=10,15,...,40 dB), where

$$S/N = 20\log_{10}\frac{\mathrm{RMS(signal)}}{\mathrm{RMS(noise)}}. \qquad (4)$$

For each analytically calculated map and for each noise level, we generated 10 different random noise distributions and performed the inverse solutions either for a single-dipole model or dual-dipoles model.

As a measure of localization accuracy, we used the distance between the recovered location(s) and original location(s) of dipole(s). In addition, we also calculated relative errors between noisy $\mathbf{V}_{\mathrm{n}}$ and fitted $\mathbf{V}_{\mathrm{f}}$ maps

TABLE I. RELATIVE POSITIONS OF SINGLE-DIPOLES*

1st icosahedron			2nd icosahedron		
N_1	ρ_B	ρ_E	N_2	ρ_B	ρ_E
1	0.340	0.078	1	0.714	0.681
2	0.340	0.418	2	0.684	0.681
3	0.210	0.482	3	0.557	0.421
4	0.400	0.388	4	0.750	0.800
5	0.210	0.494	5	0.554	0.421
6	0.340	0.441	6	0.680	0.681
7	0.340	0.683	7	0.628	0.681
8	0.210	0.719	8	0.489	0.421
9	0.000	0.748	9	0.374	0.000
10	0.340	0.692	10	0.626	0.681
11	0.210	0.732	11	0.484	0.421
12	0.210	0.840	12	0.438	0.421

* Dipoles are positioned in nodes of two icosahedrons, which have same dimension defined by circumsribed sphere radius of 0.3 and are positioned inside body surface ($\rho_B = r_p/R_B$, outer sphere with radius $R_B = R = 1$) positioned in the center of Cartesian coordinate system, and inside epicard surface ($\rho_E = ||\vec{r}_p - \vec{r}_E||/R_E$, inner sphere with radius $R_E = 0.5$) shifted by $\vec{r}_E = (0.1, -0.2, 0.3)$.

TABLE II. DUAL-DIPOLES POSITIONS**

close (0.178±0.069)		median(0.374)		distant(0.696±0.026)	
(N_1,N_2)	Δr_{12}	(N_1,N_2)	Δr_{12}	(N_1,N_2)	Δr_{12}
(1,9)	0.039	(1,1)	0.374	(9,4)	0.745
(4,12)	0.039	(2,2)	0.374	(12,1)	0.739
(2,8)	0.165	(3,3)	0.374	(9,1)	0.714
(6,11)	0.165	(4,4)	0.374	(12,4)	0.714
(4,9)	0.194	(5,5)	0.374	(9,2)	0.684
(2,9)	0.209	(6,6)	0.374	(11,4)	0.684
(4,11)	0.209	(7,7)	0.374	(8,4)	0.680
(4,8)	0.220	(8,8)	0.374	(9,6)	0.680
(6,9)	0.220	(9,9)	0.374	(11,2)	0.680
(1,11)	0.227	(10,10)	0.374	(11,1)	0.678
(2,12)	0.227	(11,11)	0.374	(12,2)	0.678
(1,5)	0.230	(12,12)	0.374	(8,1)	0.675

**Pairs (N_1,N_2) of positions from the 1st and the 2nd icosahedron defined in Table I, Δr_{12} is a distance betveen the two dipoles.

($\mathrm{RE_{fn}}$), and between analytical $\mathbf{V}_{\mathrm{a}}$ and fitted maps ($\mathrm{RE_{fa}}$):

$$\mathrm{RE_{fn}} = \frac{||\mathbf{V}_{\mathrm{f}} - \mathbf{V}_{\mathrm{n}}||_2}{||\mathbf{V}_{\mathrm{n}}||_2}, \ \mathrm{RE_{fa}} = \frac{||\mathbf{V}_{\mathrm{f}} - \mathbf{V}_{\mathrm{a}}||_2}{||\mathbf{V}_{\mathrm{a}}||_2}, \qquad (5)$$

where $||\mathbf{F}||_2 = \sqrt{\sum_i F_i^2}$ denotes 2-norm.

B. Anatomical ventricular model

In the second part of this study we simulated body surface potentials following a methodology that was presented in detail elsewhere [9], [10]. First, we positioned an anatomical model of the human ventricles (Fig. 2) in the homogeneous boundary-element torso model [11] at the hearts anatomical location documented by radiographic images. Secondly, we simulated an activation sequence using a ventricular model [13] that features an anatomically accurate geometry, with a resolution of 0.5 mm, an intramural structure with rotating anisotropy and a computational implementation of a propagation algorithm based on combining a cellular automaton with bidomain models [14], [15]. From this realistic activation sequence, we calculated extra-cardiac electric potentials employing the oblique dipole model of cardiac sources in combination with the boundary element torso model [9]. The infinite medium electric potentials $\mathbf{\Phi}_\infty$ was determined from the discretized equation

$$-4\pi\sigma_0\mathbf{\Phi}_\infty = \sigma_1 \int \frac{\nabla v_m \cdot \boldsymbol{r}}{r^3} dV + \sigma_2 \int \boldsymbol{a}\boldsymbol{a}^T \frac{\nabla v_m \cdot \boldsymbol{r}}{r^3} dV, \ (6)$$

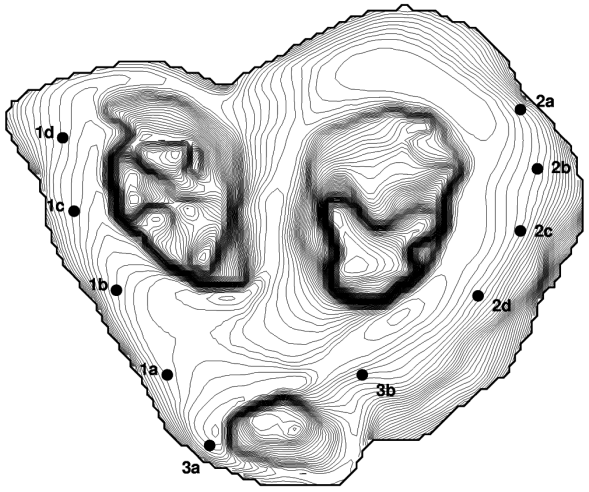


Fig. 2. Basal view of human ventricular model shown with 10 sites of accessory pathways, which are related to locations obtained with clinical studies [12]. Layers are 1 mm apart, and each is represented by smoothed contour line to achieve better rendering of shape. Accessory pathways labeled 1 are located at right anterolateral to right posterolateral region, those labeled 2 are at left posterolateral to left anterolateral region, and those labeled 3 are at right or left anteroparaseptal region (see Table III). Right ventricle is to left, left ventricle is to right, and pulmonary artery is at bottom.

where the integrals were evaluated over the ventricular volume; v_m was the transmembrane potential calculated using the propagation algorithm; σ_0 was the conductivity of the homogeneous monodomain; σ_1 and σ_2 were conductivities characterizing anisotropic myocardium; $\boldsymbol{a}$ is the local direction of the fiber axis; and $\boldsymbol{r}$ is the distance from the source point (each activated cell) to a field point. To compute the body surface potentials we used a fast forward solution [16], [17]. To account for the influence of the torsos outer boundary on localization results, we used, in addition to the standard male torso model, an individualized male torso model [9] to simulate body surface potentials generated by the ventricular model[1]. As before, the two-dipole inverse problem was solved using the standard male torso model. For each activation sequence, we simulated body surface potentials (at 117 sites on both the anterior and posterior torso) at 4 ms increments, within the first 40 ms after the onset of activation. As a measure of the accuracy of the localization, we used the localization error, defined as the three-dimensional distance between the locations of the best-fitting pair of dipoles and each of the two accessory pathways in the ventricular model.

Macchi et al. [18] and Taccardi et al. [19] pointed out that electric potentials during the initial phase of activation resemble those of two opposing dipoles oriented along the major axis and located near the ends of an elliptical wavefront of the propagated activation. To test this hypothesis, we first assessed the performance of a two-dipole generator in localizing single accessory pathways. We simulated electric potentials corresponding to sequences initiated at ten single pacing sites in the right anterolateral/posterolateral segment, in the left posterolateral/anterolateral segment, and in the right/left anteroparaseptal segment of the atrioventricular ring (Fig. 2). Using these simulated potentials, we solved the two-dipole inverse problem and specifically verified whether the positions of the reconstructed dipoles corresponded to the leading edge of the activation wavefront. The results of this study can be found in Section III-B1.

[1]We altered the geometry of the outer surface of the torso and the position of the ventricular model in the torso, but did not change the size of the ventricular model.

TABLE III. ANATOMICAL DESCRIPTION OF DUAL SITES

	Abbreviation	Anatomical description	Distance*
1a-1b	RAL	Right anterolateral	18
1a-1c	RAL	Right anterolateral/ right lateral	18
1a-1d	RAL/RPL	Right anterolateral/ right posterolateral	48
2a-2b	LPL/LL	Left posterolateral/ left lateral	11
2a-2c	LPL/LL	Left posterolateral/ left lateral	23
2a-2d	LPL/LAL	Left posterolateral/ left anterolateral	36
3a-3b	RAP/LAP	Right anteroparaseptal/ left anteroparaseptal	30
1b-2b	RAL/LPL	Right anterolateral/ left posterolateral	139

*Distance in mm measured along atrioventricular (AV) ring

Next, we solved the inverse problem using the sequences initiated at eight different pairs of pacing sites in the same segments of the AV ring (Fig. 2 and Table III). Gaussian noise at the root-mean-square (RMS) levels of 2.5 μV, 5 μV and 20 μV was added to all simulated body surface potentials. We generated ten different noise distributions for each noise level, and, with these data, the two-dipole inverse problem was solved. The results of this study can be found in Section III-B2.

To compare the accuracy of the two-dipole model inverse solution with that of the single-dipole inverse solution, we also calculated the inverse solution for the single-dipole model for the simulated data described above. For this inverse solution, we used methods described earlier [13]. When using the individualized male torso model, Gaussian noise at the RMS level of 5 μV was added to the simulated body surface potentials. The results of this study can be found in Section III-B3.

III. RESULTS

A. Conducting sphere model

Table IV and Fig. 3 display average single dipole source localization errors (Δr) and REs calculated with different lead system and noise levels, where

$$\Delta r = ||\vec{r}_f - \vec{r}_p||_2\,, \tag{7}$$

and $\vec{r}_f$ and $\vec{r}_p$ are fitted and original dipole locations, respectively. Results clearly show that for high *S/N* ratio (40 dB) location of the single current dipole almost coincide with the original source for all lead systems. The 32-Epi lead system outperforms the other two in terms of localization error. On the other hand, relative fit errors RE_{fn} are almost the same for all lead systems. Note, that fitted potential maps are more similar to the analytical maps than noisy maps ($RE_{fa} < RE_{fn}$).

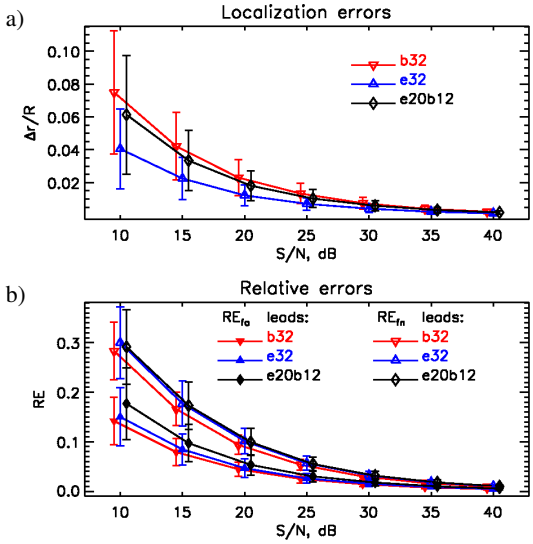


Fig. 3. a) Localization errors ($\Delta r/R$) and b) relative errors (RE_{fn} and RE_{fa}) vs. noise levels for single dipole sources.

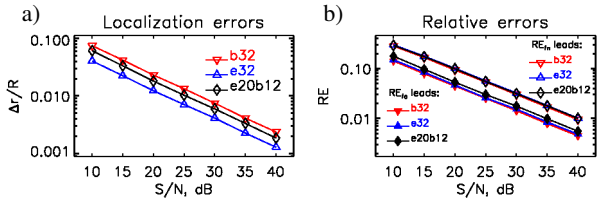


Fig. 4. Logarithmic scaling of a) localization errors and b) relative errors for single dipole sources.

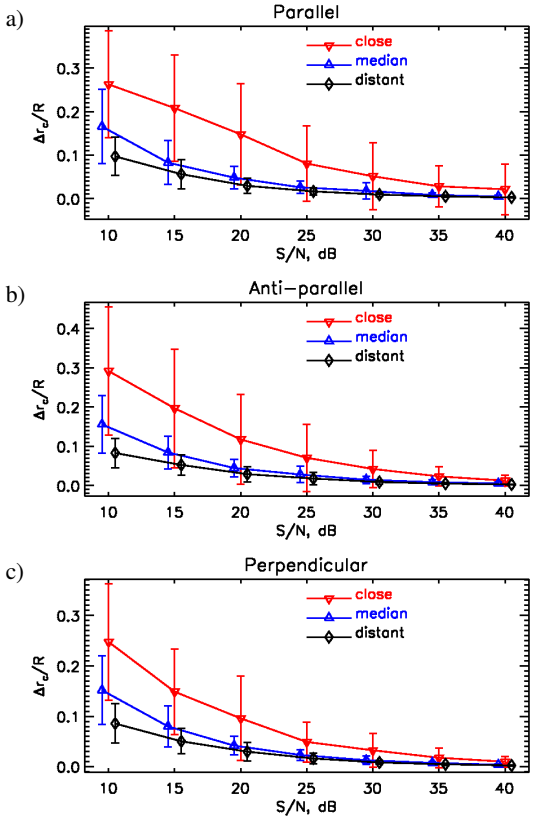


Fig. 5. Localization errors ($\Delta r_c/R$) calculated with 32-epi lead system for dual dipole sources with a) parallel, b) anti-parallel and c) perpedicular mutual directions and different distances between dipoles that forming dual sources.

Fig. 4 displays Δr and REs in logarithmic scale. We observe linear dependence of both $\log(\Delta r/R)$ and $\log(\mathrm{RE})$ vs. *S/N*. According to Eq. (4), both Δr and RE are therefore proportional to noise RMS value.

Table V displays average dual dipole source localization errors (Δr_1, Δr_2 and Δr_c) and REs calculated by 32-Epi lead system using data with different noise levels generated with dual dipole sources with different mutual distances and orientations. Localization errors are defined as distances between the recovered locations ($\vec{r}_{f1}$, $\vec{r}_{f2}$) and original locations ($\vec{r}_{p1}$, $\vec{r}_{p2}$) of both dipoles. Combined error is defined as

$$\Delta r_c = \sqrt{\Delta r_1^2 + \Delta r_2^2}, \qquad (8)$$

and it is shown in Fig. 5. Results show that distance between dipoles in the dual dipole source model plays an important role in the source localization procedure. In the presence of noise, locations of close positioned dipoles are poorly recovered. On the other hand, mutual orientation between dipoles is not so important.

B. Anatomical ventricular model

1) Localization of single accessory pathway using two-dipole model: Fig. 6a shows a typical example of localization of a single accessory pathway when the two-dipole model is used under ideal noise-free conditions. Localization of a pair of dipoles is shown at different time instants for an activation sequence initiated at the left lateral site (2c). Both reconstructed dipoles are initially in the sequence located close to the site of an accessory pathway, but later become separated by the distance that is progressively increasing with time. The distance between the leading edge of the simulated wavefront and the locations of the two corresponding, reconstructed dipoles is, on average, 3 mm (range 1-5 mm) during the first 28 ms of an activation sequence. This observation strongly supports the notion that progressive separation of the two dipoles reflects the propagation of an activation wavefront. Although the two-dipole model reflects well qualitative features of the propagated activation wavefront, it is less suitable for localizing the single sites of early activation than the single-dipole model. Fig. 7 comprehensively compares the accuracy of the localization when both models are used, with the evident superiority of the single-dipole model. The reason for the poorer performance of the two-dipole model is the fact that such a model, as explained above, localizes both leading edges of the activation wavefront, which are progressively moving away from the actual site of pre-excitation.

2) Localization of dual accessory pathways in presence of different noise levels: Fig. 6b illustrates localization results for the pair of right anterolateral/lateral accessory pathways with the two-dipole model. We can see that dipoles are clearly separated and located close to the actual locations of accessory pathways. In this specific case, the localization errors ($\pm$SD) attained a minimum 20 ms after the onset of activation (11$\pm$1 mm and 5$\pm$1 mm), see Table VI, where the first three columns summarize the localization results for the three noise levels. RMS

TABLE IV. SINGLE DIPOLE FIT RESULTS USING DIFFERENT LEAD SYSTEMS*

Noise	32-Body			32-Epi			12-Body-20-Epi		
[dB]	$\Delta r/R\pm$SD	$RE_{fn}\pm$SD	$RE_{fa}\pm$SD	$\Delta r/R\pm$SD	$RE_{fn}\pm$SD	$RE_{fa}\pm$SD	$\Delta r/R\pm$SD	$RE_{fn}\pm$SD	$RE_{fa}\pm$SD
10	0.075±0.037	0.283±0.058	0.142±0.048	0.041±0.024	0.300±0.072	0.150±0.059	0.061±0.036	0.291±0.075	0.177±0.073
20	0.023±0.011	0.094±0.019	0.045±0.014	0.012±0.006	0.101±0.026	0.047±0.019	0.018±0.009	0.099±0.027	0.053±0.021
30	0.007±0.004	0.030±0.006	0.014±0.005	0.004±0.002	0.032±0.008	0.015±0.006	0.006±0.003	0.031±0.008	0.018±0.007
40	0.002±0.001	0.009±0.002	0.004±0.002	0.001±0.001	0.010±0.003	0.005±0.002	0.002±0.001	0.010±0.003	0.006±0.002

* Averaged over 720 samples (72 single dipole sources × 10 random noise distributions) for different lead systems and noise levels.

TABLE V. DUAL DIPOLE FIT RESULTS USING 32-EPI LEAD SYSTEM**

	S/N	Parallel dipoles					Anti-parallel dipoles					Perpendicular dipoles				
	[dB]	$\Delta r_1/R$	$\Delta r_2/R$	$\Delta r_c/R\pm$SD	RE_{fn}	RE_{fa}	$\Delta r_1/R$	$\Delta r_2/R$	$\Delta r_c/R\pm$SD	RE_{fn}	RE_{fa}	$\Delta r_1/R$	$\Delta r_2/R$	$\Delta r_c/R\pm$SD	RE_{fn}	RE_{fa}
close	10	0.163	0.162	0.262±0.123	0.223	0.206	0.197	0.204	0.291±0.163	0.590	1.389	0.160	0.161	0.247±0.115	0.300	0.293
	20	0.093	0.088	0.148±0.116	0.077	0.063	0.081	0.080	0.117±0.114	0.346	0.409	0.063	0.060	0.096±0.083	0.108	0.086
	30	0.033	0.031	0.051±0.077	0.025	0.020	0.029	0.029	0.042±0.047	0.149	0.128	0.022	0.021	0.033±0.034	0.034	0.027
	40	0.013	0.013	0.021±0.058	0.008	0.006	0.009	0.008	0.012±0.013	0.050	0.038	0.007	0.007	0.010±0.010	0.011	0.008
median	10	0.116	0.101	0.166±0.086	0.199	0.177	0.122	0.086	0.156±0.074	0.363	0.345	0.117	0.086	0.152±0.068	0.249	0.229
	20	0.033	0.031	0.048±0.026	0.068	0.054	0.030	0.030	0.045±0.022	0.134	0.106	0.030	0.026	0.042±0.018	0.088	0.067
	30	0.010	0.013	0.018±0.019	0.022	0.017	0.010	0.010	0.015±0.009	0.043	0.032	0.009	0.008	0.013±0.009	0.028	0.021
	40	0.003	0.003	0.005±0.003	0.007	0.005	0.003	0.003	0.005±0.004	0.013	0.011	0.003	0.003	0.004±0.002	0.008	0.007
distant	10	0.061	0.068	0.097±0.044	0.183	0.148	0.051	0.059	0.083±0.037	0.224	0.183	0.055	0.059	0.086±0.039	0.199	0.160
	20	0.018	0.021	0.029±0.018	0.059	0.047	0.017	0.021	0.029±0.019	0.075	0.057	0.018	0.021	0.030±0.018	0.065	0.051
	30	0.006	0.006	0.009±0.005	0.019	0.015	0.005	0.006	0.009±0.005	0.024	0.018	0.005	0.006	0.008±0.004	0.021	0.016
	40	0.002	0.002	0.003±0.001	0.006	0.005	0.002	0.002	0.003±0.001	0.007	0.006	0.002	0.002	0.003±0.001	0.007	0.005

** Averaged over 360 samples (36 dual dipole sources × 10 random noise distributions) for different *S/N* and groups of distances and orientations between two single dipoles that forming dual dipole sources.

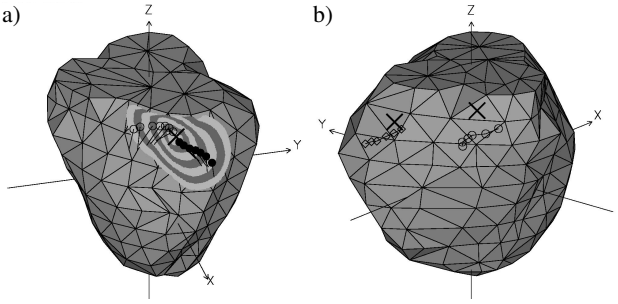


Fig. 6. a) Localization of left-lateral single accessory pathway. Inverse solutions were performed under no-noise conditions. (×) Onset of accessory pathway; reconstructed positions: (○) first and (●) second dipoles, where short lines indicate directions of reconstructed dipole moments. Alternating light and dark gray zones represent projection of activation isochronal surface on epicardial surface, 4-28 ms after onset of activation.
b) Localization of right anterolateral/right lateral (case 1a-1c) dual accessory pathway using two-dipole model. Positions of two dipoles reconstructed at different time instants, 16-36 ms after onset of activation, are shown with epicardial surface. Inverse solutions were obtained with Gaussian RMS noise of 5 μV added to simulated body surface potentials. (×) Onset of each accessory pathway; reconstructed dipoles: (○) first and (◇) second dipoles.

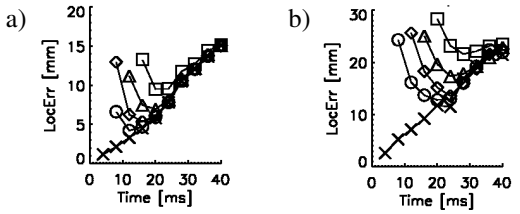


Fig. 7. Localization errors a) for a single dipole source and b) for a two-dipole source model averaged over 8 single accessory pathways, 4-40 ms after onset in presence of different noise levels: (×) no noise; (○) RMS noise level of 2.5 mV; (◇) 5 mV; (△) 10 mV; (□) 20 mV.

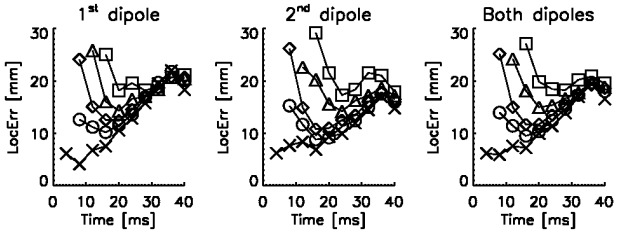


Fig. 8. Localization errors for dual accessory pathway with two-dipole source in presence of different noise levels: (×) no noise; (○) RMS noise level of 2.5 mV; (◇) 5 mV; (△) 10 mV; (□) 20 mV.

levels of the simulated body surface potentials were, at 20 ms after the onset of activation, between 104 μV (case 2a-2b) and 164 μV (case 1a-1c). We found that two-dipole localization errors for typical measuring conditions (RMS noise level of 5 μV) reached a minimum between 12 ms and 24 ms after the onset of activation (Fig. 8). For such measuring conditions, the average localization errors were between 5 and 21 mm (12±6/11±6 mm at 20 ms for the first/second dipole, respectively). Localization errors were, on average, smaller for the pairs of accessory pathways located on the right side (8 mm for cases 1a-1b, 1a-1c, 1a-1d) than for those located on the left side (14 mm for cases 2a-2b, 2a-2c, 2a-2d).

3) Localization of dual accessory pathways in the presence of modeling errors: Localization errors due to inaccuracies in rendering individualized torso boundaries (i.e. in the presence of modeling errors) were, in general, larger than those due to noise and reached their minimum between 16 and 28 ms after the onset (the last column in Table VI). The errors were, on average, in the range of 11-39 mm (24±18/30±13 mm at 20 ms) for the body surface potentials. Visual inspection of inversely calculated dipole positions revealed that they were often distal to the actual locations of accessory pathways and sometimes fell outside anatomically plausible (ventricular) region.

TABLE VI. ACCURACY IN LOCALIZING DUAL ACCESSORY PATHWAYS IN PRESENCE OF DIFFERENT NOISE SOURCES*

Abbr.	2.5 μV first / second dipole (time)	5 μV first / second dipole (time)	20 μV first / second dipole (time)	Modeling noise & 5 μV first / second dipole (time)
1a-1b	7±2 / 5± 2mm (16ms)	9± 3 / 7±3mm (16ms)$^{10\%}$	13± 5 / 18± 7mm (20ms)	22±10 / 20± 5mm (24ms)
1a-1c	10±1 / 5±0.4mm (20ms)	11± 1 / 5±1mm (20ms)	15± 3 / 12± 3mm (28ms)$^{10\%}$	23± 3 / 22± 2mm (28ms)
1a-1d	9±5 / 4± 2mm (8 ms)	11± 6 / 5±1mm (12ms)	20± 7 / 8± 4mm (20ms)	17± 1 / 27± 4mm (28ms)
2a-2b	11±3 / 9± 4mm (16ms)$^{20\%}$	21±10 / 11±4mm (24ms)$^{20\%}$	17± 8 / 20± 7mm (28ms)$^{10\%}$	39± 9 / 18±10mm (16ms)$^{50\%}$
2a-2c	12±5 / 12± 10mm (16ms)$^{30\%}$	19±12 / 8±6mm (20ms)$^{30\%}$	19± 5 / 21±10mm (28ms)$^{20\%}$	18± 9 / 36±18mm (16ms)$^{50\%}$
2a-2d	10±4 / 11± 6mm (16ms)$^{40\%}$	14± 9 / 13±5mm (16ms)$^{30\%}$	22±11 / 19±10mm (24ms)$^{30\%}$	19±12 / 34±16mm (16ms)$^{50\%}$
1b-2b	5±1 / 6± 1mm (12ms)	7± 2 / 8±3mm (16ms)	13± 4 / 19±10mm (20ms)	16± 2 / 11± 4mm (16ms)
3a-3b	10±2 / 6± 3mm (20ms)	11± 2 / 10±6mm (20ms)	15± 4 / 11± 8mm (32ms)$^{20\%}$	11± 5 / 23± 7mm (28ms)$^{10\%}$
All sites	10±4 / 9± 7mm (16ms)$^{11\%}$	12± 6 / 11±6mm (20ms)$^{10\%}$	18± 8 / 18± 9mm (20ms)$^{12\%}$	24±18 / 30±13mm (28ms)$^{9\%}$

*Three-dimensional distances (±SD) between each of two accessory pathways and two reconstructed dipoles are shown. In the first three columns standard torso model was used for the forward solution with different Gaussian noise RMS level of 2.5, 5 and 20 μV. In the last column the individualized torso model was used for the forward problem solution with Gaussian noise RMS level of 5 μV. The standard torso model was used for the inverse problem solution in all cases. See Table III for anatomical descriptions of accessory pathways. Label N% indicates rejected inverse solutions as percentage of total number of inverse solutions. We rejected all solutions for which the magnitude of the stronger of the two dipoles exceeded the weaker dipole by a factor of 5.

IV. CONCLUSIONS

In this simulation study, we revisited a well-known bio-electromagnetic model that substitutes true cardiac sources with idealized equivalent single and dual dipole sources. We constructed a simplified analytically solvable source and volume conductor model for evaluation electrocardiographic inverse problem solutions. Both fitted single current dipoles and dual current dipoles virtually coincided with the original source for high S/N ratios. In the presence of higher noise we found that i) lead systems positioned closer to sources are more efficient and ii) dual current dipole location recovery is sensitive to the distance between the original dipoles.

Our anatomical model of the human ventricles provided us with means to investigate quantitatively the performance of such a model for more realistic electrocardiographic inverse solutions. The results of this study demonstrate that a source model consisting of two dipoles embedded in the realistically shaped torso volume conductor model could be useful in localizing dual accessory pathways providing that i) we know the torso geometry of a given patient ii) we have a priori knowledge of the presence of dual accessory pathways. Both aspects need to be determined non-invasively to be of clinical use.

REFERENCES

[1] P. Savard, F. A. Roberge, and R. M. Gulrajani, "A simulation study of the single moving dipole representation of cardiac electrical activity," *IEEE Trans. Biomed. Eng.*, vol. BME-29, pp. 700–707, 1982.

[2] R. M. Gulrajani, F. A. Roberge, and P. Savard, "Moving dipole inverse ECG and EEG solutions," *IEEE Trans. Biomed. Eng.*, vol. BME-31, pp. 903–910, 1984.

[3] T. F. Oostendorp and A. van Oosterom, "Source parameter estimation in inhomogeneous volume conductors of arbitrary shape," *IEEE Trans. Biomed. Eng.*, vol. BME-36, no. 3, pp. 382–391, 1989.

[4] J. de Guise, R. M. Gulrajani, P. Savard, and R. G. F. A. Roberge, "Inverse recovery of two moving dipoles from simulated surface potential distributions on a realistic human torso model," *IEEE Trans. Biomed. Eng.*, vol. BME–32, pp. 126–135, 1985.

[5] J. H. McClelland, K. J. Beckman, X. Wang, H. A. Hazlitt, M. I. Prior, N. Twindale, K. P. Moulton, C. A. Roman, R. Lazzara, and W. M. Jackman, "Radiofrequency ablation elucidates accessory pathway anatomy," *Circulation*, vol. 84, pp. II–24, 1991.

[6] I. Singer, *Intervantional electrophysiology*. Baltimore: Williams and Wilkins, 1997.

[7] D. Yao, "Electric potential produced by a dipole in a homogeneous conducting sphere," *IEEE Trans. Biomed. Eng.*, vol. BME-44, pp. 964–966, 2000.

[8] D. A. Brody, F. H. Terry, and R. E. Ideker, "Eccentric dipole in a spherical medium: Generalized expressio n for surface potentials," *IEEE Trans. Biomed. Eng.*, vol. BME-20, pp. 141–143, 1973.

[9] R. Hren, G. Stroink, and B. M. Horáček, "Spatial resolution of body surface potential maps and magnetic field maps: A simulation study applied to the identification of ventricular preexcitation sites," *Med. Biol. Eng. Comp.*, vol. 36, pp. 145–157, 1998.

[10] V. Jazbinšek, R. Hren, G. Stroink, B. M. Horáček, and Z. Trontelj, "Value and limitations of an inverse solution for two equivalent dipoles in localising dual accessory pathways," *Med. Biol. Eng. Comp.*, vol. 41, pp. 133–140, 2003.

[11] B. M. Horáček, "Numerical models of inhomogeneous human torso," *Adv. Cardiol.*, vol. 10, pp. 51–57, 1974.

[12] P. Iturralde, M. Guevara-Valdivia, L. Rodríguez-Chávez, A. Medeiros, and L. Colin, "Radiofrequency ablation of multiple accessory pathways," *Europace*, vol. 4, pp. 273–280, 2002.

[13] R. Hren, G. Stroink, and B. M. Horáček, "Accuracy of single-dipole inverse solution when localising ventricular pre-excitation sites: simulation study," *Med. Biol. Eng. Comp.*, vol. 36, pp. 323–329, 1998.

[14] L. J. Leon and B. M. Horáček, "Computer model of excitation and recovery in the anisotropic myocardium. I. rectangular and cubic arrays of excitable elements," *J. Electrocardiol.*, vol. 24, pp. 1–15, 1991.

[15] J. Nenonen, J. A. Edens, L. J. Leon, and B. M. Horáček, "Computer model of excitation and recovery in the anisotropic myocardium. i. rectangular and cubic arrays of excitable elements," in *Computers in Cardiology*, A. Murray and R. Arzbacher, Eds. Los Alamitos: IEEE Computer Society, 1991, pp. 545–548.

[16] C. J. Purcell and G. Stroink, "Moving dipole inverse solutions using realistic torso models," *IEEE Trans. Biomed. Engr.*, vol. BME-38, pp. 82–84, 1991.

[17] J. Nenonen, C. J. Purcell, B. M. Horacek, G. Stroink, and T. Katila, "Magnetocardiographic functional localization using a current dipole in a realistic torso," *IEEE Trans. Biomed. Eng.*, vol. BME-38, no. 7, pp. 658–664, 1991.

[18] E. Macchi, G. Arisi, and B. Taccardi, "Identification of ectopic ventricular foci by means of intracavitary potential mapping: a proposed method," *Acta Cardiol.*, vol. 47, pp. 421–433, 1992.

[19] B. Taccardi, E. Macchi, R. L. Lux, P. R. Ershler, S. Spaggiari, S. Baruffi, and Y. Vyhmeister, "Effect of myocardial fiber direction on epicardial potentials," *Circulation*, vol. 90, pp. 3076–3090, 1994.

Automatic knot fastener Cor-Knot and United States Pharmacopeial Convention Class I knot pull-apart force safety regulations

B.M. Gersak*, J.S. Sauer*
* LSI Solutions, Victor, New York, United States of America
blaz.gersak@maat.si, jsauer@lsisolutions.com

Abstract

Objective was to evaluate security of sutures fixated with automatic knot fastener Cor-Knot and compare to United States Pharmacopeial Convention (USP) regulations for use in minimally invasive surgery.

In vitro we measured the knot pull-apart force (KPAF) for two 2-0 B.Braun PremiCron suture lots - Lot#1-9364 (24 sutures); Lot#110095 (24 sutures), fixated using Cor-Knot devices.

USP Class I minimum average 2-0 suture KPAF is defined as 28.24 newtons. For Lot#1-9364 we measured 46.84 ± 1.97 N (mean ± std dev); for Lot#110095 it was 54.08 ± 2.57 N. There were no statistically significant differences between the two lots (F-test, p=0.102). KPAF for all 48 tested knots was higher than USP Class I. Minimal KPAF - 39.27 N (suture P/N M0027845; Lot#110095) is approximately 11 N above the USP recommendation.

To conclude, Cor-Knot-fastened 2-0 B.Braun PremiCron sutures showed a much higher KPAF compared to the USP regulation recommendations.

I. INTRODUCTION

In minimally-invasive-surgery procedures, although varying from one surgeon to another based on experience, the time required to manually make each individual suture knot is significantly longer than in open-surgery procedures, mainly due to a distorted, restricted field of view and limitations in the available range of instrument and hand movements [1]. For the purpose of speeding up the process, thus reducing the operation time, numerous devices for automatic suturing and/or knotting have been or are being developed.

The device used in this study (Cor-Knot by LSI-Solutions), is an automatic knot-making device, which utilizes small titanium rings to make "metal knots". Each knot is made by crimping a titanium ring, after the two ends of the suture loop have been routed through it. The friction between the crimped ring and the suture ends holds the "metal knot" in place, eliminating the need to manually tie several knots to prevent the suture loop from coming loose. After the ring has been crimped in place, the device also cuts the suture ends at a safe distance from the ring, doing the precise cutting work by itself, so assisting the surgeon.

To ensure the safety of newly-developed medical devices, all United-Stated-based manufacturers must comply with United States Pharmacopeial Convention (USP) regulations. The USP Class I knot pull-apart force safety regulations apply to devices the purpose of which is to fixate surgical suture loops in place (knot-fastener devices), Cor-Knot being an example of such a device.

II. OBJECTIVE

In our study, we tested the knot pull-apart force of knots made with the titanium fastener Cor-Knot. The measured results have been compared to USP Class I safety regulations for use in human minimally invasive general and cardiovascular surgery, which recommend a minimum average 2-0 suture (a suture with a diameter of 0.3 mm, used in minimally-invasive cardiac surgery) knot-pull tensile strength (doubled for tests with suture loops) to be 28.24 newtons.

III. METHODS

We used two suture lots of 2-0 B.Braun PremiCron sutures: Lot #1-9364 (24 sutures) and Lot #110095 (24 sutures). The sutures were fastened with the titanium fastener Cor-Knot, which fixates the suture by crimping a titanium fastener around two suture tails, followed by cutting the two tails at a length approximately 1.5 mm away from the titanium fastener. The pull-apart force was measured for both suture lots, using a Lloyd Instruments Chatillon LRX Plus tensile testing machine using a 2.5 kN Load Cell.

To prepare the samples for the knot pull-apart test, two suture tails were loaded through the Cor-Knot titanium fastener in the Cor-Knot device tip using a Quick-Load-Unit wire snare in routine fashion [1]. With the suture loop placed around a 1.25 inch diameter sizing cylinder, the Cor-Knot device purple lever was squeezed to crimp the fastener, trim the suture, and release the fastener from the device [2]. In order to perform the knot pull-apart test, each suture loop sample was loaded into the Chatillon LRX Plus with one side of the loop placed around an upper hook and the opposite side of the loop placed around a lower hook. The lower hook was affixed to a platform on the base of the machine while the upper hook was affixed to a crosshead that was driven upwards to apply tension to the specimen. The test setup is shown in Figure 1.

The Chatillon LRX Plus was set to move the heads apart at 5.5 inches per minute. Throughout the experiment, we collected the current force data. For the knot-breaking force variable we took the highest measured force for each suture [3].

IV. Results

The tensile strength of Lot #1-9364 was 46.84 ± 1.97 N (mean ± std dev) – min. 41.46 N, max. 51.28 N, and for Lot #110095, it was 54.08 ± 2.57 N (mean ± std dev) – min. 39.27 N, max. 50.36 N, respectively. There were no statistically significant differences between the two lots (F-test, $p = 0.102$).

Comparing tensile strength for both lots fastened with Cor-Knot with the USP Class I safety regulations, the tensile strength of knots fastened with Cor-Knot showed higher values for all 48 individual tested knots [Figure 2]. The minimal tensile strength measured was with suture P/N M0027845 of Lot #110095, and was 39.27 N, which is 11.03 N above the USP-recommended strength of 28.24 N.

V. Conclusions

Cor-Knot titanium-fastened 2-0 B.Braun PremiCron sutures of two tested lots showed a much higher knot pull-apart force compared to the USP Class I safety regulations-recommended minimal average knot-pull tensile strength. The use of Cor-Knot is safe and reliable, and may be recommended for use in human minimally invasive general and cardiovascular surgery as a routine knot fastener, especially in combination with endoscopic surgery.

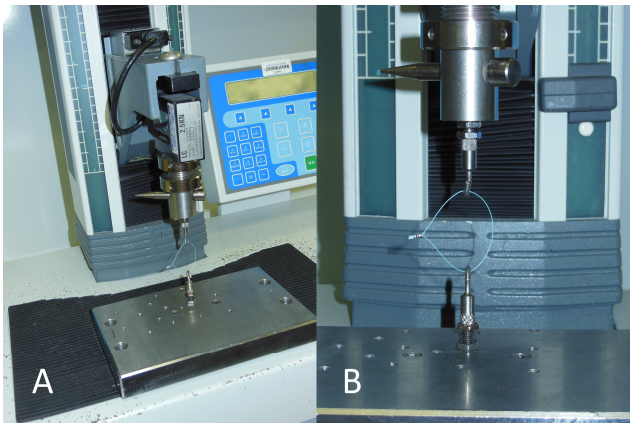


Figure 1. The test setup with a Chatillon LRX Plus tensile testing machine using a 2.5 kN Load Cell (A); a close-up photograph of the suture loop, fixated to the upper and lower hooks - on the machine crosshead and on the base platform, respectively (B).

Acknowledgment

Authors would like to thank Heather R. Gorea for help with experiment execution.

References

[1] B. Gersak, M. Sostaric, J.M. Kalisnik and R. Blumauer, "The preferable use of port access surgical technique for right and left atrial procedures," The heart surgery forum, vol. 8, pp. E354-63. 2005.

[2] C.L. Wilshire, J.W. Kaufer, H.R. Gorea, J.S. Sauer and P.A. Knight, "Healing study of porcine heart transapical wounds closed using a remote automated suturing technology," Innovations (Phila), vol. 8, pp. 50-5. January-February 2013.

[3] B.M. Gersak and J.S. Sauer, "In vitro testing of titanium knot fastener used in cardiac and general surgery with pull apart force," J Cardiothorac Surg, vol. 8, suppl. 1, pp. 64, September 2013.

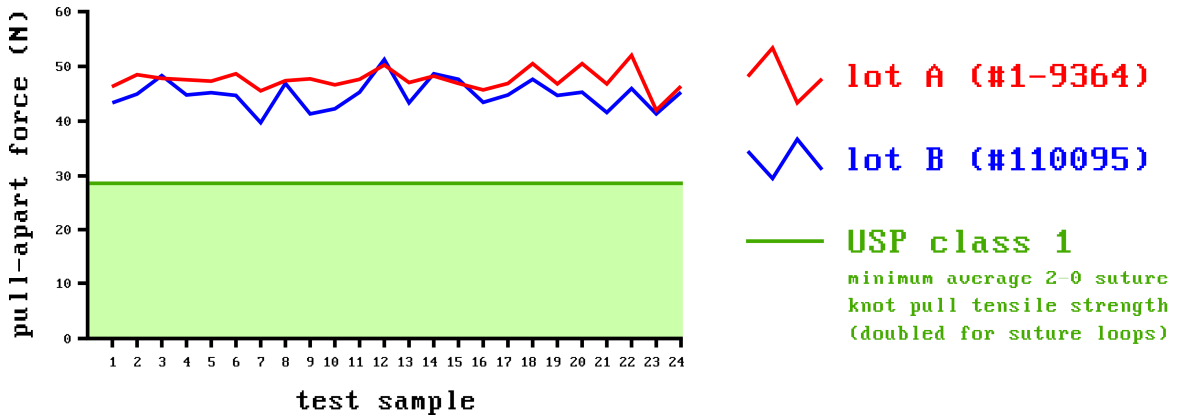


Figure 2. Pull-apart force values for individual knot samples of lots #1-9364 and #110095 in comparison to the USP Class I recommended minimum average value

Control and Optimization of Complex Biological Systems

Draško Tomić, * Božica Pernarić, **

*Hewlett-Packard, Zagreb, Croatia, drasko.tomic@hp.com

** Spec. Vet. Ambulante Marković, Zagreb, Croatia, spec.vet.amb.markovic@inet.hr

Abstract – Healthy and stable state of complex biological systems lies in the narrow region between chaos and frozen state. The cause of chaos is often small initial change in starting conditions, quickly spreading and leading to system wide changes. For example, few microbial pathogens can quickly lead the host to the serious illness and even death. By the definition, an attractor in a certain dynamical system is a set of physical properties toward which this system tends to evolve, regardless of the starting conditions. It is not easy to control chaos and move system from attractor back to the stable state. This is even more difficult in case of hyperchaos with more attractors. Existing control strategies are passive; modify control parameters and wait until system lends in a future, hopefully stable state. At there is a clear need for more efficient control strategy, we propose the optimization of the matrix representing an evolution operator of the system under the control. Matrix elements are system state functions, and function variables are control parameters. Eventually, optimizing this matrix with respect to some goals can help to drive system back to stability. On the case of Samoyed dog lady attacked by methicillin-resistant and multi-species bacterial pathogens, we were able to identify the chaotic system with at least three attractors; two of them leading the system into the illness, and one driving it back to the healthy state. Consecutively, we established the matrix of its evolution operator and discussed some ways of the matrix optimization, in a hope our approach might help to develop therapies that are more efficient. Not of the less importance, we hope our experience with integrative therapy including standard therapeutics, herbal extracts and homeopathic remedies could help a clinician confronted with a similar case.

Key words: complex biological systems, chaos control, evolution operator, attractor, matrix optimization, integrative therapy

I. INTRODUCTION

Complex biological systems are dynamic and adaptive systems consisting of billions of cells interacting at multiple spatial and temporal scales.

Given that relatively small initial changes can eventually cause great and difficult to predict changes, we can say that the complex biological systems are also highly nonlinear dynamic systems, and small initial changes are starting conditions. However, there are changes that we call diseases, and that may lead complex biological system into undesirable state, from which he is not able to escape without intervention from outside, i.e. therapy. Speaking in the framework of chaos theory, system is in the trapping region of chaos, and will remain there until eventually external stimulus while help him to escape from there. By analyzing dynamic systems captured into the chaos, one can find an amazing fact that there is a sub-region within space, seeming to attract the system there. This phenomenon is of the outmost importance in the theory of chaos, and its name is strange attractor. Lorenz attractor [1] is the most famous example of such phenomenon, because not only it was the first, but also for his interesting shape resembling butterfly wings, as can be seen from Fig. 1:

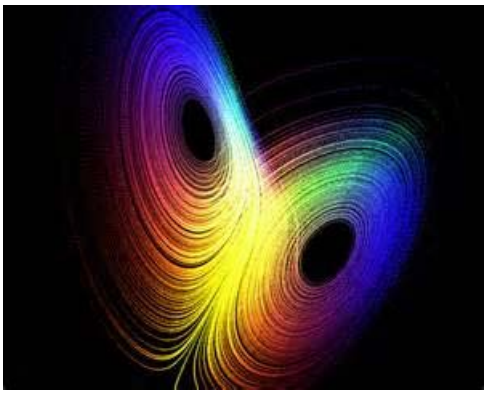


Fig.1: Lorenz strange attractor

On the other side, and despite of the fast development of new drugs, now day is not easy to fight even simple, non-complicated bacterial infection. The reason might seem to be obvious, i.e. ever increasing bacterial resistance to antibiotics and genome transfer of this resistance among various bacterial strains and even species.

However, many bacterial species developed additional surviving strategies, like ability to penetrate deep in the tissues, autonomous movement, and production of toxins able to damage not only host, but also other competitive bacteria. Therefore, the clinician treating bacterial infection eligible for up to twenty potential drugs may need to provide answers on many other questions; Should I treat the infection or not? Which is the best drug or combination for that pathogen and patient? Which dosage is optimal? Should we give the drugs together, or at some interval? Should the drug dose and interval change over time? How long administration can take place? What is the metric of success? In case of the host invaded by several bacterial species, answering these questions is even more difficult, and the complexity of therapy might be far beyond standard therapeutic protocols. Clearly, the treatment parameter space available is huge and the options virtually infinite. That is where mathematical modelling may help. By simulating millions of treatment combinations in a fraction of the time, one can help the clinician to make better decisions that have a greater chance of success. An attempt to develop mathematical models of therapies is not so frequent, because they appear complicated to model, and because of ever evolving complex, dynamical system. From our work on real case described in the next session, has come a clear conclusion: biological system under simultaneous attack of various microbial pathogens is indeed complicated and complex, but incomprehensible it is not. However, it requires a lot of collaboration between clinician, microbiologists and patient's family, is time consuming, and expensive. While development of new and more effective drugs remains the dominant goal in microbiology, clinician able to simulate millions of treatment combinations in a fraction of the time might make better decisions that have a greater chance of success. Moreover, it does not mean that mathematicians and their models will replace clinicians. It will allow placing the complete system in an integrated framework, from which it will be possible to examine the impact that change of one or several components has on the whole system. As it will be clearer in the next section describing a real case, the infection of complex biological system with multi-species bacterial pathogens is very hard to control. Failure to integrate mathematical and computational models into biology and therapy, will inevitably limit our progress with diseases of such complexity. Clinicians need to use advanced quantitative methods to integrate known data with underlying evolutionary-first principles to make predictions about the unknown outcome. By doing so one can develop personalized therapeutic strategies for complex diseases that are as dynamic and adaptive as host they are attacking. Besides, mathematical theory of complex nonlinear systems, aka chaos theory [2], is already available, and can help us to analyze and control such systems. Among several approaches available in chaos theory, one is to establish spectral evolution operator of the system analyzed [3].

In addition to physics, spectral analysis of complex systems has been successful in the analysis and optimization of human walk [4] and high performance computing [5]. For this reason, in this work we investigated spectral analysis applicability on complex biological systems infected by multi-species bacterial pathogens.

II. CASE STUDY

Mini Vini, 8 years old and 23 kilogram weighing female Samoyed dog lady came to the referring veterinarian because of the frequent urination, i.e. pollakiuria and blood in urine, hematuria. The dog already had a history of urinary tract infection, caused by Escherichia Coli and Staphylococcus spp., cured with amoxicillin/clavulanic acid and enrofloxacin antibiotics. Besides, there was the history of the serious diarrhea, probably caused by an excessive use of antibiotics. The current diet was a commercial vet food, vaccinations were current, and the dog was not receiving any medication. Ultrasound examination of bladder revealed urolithiasis and formations on bladder walls, as on Fig. 2:

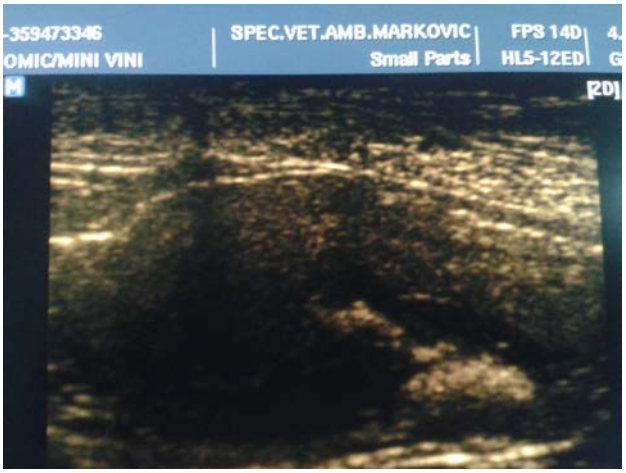


Fig 2: Ultrasound image of bladder

Urine analysis showed ph value was high and up to nine, with high number of leukocytes and erythrocytes. Blood analysis revealed anemia. Subsequently, antibiotic therapy with amoxicillin/clavulanic acid followed, until urine values came to normal values. Besides, as without the cystoscopy it was not possible to define the cause of formations on bladder walls, we accompanied the antibiotic therapy with an extract from Uncaria tomentosa herb, supposed to have some anti-cancer properties. About two months later, ultrasound examination showed no changes on bladder walls. However, hematuria and pollakiuria repeated several weeks later, with similar values of urine found at the first examination.

This time urine microbial analysis revealed Staphylococcus pseudintermedius. Because of the

possible microbial resistance on antibiotics, dog owner decided to start an alternative therapy with Arctium lappa herb, known for his in vitro activity against Staphylococcal species [6]. We accompanied the therapy by struvite prescription diet and DL-Methionine, urine acidifier. About two months later, urine control parameters came to normal values, and urine sediment analysis under microscope showed crystals dissolving. For the next three months, patient continued with the same diet and acidifier, accompanied with Rowatinex, herbal remedy known to be effective against cristalliuria [7]. Several months passed before the next UTI symptoms repeated, and this time microbial analysis of urine revealed E. coli. Repeated therapy with amoxicillin/clavulanic acid followed. Besides, and to avoid the recurrence of UTI, we accompanied the antibiotic therapy with immunotherapy, i.e. with Uro-Vaxom, known to significantly suppress the recurrence of UTI caused by E. coli. Because E. coli was still present on the subsequent microbial urine analysis, we continued therapy with ciprofloxacin. Following the next microbial urine analysis, E. coli was not present, however this time Enterococcus faecalis and Staphylococcus pseudintermedius, both resistant to all approved veterinarian antibiotics, showed their presence. Besides, ultrasound examination of the bladder revealed many urine stones, and the blood analysis showed extremely high values of triglycerides and alkaline phosphatase values. For this reasons, we discontinued further antibiotic therapy and struvite prescription diet, and started combined therapy with the preparations from Arctium lappa, Uncaria tomentosa, Cinnamonum zeylanicum, Hydrastis canadiensis, Origanum compactum and Punica granatum herbs instead. According to [9], [10], [11] and [12], these herbs express in-vitro antimicrobial activity against Enterococcus and Staphylococcus species. We also accompanied the therapy with herbal and homeopathic remedies from AllergicPet, which according to the producer, can help dissolve bladder stones and fight UTI. About two months after starting this alternative therapy, urine analysis showed no presence of Enterococcus, and ultrasound bladder examination showed massive dissolution of struvite stones. However, Staphylococcus was still present, but this time showing no antibiotic resistance. As urine ph value was above seven, and Uva Ursi herb is supposed to have antimicrobial activity above this value, we started the therapy with this herb, accompanied with herbal tea from Betula pendula [13], until UTI symptoms gone. Subsequent bladder ultrasound examination found no crystals, however mud was still there. After discontinuance of Uva Ursi therapy, which might cause serious side effect if applied for more than a two weeks, UTI symptoms returned in a few days, and subsequent urine analysis revealed E. coli again. Antibiogram showed its sensitivity to a broad spectrum of antibiotics, including amoxicillin/clavulanic acid.

Besides, and as the intake of Uva Ursi should not be given more than a five times per year, therapy with amoxicillin/clavulanic-acid followed. At the end of this antibiotic therapy urine analyses showed it sterile. However, UTI symptoms returned only few days after. For this reason, we started therapy with Uva Ursi again. As at this time urine was acidic, we accompanied therapy with Basenpulver, which might increase the alkalinity of urine. Despite our expectations, E. coli showed no sensitivity to Uva Ursi, and UTI symptoms worsened. Therefore, we discontinued its further use, and tried therapy with D-mannose, at dosage of 2 grams per every 4 hours. D-mannose is non-digestible synthetic sugar, therefore completely flushed out several hours after the intake. However, its chemical structure is similar to those of oligosaccharide, the component of protective glycoprotein plagues on bladder wall. For this reason, E. coli might stick to D-mannose sugar molecules and host is able to flush it from bladder by urination. Very recently, pilot study proved the efficiency of D-Mannose [14]. After several days of the treatment with D-Mannose, urine dipstick values returned to normal values, however UTI symptoms returned soon after discontinuing the intake of D-mannose. Therefore, we applied the therapy with amoxicillin/clavulanic acid and subsequent microbial analysis showed sterile urine. Only several days later UTI symptoms recurred. In compliance with antibiogram, new therapy cycle with sulfametoxazole/trimetoprim antibiotic followed. At the same time, and to prevent possible UTI recurrence, we started a long-term supplemental therapy with d-mannose, forskolin and Cranberry fruit extract. E. coli is able to attach and penetrate in bladder wall, thus hiding from antimicrobial agents. According to [15], forskolin, an extract from Coleus forskohlii herb, can elevate host cAMP levels, thus forcing pathogenic bacteria out of bladder epithelial cells. Proanthocyanidin is an active component present in Cranberry fruit and can prevent the adhesion of E. coli to bladder wall [16]. After finishing the antibiotic therapy, urine analysis revealed it sterile, without leukocytes and leukocyte esterase present. At the time of the submitting the revised version of this paper, urine was sterile for a several weeks, and we were proceeding with supplemental therapy with Uro-Vaxom, D-mannose, Cranberry and forskolin therapy, all in order to prevent E. coli recurrence. On the other side, mud in the bladder was still present. Therefore, and in order to prevent a possible recurrence of recurrent Staphylococcus and Enterococcus based UTI, we are continuing to apply herbal preparations of Arctium lappa, Cinnamonum zeylanicum, Origanum majorana and Hydrastis canadiensis. Besides, and in in order to get completely free of mud, we are continuing to give homeopathic remedy for struvite stones dissolution, plus DL-Methionine and vitamin C.

III. THE CONTROL OF DYNAMIC SYSTEM

During the therapy course, periodical weekly analasys of urine with dipstic took place. Among the ten urine parameters available on diagnostic dipstick, we concentrated of the following, relevant for UTI:

- Number of leukocites in urine, abbreviated LEU, ranging from 0 to 500 per microliter; no leukocites is normal value.
- Acidicy of urine, abbreviated ph, ranging from 5 to 9; expected values are in range from 6 to 7.
- Protein in urine, abbreviated PRO, ranging from 0 to 500 mg/L; no proteins in urine is normal value.
- Urine specific gravity, abbreviated SG, ranging from 1000 to 1030; expected values are in a range from 1015 to 1025.
- Number of erythrocytes in urine, abbreviated ERY, ranging from 0 to 250 per microliter; no erythrocytes are normal value.

Besides, monthly periodical analysis of urine on microbes took place. The normal value is zero number of microbes, the significant number of microbes is at $>10^3$, and high is as $>10^6$ microbes in microliter of urine. When microbes are present in urine, usually antibiogram with about ten or more antibiotics takes place, and microbes tested on their sensitivity against specific antibiotics. We also performed the blood tests, either periodically or on demand. There are several dozens of blood parameters significant for the severity of the infection and the side effects of therapy. For the sake of simplicity, we will not consider these parameters, however for the goal of our paper it is important to note that together with urine parameters, number of state parameters can readily jump over 50. Making together with available control parameters, i.e. antibiotics, herbal and homeopathic medications active against one or more microbe species, one quickly confronts with a very complex dynamic system. During a therapy, we noticed that a system under the microbial attack moves to two regions, both with high number of leukocytes in urine. However, under the attack of Escherichia coli and Enterococcus faecalis urine ph was acidic, while under the attack of Staphylococcus species was alkaline. By plotting ph values versus leukocyte numbers, we realize the existence of two regions. We encircled them with blue, as can be seen in Fig. 3:

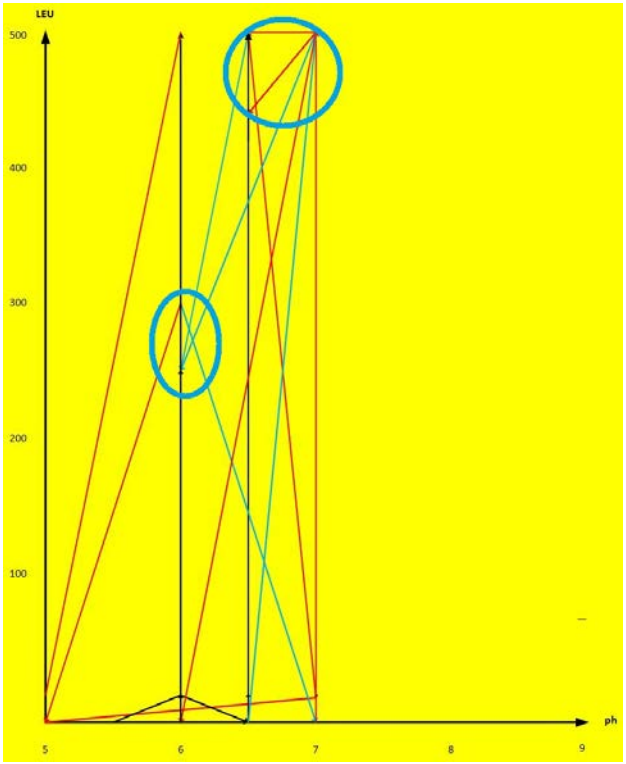


Fig. 3: Number of leukocytes versus ph urine values

In the former figure, lines signify the development of the system from one state to another. Moreover and for the sake of visibility, we denoted with red lines the cases when number of leukocytes was rising and with blue when it was failing over the time. Finally, with black lines we denote cases where red and blue lines are connecting same states. What we found is that leukocyte number correlates with severity of the infection. For example, with number of leukocytes approaching 500, additional symptoms like hematuria and pollakiuria quickly take place. In addition, ph and leukocyte values change very quickly, and there are two regions where they tend to reside, which we cycled in blue. Furthermore, we realized several phenomena that support our assumption that the system under observation is chaotic. It moves quickly from one state to another, it is not possible to predict which state it will reach, and even small elevation of ph value above 6.5 starts the process of massive struvite stones dissolution, thus releasing a huge number of bacteria cocooned in them. Besides, and as states are grouped within two regions encircled blue in Fig. 3., we might assume that this system has at least two attractors: One due to the pathogens E. Coli and Enterococcus Faecalis, another due to the Staphylococcus intermedius. Besides, it is quite possible that at least one additional attractor due to the forces keeping the system in the healthy state exists. For the reasons explained, we assume that system observed is chaotic, and because there are several regions of attraction, we suppose the presence of hyperchaos [17].

The current strategies of controlling hyperchaos in systems with even low number of attractors are passive; change control parameters, and wait until system transit from the current into the next state. If the next state is also the stable state, then we can say control was successful. Otherwise, one needs to change control parameters again, hoping that system will eventually land into the stable state next time. It is clear that passive control strategies are not appropriate in many cases, especially in those when right control can made a narrow line between life and the death. Therefore, we need more advanced control strategies for chaotic and especially hyper-chaotic systems.

IV.CALCULLATING MICROBES

During the therapy course, we focused on three groups of state parameters: first group of parameters obtained by reading urine dipstick, second group obtained by performing microbial analysis of urine and antibiogram, and the last group of parameters resulting from blood analysis. Besides, we regularly checked the quantity and state of urine stones via ultrasound and under the microscope. As therapeutic, i.e. control parameters, we used several antibiotics according to antibiograms, herbal preparations with in-vitro proven activity against bacteria found, D-Mannose sugar, struvite prescription food low in phosphates, proteins and magnesium, but high in fat. Additionally, we applied Uro-Vaxom, Rowatinex, herbal and homeopathic remedies against stones and UTI from Allergic Pet, vitamin C and DL-Methionine as urine acidifiers and Basenpulver, which alkalizes urine. All state and control parameters are viewable from Fig. 4:

State parameters	**Abb.**	**Control parameters**	**Abb.**
Specific gravity	SG	Ciprofloxacin	CFX
Acidity	PH	Amoxicillin	AMX
Leukocytes	LEU	Enrofloxacin	EFX
Nitrites	NIT	Sulfametoxazole/ trimetoprim	SMX
Proteins	PRO	Cranberry	CB
Blood in urine	BL	Uncaria tomentosa	UT
Urine stones	UST	Arctium lappa	AL
Triglycerides	TG	Cinnamon zeylanicum	CZ
Erythrocytes	ERY	Rosa canina	RC
Alkaline phosphatase	AP	Coleus forskohlii	CF
E. coli	EC	Origanum compactum	OC
Enterococcus faecalis	EF	Hydrastis canadiensis	HC
Enterococcus faecalis resistant	EFR	Uva Ursi	UU
Staphylococcus pseudintermedius	S	Punica granatum	PG
Staphylococcus pseudintermedius resistant	SR	Betula pendula	BP
		D-Mannose	DM
		Struvite prescription diet	SD
		Uro-Vaxom	UX
		Rowatinex	RW
		Homeopathic remedy against stones	HR1
		Homeopathic remedy against UTI	HR2
		Vitamin C	UD1
		DL Methionine	UD2
		Basenpulver	UA

Fig. 4: State and control parameters

Further, we introduce the evolution operator of the system, with functions describing status parameter changes over the time. As a response to the therapy, state parameters, with control parameters as variables, change over the time, and the system advances through the time space. In [3], there is a mathematical definition of evolution operator and its discrete presentation in form of iterated map given. Here, we express the intermediate form of the evolution operator for this case and for some parameters, as can be seen on Fig. 4:

	SG	PH	LEU	NIT	PRO	BL	UST	TG	ERY	EC	EF	EFR	S	SR
CFX	x	x	x	x	x	x		x	x	x	x		x	
AMX		x	x	x	x	x			x	x	x		x	
EFX		x	x	x	x	x			x	x	x		x	
SMX		x	x	x	x	x			x	x				
CB			x	x	x	x			x	x				
UT			x		x	x			x		x	x	x	x
AL			x		x	x			x		x	x	x	x
CZ			x	x	x	x			x					
RC			x	x	x	x			x	x				
CF					x	x			x	x				
OC			x	x	x	x			x	x	x	x	x	x
HC			x	x	x	x			x	x	x	x	x	
UU			x	x	x	x			x	x				
PG											x	x	x	x
BP			x		x	x			x	x	x	x	x	x
DM			x	x	x					x				
SD	x					x		x	x					
UX			x	x	x	x			x	x				
RW	x					x	x		x					
HR1	x					x	x		x					
HR2			x	x	x					x	x	x	x	x
UD1	x	x					x							
UD2	x	x					x							
UA	x	x					x							

Fig 4: The evolution operator of UTI

State parameters are nonlinear functions of control parameters. As functions are unknown, we denote them with **x**. By setting up these functions, one can establish the algebraic form of evolution operator, and finally perform an optimization with respect to the given goals. In case of UTI case described here, there are two goals. One is to get completely free of pathogens, and another, not of less importance, is to get rid of struvite crystals, which are reservoirs of pathogens on a long-term basis. If optimization is successful, values of control parameters guarantee the most efficient therapy. In addition, the most efficient therapy means to delete pathogens as fast as possible, without introducing any serious side effects. However, the functions mentioned before are of very high complexity; good example could be analytically modeling an activity of Coleus forskohlii, supposed to wipe out pathogen from bladder epithelia.

V. CONCLUSION AND FUTURE WORK

The scope of this paper is not to find out the optimal therapy for this case, but to establish directions for the future work. The work to perform this might seem to be enormous, however disserves our full attention, both because bacteria become more resistant on therapeutics than ever, and because the costs of developing new antibiotics are sky rocketing. On the other side, sophisticated biomechanical models are developing at an ever-increasing rate, thus enabling us to simulate on Supercomputers, even at a fine scale, quite complex biological systems. However, some questions still wait for answer; How long will our optimization process execute? How big is an error between optimal and optimized solution? The final, but by far not easier question remains: Is there an algorithm able to perform an optimization of my problem available at all? Now, one of the most promising ways seems to be the semidefinite optimization of the spectrum of the matrix representing evolution operator. For example, this approach was successful in optimizing human walk [4] and speeding-up Supercomputer [5]. Therefore, it is likely to expect that the same or slightly modified approach will help us to improve complex therapy protocols as well. Lastly, we hope our integrative therapy approach, i.e. the combination of the standard, herbal and homeopathic medicine can help clinicians to overwhelm similar cases and provide a cure to the severe UTI diseases.

VI. ACKNOWLEDGMENT

We would like to thanks Boris Habrun and Gordan Kompes, both from Croatian Veterinary Institute, for their fruitful advices regarding the treatment process.

We would also like to thank lovely Samoyed dog Mini Vini, for her enormous patience, collaboration, and for her faith we will cure her illness.

REFERENCES

[1] http://en.wikipedia.org/wiki/Lorenz_system

[2] R. L. Devanay, “An Introduction to Chaotic Dynamical Systems, 2nd Edition, Amazon publishing.

[3] D. Tomic, “Spectral performance evaluation of parallel processing systems”, J. Chaos, Solitons and Fractals, 13 (1) (2002) 25-38.

[4] D. Tomic, O. Muftic, S. Baksa, “Computer 3D Spectral Analysis of Human Movements”, In. Proc. of the Int. Conf. on Computer-Aided Ergonomics, Kosice, Slovakia, 2005.

[5] D. Tomic, L.Gjenero, E. Imamagic, “Semidefinite optimization of High Performance Linpack on heterogeneous cluster”, 2013. Proceedings of the 36tth International Convention MIPRO, Opatija, pp. 157 – 162.

[6] J. V. Pereira, D. C. B. Bergamo, J. O. Pereira, S. de Castro Franca, R. C. L. Pietro, Y. T. C. Sylva-Sousa, “Antimicrobial Activity of Arctium lappa Constituents Against Microorganisms Commonly Found in Endodontic Infections”, Braz. Dent. J. 16(3), pp. 192-196, 2005.

[7] Djaladat H., Mahouri K., Khalifeh Shoostary F., Ahmadieh A., “Effect of Rowatinex on calculus clearance after extracorporeal shock vawe lithotripsy”, Urol J. 2009 Winter;6(1):9-13

[8] K. Kim et all, “A Prospective Multi-center Trial of Escherichia Coli Extract for the Prophylactic Treatment of Patients with Chronically Recurrent Cystitis”, J Korean Med Sci 2010; 25; pp. 435-439.

[9] Daniel R. Herrera, Lidia Y. Tay, Eluise C. Rezende, Vitoldo A. Kozlowski Jr. and Elisabete B. dos Santos, “In vitro antimicrobial activity of phytotherapic Uncaria Tomentosa against endodontic pathogens”, Journal of Oral Science, Vol. 52, No. 3, 473-476, 2010

[10] B. Joshi, S. Lekhak, A. Sharma, “Antibacterial Property of Different Medicinal Plants: Ocinum sanctum, Cinnamonum zeylanicum, Xanthosylum armatum and Origanum majorana”, Kathmandu University Journal of Science, Engineering and Technology, Vol. 5, No I, January, 2009, pp 143 –

[11] K. A. Ettefagh, J. T. Burns, H. A. Junio, G. W. Kaatz, N. B. Cech, “Goldenseal (Hydrastis canadensis L.) extracts synergistically enhance the antibacterial activity of berberine via efflux pump inhibition”, Planta Med. 2011 May ; 77(8), pp. 835-840.

[12] T. Dadgar et all, “Antibacterial Activity of Certain Iranian Medicinal Plants Against Methicillin-Resistant and Sensitive Staphylococcus aureus”, Asian Journal of Plant Sciences 5 (5): pp 861 – 866, 2006.

[13] S. Ljubisa, C. Ivana, B. Bojana, M. Aleksandra, S. Marijana and P. Dragana, “Antimicrobial Activity of Plants Extracts from Serbia”, Journal of Institute for Food Technology, Serbia, Vol. 36, Issue 1-2, pp. 1-6.

[14] D Mannose in Recurrent Urinary Tract Infection, Clinical trial, ID: NCT01808755

[15] D. S. Eto, M. A. Mulvey, “Flushing bacteria out of the bladder”, Nature Medicine 13, pp. 531-532.

[16] R.Raz, B. Chazan, M. Dan, “Cranberry Juice and Urinary Tract Infection”, L. Clinical Infectious Diseases, Vol. 38, Issue 10, pp. 1413-1419.

[17] G. M. Mahmoud, E. E. Mahmoud, A. A. Arafa, “Controlling Hyperchaotic complex systems with unknown parameters based on adaptive passive method”, Chin. Phys. B, Vol. 22, No 6, 2013.

Classification of benign and malignant masses in breast mammograms

A. Šerifović-Trbalić*, A. Trbalić**, D. Demirović*, N. Prljača* and P.C. Cattin***
* Faculty of Electrical Engineering, University of Tuzla, Tuzla, Bosnia and Herzegovina
** Drvodom d.o.o, Tuzla, Bosnia and Herzegovina
*** Medical Image Analysis Center (MIAC), University of Basel, Basel, Switzerland
amira.serifovic@untz.ba

Abstract - An accurate and efficient computer-aided mammography diagnosis system plays an important role as a second opinion to assist radiologists. Finding an accurate and robust computer-aided diagnosis system for classification of the abnormalities in the mammograms as malignant or benign still remains a challenge in the digital mammography. In this paper, a fully autonomous classification system is presented and it consists of the three stages. The input Regions of Interest (ROIs) are obtained using an efficient Otsu's N thresholding and further subjected to a number of preprocessing stages. After preprocessing stage, from the ROIs, a group of 32 Zernike moments with different orders and iterations have been extracted. These moments have been applied to the neural network classifier. The experimental results show that the proposed algorithm is efficient comparing to the ground truth table given in the Mammography Image Analysis Society (MIAS) database.

Keywords: computer-aided diagnosis, mammography, segmentation, neural network, Zernike moments

I. INTRODUCTION

Breast cancer represents the major cause of death by cancer among women. Cancer is a disease in which cells in a part of the body become abnormal and form more cells in an uncontrolled way. In most cases, the cancer cells form a tumor. With breast cancer, the cancer originates in the tissues that make up the breasts. They may also invade nearby tissues and spread to lymph nodes and other parts of the body.

Breast cancer tumors typically do not produce symptoms until they are relatively large at advanced stages. Most breast cancer tumors at earlier stages are benign and not yet malignant and life-threatening. Thus, the early detection of the breast cancer masses can increase the survival rate for patients. Mammography has been one of the most reliable methods for early detection and diagnosis of the breast cancer, typically through detection of characteristic masses and/or microcalcifications (such as tiny deposits of calcium). Although mammography is a sensitive procedure for detecting breast cancer, it lacks high diagnostic accuracy and it is difficult to interpret [1].

Therefore, an accurate and efficient computer-aided mammography diagnosis system plays an important role as a second opinion to assist radiologists. Finding an accurate and robust computer-aided diagnosis system for the classification of the abnormalities in the mammograms still remains a challenge.

In the past two decades, a lot of research has been invested in computer-aided analysis and diagnosis of breast mammograms. Several computer-aided diagnosis (CAD) systems have been developed for the detection and classification of breast masses in mammograms using various image features such as shape, fractal dimension, image moments and laws texture features, e.g. [2 – 7].

Polakowski [2] used shape, size, contrast, and Laws texture features for the classification of masses using a neural network classifier. Brzakovic [3] proposed a classifier that employed four features: size, shape, edge distance variation, and edge intensity variation. Classification of masses using fractal dimension features was proposed in [4]. Rangayyan [5] utilized the "acutance" feature to discriminate between benign and malignant tumors based on averaged gray-level differences between the inside and outside of a tumor boundary. In [6] different combinations of shape features including compactness, Fourier descriptors, moments, and chord-length statistics were examined. Parabolic modeling method to characterize a tumor's boundary was proposed in [7].

Mass classification problems were investigated in [8,9] with different Gabor features and successive enhancement learning based weighted support sector machine. The Gabor feature extraction technique is used to extract multi-scale and multi-orientation texture features. Masses in mammogram are classified using Extreme Learning Machine (ELM) based classifier as proposed in [10]. Automatic mass classification into benign and malignant was presented in [11] based on the features extracted from the contourlet coefficients using non subsampled contourlet transform. The mammogram images are classified by using a classifier based on support vector machine. Discrete wavelet transforms decomposition and an artificial neural network classifier is used for malignant/benign masses classification of region of interest in mammograms is proposed in [12].

A classification using multilayer perceptron network with a novel learning rule was developed in [13]. The features include circularity, Zernike moments, contrast

and average gray level. Decision making is performed in both feature extraction and classification stage.

Classification of breast masses as proposed in [14] is based on shape and texture descriptors, i.e. masses and non-masses. The shape descriptor includes eccentricity, circularity and convexity. Artificial neural network is then used to categorize the masses, which performs benign-malignant classification on region of interest, as explained in [15]. The textural features used for characterizing the masses are mean, standard deviation, entropy, skewness, kurtosis and uniformity. Support vector machine (SVM) and wavelet decomposition was investigated for classifying the masses in [16]. Decision making is performed in both stages: in feature extraction stage and classification stage.

In this paper we mainly concentrate on the recognition of two general masses: benign masses and malignant masses. The input ROIs are obtained from the digital mammograms using Otsu's segmentation algorithm. In order to classify types of masses present in the ROIs, we use 32 Zernike moments as features. In the last stage of algorithm, segmented masses are classified using resilient backpropagation neural networks.

II. MATERIALS AND METHODS

Fig.1 presents the block diagram of the proposed method for the classification of benign and malignant masses in breast mammograms. It involves four major stages. First, image preprocessing of the digitized mammogram is used to suppress noise and improve the contrast of the image. Second, image segmentation is used to identify the masses in mammograms, i.e. to detect the Regions of Interest (ROI). In the third phase, features are extracted and selected for classification of mass. Finally, the classification of masses is conducted.

A. Image preprocessing

Prior to the image segmentation it is necessary to improve the image quality using appropriate noise removal and contrast enhancement techniques. Digitization noises such as straight lines present in the majority of acquired mammogram images are filtered using a two-dimensional (2D) median filter with a 3×3 neighborhood. The edges of the images, however, are replaced by zeros. Image acquisition can induce different artifacts, background noise, scratches that could influence the reliability of the segmentation process.

Most of the mammograms have non-uniform background and very little contrast in the breast tissue region. So image enhancement is required before segmentation. As enhancement technique was selected *contrast-limited adaptive histogram equalization* (CLAHE) [17]. CLAHE operates on small regions in the image, called *tiles*, rather than the entire image. Each tile's contrast is enhanced, so that the histogram of the output region approximately matches a specified histogram distribution. The neighboring tiles are then combined using bilinear interpolation to eliminate artificially induced boundaries. The contrast, especially in homogeneous areas, can be limited to avoid amplifying any noise that might be present in the image.

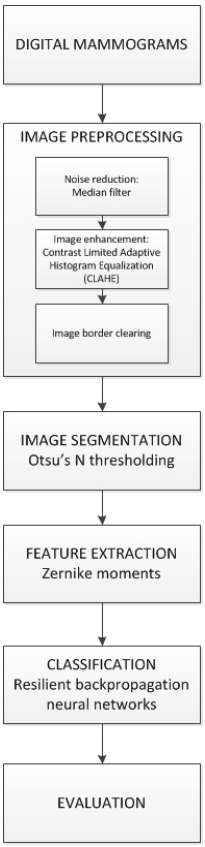


Figure 1. Block diagram of the proposed method

After image enhancement, the automatic thresholding and image border cleaning were applied to suppress the light structures connected to the image border and thus to identify the foreground in the image. The labels on the image as well as the pectoral muscle are removed, since they can disturb the detection of the breast cancer and reduce the rate of accuracy of the CAD system.

B. Segmentation

The objective of the second processing stage is to identify the masses in mammograms. Generally, the identification of different masses is a very difficult task due to rich variations of gray level intensities in mammograms. In this paper we have used Otsu's N segmentation algorithm for segmentation of digital mammogram images. This algorithm represents an optimum global thresholding method proposed by [18] and later modified by [19] to threshold the image with multiple thresholds.

A multiclass Otsu's thresholding has been used to find the Region of Interest (ROI). Otsu's thresholding is a nonparametric and unsupervised method of automatic threshold selection for the image segmentation [19]. It is a simple procedure that utilizes the zeroth- and first-order moments of the graylevel histograms and it maximizes the between class variances, while minimizing the within class variances. Otsu's N segmentation can be described as follows.

Let the pixels of the input image be represented in L gray intensity levels $[0,1,...,L-1]$. The number of pixels

at level i is denoted by n_i . In order to simplify the discussion, the gray-level histogram is normalized by the total number of image pixels.

Suppose that N thresholds, k_j , $j = 1,...,N$, have been selected and they were used to threshold the image into $N+1$ classes, $C_1, C_2, ..., C_{N+1}$. Thus, class C_1 consists of pixels with intensity values in the range $[0, k_1]$, class C_2 consists of pixels with intensity values in the range $[k_1, k_2 - 1]$, etc. Using this thresholding, the probability, P_j , that a pixel is assigned to class C_j is given by the cumulative sum,

$$P_j = \sum_{i \in C_j} p_i \,, \tag{1}$$

The mean intensity values of the pixels assigned to class C_j is given by

$$\mu_j = \frac{1}{P_j} \sum_{i \in C_j} \quad , \tag{2}$$

while the global mean is given by

$$\mu_T = \sum_{i=0}^{L-1} i p_i . \tag{3}$$

Using discriminant analysis, Otsu defined the between-class variance of the thresholded image as

$$\sigma_B^2 = \sum_{j=1}^{N} P_j \left(\mu_j - \mu_T \right)^2 . \tag{4}$$

The optimal thresholds $\{k_1^*, k_2^*, ..., k_N^*\}$ are chosen by maximizing σ_B^2 as follows:

$$\begin{aligned} &\{k_1^*, k_2^*, ..., k_N^*\} = \text{Arg Max } \{\sigma_B^2(k_1, k_2, ..., k_N)\}, \\ &0 \le k_1 < ... < k_N < L. \end{aligned} \tag{5}$$

Fig. 2 shows typical mammogram and the output of each step during preprocessing and segmentation.

In order to solve the dependency problem of Zernike moments on translation, ROIs obtained after segmentation are centered by translation of the centroid of the mass into the center of the image ROI.

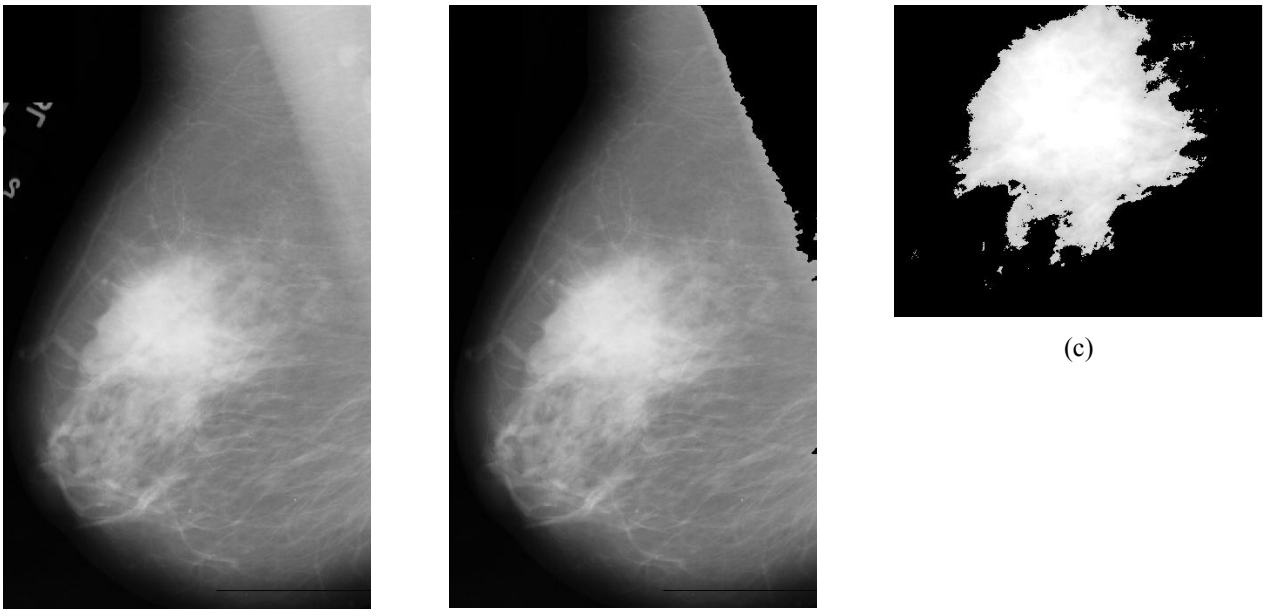
C. *Feature extraction*

Zernike moments represent a set of descriptors obtained using complex kernel functions based on Zernike polynomials. The main characteristic of these moments is their ability to describe a shape with minimum information redundancy (due to their orthogonality property), the robustness in noisy environments, and the fact that these moments are invariant to an arbitrary rotation of the describing shape.

The computation of Zernike moments from an input image includes three steps: computation of radial polynomials, computation of Zernike basis functions and computation of Zernike moments by projecting the image onto the Zernike basis functions [20 – 23]. For a digital image, the Zernike moment of order n , with repetition m is given by

$$Z_{nm} = \frac{n+1}{\pi} \sum_x \sum_y f(x,y) [V_{nm}(x,y)]^* , \tag{6}$$

where $x^2 + y^2 \le 1$. To compute Zernike moments of an image, the center of the image is taken as the origin and the pixel coordinates must be mapped to the range of the unit circle. The part of the original image inside the unit circle can be reconstructed with an arbitrary precision



(a) (b) (c)

Figure 2. Different segmentation steps. (a) Original Image. (b) Image after marking foreground. (c) Obtained ROI.

using

$$f(x,y) = \lim_{N\to\infty} \sum_{n=0}^{N} \sum_{m} A_{nm} V_{nm}(x,y), \quad (7)$$

where the second sum is taken over all $|m| \le n$, such that $n-|m|$ is even. $V_{nm}(x,y)$ is 2D Zernike basis function

$$V_{nm}(x,y) = V_{nm}(r,\theta) = R_{nm}(\rho) e^{jm\theta}, \quad (8)$$

where $\rho = \sqrt{x^2+y^2}$ is the length of the vector from the origin to a pixel (x,y), $\theta = \arctan(y/x)$ is the angle between the *x*-axis and the vector ρ in the counter-clockwise direction, and $R_{nm}(\rho)$ is a radial polynomial defined as follows

$$R_{nm}(\rho) = \sum_{s=0}^{\frac{n-|m|}{2}} (-1)^s \frac{(n-s)!}{s!\left(\frac{n+|m|}{2}-s\right)!\left(\frac{n-|m|}{2}-s\right)!} \rho^{n-2s} \quad (9)$$

High order Zernike moments have a large computational complexity and represent high sensitivity to noise. But with careful selection, these moments can be better descriptors of shape and density characteristics than the low-order Zernike moments. As in [23], a group of 32 low-order moments which satisfy the following conditions have been used:

$$Features = \left\{ Z_{n,m} \right\} \forall \begin{cases} 3 \le n \le 10 \\ |m| \le n \\ n-|m| = 2k \\ k \in \mathrm{N} \end{cases} \quad (10)$$

D. Classification

Neural networks have been successfully applied in many pattern recognition and machine learning systems. According to literature, the backpropagation algorithm is the most popular method for supervised training of multilayer neural networks. It employs gradient descent to attempt to minimize the squared error between the network output values and the target values for these outputs.

In the backpropagation neural networks, during training, the error must be propagated from the output layer back to hidden layer in order to perform the learning of the input-to-hidden weights. Multilayer networks typically use sigmoid transfer functions in the hidden layers. Sigmoid functions are characterized by the fact that their slope must approach zero as the input gets large. This causes a problem when using steepest descent to train a multilayer network with sigmoid functions, since the gradient can have a very small magnitude, and therefore cause small changes in the weights and biases, even though the weights and biases are far from their optimal values. Therefore, it is desirable to use resilient backpropagation training algorithm to eliminate these negative effects of the magnitudes of the partial derivatives. Only the sign of the derivative is used to determine the direction of the weight update, while the size of the weight change is determined by a separate update parameter. The weight update follows a simple rule: if the derivative is positive, which means an error increasing, the weight value is decreased by its update value, and if the derivative is negative, the weight value is increased by its update value.

The resilient backpropagation algorithm is generally much faster than the standard steepest descent algorithm and it requires only a modest increase in memory consumptions.

In this research, the input images were split in three sets intended for training, validation and testing of the neural network. The number of hidden layers and their nodes are being changed until the best network topology, which yields the best accuracy, is found. The final neural network consists of one input, one hidden and one output layer. The input and hidden layers have 32 nodes which match to the number of utilized features. The sigmoid function has been utilized as the activation function of all internal nodes. The output layer has one node which uses linear activation function.

III. EXPERIMENTS AND RESULTS

In this paper, the digital mammograms used for experimentation were taken from the Mammography Image Analysis Society (MIAS) database. It contains 322 digital mammograms belonging to the left and right breast of 161 different women. The acquired mammogram images are classified into three major classes: malignant, benign and normal, all of which are subdivided into five categories as follows: circumscribed masses, speculated masses, ill-defined masses, architecturally distorted masses and asymmetrical masses. From the viewpoint of breast tissue, these images have been categorized into three different classes: fatty, fatty-glandular, and dense-glandular. The images are digitized at 200 micron/pixel and padded in order to obtain all images with a size of 1024 × 1024 pixels. The digital mammograms are gray-scale images with an 8 bit pixel depth.

Through inspection of the segmented masses, it was noticed that the best result is for round-shape masses that have well-defined edges and they are segmented almost precisely. For non-circular cases, most false detections occurred for calcifications which consist of very small spots in mammograms, and speculated masses which are not segmented quite well.

Receiver operating characteristic (ROC) analysis was conducted to evaluate the classification system performance. The following terms, beside the area under ROC curve, were used as benchmark functions of proposed system:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (11)$$

$$Sensitivity = TPR = \frac{TP}{TP+FN} \quad (12)$$

$$Specificity = \frac{TN}{TN+FP} \quad (13)$$

where FP, FN, TP and TN denote false-positive, false-negative, true-positive and true-negative answers, respectively. Fig. 3 illustrates the ROCs of the system, separately for each type of masses.

The system yielded a malignant mass versus benign mass classification accuracy represented as an area (A_z) under ROC curve of 0.8920. The proposed approach tested on the MIAS database, resulted in 90.4 % success rate for classification malignant from benign masses using resilient backpropagation neural network. The sensitivity was 0.9216 and specificity was 0.9608.

It is a difficult task to critically compare the presented approach with various methods present in the literature due to several reasons. For example, the mammogram database used for the evaluation of the methods is different, the sample mammograms used for the experiments are different or a different experimental setup is used [8]. Table 1 shows a performance comparison of our method with other techniques on the basis of sensitivity, specificity and accuracy. The quantities that are not reported in the corresponding research works are indicated with dash symbol. The comparisons with the other classification systems published in the literature have confirmed the potential of the proposed method for further investigation.

IV. CONLUSION

In this paper, a novel computer-aided diagnosis system has been presented for mass detection and classification in breast mammography images. The proposed system is a fully autonomous. Input ROIs are obtained from the digital mammograms using image preprocessing and Otsu's N thresholding. After the first stage, a group of 32 Zernike moments with different orders and iterations are extracted from the ROIs. These moments have been applied to the resilient backpropagation neural network classifier.

Experimental results show that the proposed technique is able to classify efficiently malignant and benign mass types of abnormalities in digital mammograms.

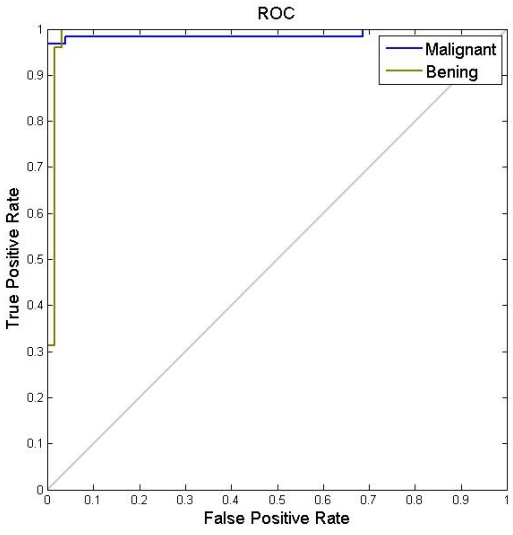


Figure 3. ROCs of the system

TABLE I. COMPARISON WITH STATE-OF-THE-ART METHODS BASED ON SENSITIVITY, SPECIFICITY AND ACCURACY

Author's name	Sensitivity	Specificity	Accuracy (%)
Menut et al. [4]	-	---	76.0
Hussain [9]	-	---	99.0
Gorgel et al. [16]	0.9470	0.7140	84.8
Islam et al. [15]	0.9091	0.8371	---
Proposed method	0.9216	0.9608	90.4

Nevertheless, there is the need to perform tests with a larger database and more complex cases in order to obtain a more precise behavior pattern.

ACKNOWLEDGMENT

This work has been supported by the SCOPES program of the Swiss National Science Foundation (http://www.snf.ch)

REFERENCES

[1] K.H. Miao and G.J. Miao, "Mammographic Diagnosis for Breast Cancer Biopsy Predictions Using Neural Network Classification Model and Receiver Operating Characteristic (ROC) Curve Evaluation", Journal of Selected Area in Bioinformatics (JBIO), vol. 3, issue 9, 2013

[2] W. Polakowski, D. Cournoyer, S. Rogers, M. DeSimio, and D. Ruck, "Computer-aided breast cancer detection and diagnosis of masses using difference of Gaussians and derivative-based feature saliency," IEEE Transactions on Medical Imaging, vol. 16, no. 6, pp. 811 – 819, 1997

[3] D. Brzakovic, X. Luo, and P. Brzakovic, "An approach to automated detection of tumors in mammograms," IEEE Transactions on Medical Imaging, vol. 9, no. 3, pp. 233 – 241, 1990

[4] R. Rangayyan, N. El-Faramawy, and J. Desautels, O. Alim, "Measures of acutance and shape for classification of breast tumors," IEEE Transactions on Medical Imaging, vol. 16, no. 6, pp. 799 –810, 1997

[5] N. El-Faramawy, R. Rangayyan, J. Desautels, and O. Alim, "Shape factors for analysis of breast tumors in mammograms," in Proceedings Canadian Conference on Electrical and Computer Engineering, Calgray, AB, Canada, pp. 355 – 358, 1996

[6] O. Menut, R. Rangayyan, and J. Desautels, "Classification of breast tumors via parabolic modeling of their contours," in Proceedings 1997 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing, Victoria, BC, Canada, pp. 1002 – 1005, 1997

[7] T. Matsubara, H. Fujita, S. Kasai, M. Goto, Y. Tani, T. Hara, and T. Endo, "Development of new schemes for detection and analysis of mammographic masses," in Proceedings Intelligent Information Systems. IIS'97, Grand Bahamas Island, Bahamas, pp. 63-66, 1997

[8] M. Hussain, S. Khan, G. Muhammad and G. Bebis, "A Comparison of different gabor features for mass classification in mammography". Proceedings of the 8th International Conference on Signal Image Technology and Internet Based Systems, IEEE Xplore Press, Naples, pp: 142-148, 2012

[9] M. Hussain, "False Positive Reduction using Gabor Feature Subset Selection", 2013 International Conference on Information Science and Application (ICISA), 2013

[10] G. Vani, R. Savitha and N. Sundararajan, "Classification of abnormalities in digitized mammograms using extreme learning machine". Proceedings of the 11th International Conference on Control Automation Robotics and Vision, IEEE Xplore Press, Singapore, pp: 2114-2117, 2010

[11] J.S.L. Jasmine, S. Baskaran and A. Govardhan, “An automated Mass classification system in digital mammograms using contourlet transform and support vector machine”. Int. J. Comp. Appl., 31: 54-54, 2011

[12] M. Fraschini, “Mammographic masses classification: Novel and simple signal analysis method”. Proceedings of the Electronics Letters, IEEE Xplore Press, pp: 14-15, 2011

[13] F. Saki, A. Tahmasbi and S.B. Shokouhi, “A novel opposition based classifier for mass diagnosis in mammography images”. Proceedings of the 17th Iranian Conference of Biomedical Engineering, IEEE Xplore Press, Isfahan, pp: 1-4, 2010

[14] L.D.O. Martins, G.B. Junior, A.C. Silva, A.C.D. Paiva and M. Gattass, “Detection of masses in digital mammograms using k-means and support vector machine”. Elect. Lett. Comp. Vis. Image Anal., 2009

[15] M.J. Islam, M. Ahmad and M.A. Sid-Ahmed, “An efficient automatic mass classification method in digitized mammograms using artificial neural network”. Int. J. Artif. Intell. Applic., 1: 6-6, 2010

[16] P. Gorgel, A. Sertbas, N. Kilic, O.N. Ucan and O. Osman, “Mammographic mass classification using wavelet based support vector machine”. J. Elect. Elect. Eng., 9: 867-875, 2009

[17] Zuiderveld, Karel. "Contrast Limited Adaptive Histograph Equalization." *Graphic Gems IV*. San Diego: Academic Press Professional, 474–485, 1994

[18] N. Otsu, “A threshold selection method from gray-level histogram,” IEEE Transactions on System Man Cybernetics, Vol. SMC-9, No. 1, pp. 62-66. 1979

[19] S. Deepa, V. SubbiahBharathi, Efficient ROI segmentation of Digital Mammogram images using Otsu’s N thresholding method, International Journal of Engineering Research & Technology, vol. 2 - issue 1, 2013

[20] W. Wang, J. E. Mottershead, and C. Mares, “Mode-shape recognition and finite element model updating using the zernike moment. descriptor,” J. Mechanical Systems and Signal Processing, vol. 23, pp. 2088–2112, 2009

[21] S. K. Hwang, W. Y. Kim, “A novel approach to the fast computation of zernike moments,” J. Pattern Recognition, vol 39, pp. 2065–2076, 2006

[22] A. Tahmasbi, F. Saki, S. B. Shokouhi, “An Effective Breast Mass Diagnosis System using Zernike Moments,” in Proc. IEEE, 17th Iranian Conf. on Biomedical Engineering (ICBME’2010), Isfahan, Iran, pp. 1-4, 2010

[23] A. Tahmasbi, F. Saki, S. B. Shokouhi, “Classification of benign and malignant masses based on Zernike moments,” J. Computers in Biology and Medicine, vol. 41, no. 8, pp. 726-735, 2011

CHO cell viability in CO_2 saturated media

A. Štefanič*, J. Jelenc**, A. Osterman***, M. Dular***, D. Miklavčič* and A. Maček Lebar*

* University of Ljubljana, Faculty of Electrical Engineering, Ljubljana, Slovenia

** Iskra Medical d.o.o., Ljubljana, Slovenia

*** University of Ljubljana, Faculty of Mechanical Engineering, Ljubljana, Slovenia

jure@iskramedical.eu

Abstract – ***Thermal food preservation is a standard technique for reducing the microbial levels in food. The use of high pressure carbon dioxide (HPCD) has been proposed as alternative "cold pasteurization". It is known that excess CO_2 exerts inhibitory effects on cell growth and even cell death. For cell disruption and consequently cell death we can also use a method of a rapid decompression from a pressure vessel, known as a "cell bomb". In this study we have evaluated the effects of CO_2 on CHO cell growth and effects of a rapid decompression from relatively low pressure, 500 kPa. A rapid pressure decompression had no effect on cell growth. When CHO cells were exposed to the medium with dissolved CO_2 we noticed a small decrease in cell viability in exposures longer than 45 minutes.***

Key words: pasteurization, cell viability, CO_2

I. INTRODUCTION

Thermal food preservation is an old but almost standard technique for reducing the microbial count of foods. However, it has undesirable effects on organoleptic and nutritious quality of food, especially on its vitamin and mineral content.

Therefore nonthermal processing alternatives have been proposed, like electroporation and high hydrostatic pressure. Irreversible electroporation occurs when high electric field applied results in leakage of cellular components, which leads to cell death [2]. Preservation of food by electroporation maintains color and flavor, and the antioxidant levels are unaffected. During the last few years microbial deactivation using electric pulses has been proved in a variety of foods and beverages [3]. The effectiveness of the procedure depends on several parameters, e.g., pulse amplitude, duration, frequency and polarity, and temperature [4], that still have to be optimized. Although high hydrostatic pressure method maintain the quality of food as well as electroporation, it also has some limitations (the occurrence of pressure resistant vegetative bacteria after successive pressure treatments, the large investment costs, undefined regulatory and product safety related issues) that obstructing widespread use of the technology [1].

Interesting method for cell disruption using high hydrostatic pressure is also rapid decompression or the "cell bomb" method. In this process, cells are placed under high pressure (usually nitrogen or other inert gas up to about 170 MPa) within appropriate pressure chamber. The pressure is rapidly released. The rapid pressure drop causes the dissolved gas to be released as bubbles that ultimately lyse the cell [3].

Relatively new non-thermal pasteurization technique is the use of high-pressure carbon dioxide. In this process, food is contacted with either sub-or supercritical CO_2 for a certain amount of time in the series or continuous manner. Under such conditions (Tc = 31,1°C in Pc = 7,38 MPa) CO_2 has the unique ability to diffuse through solids like a gas, and dissolve materials like a liquid. Additionally, it can readily change in density upon minor changes in temperature or pressure. Using this technique gas (CO_2) pressure applied for preservation purposes is much lower (generally < 20 MPa) as compared to the high hydrostatic pressure method. Although a number of theories have been proposed to explain the bacteriostatic action of pressurized CO_2, the exact inactivation mechanisms are still under consideration [1].

Due to promising effects of CO_2 alone, it is also a good candidate for combined treatments, particularly in combination with ultrasound where it can produce additional cavitation seeds. For evaluation of combined treatment effects, it is important to be familiar with the effects of individual therapies. For this reason, we focused in the present study on the effects of CO_2 saturated medium on CHO cells. We evaluated the concentration of CO_2 in CO_2 saturated cell media, exposed CHO cells in CO_2 saturated medium for various time intervals and followed their 24h viability. We also tested »cell bomb« that was constructed using relatively low CO_2 pressure, i.e. 500 kPa.

II. METHODS

A. The cells

Ham's nutrient mixtures (F-12 HAM), trypsin/EDTA, fetal bovine serum (FBS), L-glutamine, crystal violet, phosphate buffered saline (PBS), antibiotics crystacillin and gentamicin were purchased from PAA Laboratories (PAA Laboratories, Pasching, Austria). Chinese hamster ovary cells CHO (European Collection of Cell Cultures, Salisbury, UK) were cultured in the humidified atmosphere at 37°C and 5% CO_2 in an incubator (Kambič, Semič, Slovenia). Cell growth media were supplemented with 10% fetal bovine serum (FBS), antibiotics (gentamicin, crystacillin) and L-glutamine. Cell growth medium was prepared at least 24 hours before experiments.

After adherent cell line reached 80-90% confluence in 150 cm^2 culture flask (TPP, Trasadingen, Switzerland), cells were carefully rinsed with 10 ml of Phosphate Buffered Saline (PBS) and detached using 10 ml of 0.25% trypsin/EDTA. Trypsin was present in the flask for 30 seconds. After trypsin solution was removed, cells were harvested and centrifuged (1000 rpm, 5 min, 4°C), and then re-suspended at concentration of $4*10^6$ cells/ml. Before each experiment cell suspension was carefully mixed by repeated pipetting until homogeneous suspension was obtained.

B. Preparation of a liquid saturated with CO_2

Pressurized carbon dioxide (CO_2) gas at 2 MPa was purchased from Messer Slovenia (Messer, Ljubljana, Slovenia). Pressure regulator valve kept output pressure in tubing at 500 kPa. Tubing was connected to a 0.5 liter plastic bottle, with a drilled hole on the "stopper", where the tubing was inserted into the bottle. A 75 ml of growth medium or water was added into the bottle. CO_2 gas was introduced into the liquid by increasing CO_2 partial pressure of air above liquid using compressed gas at 500 kPa for 300 s. The growth medium was slowly depressurized to ambient pressure for 10 seconds, and was used in experiments immediately after.

A CO_2 dissolved in water forms carbonic acid. Therefore after CO_2 pressurization of cell media a drop of pH is present that is exhibited as a color transition of pH indicator Phenol Red. A pH decrease was measured using SG2 pH meter (Mettler Toledo, Greifensee, Switzerland).

C. Concentration of CO_2 in the liquid

The concentration of dissolved gas in the liquid was determined by measuring system (Figure 1) adopted the modified Van Slyke manometric method. The method is not gas selective and it is easy to use.

The measuring system consists of a cylindrical container in which there is airtight rubber bladder, which is connected with the exterior through the tube. Cylindrical vessel concludes with a narrower dipstick, which closes above with the sealing screw. Before use rubber bladder is inflated and sealed. Then through the dipstick the cylindrical container is filed with a known volume of liquid and then the system is hermetically sealed. The pressure of the bladder is balanced with air pressure and the initial fluid level on the dipstick is marked. The air is sucked from the bladder with the vacuum pump; therefore the pressure inside the measuring system is reduced. Reduced partial gas pressure results in the release of gases from liquid. Gases are removed from the liquid until they reach a balance defined by Henry's Law, between the partial pressure of the gases above the liquid and the concentration of gases dissolved in the liquid. The speed of removing gas from the liquid is accelerated by shaking the device. After switching off the pump and equalizing the pressure in the bladder to the air pressure, the extracted gases accumulate on the top of the measuring tube. Then the volume of liquid displaced with the gases is measured. Experience has shown that the above procedure should be repeated three to four times. By knowing the initial volume of liquid and assuming the gas density a concentration of dissolved gases in the liquid can be calculated.

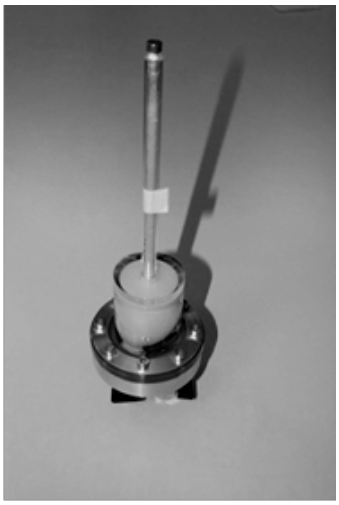


Figure 1: A measuring system for measuring the concentration of dissolves gases by the modified method of Van Slyke.

The concentration of CO_2 gas, which is dissolved in the medium F-12 HAM, after 300 second exposure to the bottle at a positive pressure of 500 kPa can be calculated according to Henry's Law:

$$p = k_H * c, \quad (1)$$

where p is the gas pressure above the solution, c is the concentration of gas and k_H is the Henry's constant, which depends on the solute, the solvent and the temperature (T).

Temperature dependence is described by the equation

$$k_H(T) = k_H(T^0)e^{[-C\left(\frac{1}{T}-\frac{1}{T^0}\right)]}, \quad (2)$$

where $k_H(T^0)$is the Henry's constant at room temperature (25 °C) and at 29.41 atm/mole/L, T_0 is room temperature (25 °C) and c is a constant for the CO_2 that is 2400 K [4].

D. CO_2 Cell bomb

Plastic bottle (Kokta, Droga Kolinska, Slovenia) filled with 20 ml of the cell suspension with a density of $4 * 10^4$ cells/ml was exposed to CO_2 overpressure of 500 kPa for 300 seconds. The bottle was connected through the control valve, which reduced the output pressure to 500 kPa, to the cylinder with compressed CO_2 with overpressure of 2 MPa. After the 300 second exposure a decompression is performed in two ways: by the slow opening of the bottle (15 seconds) and by the rapid opening of the bottle (< 1 second).

E. Cells in CO_2 saturated media

The PCR microtube (Sorenson BioDcience, Salt Lake City, UT, USA) was poured with 180 µl of cell culture medium or cell culture medium with added CO_2 gas. Then 60 µl of cell suspension with a concentration of $4 * 10^6$ cells/ml is added to the cup, so the final concentration of the cells in 240 µl of cell suspension was 10^6 cells/ml. Cells were exposed to a medium with added CO_2 up to 60 minutes. After the exposure the 80 µl of the

sample was planted on the plate with 96 wells and 20 ml of cell medium F-12 was added to the sample. Cells were incubated for 22 hours and then checked for their survival.

Control exposure was done in normal cell medium and served as a reference for normalization of the results.

Exposure of the cells to a medium with added CO_2 was carried out in four independently repeated experiments.

F. 24 hour cell viability

Cell viability was tested by means of CellTiter 96 AQueous One Solution Assay (MTS). After 22 hour incubation at 37°C and 5% CO_2 in an incubator (Kambič, Semič, Slovenia) 20µl of MTS reagent was added to each well. The cell plate was incubated for another 3 hours in an incubator. The absorbance was measured using micro plate reader Infinite M200 (Tecan, Männedorf, Switzerland). Cell viability was defined as:

$$x = \frac{OD_v - OD_{oz}}{OD_k - OD_{oz}} * 100\% \quad (3)$$

where x is percentage of viable cells, OD_v absorption value of treated cells, OD_k absorption value of control cells (100 % viability) and OD_{oz} absorption value of HAM F-12 media.

All results are reported as a mean value of 4 experiments. The spread of the data is given by standard deviation. The difference between the experimental and control group was analyzed using SigmaStat 11 software (Systat Software, Chicago, IL, USA) and determined with the use of t-test. The statistically significant difference was indicated by $p < 0.05$.

III. RESULTS

A. Concentration of CO_2 in distilled water and in CO_2 saturated HAM media

Using the system for measuring the concentration of dissolved gas in the liquid by the modified Van Slyke method the concentration of dissolved gases in the tap water, deionized water and in the cell medium HAM F-12 was measured. The results are presented as an average value of three independent experiments (Figure 2).

The average concentration of dissolved gases in tap water is 30.3 mg/l, in deionized water is 18.5 mg/l and in the cell medium HAM F-12 is 26.7 mg/l. The average density of the air at room temperature, $\rho = 1.184$ mg/l, was used as the density of the gas in the calculation of a concentration of dissolved gases in the liquid.

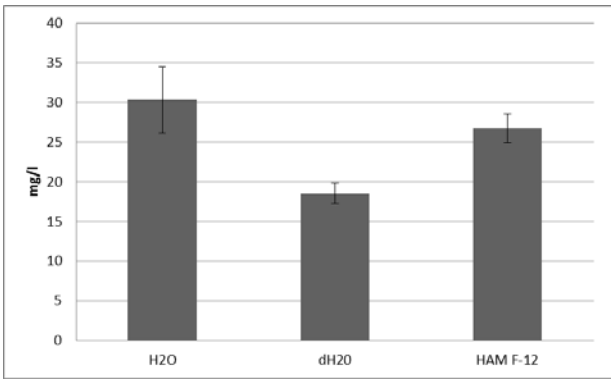


Figure 2: The concentration of dissolved air in tap water (H2O), deionized water (dH2O) and cell medium (HAM F-12) in miligrams per liter.

Cell culture medium was exposed to increased CO_2 partial pressure for 300 seconds. For comparison the same procedure was applied to deionized water as well. In these two cases we measured the volume of liquid that was displaced with the air and CO_2 gas. Therefore this volume was reduced with the volume of liquid that was displaced with the air alone. An average density of CO_2 at room temperature, $\rho = 1.842$ mg/l [8], was used as the density of the gas in the calculation of a concentration of dissolved CO_2 in the liquid. The results are presented as an average value of four or five independent experiments (Figure 3).

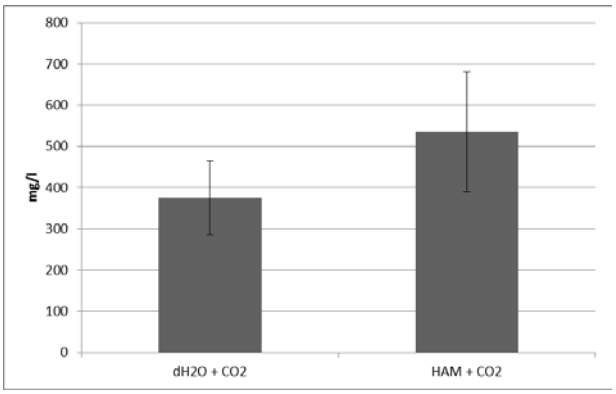


Figure 3: The concentratin of dissolved CO_2 gas in deionized water (dH2O + CO2) and in cell medium (HAM + CO2) in miligrams per liter.

The average concentration of dissolved CO_2 in deionized water is 375.2 mg/l while in the cell medium HAM F-12 is 535.1 mg/l.

B. pH of CO_2 saturated HAM media

After the cell medium was saturated with CO_2 it was closed in the PCR tube. The pH dropped from 7.6 to 5.85. After 30 minutes it stabilized at 5.93 (Figure 4).

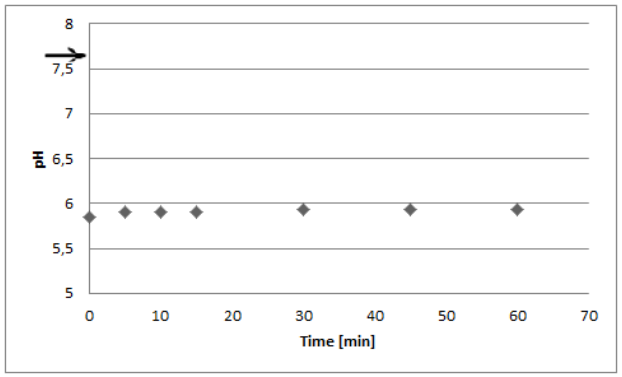


Figure 4: pH of the cell medium after saturation with CO_2. The arrow indicates the reference value of pH of the medium, i.e. 7.6.

C. *Cell viability in CO_2 saturated medium*

The cells in the PCR tube were exposed to CO_2 saturated medium. The PCR tube was closed during the exposure, therefore the reduced level of pH was maintained, as shown in figure 4. Cell viability test showed a reduction of viable cells after 60 minutes of exposure to CO_2 saturated medium (Figure 5).

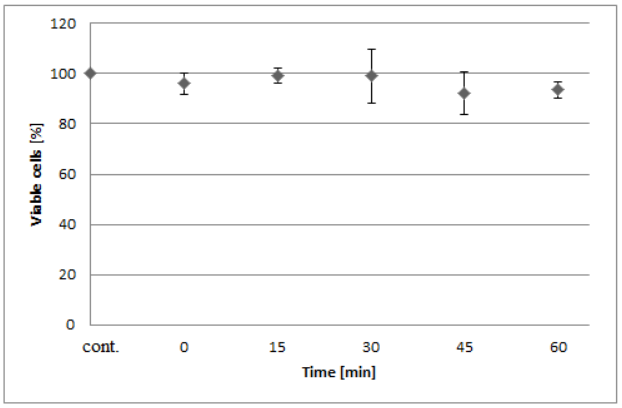


Figure 5: The percentage of viable cells exposed to CO_2 saturated medium. The cell viability was normalized to the control.

D. *CO_2 Cell bomb*

Immediately after decompression the cells were observed under a microscope (Zeiss AxioVert 200) at 400x magnification. There were no visible changes of cell morphology (Figure 6).

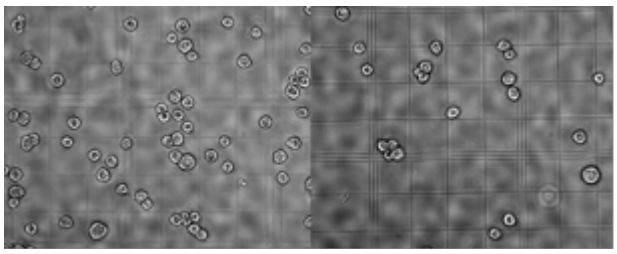


Figure 6: Cells after fast (left) and slow (right) decompression. There is no visible damage on cells.

Cell viability test showed no difference between the fast and slow decompression (Figure 7).

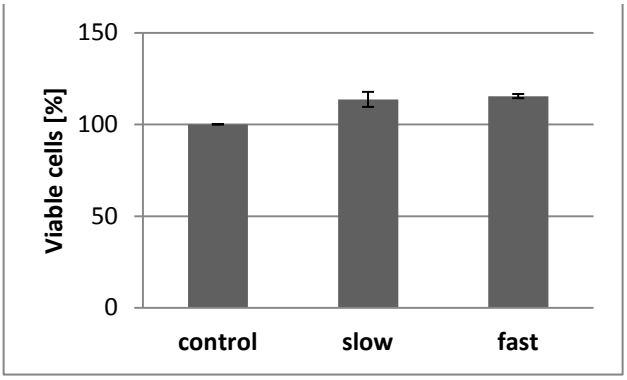


Figure 7: The percentage of viable cells after slow and fast decompression. The cell viability was normalized to the control.

IV. DISCUSSION

The use of CO_2 for preserving food started in 1939 with the study by Brown et al. [9] where apple cider was carbonated and microbial inactivation changes were record. Unlike toxic chemicals, CO_2 has less impact on the environment, leaves no residue on the food and is recyclable. Therefore its effects are still under consideration, especially in combination with some others nonthermal food preservation methods like electroporation and sonoporation.

In the present study, we analyzed the effects of CO_2 saturated medium on CHO cells. CO_2 is soluble in water and under increased partial CO_2 pressure the gas solubility even increases. The average concentration of dissolved CO_2 in the cell medium HAM F-12 after increasing CO_2 partial pressure above the medium to 500 kPa for 300 s was 535.1 mg/l. In comparison, the average concentration of dissolved CO_2 in deionized water under the same conditions was 375.2 mg/l.

CO_2 reacts with H_2O and forms carbonic acid. It causes a pH cell medium decrease from 7.6 to 5.8. We examined the cell survival in CO_2 saturated media over a period of 60 minutes. The results show that 45 minute exposure of cells to the medium with added CO_2 does not affect the cell survival. Longer exposure slightly reduced the proportion of viable cells.

Slow and fast decompression at relatively low pressure of 500 kPa, has no adverse effect on the cells. We assume that a small increase in the number of cells, which was detected in the slow and fast decompression, is the result of a slightly different handling of the cells in the control and the cells during the experiment.

According to presented results it can be concluded that short-term exposure of the CHO cells to cell medium with an increased concentration of CO_2 up to 535.1 mg/l has no adverse effect on CHO cells.

V. CONCLUSION

CO_2 addition is an alternative technique for reducing the microbial levels in food. Additional research is necessary for CO_2 alone or in combination with other agents to achieve levels of microbial deactivation known in thermal food preservation.

ACKNOWLEDGMENT

This work was supported by European Regional Development Fund (Biomedical Engineering Competence Center, Slovenia) and European Social Found.

REFERENCES

[1] L.G. Gonzalez, A.H. Geeraerd, S. Spilimbergo, K. Elst, L. Van Ginneken, J. Debevere, J.F.Van Impe, F. Devlieghere: High pressure carbon dioxide inactivation of microorganisms in foods: The past, the present and the future, International Journal of Food Microbiology 117 (2007) 1 – 28.

[2] Haberl S, Miklavčič D, Serša G, Frey W, Rubinsky B. Cell membrane electroporation – Part 2: The applications. IEEE Electr. Insul. M. 29(1): 29-37, 2013.

[3] R.J. Simpson, "Disruption of Cultured Cells by Nitrogen Cavitation", Cold Spring Harb Protoc, vol. 2010, no. 11, p. pdb. prot 5513, Nov. 2010.

[4] G. Saulis, "Electroporation of cell membranes: The fundamental effects of pulsed electric fields in food processing," Food Eng. Rev., vol. 2, no. 2, pp. 52–73, 2010.

[5] M. Ajtnik: Študija vpliva superkritičnega ogljikovega dioksida na preživetje kvasovk saccharomyces cerevisiae, Univerza v Mariboru, Fakulteta za kemijo in kemijsko tehnologija, 2009.

[6] T. Takahashi, C.S. Sutherland, C. Sweeney, A. Poisson, N. Metzl, B. Tilbrook, N. Bates, R. Wanninkhof: "Global sea–air CO2 flux based on climatological surface ocean pCO2, and seasonal biological and temperature effects". *Deep Sea Research Part II: Topical Studies in Oceanography* 49 (9–10): 1601– 1622 (2002).

[7] http://www.engineeringtoolbox.com/air-density-specific-weight-d_600.html

[8] http://www.engineeringtoolbox.com/gas-density-d_158.html

[9] H.D. Brown, C. Fitzgerald, F. Neuman. Preserving cider by carbonation. Am. Soc. Of Hortic. Sci. 36, 371 - 373, 1939.

Automatic cell counter for cell viability estimation

Jasna Lojk[1], Luka Šajn[2], Uroš Čibej[2], Mojca Pavlin[1]
[1]Faculty of electrical engineering, University of Ljubljana, Ljubljana, Slovenia
[2]Faculty of computer and information science, University of Ljubljana, Ljubljana, Slovenia
email: mojca.pavlin@fe.uni-lj.si

Abstract - Despite several methods that exist in different fields of life sciences, certain biotechnological applications still require microscopic analysis of the samples and in many instances, counting of cells. Some of those are drug delivery, transfection or analysis of mechanism fluorescent probes are used to detect cell viability, efficiency of a specific drug delivery or some other effect. For analysis and quantification of these results it is necessary to either manually or automatically count and analyze microscope images. However, in everyday use many researchers still count cells manually since existing solutions require either some specific knowledge of computer vision and/or manual fine tuning of various parameters.

Here we present a new software solution (named CellCounter) for automatic and semi-automatic cell counting of fluorescent microscopic images. This application is specifically designed for counting fluorescently stained cells. The program enables counting of cell nuclei or cell cytoplasm stained with different fluorescent stained. This simplifies image analysis for several biotechnological applications where fluorescent microscopy is used. We present results and validate the presented automatic cell counting program for cell viability application.

We give empirical results showing the efficiency of the proposed solution by comparing manual counts with the results returned by automated counting. We also show how the results can be further improved by combining manual and automated counts.

I. INTRODUCTION

Cell counting is one of the most fundamental tools in biology, medicine, and other life sciences. Several methods exist for quantification of different phenomena, but substantial part uses fluorescent dyes in combination with various protocols in order to analyze different processes. Fluorescently stained cells can be analyzed using flow cytometry[1], spectrofluorometry[2] and/or fluorescent microscopy[3]. There are of course other existing solutions for counting cells without the use of florescence, like counting with counting chambers (hemocytometer)[4], but those can be time consuming and also require detachment of the cells, which can result in their loss and damage. Fluorescent microscopy is also a standard method for visualization and quantification of different phenomena and many researchers rely on microscopic images and image analysis tools to obtain reliable quantitative data. Cell counting using microscopic images can be automated, and several semi-automatic and fully automatic solutions have been developed. In practice, however, such solutions require some additional input regarding various image parameters, such as thresholds, approximate nuclei distances, contrast settings, etc. These parameters require specific knowledge of image processing, which is not common among life-science experts.

Currently, several programs for cell counting are already available. From many commercial solutions, such as MetaMorph [5] (Molecula Devices, Downingtoetn, PA, USA), BioQuant [6] (Image Analysis Corporation, Nashville, TN, USA), Image-Pro [7] (Media Cybernetics, Bethesda, MD, USA) and SynenTec [8], to free software such as ITCN ImageJ Plugin [9] (US National Institute of Health, Bethesda, MD; USA), CellProfiler [10], UTHSCSA ImageTool (University of Texas Health Science Center, San Antorio, TX, USA) [11] and CellC [12]. However, all of the existing solutions require some additional input or even some image pre-processing of the raw images, which is not only time-consuming but can also be also less user-friendly for life scientists [13].

Besides the human factors, full automatisation of cell counting faces several technical obstacles. Some of the most common problems, are: i) a wide variety of assays and cell types resulting in a range of objects which are hard to generalize, ii) this can be even more problematic due to uneven illumination and other equipment related factors, such as electronic and/or optical noise that lead to images having variable contrast and quality [14], iii) cells can have a varying density in the cell culture and can overlap [15], iv) extracellular debris and internal cell structures can interfere with the recognition algorithm, giving false results and v) depending on the microscopy technique used, images can have low contrast therefore detection of individual nuclei in multinucleated cells is difficult.

In this article we present a novel structured method for effective cell counting which does not require any specific knowledge of image processing. The method is specifically designed for counting fluorescently stained cells, that are used in many applications from assays for measuring cell viability [16] to determination of transfection efficiency [17–19]. We also show that the presented solution has been already intensively implemented for cell counting of different sets of experiments and it has drastically improved the typical workflow of researchers, as it is shown in our empirical evaluation.

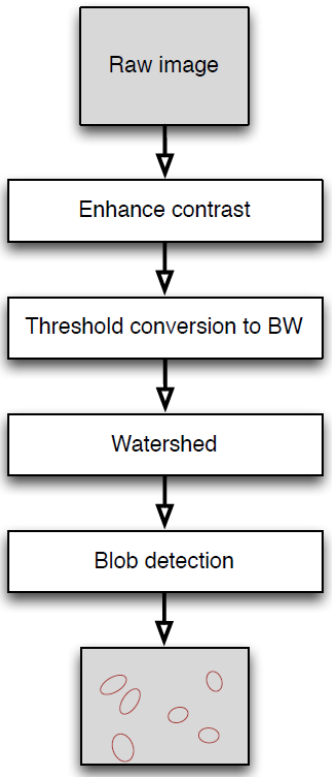


Figure 1. Schematic overview of the cell-counting method

II. MATERIALS AND METHODS

A. *Algorithm used*

The presented program was written in C\# and uses an open source C\# framework designed for developers and researchers in the fields of Computer Vision and Artificial Intelligence (AForge.NET [20]). The algorithm used in the program can be viewed as a simple dataflow of images passing through various image processing algorithms. The sequence of these algorithms was determined empirically by trying to overcome as many problems on a sample set of images as possible. The schematic overview of our method is shown in Fig. 1 and will be described in detail below.

The algorithm first addressed the problem of the image contrast, which is can be really low on microscopic images and not necessarily distributed evenly throughout the image due to uneven illumination or other optical problems. The contrast of the image is thus improved using the Contrast Limited Adaptive Histogram Equalization (CLAHE) algorithm [21–23]. It differs from ordinary histogram equalization in the respect that the adaptive method computes several histograms, each corresponding to a distinct section of the image, and uses them to redistribute the brightness values of the image. It is therefore suitable for improving the local contrast of an image and bringing out more detail. In our solution, the image is cut into 64 (8 × 8) disjoint sections and histogram equalization is applied on each part independently. Within each image part, local maximum and minimum pixel intensity are calculated and the intensities are proportionally adjusted to the scale ranging from 0 to 255. The individual parts are then recomposed again using interpolation at image borders. The final image has better contrast and can reveal more local information (Fig. 2).

Next step is converting grey-scale images to black-white images where the object borders are defined using a threshold algorithm. In this step, Otsu Threshold algorithm [24], [25] was used, as it gave slightly better results compared to other tested algorithms (Huang, Renyi entropy method, and other). In this algorithm, the threshold value t is defined as the value that gives the minimal intra-class variance defined by:

$$\omega(t)\,\sigma_1^2(t) + \omega_2(t)\sigma_2^2(t)$$

where ω_1 (ω_2) is the probability that the pixel value in the image is lower (higher) than t and σ_1^2 (σ_2^2) is the variance of the pixel values lower (higher) than t.

In cell cultures, cells tend to grow in close proximity, touch or even overlap, resulting in connected objects on the images. To split the overlapping (touching) objects, the Watershed algorithm [26] was applied. The image is eroded on the watershed lines, which usually represent exactly the borders between different cells. This enables the detection of individual cells, even multinucleated and overlapping cells (Fig. 2). Next, the connected regions are counted using classical blob detection algorithm [27], which counts the objects separated by black background.

Although the counting algorithm is robust, as we will show in the following sections, the miscounts can still occur on more problematic images. To correct erroneously counted objects, a function was added to the user interface that enables the user to add or remove objects from the count with a mouse click. This will be

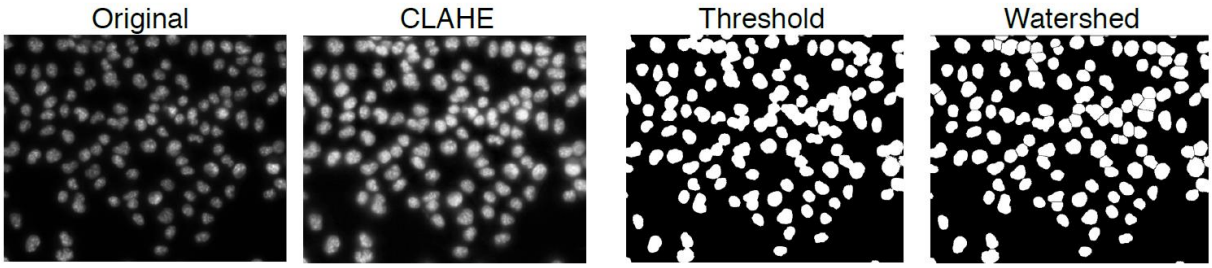


Figure 2. A sequence of images produced by the program on a sample image.

referred as a semi-automatic counting henceforth.

B. Cell culturing

The experiments used for program validation were done on Chinese hamster ovary cells (CHO) grown in Ham's tissue culture medium for mammalian cells (HAM) supplemented with 10% fetal bovine serum (FBS) at 37°C in 5% CO_2-enriched air at saturation humidity. All experiments were performed on 24 h old cell cultures in exponential growth phase.

C. Nanoparticles and cell viability

The cell viability experiment was performed as described previously [12]. Briefly, cells were incubated with increasing concentration of polycationic polymer coated magnetic nanoparticles for 24 h and stained with two fluorescent dyes; Hoechst 33342, which stained all cell nuclei, and propidium iodide (PI), which differentially stained only dead cells. At least 15 visual fields at 200 × magnification were taken of each sample for each used fluorescent dye using a fluorescent microscope (Zeiss 200, Axiovert, Germany). The images were recorded by MetaMorph imaging system software (Visitron, Germany) and saved in TIF format.

D. Cell counting

The algorithm was developed and optimized based on microscopic images obtained from several different types of experiments and experimental repeats, but for evaluation of the program efficiency, one random experiment was selected.

Cells on the images of the selected experiment were counted using only CellCounter (automatic), by manually correcting CellCounter's results (semi-automatic) and by manually counting using ImageJ software (manual) (v1.45s, National Institute of Health, USA). Three people counted the experiment independently to determine inter-personal error. Counting time was noted for each counting step. Results are expressed as mean ± standard error (SE).

III. Results and Discussion

We have developed an algorithm for automatic cell counting of fluorescently stained cells. As an example to evaluate the program, cells were stained with blue fluorescent dye Hoechst and with red fluorescent dye PI. The program can also be used to count cells or objects labeled with other fluorescent markers, like green fluorescent protein (GFP) [13], [14].

The algorithm of our program works by improving the contrast of bright objects on the images (in our case fluorescently stained nuclei and/or cytoplasm) and converts the obtained grey-scale images to black-white images on which the automatic counting is applied. As such, the program does not require images with objects that have only clear borders and was found quite flexible/robust in determining accurate numbers of cells even if the quality of images is not ideal. As seen from Fig. 3, the program is able to count cells on out of focus images with relatively small errors. This allows less accurate and thus faster acquisition of images for counting. Also, because the algorithm does not rely on size or shape of the objects, the program can be applied to fluorescent images of any appropriately stained cell line, as already confirmed for several different cell types (results not shown).

Cells of the same experiment were stained with Hoechst, which stains all cell nuclei, thus giving us the total number of the cells (Fig. 4), and with PI, which only stains dead or damaged cells (Fig. 5, Fig. 6). This combination of dyes can be used to determine cell viability [12]. To obtain the number of viable cells, which was the purpose of the experiment, the number of dead cells was subtracted from the number of all cells (Fig. 7).

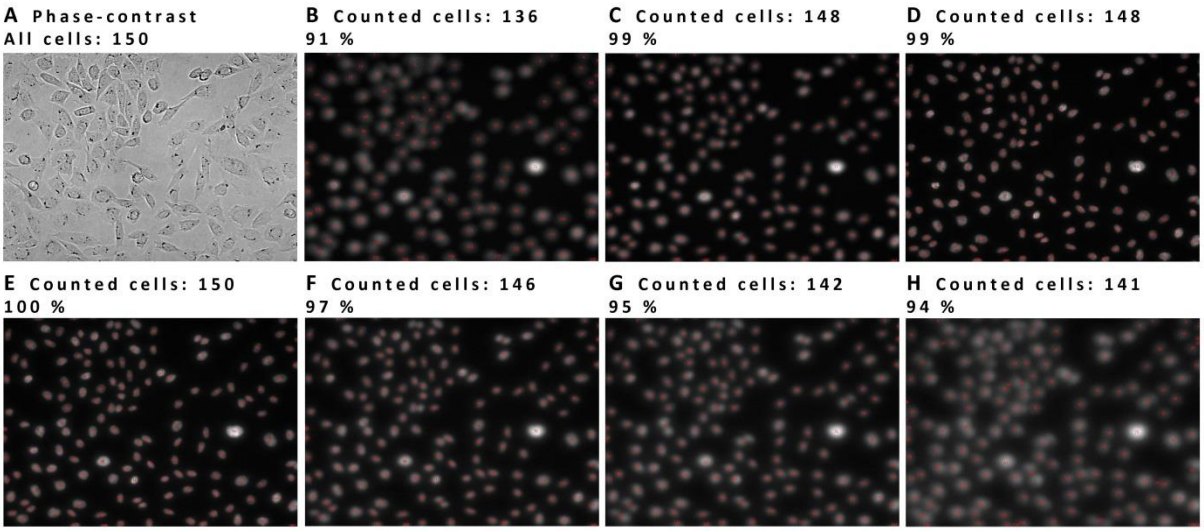


Figure 3. A demonstration of the robustness of the program. The figures are images of Hoechst stained nuclei on the same visual field obtained on different out of focus planes, under (B, C, D) or above (F, G, H) the plane in focus (E). The total number of cells was obtained by manually counting the cells on phase contrast image. Above each image are given the total number of automatically counted cells and the percentage of the cells compared to manual count. The fluorescent image in focus (E) gave a 100% counting accuracy.

All images were counted by three users to obtain the inter-personal error and the differences in obtained counting results for all three applied counting methods; manual, semi-automatic and automatic counting.

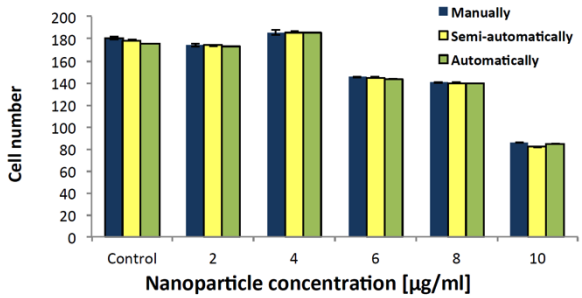


Figure 4. Comparison of counting results obtained by counting fluorescently stained cell nuclei to determine the total number of CHO cells after exposure to increasing concentration of nanoparticles for 24 h. Three counting methods were used: manually by counting Hoechst positive cells using ImageJ software, semi-automatically by correcting CellCounter counts and automatically by using CellCounter only. The results presented the average percentage of viability for separate samples compared to non-exposed control obtained by independent counting of the same experiment by three users with corresponding standard error (N=3).

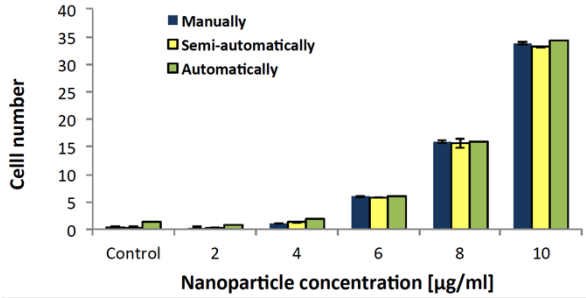


Figure 5: Comparison of counting results obtained by counting fluorescently PI stained cell to determine the number of dead CHO cells as a measure of cytotoxicity of increasing concentration of nanoparticles for 24 h. Three counting methods were used: manually by counting PI positive cells using ImageJ software, semi-automatically by correcting CellCounter counts and automatically by using CellCounter only. The results presented are average cell numbers for separate samples obtained by independent counting of the same experiment by three users with corresponding standard error (N=3).

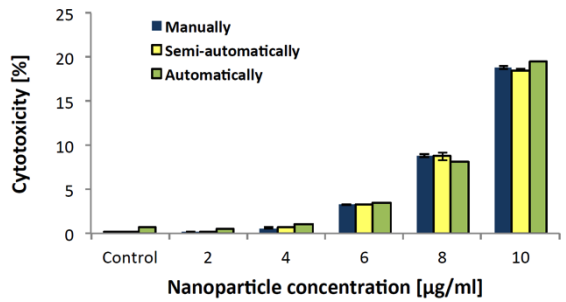


Figure 6. Comparison of counting results obtained by counting fluorescently stained cell nuclei to determine the number of dead CHO cells as a measure of cytotoxicity of increasing concentration of nanoparticles for 24 h. Three counting methods were used: manually by counting PI positive cells using ImageJ software, semi-automatically by correcting CellCounter counts and automatically by using CellCounter only. The results presented the average percentage of toxicity for separate samples compared to the number of all cells in non-exposed control obtained by independent counting of the same experiment by three users with corresponding standard error (N=3).

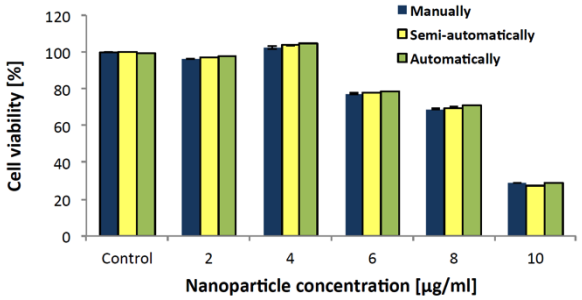


Figure 7. Comparison of counting results obtained by counting fluorescently stained cell nuclei (stained with Hoechst) reduced by the number of dead cells (stained with PI) to determine the viability of CHO cells exposed to increasing concentration of nanoparticles for 24 h. Three counting methods were used: manually by counting Hoechst positive cells using ImageJ software, semi-automatically by correcting CellCounter counts and automatically by using CellCounter only. The results presented are average cell numbers for separate samples obtained by independent counting of the same experiment by three users with corresponding standard error (N=3).

The counting results for all three counting methods showed reasonably good correlation and small standard errors, indicating also a relatively low inter-personal error as shown in Figs. 4-7. The differences between manual or semi-automatic counting and automatic counting are mostly due to two or more objects in substantial contact being counted as one or due to objects positioned predominantly outside of the margin of the image. Generally, we observed better performance in images with fewer and randomly scattered objects. On the other hand, the program had problems recognizing the right objects on images with low contrast, either images with no cells or with dimly stained cells. Still, the obtained differences are small compared to the expected deviations between repeated experiments, indicating the algorithm is robust enough and can be used as a helpful aid for image analysis.

Simultaneously, time required for counting with each counting method (automatic, semi-automatic or manual) was recorded for all three users. As seen from Fig. 8, the

program significantly reduces the time necessary to count the acquired images from hours (manual counting) to minutes (semi-automatic counting) or even seconds (automatic counting), depending on the required accuracy of the results.

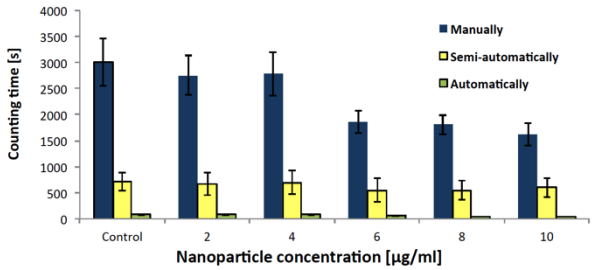


Figure 8. Comparison of counting time using three counting methods: manually by counting Hoechst and PI positive cells using ImageJ software, semi-automatically by correcting CellCounter counts and automatically by using CellCounter only. The results presented are average counting times for separate samples of three independent users with corresponding standard error (N = 3).

Moreover, the requirements of installation of CellCounter program are minimal, the interface is user friendly (Fig. 9), requiring no additional image manipulation or tuning of parameters to get reliable results (RAW images as obtained from the microscope can be used). Images can be loaded through the interface or by simple drag and drop action. The program counts the cells automatically as soon as the user selects the image, and allows simple corrections to the program count by adding or removing hits on the selected image with simple mouse clicks. The program also enables to skip a non-representative image. The results are automatically ordered and can be easily copied to clipboard to transfer them to the editing program of choice.

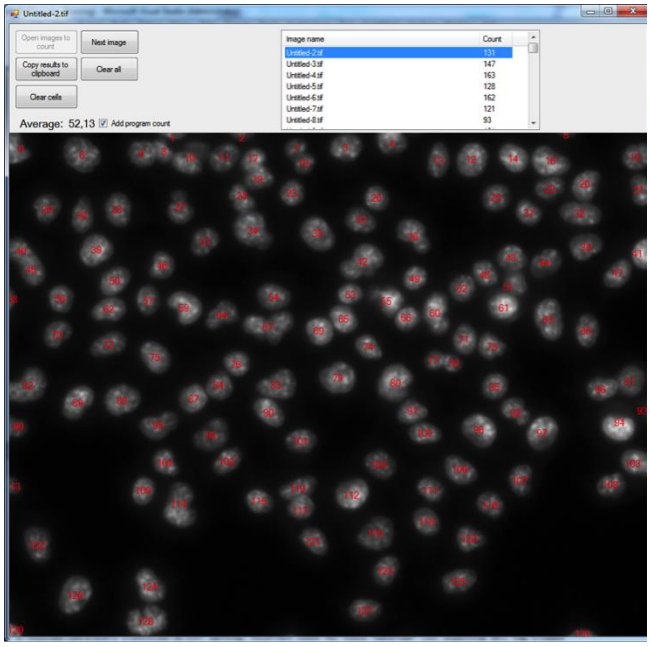


Figure 9. A screenshot of the CellCounter program interface showing the result of one automatic count. Each object (cell) is labeled so the user can quickly assert the accuracy of the counting event.

IV. Conclusions

In this article, we present an alternative solution for automatic and semi-automatic counting of cells on fluorescent microscopic images. As we showed in this paper, the program enables consistent, robust, fast and adequately accurate determination of fluorescent cells and can therefore be applied to a range of different applications in different fields of life sciences where fluorescent labeling is used for quantification of different phenomena.

Acknowledgment

The authors would like to thank David Karlaš for the help with the program and Klemen Strojan and Maja Jankovec for their help with the manual cell counting. This work was supported by Slovenian Research Agency within project J4-4324, young researchers program and MRIC UL IP-0510 Infrastructure program.

References

[1] M. Rieseberg, C. Kasper, K. F. Reardon, and T. Scheper, “Flow cytometry in biotechnology,” *Appl. Microbiol. Biotechnol.*, vol. 56, no. 3–4, pp. 350–360, Aug. 2001.
[2] J. R. Lakowicz, *Principles of Fluorescence Spectroscopy*, 3rd ed. Springer, 2006.
[3] D. J. Stephens and V. J. Allan, “Light Microscopy Techniques for Live Cell Imaging,” *Science*, vol. 300, no. 5616, pp. 82–86, Apr. 2003.
[4] F. J. Dein, A. Wilson, D. Fischer, and P. Langenberg, “Avian Leucocyte Counting Using the Hemocytometer,” *J. Zoo Wildl. Med.*, vol. 25, no. 3, pp. 432–437, Sep. 1994.
[5] Metamorph - http://www.moleculardevices.com/products/software/meta-imaging-series/metamorph.html
[6] BioQuant - http://lifescience.bioquant.com/
[7] Image-Pro - http://www.mediacy.com/index.aspx?page=IPP
[8] SynenTec - http://www.synentec.com/
[9] S. R. Gallagher, “Digital Image Processing and Analysis with ImageJ,” in *Current Protocols Essential Laboratory Techniques*, John Wiley & Sons, Inc., 2008.
[10] A. E. Carpenter, T. R. Jones, M. R. Lamprecht, C. Clarke, I. H. Kang, O. Friman, D. A. Guertin, J. H. Chang, R. A. Lindquist, J. Moffat, P. Golland, and D. M. Sabatini, “CellProfiler: image analysis software for identifying and quantifying cell phenotypes,” *Genome Biol.*, vol. 7, no. 10, p. R100, 2006.
[11] “UTHSCSA ImageTool.” [Online]. Available: http://compdent.uthscsa.edu/dig/itdesc.html. [Accessed: 02-Apr-2014].
[12] J. Selinummi, J. Seppälä, O. Yli-Harja, and J. A. Puhakka, “Software for quantification of labeled

bacteria from digital microscope images by automated image analysis," *BioTechniques*, vol. 39, no. 6, pp. 859–863, Dec. 2005.
[13] J. Byun, M. R. Verardo, B. Sumengen, G. P. Lewis, B. S. Manjunath, and S. K. Fisher, "Automated tool for the detection of cell nuclei in digital microscopic images: application to retinal images," *Mol. Vis.*, vol. 12, pp. 949–960, 2006.
[14] R. M. Haralick and L. G. Shapiro, "Image segmentation techniques," *Comput. Vis. Graph. Image Process.*, vol. 29, no. 1, pp. 100–132, Jan. 1985.
[15] K. Li, E. D. Miller, L. E. Weiss, P. G. Campbell, and T. Kanade, "Online tracking of migrating and proliferating cells imaged with phase-contrast microscopy," in *IN: PROC. IEEE CONF. COMP. VISION AND PATT. RECOG. WORKSHOP, P. 65. IEEE COMPUTER SOCIETY PRESS, LOS ALAMITOS*, 2006, p. 65.
[16] V. B. Bregar, J. Lojk, V. Suštar, P. Veranič, and M. Pavlin, "Visualization of internalization of functionalized cobalt ferrite nanoparticles and their intracellular fate," *Int. J. Nanomedicine*, vol. 8, pp. 919–931, 2013.
[17] M. Kanduser, D. Miklavcic, and M. Pavlin, "Mechanisms involved in gene electrotransfer using high- and low-voltage pulses--an in vitro study," pp. 265–271, Feb. 2009.
[18] T. Marš, M. Stražišar, N. Kotnik, K. Pegan, J. Lojk, Z. Grubič, and M. Pavlin, "Electrotransfection and lipofection show comparable efficiency for in-vitro gene therapy of primary human myoblasts," *unpublished*, 2014.
[19] I. Marjanovic, M. Kandušer, D. Miklavčič, M. Manček Keber, and M. Pavlin, "Comparison of flow cytometry, fluorescence microscopy and spectrofluorometry for analysis of gene electrotransfer efficiency," *unpublished*, 2014.
[20] "AForge.NET Framework." [Online]. Available: http://www.aforgenet.com/aforge/framework/. [Accessed: 02-Apr-2014].
[21] S. A. Ahmad, "Analysis of Compound Enhancement Algorithms (CEA) based on Adaptive Histogram Equalization (AHE) on Intra-oral Dental Radiographs Images," *Int. J. New Comput. Archit. Their Appl. IJNCAA*, vol. 4, pp. 902–916, 2011.
[22] H. D. Cheng and X. J. Shi, "A simple and effective histogram equalization approach to image enhancement," *Digit. Signal Process.*, vol. 14, no. 2, pp. 158–170, Mar. 2004.
[23] K. Zuiderveld, "Contrast limited adaptive histogram equalization," pp. 474–485, 1994.
[24] N. Otsu, "A Threshold Selection Method from Gray-Level Histograms," *Syst. Man Cybern. IEEE Trans. On*, vol. 9, pp. 62–66, 1979.
[25] X. Xu, S. Xu, L. Jin, and E. Song, "Characteristic analysis of Otsu threshold and its applications," *Pattern Recognit. Lett.*, vol. 32, no. 7, pp. 956–961, May 2011.
[26] L. Vincent and P. Soille, "Watersheds in digital spaces: an efficient algorithm based on immersion simulations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, no. 6, pp. 583–598, Jun. 1991.
[27] F. Chang, C.-J. Chen, and C.-J. Lu, "A linear-time component-labeling algorithm using contour tracing technique," *Comput. Vis. Image Underst.*, vol. 93, no. 2, pp. 206–220, Feb. 2004.

Morphological Changes of Pressure Pulses in Oscillometric Non-Invasive Blood Pressure Measurements

V. Avbelj
Department of Communication Systems, Jožef Stefan Institute, Ljubljana, Slovenia
viktor.avbelj@ijs.si

Abstract – The oscillometric method is the most widely used method for automatic non-invasive blood pressure measurements. An oscillometric device computes systolic and diastolic blood pressures from pressure oscillations in the cuff during slow deflation. Normally only the envelope of the pressure oscillations in the cuff is used. A closer look at these oscillations shows that the morphology of the oscillation pulses is changing during the deflation. The evolution of morphology is different from person to person indicating that further information about cardiovascular status is present in the data morphology. We present a set of morphology characteristics of pressure oscillations in a small group of adults. Greater knowledge of morphology could lead to a better accuracy of blood pressure monitoring in the medical and home use.

I. INTRODUCTION

A stethoscope in hands of a doctor together with an inflatable cuff and a manometer are the requisites for non-invasive determination of blood pressure parameters by the method of Riva-Rocci and Korotkoff, which is still the reference non-invasive method (auscultatory method). An automatic blood pressure meter is one of the most popular medical instruments used at home. These devices mostly use oscillometric method, where the blood pressure parameters are estimated from small oscillations of the pressure in the cuff during its deflation. Nowadays, such oscillometric devices are replacing the conventional auscultatory method also in the clinical environment.

During an oscillometric measurement, first the cuff on the arm is inflated until the blood circulation stops. Many devices also measure the small pressure oscillations during the inflation to stop inflating the cuff with regard to individual blood pressure. After that, the device slowly deflates the cuff and measures the pressure. It is recommended that the deflation rate of 2-3 mmHg/s is used [1]. The amplitude of small pressure oscillations first rises and then falls. At the end, the cuff is completely deflated at a fast rate (Figure 1). Currently available automated blood pressure measuring devices use oscillometric method based on empirically derived algorithms (patent protected) to calculate the systolic (SBP) and diastolic (DBP) blood pressure [2]. There are two basic algorithms for calculation of the SBP and the DBP from the oscillometric measurement.

The study was supported in part by the Slovenian Research Agency under grant P2-0095.

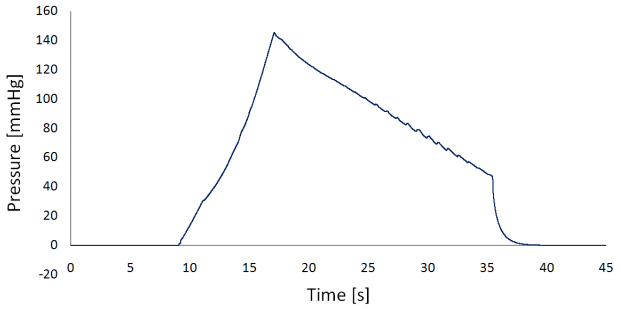


Figure 1. Pressure in the cuff during the complete measurement cycle. Pay attention to small oscillations during the deflation portion of the procedure.

The first is based on the determination of the pressure at the peak of the oscillometric envelope. This peak emerges around the mean arterial pressure (MAP). From the MAP, the SBP and the DBP are computed, in simple cases, as empirically defined fixed ratios relative to the MAP. For example, if the ratio for DBP was defined as 0.7 the DBP value is determined as pressure at the time, where the amplitude of the oscillometric envelope is reduced to 0.7 of its' maximal value. The second algorithm is slope-based where the inflection points of the oscillometric envelope define the SBP and the DBP [3]. The SBP is the pressure where the derivative of the envelope has a maximum and the DBP where the derivative has a minimum. In this case, the determination of the peak value of the envelope is not needed.

In a study of Jazbinsek et al. [4] authors compared oscillations of the pressure in the cuff with the Korotkoff sounds obtained from a microphone implanted in the cuff. They found that the higher frequency portion of the microphone signal can be directly related to the conventional auscultatory method and that the shape of this signal can be used for automatic determination of the SBP and the DBP.

The oscillometric envelope of real measurements is rarely a monotonically rising or falling function; therefore, the determination of the peak of the envelope and the derivatives is not a simple task (Figure 2). To solve this problem, the known algorithms attempt to smooth the oscillometric envelope appropriately. Some do this in the time-domain, for instance by filtering, others in the frequency-domain. The result is much smoother envelope of the oscillometric signal that enables "correct"

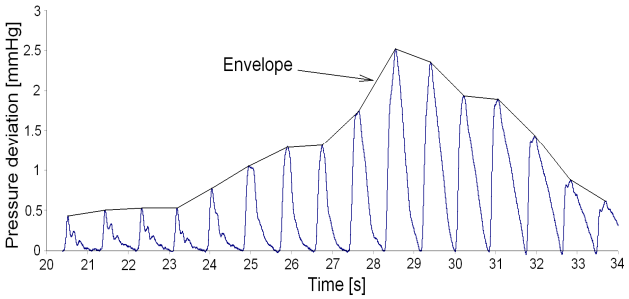


Figure 2. Raw oscillometric signal presented as zoomed pressure deviations from the down-sloping deflation pressure shown in Figure 1. An example of an envelope is shown as lines connecting the peak values of the oscillometric signal.

determination of the peak and the derivative of the signal. However, any process of smoothing definitely throws away the high frequency information of the raw signal.

In the study of Van Moer et al. [5] authors proposed a method based on a frequency-domain signal-processing approach. They linearize the oscillometric blood-pressure measurements by removing some spectral components of the raw signal and thus obtain a smooth envelope. The pressure at the peak of the envelope is declared as the MAP from which the SBP and the DBP were computed as fixed ratios to the MAP (0.52 and 0.7 respectively). Comparing the SBP and the DBP to the values from commercial non-invasive blood pressure meter (NIBP) and to the values obtained by the golden standard auscultatory method, authors concluded that their linearization is more capable of avoiding the extreme values since the procedure is less sensitive to outliers, noise, and leakage effects in comparison to the results from NIBP.

Is it worth throwing away high frequency components of the oscillometric signal? In this paper, we observe the oscillometric signal in time-domain, trying to preserve all frequency components. The main objective is to determine how the morphological changes in the oscillometric signal which occur during the deflation of the cuff could be used in blood-pressure measurements.

II. METHODS

A commercial and approved NIBP (Omron M6, HEM-7001-E) [6] was adapted with an additional pressure transducer and custom designed amplifier connected with a glass Y shaped piece to the rubber tube near the cuff. During blood pressure measurements the amplified signal was sampled with a digital oscilloscope (PicoScope 4424). The resolution was enhanced by oversampling (19.6 kHz) filtering and decimation to a final 100 Hz low-pass filtered output signal that was stored on a laptop computer for further processing.

A typical measurement is shown in Figure 1. At first, a fast inflation can be observed at a constant rate up to 30 mmHg when the adaptive inflation takes place up to 145 mmHg where the deflation begins. After some transitional instability of the pressure during the first seconds of the deflation, the deflation rate is constant (4.9 mmHg/s) till the end of the measurement when a fast deflation occurs. We made 40 measurements of blood pressure with 5 subjects. Each subject was measured once, except the one that was measured on 5 consecutive days with up to 8 measurements per day. We present here only the measurements of two subjects, where the first had low blood pressure values and the second had high. The repetitive measurements (36) of one subject will serve for the analysis of repeatability of the procedures in the future research.

To compare the pressure oscillation waveforms, the signal of the deflation part was segmented beat by beat. The segments were superimposed as shown in Figure 3. The synchronization point was set in the middle of the first upslope portion of each segment. The start time 0 s of each segment was set to 76.5 ms before the synchronization point. Thus the whole beginning portion of the upslope signal could be seen. The pressure deviation at start time was set to 0 mmHg (clamping).

III. RESULTS

Thorough examination of the pressure pulses of a person with a low blood pressure (NIBP values: SBP - 93 mmHg, DBP - 62 mmHg) shows the characteristics listed below. At Beat 1, which is the ninth beat (-9 in Figure 3) before the beat with the maximum pressure pulse M, three distinct waves can be seen with maximal amplitude of around 0.4 mmHg. Next three pulses (-8, -7 and -6) have also three distinct waves but with a little bit larger amplitudes. Pulse -5 has three waves, with the first much larger than the following two, and with the amplitude of the first wave 0.75 mmHg. Pulse -4 has the first two waves nearly fused and after the second wave, the signal falls down to the third wave.

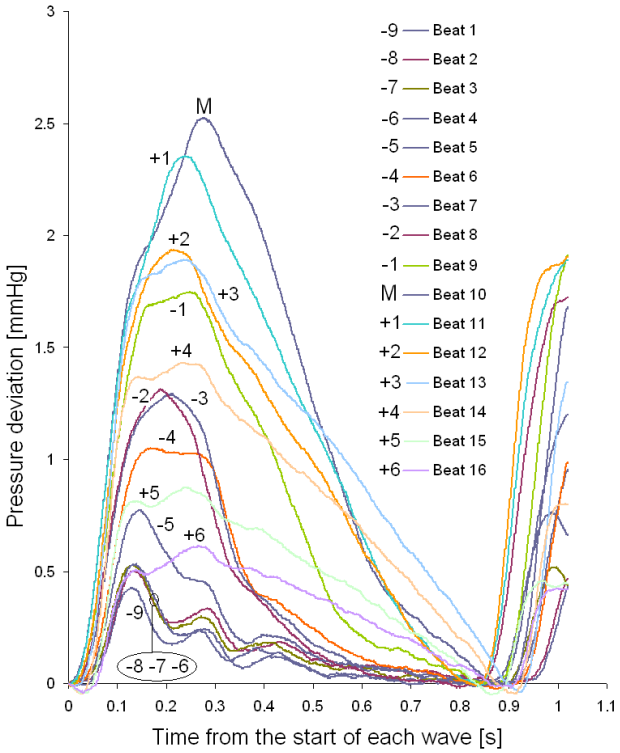


Figure 3. Aligned pressure pulses during deflation of a cuff. The pressure pulse with maximal amplitude that occurs in beat 10 is denoted by M. The previous beat is denoted by -1 and so on. The signals are taken from the measurement presented in Figure 1.

Pulses -3 and -2 are very similar with only a single peak, which is probably a fusion of the first two waves, with a fast down-slope up to time 0.35 s, where the slope changes. Pulse -1 is also modulated by three waves and has a fast down-slope that reaches the half amplitude around 0.45 s. The next pulse M has the maximal amplitude and also a fast down-slope that reaches the half of amplitude around 0.5 s. Pulse +1 with some lower amplitude, has lower down-slope rate as pulse M, but reaches the half of amplitude value at the same time as the pulse M. Finally, last five pulses (+2, +3, +4, +5 and +6) have similar morphologies and progressively lower amplitudes and down-slope constants. The largest difference of the pulses before M in comparison to the pulses after M is in the down-slope portion of the signal. In fact, the morphology of the pulses determines if a pulse is before or after the pulse M.

Another example from a person with a high blood pressure (NIBP values: SBP - 144 mmHg, DBP - 90 mmHg) is shown in Figure 4. As in Figure 3, the first two pulses (-14 and -13) in Figure 4 have distinct three waves but they are more fused. Again the pulses before M, up to the pulse -5, have fast down-slope portion. After that the down-slope portion needs progressively more time to reach the baseline level. In distinction to the previous example, here three pulses before M and one pulse after M have nearly the same amplitude as pulse M but again progressively needs more time to reach the baseline level as one observes the pulses from -3 to +1. A significant distinction between pulses -1 and M is shown in Figure 5. In pulse M, there is a slope transition point at 0.32 s where the pressure begins to fall with a slower rate.

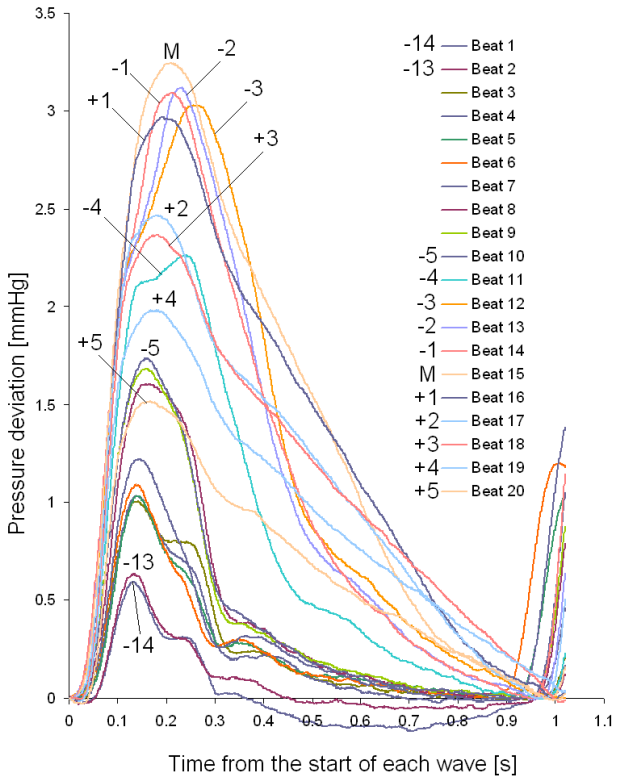


Figure 4. Aligned pressure pulses during deflation of a cuff from a person with a high SBP and DBP. The pressure pulse with maximal amplitude is denoted by M.

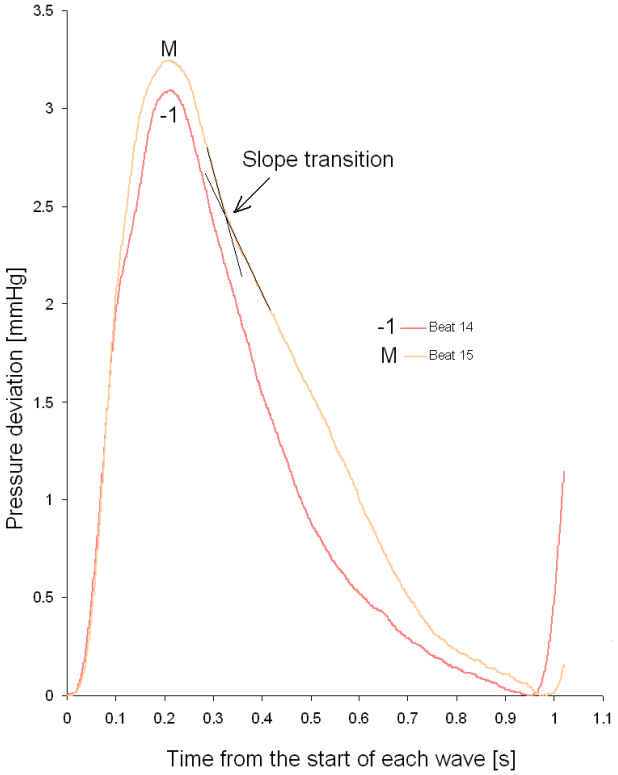


Figure 5. Different morphologies of the two successive pressure pulses (-1 and M) with a slope transition in pulse M.

Note that pulses before M, starting with pulse -1, have no such transitions so close to the peak of the pulse. We can conclude that the morphology of pulses holds additional information as in the previous case from Figure 3.

IV. Discussion and Conclusions

Today, oscillometric blood-pressure devices heavily depend on the envelope of the oscillometric signal. The smoothing of the envelope can produce better results on average, but still there are many failed results – outliers. To the best of our knowledge the presented study includes, for the first time, all the information of the oscillometric signal with no attempt to smooth out the envelope. A thorough examination of the measured signals with the proposed approach can enhance the accuracy of future oscillometric blood-pressure devices. We believe that losing information by smoothing the measured signals does not pay off.

In the future research of "what is going on" under the cuff during the measurement, it would be of great help to have an ECG signal and a respiration signal measured at the same time. A lot of previous research work was published that is related to applications of body sensors. In a previous work [7], it was shown in an animal study, that vessel clamps, equipped with force sensors, can provide an ECG signal with additional information about the blood-pressure inside the vessel. However, additional sensors should be unobtrusive not to further influence the blood pressure measurements.

A lot of studies and technological developments have been done elsewhere to improve the acceptance of body

sensor technology. Modern ECG sensors are non obtrusive, multifunctional, lightweight and battery powered wireless devices presented in our previous papers [8-11]. The problem of synchronization of independent wireless sensors was addressed in [12] and in another study [13], where the question if the heart rate variability could be assessed through blood pressure measurements is discussed. Cuff sensors can use a smart phone for the acquisition device as described in [14]. The positions of the electrodes should be carefully selected [15], so with the use of three pairs of electrodes, also the standard 12-lead ECG could be synthesized [16].

It is known that the procedures, performed during blood pressure measurements can influence the value of the blood pressure (white coat syndrome). Also, an occlusion of arteries in the arm by the cuff can do the same. In future validations of the NIMBs, the precise measurements and analyses [17] of physiologic parameters are needed with high resolution and multichannel devices [18].

We have shown that contemporary NIMB devices for measurements of blood pressure do not take full advantage of the available signal. The details in the morphology of oscillometric signal designate the vicinity of the maximal pulse wave so the envelope is not needed. If we have both, the envelope and the evolution of the details in the pulse waves, the pressure values could be determined with higher reliability. We analysed only the measurements of 5 subjects, therefore definite conclusions will be possible after an extended analysis of measurements from many different people. It is also possible that the details in morphology have additional information about the status of cardiovascular system, but this has yet to be determined.

Further efforts will be needed to improve the outcome of this routine and broadly accepted non-invasive blood pressure measurement procedure that could provide, possibly with added sensors on the cuff or body, higher value of the diagnostic outcome.

REFERENCES

[1] T.G. Pickering, J.E. Hall, L.J. Appel, B.E. Falkner, J. Graves, M.N. Hill, D.W. Jones, T. Kurtz, S.G. Sheps, E.J. Roccella, "Recommendations for blood pressure measurement in humans and experimental animals: part 1: blood pressure measurement in humans: a statement for professionals from the Subcommittee of Professional and Public Education of the American Heart Association Council on High Blood Pressure Research," Circulation, vol. 111, no. 5, pp. 697-716, 2005.

[2] G. Geršak, "Acquiring the oscillometric envelopes for non-invasive blood pressure measurements," Elektrotehniški vestnik, vol. 76, no. 3, pp. 97-102, 2009. In Slovenian.

[3] A. Sapinski, "Theoretical basis for a proposed standard algorithm of blood pressure measurements by the sphygmo-oscillometric method," J. Clin. Eng., vol. 22, no. 3, pp. 171–174, 1997.

[4] V. Jazbinsek, J. Luznik, Z. Trontelj, "Non-invasive blood pressure measurements: Separation of the arterial pressure oscillometric waveform from the deflation using digital filtering," The 3rd European Medical and Biomedical Engineering Conference, EMBEC'05, Prague, Czech, Nov. 20-25, 2005, pp. 3171-3174.

[5] W. Van Moer, L. Lauwers, D. Schoors, K. Barbe, "Linearizing Oscillometric Blood-Pressure Measurements: (Non)Sense?," IEEE Trans Instrum Meas., vol. 60, no. 4, pp. 1267-1275, 2011.

[6] S. Altunkan, N. Iliman, N. Kayaturk, E. Altunkan, "Validation of the Omron M6 (HEM-7001-E) upper-arm blood pressure measuring device according to the International Protocol in adults and obese adults," Blood Press Monit, vol. 12, no. 4, pp. 219-225, 2007.

[7] B. Gersak, R. Trobec, I. Krisch, and M. Psenicnik, "Loss of endothelium-mediated vascular relaxation as a response to various clamping pressures," European Journal of Cardio-thoracic Surgery, vol. 10, no. 8, pp. 684-689, 1996.

[8] R. Trobec, V. Avbelj, and A. Rashkovska, "Multi-functional Wireless Body Sensor", Proceedings of GLOBAL HEALTH 2012: The First Internacional Conference on Global Health Challenges, Venice, Italy, Oct. 21-26, 2012, pp. 71-74.

[9] R. Trobec, A. Rashkovska, and V. Avbelj, "Two proximal skin electrodes - A respiration rate body sensor," Sensors, vol. 12, no. 10, pp. 13813-13828, 2012.

[10] R. Trobec, V. Avbelj, and A. Rashkovska, "Multi-functionality of Wireless Body Sensors", Transactions on Internet Reserch, vol. 10, no. 1, 2014.

[11] K. Bregar and V. Avbelj, "Multi-Functional Wireless Body Sensor Analysis of Autonomy," Proc. of 36th International Convention MIPRO 2013, Opatija, Croatia, May 20-24, 2013, pp. 346-349.

[12] R. Trobec, M. Depolli, and V. Avbelj, "Wireless network of bipolar body electrodes," Proceedings of the 7th International Conference on Wireless On-demand Network Systems and Services, WONS 2010, Kranjska Gora, Slovenia, Feb. 3-5, 2010, pp. 145–149.

[13] A. Šmigoc, V. Avbelj, T. Jarm, "Applicability of continuous measurements of blood pressure for heart rate variability analysis," Proceedings of the 21th International Electrotechnical and Computer Science Conference, ERK 2012, Portorož, Slovenia, Sep. 17-19, 2012, pp. 191-194.

[14] A. Rashkovska, I. Tomašić, and R. Trobec, "A Telemedicine application: ECG data from wireless body sensors on a Smartphone," Proc. of 34th International Convention MIPRO 2011, Opatija, Croatia, May 26-30, 2011, pp. 262-265.

[15] I. Tomasic, S. Frljak, and R. Trobec, "Estimating the universal positions of wireless body electrodes for measuring cardiac electrical activity," IEEE Transactions on Biomedical Engineering, vol. 60, no. 12, pp. 3368-3374, 2013.

[16] R. Trobec and I. Tomašić, "Synthesis of the 12-lead electrocardiogram from differential leads," IEEE Trans. Inf. Technol. Biomed., vol. 15, pp. 615-621, 2011.

[17] V. Avbelj and R. Trobec, "A closer look at electrocardiographic P waves before and during spontaneous cardioinhibitory syncope," International Journal of Cardiology, vol. 166, pp. e59-e61, 2013.

[18] R. Trobec, "Computer analysis of multichannel ECG," Computers in Biology and Medicine, vol. 33, no. 3, pp. 215-226, 2003.

Personalized guiding templates for pedicle screw placement

U. Tominc*, M. Vesel*, S. Al Mawed*, M. Dobravec*, M. Jug*, S. Herman*, D. Kreuh**

* University Medical Center Ljubljana, Department of Traumatology, Zaloška 7, 1000 Ljubljana
Email: uros.tominc@kclj.si
** Ekliptik d.o.o, Tehnološki park, Teslova ulica 30, 1000 Ljubljana

ABSTRACT

INTRODUCTION: Pedicle screw fixation is a standard procedure of spinal instrumentation. The aim of this work was to evaluate accuracy and safety of using personalized guiding templates for pedicle screw placement on thoracic spine.

METHODS AND RESULTS: We performed a computer tomography (CT) scan of the spine specimen to acquire virtual 3D model of the spine. Computer software (EBS™, Ekliptik, Slovenija) was used to design drill templates according to pedicle trajectories. Drill templates were then printed out of the biocompatible polymer using 3D printing technology. Templates were placed on the anatomic specimen vertebras anchoring at three sites – on the lamina at the base of the superior articular process on both sides and at the tip of the spinous process. Holes were drilled, and 3,5 mm screws inserted from T1 do T7 (7 pairs) using custom template for each vertebrae. CT scan was obtained after screw placement to evaluate screw positions. 13 screws were inside of pedicle trajectory without violation of pedicle wall with the tip inside of the vertebral body. One screw perforated medial pedicle wall.

CONCLUSION: The potential use of such a navigational template to insert thoracic pedicle screws in spinal instrumentation is promising. This novel method could improve the accuracy of pedicle screw insertion and reduce the operating time and radiation exposure of spinal fixation surgery.

I. INTRODUCTION

Pedicle screw fixation is a frequent procedure of spinal surgery. Accurate screw placement is essential to avoid injury to the adjacent structures - the vessels, nerves, and viscera[1]. Pedicle screw placement is complex and technically demanding procedure with a steep learning curve[2]. There is limited visibility of spinal anatomical landmarks during surgery; so, it is important to gain a conceptualization of anatomy of the spinal structures that are hidden from direct view[3]. The accuracy with which screws are inserted in the pedicle has a direct effect on the surgical outcome[3]. The basis for pedicle screw placement is freehand technique where screws are placed according to the anatomical landmarks[2]. Multiple methods for improving the accuracy of pedicle screw placement were developed. Some examples of currently available guided screw insertion include, but are not limited to: intraoperative fluoroscopy, fluoroscopic and CT-guided computer-assisted surgery (CAS), pedicle probes that guide pedicle screw insertion using ultrasound or tissue conduction measurements, and electrophysiological monitoring techniques. According to the literature the percentage of misplaced pedicular screws is as high as 21% in freehand technique and as low as 3% in CT guided screw placement[4].

In the last 5-6 years the idea of personalised guiding templates for pedicle screw placement has evolved[6-8]. These templates are designed (using preoperative CT) to fit in a unique position on the individual's bone, and they have carefully designed holes to guide the drill through a pre-planned trajectory[7-9]. A rapid prototyping device is used as a three-dimensional printer to produce the shape of reference areas on the bone surface. One of the main advantage of this method is significantly reduced time of the usage of ionizing radiation during surgery[10-13]. It also has economical benefits in comparison to computer assisted surgery which is expensive and has high depreciation costs[5-13].

The aim of this work was to evaluate accuracy and safety of using personalized guiding templates for pedicle screw placement on the thoracic spine.

II. METHODS AND RESULTS

From the Institute of Anatomy of the Medical Faculty of Ljubljana we obtained one cadaveric spine specimen for the purpose of this study. We performed a helical computer tomography (CT) scan of the spine specimen to acquire virtual 3D model of thoracic spine. Specialized computer software for preopertive planing (EBS^{TM}, Ekliptik, Slovenija) was then used to design laminar drill templates according to predetermined pedicle trajectories (Figure 1).

Templates were specially designed to fit and lock on the lamina during the procedure. The templates fit in a unique position on the individual vertebrae, and they have carefully designed holes to guide the drill through a pre-planned trajectory. Drill templates for each vertebrae (T1-T7) were then printed out of the biocompatible polymer using 3D printing technology. The templates have a special metallic drill sleeve with inner hole corresponding to drill bit diameter. These metallic sleeves also prevent formation of debris from the template itself when drilling.

Cadaveric spine specimen was then prepared for the procedure – skin, subcutaneous tissues and fascia were incised, erector muscles were striped periosteally and retracted preserving interspinal and supraspinal ligaments. 3D drill guide templates were placed on the vertebras anchoring at three sites – on the lamina at the base of the superior articular process on both sides and at the tip of the spinous process (Figure 2). Holes were then drilled with power drill (Figure 3), probed and 3,5 mm steel screws were inserted. Screws were inserted from the first to seventh thoracic vertebrae (7 pairs) using custom template for each vertebrae.

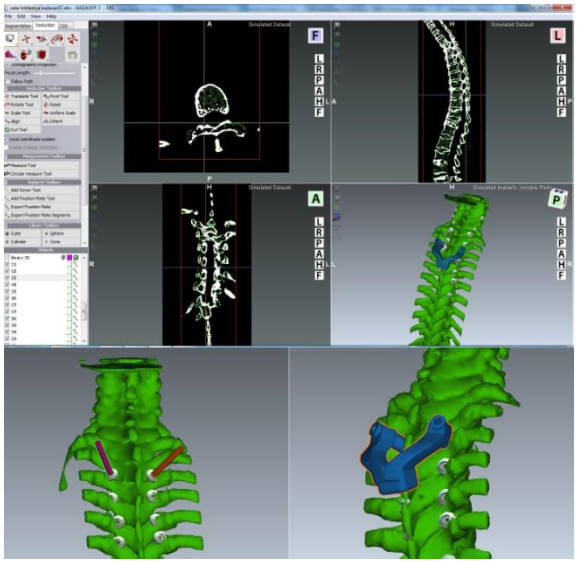


Figure 1. Designing templates using virtual 3D model of the spine.

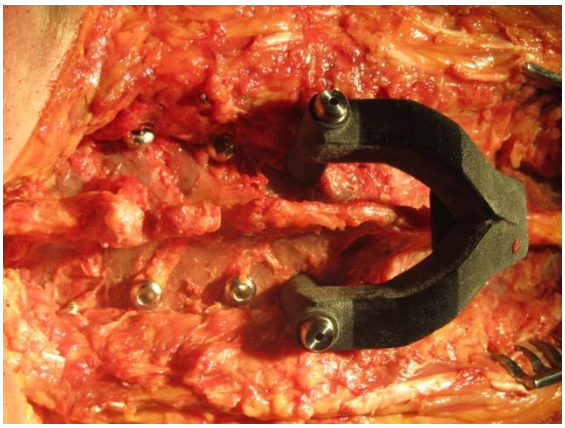


Figure 2. Template fitted on third thoracic vertebrae (T3)

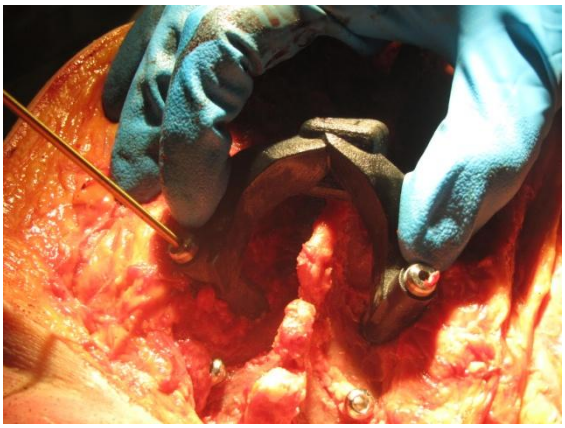


Figure 3. Template fitted on third thoracic vertebrae (T3)

CT scan was again performed (Figure 4) after screw placement to evaluate screw positions. Analysis showed that 13 out of 14 (92,9%) screws were inside of pedicle trajectories without violation of pedicle wall with the tip inside of the vertebral body. One screw violated medial pedicle wall in an "in-out-in" fashion.

III. CONCLUSSION

The potential use of such a navigational template to insert thoracic pedicle screws in spinal instrumentation is promising. When comparing it to a freehand technique, which has a 69% to 94% accuracy rate[14], this novel method could improve



Figure 4. CT scan after screw placement.

the accuracy of pedicle screw insertion. In comparison to fluoroscopic and CT guided pedicle screw insertion it could reduce the operating time and radiation exposure of the patient and the surgical team[10,12]. This method is also economically favorable in comparison to computer assisted navigation. As the freehand technique has a very steep learning curve[2], with this method a learning curve was not obvious.

Downsides of this method are relatively lengthy manufacturing process of the templates and possible manufacturing faults[10]. For the designing of the templates the surgeon needs to have computer skills and has to be familiar with the preoperative planning software to minimize possible human factor faults during production of the templates. This method also can not be used in minimally invasive spinal surgery, because good exposure of the posterior spinal elements is essential for correct placement of the templates[5-8].

REFERENCES

[1] Sugawara T, Higashiyama N, Kaneyama S, Takabatake M, Watanabe N, Uchida F, Sumi M, Mizoi K. Multistep pedicle screw insertion procedure with patient-specific lamina fit-and-lock templates for the thoracic spine: clinical article. J Neurosurg Spine. 2013 Aug;19(2):185-90.

[2] Gonzalvo A, Fitt G, Liew S, de la Harpe D, Turner P, Ton L, Rogers MA, Wilde PH. The learning curve of pedicle screw placement: how many screws are enough? Spine (Phila Pa 1976). 2009 Oct 1;34(21):E761-5.

[3] Manbachi A, Cobbold RS, Ginsberg HJ. Guided pedicle screw insertion:techniques and training. Spine J. 2014 Jan;14(1):165-79.

[4] Shin BJ, James AR, Njoku IU, Härtl R. Pedicle screw navigation: a systematic review and meta-analysis of perforation risk for computer-navigated versus freehand insertion. J Neurosurg Spine. 2012 Aug;17(2):113-22.

[5] Lu S, Xu YQ, Zhang YZ, Li YB, Xie L, Shi JH, Guo H, Chen GP, Chen YB. A novel computer-assisted drill guide template for lumbar pedicle screw placement: a cadaveric and clinical study. Int J Med Robot. 2009 Jun;5(2):184-91.

[6] Lu S, Xu YQ, Zhang YZ, Li YB, Shi JH, Chen GP, Chen YB. Rapid prototyping drill guide template for lumbar pedicle screw placement. Chin J Traumatol. 2009 Jun;12(3):177-80.

[7] Ryken TC, Kim J, Owen BD, Christensen GE, Reinhardt JM. Engineering patient-specific drill templates and bioabsorbable posterior cervical plates: a feasibility study. J Neurosurg Spine. 2009 Feb;10(2):129-32.

[8] Ryken TC, Owen BD, Christensen GE, Reinhardt JM. Image-based drill templates for cervical pedicle screw placement. J Neurosurg Spine. 2009 Jan;10(1):21-6.

[9] Owen BD, Christensen GE, Reinhardt JM, Ryken TC. Rapid prototype patient-specific drill template for cervical pedicle screw placement. Comput Aided Surg. 2007 Sep;12(5):303-8.

[10] Merc M, Drstvensek I, Vogrin M, Brajlih T, Recnik G. A multi-level rapid prototyping drill guide template reduces the perforation risk of pedicle screw placement in the lumbar and sacral spine. Arch Orthop Trauma Surg. 2013 Jul;133(7):893-9.

[11] Ferrari V, Parchi P, Condino S, Carbone M, Baluganti A, Ferrari M, Mosca F, Lisanti M. An optimal design for patient-specific templates for pedicle spine screws placement. Int J Med Robot. 2013 Sep;9(3):298-304.

[12] Lu S, Zhang YZ, Wang Z, Shi JH, Chen YB, Xu XM, Xu YQ. Accuracy and efficacy of thoracic pedicle screws in scoliosis with patient-specific drill template. Med Biol Eng Comput. 2012 Jul;50(7):751-8.

[13] Ma T, Xu YQ, Cheng YB, Jiang MY, Xu XM, Xie L, Lu S. A novel computer-assisted drill guide template for thoracic pedicle screw placement: a cadaveric study. Arch Orthop Trauma Surg. 2012 Jan;132(1):65-72.

[14] Gelalis ID, Paschos NK, Pakos EE, Politis AN, Arnaoutoglou CM, Karageorgos AC, Ploumis A, Xenakis TA. Accuracy of pedicle screw placement: a systematic review of prospective in vivo studies comparing free hand, fluoroscopy guidance and navigation techniques. Eur Spine J. 2012 Feb;21(2):247-55

iHELP emergency care network

Andraž Ogorevc*, Borut Lončarevič **, and iHELP TEAM***
* MIDS d.o.o., Ljubljana, Slovenia
**iHELP WORLD LTD, London, UK
info@ihelp-world.com

Abstract - The idea for the iHELP project originates in 2011. Disturbing statistics related to cardiovascular disease (i.e. 224 people a day die from sudden cardiac arrest in UK) made us think about how to enable the general public with an accessible and above all faster assistance, which is essential in order to ensure survival. The aim of the iHELP project is to have the largest rescue network that will be present locally and globally, and thus offer faster emergency help! We want to reduce the current high rate of mortality in the event of sudden cardiac arrest; increase the awareness about possible ways to help victims of cardiac arrest; assist local communities in setting up the defibrillators; and help business partners save their customers and employees, because this way whole society can save a lot of money. iHELP project has set a goal to place around 20,000 Automatic External Defibrillators (AED) in the world in five years, and by doing so we will enable a greater probability of survival in the event of sudden cardiac arrest.

I. INTRODUCTION

Cardiovascular Diseases (CVD) are one of the main problems of modern society. Sudden cardiac arrest is the leading cause of death in the world. 74% of cardiac arrests happen in front of laic eyewitnesses [1]. In Europe, cardiovascular diseases cause 45% of deaths among people aged 75 years or younger. Emergency Medical Service (EMS) treated 38 cases of cardiac arrest per year per 100,000 residents outside the hospital [2]. According to the experts [3] there are 5 to 6 cardiac arrests per day in Slovenia, on average that is around 2,000 per year. At sudden cardiac arrest rapid response is crucial, because after 4 to 5 minutes without the implementation of Basic Life Support (BLS) irreparable damage to the brain occurs [4].

So-called chain survival [5], which provides a useful metaphor fort the elements of the Embedded Control Concepts (ECC) systems, is composed of five links, which must be continuously connected to each other. The first link is intended for early identification and quick 112 call. In the second link we buy time with high-quality Cardiopulmonary Resuscitation (CPR). The third link is intended for early defibrillation, where caused heart rhythm disorder requires electric shock(s). The fourth link in the chain is intended for keeping the heart and brain working. The latter carry out specially trained teams of EMS. The last link is integrated post-cardiac arrest care.

Resuscitation or implementation of the CPR by the laity and early defibrillation saves lives. With implementation of CPR we follow the prescribed algorithm that determines the type and sequence of actions [6].

More than 17.3 million people worldwide die of cardiovascular diseases every year (in Europe, more than 4.3 million of which more than 700,000 deaths are result of sudden cardiac arrest). The number of people, who die from CVDs, mainly from heart disease and stroke, will increase to reach 23.3 million by 2030 [7]. Cardiovascular diseases are the leading cause of death in the world (they are responsible for 48% of all deaths worldwide, *2008*) [8]. The number is so high mainly because of the short time frame in which the cardiac arrest victim can still be saved and most of those who witness cardiac arrest either cannot or dare not help.

II. MOTIVATION

On the basis of the regulation [9] must the EMS unit from receiving the call (directed questions, first aid instruction) until the time they leave the starting location, take two minutes or less. Before mentioned healthcare organizations standards are overviewed by their internal control, which evaluates if the EMS team performance is successful. The EMS team in the vehicle should consist of a three-member team with a doctor, equipped with a monitor defibrillator, a crash kit and oxygen tank.

After the departure of intervention unit towards the scene, time for the journey and the preparation of the team to help the injured is needed. After adding the time of all procedures we soon discover that the intervention in less than 10 minutes is a great success. Because of this finding, we ask ourselves whether there is a way to improve the response time of the implementation of basic resuscitation procedures in the event of sudden cardiac arrest.

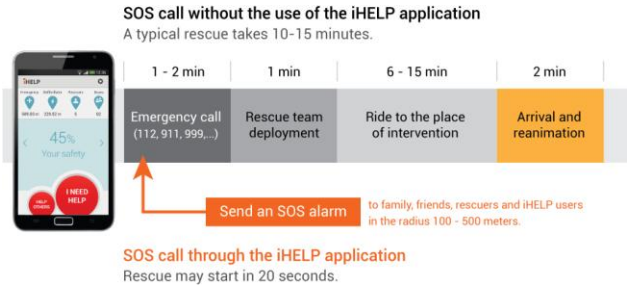


Figure 1. SOS call protocol [10]

Imagine that you are walking down the street with a friend. Suddenly your friend collapses due to a cardiac arrest. You call for help. Passers-by (lay people) who are afraid to help, because they do not know the basics of first aid, CPR, let alone how to use a defibrillator (What is a defibrillator and where are they?) run to you. You call emergency medical assistance at 112. Due to the traffic jam, ambulance will arrive after about 8-15 minutes, unfortunately your friend only has four minutes before his brain cells will start to die and as result irreversible damage will occur. What would you do?!

In most cases, you would not know what to do and this is one of the main reasons why 33 people die due to cardiac arrest worldwide every minute, (there are 7% chances of surviving sudden cardiac arrest). Statistics show that according to best scenario, rescuing takes about 8-15 minutes or more, depending on LOCATION, weather and time (traffic jams). With our solution, the arrival of rescuers can be in less than 20 seconds after we activate iHELP rescue network: family members, friends, Emergency Medical Service (EMS – 112/999), rescue workers in the field and iHELP users or laymen who are in the immediate vicinity of the unfortunate event. Due to the rapid arrival of help, the probability of survival rises up to 70% in the case of sudden cardiac arrest!

III. iHELP PLATFORM AND MOBILE APLICATION

iHELP mobile application offers users a quick and easy sending of multiple SOS alarm messages to family members, friends, professional rescuers and all users of iHELP mobile application within a radius of 300 meters or more from the victim. The application provides first aid training, overview of the nearest defibrillators, faster communication between family members in occurrence of critical events, reminder for taking pills. Application also calculates the safety of the user's at specific location; how far away are the emergency rooms, police and fire station, it shows the distance to the defibrillator, how many users are in the immediate vicinity and how far away are the family members or friends. With this information we can calculate the probability of response and arrival of help in a specific time! The great advantage of the iHELP system is that we inform the patient about the state of response for help: how many people have received SOS alarm message and how many have responded to help, and when will they arrive!

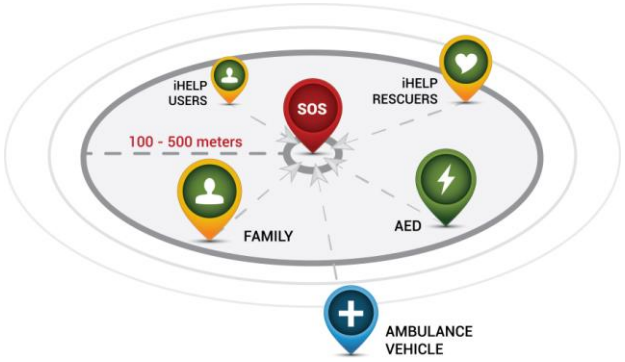


Figure 2. iHELP users

A. *iHELP Bracelet*

iHELP bracelet, which will work in situations where the user does not have a phone at hand, can easily send SOS message to family members, friends or professional rescuers. With holding down the button on the bracelet for three seconds, the Bluetooth on bracelet is activated and it sends wireless SOS signal to the mobile phone, which must be within a radius of 20 meters. Then iHELP mobile application automatically sends SOS message (geo-location of the injured person and all necessary information) to responsible person or rescuers (family members, friends, or to the SOS centre, if the client has a signed contract).

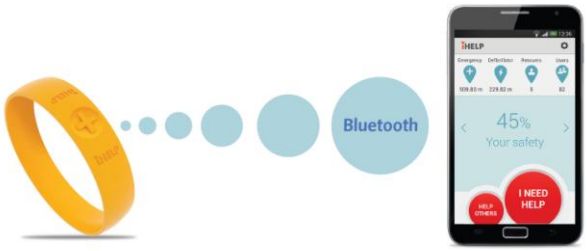


Figure 3. Bluetooth connection

With new version of the bracelet equipped with special health sensors (measuring heart rate, body temperature, the level of oxygen in the blood and measuring of G-force), which will recognise the problem of "sudden cardiac arrest". Bracelet will than automatically send a SOS signal to the victim's phone and in short time iHELP application will send SOS notifications to emergency contacts (ICE), professional rescuers and to all iHELP users in vicinity of the victim.

B. *Build a Care Network*

iHELP project is built in three steps:

Step 1: Create strong movement (community) through iHELP mobile application. Invite volunteers and

professional rescuers, coordinators, ambassadors and non-profit organizations to the iHELP project.

Step 2: establish contacts with strategic partners (socially responsible companies / cities / countries), invite to the story Heart Friendly Companies (Cities).

Step 3: Establish a marketing network and build users brand trust. iHELP – SAFETY IN THE PALM OF YOUR HAND!

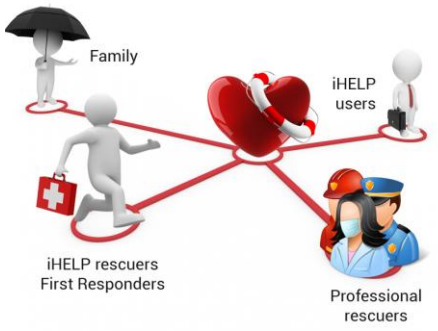


Figure 4. Care network

To create a strong movement we propose a significant involvement of volunteers in modern m-Health system. This aspect was underestimated in all previous investigations. Based on recent positive practical experiences and evidences, e.g. immediate active assistance in life threatening situations by medical events or natural disasters, we expect that the users of iHELP services can considerably benefit from this social dimension. The amount of people ready to cooperate on volunteer basis is surprisingly high, for example in Slovenia, there are 150.000 members of fireman organization, which represent 7.5 % of population, 10.000 retired individuals able to provide first aid. There are also other professionals who are able to offer first aid: 33.000 medical staff and 16.000 members of police and army staff. We will analyse the diversity of volunteering motives in two different EU regions (central Europe and west Europe) and find the optimal solution to increase the common benefits of people in need and volunteers.

We will also consider inclusion of other technologies, such Telemedicine application [11] and The Confidence system [12], into our intelligent platform for saving lives. Telemedicine application provides a comfortable option for telemonitoring of the heart activity of cardiac patients during their normal activities. It also helps prevent life threatening events by transferring most important results and alarms directly to medical personal.
The Confidence is ubiquitous care system, which was designed to support independent living of elderly, by detecting falls and unusual movement that may indicate a health problem.
The combination of these two technologies and iHELP application could mean a significant improvement of survival rate, homecare, overall health care system and it would reduce its expenses at the same time. We must take into consideration that sudden cardiac arrest is leading cause of death in modern society and that European population is aging rapidly, threatening to overwhelm the society's capacity for taking care of its elderly members.

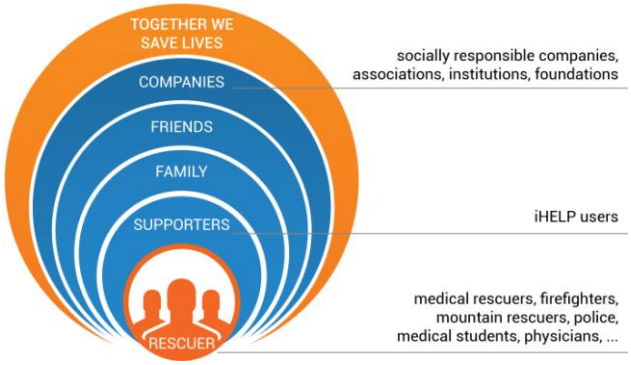


Figure 5. Community

IV. CONSLUSION

iHELP project was based on a worrying statistics of cardiovascular diseases and with death related to it. With iHELP project we wanted to connect people in interpersonal volunteer help, but now it is evolving into something much greater. iHELP project is creating a new kind of social network, which is not taking place only in the virtual world but also in real life. Its users can save lives and theirs lives can be saved by other users, their family, close friends, volunteers, professional medical staff, etc. The project combines Information and Communication Technologies (ICT) knowledge with basic human needs and natural survival instinct. Its development and usage are based in a way that encourages preventive actions and at the same offer iHELP users extra security in cases of severe illnesses, cardiac arrests or any other injuries. iHELP project strengthens bonds inside of community, develops interpersonal solidarity and creates new social norms.

REFERENCES

[1] R. Rajapakse. "Seznanjenost prebivalcev Slovenije z znaki srčnega zastoja in temeljnimi postopki oživljanja."M.A. thesis, Univerza v Ljubljani, Slovenija, 2008.

[2] J.P. Nolan, J. Soar, D.A. Zideman, D. Biarent, L.L. Bossaert, C. Deakin, et al. (2010). "European Resuscitation Council Guidelines for Resuscitation 2010. Section 1. Executive Summary." *Resuscitation.* [On-line]. 81(10), pp. 1219-76. Available: https://www.erc.edu/index.php/doclibrary/en/209/1/, [Mar. 3, 2014].

[3] T. Ploj, P. Gradišek. "Uporaba avtomatičnega defibrilatorja." in *Prva pomoč*, 1st ed. U. Ahčan. Ljubljana: Rdeči križ Slovenije, 2006, pp. 76-80.

[4] G.H. Bardy, K.L. Lee, D.B. Mark, J.E. Poole, W.D. Toff, A.M. Tonkin, et al. (2008, April). "Home use of automated defibrillators for sudden cardiac arrest." *The New England Journal of Medicine.* [On-line]. 17(4), pp. 1793-804. Available: http://www.nejm.org/doi/full/10.1056/NEJMoa0801651#t=articleTop, [Mar. 3, 2014].

[5] R.W. Koster, M.A. Baubin, L.L. Bossaert, A. Caballero, P. Cassan, C. Granja, et al. (2010). "European Resuscitation Council

Guidelines for Resuscitation 2010, Section 2. Adult basic life support and use of automated external defibrillators.“ *Resuscitation.* [On-line]. 81(10), pp. 1277-92.

[6] American Heart Association. ”Chain of Survival.” Internet: http://www.heart.org/HEARTORG/, Feb. 28, 2014 [Mar. 3, 2014].

[7] World Healt Organization. ”Cardiovascular diseases (CVDs).” Internet: http://www.who.int/mediacentre/factsheets/fs317/en/, March 2013 [Mar. 3, 2014].

[8] World Healt Organization. ”Global status report on noncommunicable diseases.“ Internet: http://www.who.int/nmh/publications/ncd_report2010/en/, April 2011 [Mar. 3, 2014].

[9] Uradni list Republike Slovenije. “Uredba o organizaciji in delovanju sistema, opazovanja, obveščanja in alarmiranja.” Internet: https://www.uradni-list.si/1/content?id=5669&part=&highlight=uredba+o+organizacij i+in+delovanju+sistema+opazovanja#!/Uredba-o-organizaciji-in-delovanju-sistema-opazovanja-obvescanja-in-alarmiranja, Jul. 25, 1997 [Mar. 3, 2014].

[10] Uradni list Republike Slovenije. “Pravilnik o službi nujne medicinske pomoči.” Internet: http://www.uradni-list.si/1/objava.jsp?urlid=200757&stevilka=3039, Jun. 29, 2007 [Mar. 3, 2014].

[11] A. Rashkovska, I. Tomasic, R. Trobec. “A Telemedicine application: ECG data from wireless body sensors on a Smartphone“, in *MIPRO 2011 - Proceedings of 34th International Convention on Information and Communication Technology, Electronics and Microelectronics*, 2011, pp. 262-265.

[12] M. Luštrek, B. Kaluža, B. Cvetković, E. Dovgan, H. Gjoreski, V. Mirchevska, and M. Gams. "Confidence: ubiquitous care system to support independent living", in *DEMO at European Conference on Artificial Intelligence*, 2012, pp. 1013-1014.

Graph-Based Diagnosis and Treatment of Neuronal Communication Disorders

Mladen Veletic*†
*Faculty of Electrical Engineering
University of Banja Luka (UNIBL), Banja Luka, Bosnia and Herzegovina
†Dept. of Electronics and Telecommunications
Norwegian University of Science and Technology (NTNU), Trondheim, Norway

***Abstract*—The report issued by Alzheimer's Disease International in 2009 estimated that 35.6 million people worldwide would be suffering from the disease in 2010, and the number was estimated to 115.4 million in 2050. Development of safe and effective disease treatments is, therefore, an urgent matter. Understanding the physical mechanisms underlying the neuronal network could enable scientists to invent new treatment strategies and techniques. Development of different invasive and non-invasive techniques of stimulating the region of the brain, that is malfunctioning owing to a neurodegenerative disease from the outside of- or inside the body, became plausible considering the recent findings in biology and neuroscience. This paper does not delve with any specific treatment fashion in terms of technique being deployed, but discusses one possible strategy formation of controlling the neuronal communication, which might be applicable in general. Specific graph theory concepts are deployed and integrated from a theoretical neuroscience perspective, with the main motivation of altering the performance of neuronal communication in order to slow down, halt or reverse mental disorders, such as Alzheimer's, depression, schizophrenia, autism etc., that reduce quality of life.**

I. Introduction

Neuronal nano-communication between neurons underlies all the processes taking place in nervous system. Neurons are specialized cells that generate electrical pulses called action potentials (APs), or spikes, in response to electrical or chemical inputs (usually) collected by compartments called dendrites. Generated response is then transmitted down through the nerve fiber called axon, that serves as communication channel connecting the neurons. The difference in electrical potential between the interior of a neuron and the surrounding extracellular medium, hyperpolarization and depolarization, are what forms the AP pulse shape. Hyperpolarization makes the neuron's membrane potential more negative due to the positively charged ions (Na^+, Ca^{2+}, and K^+) flowing out, or negatively charged ions (Cl^-) flowing in. Conversely, depolarization makes the membrane less negative. A structure that further allows a neuron to pass a message generated, i.e. an electrical pulse, to another cell is called the synapse. Synapses transfer signals from one neuron to another through neurotransmitter molecules released into the synaptic cleft. When an AP reaches the presynaptic membrane, it will depolarize, leading to opening of voltage controlled calcium ion-gates (VCCGs) and an influx of extracellular calcium ions, Ca^{2+}, into the presynaptic terminal. A rise in intracellular calcium concentration, hereinafter referred to $[Ca^{2+}]_i$, from its resting concentration increases the probability that vesicles (small bubbles within a cell that store various neurotransmitters) docked onto the neuronal membrane releases neurotransmitters enabling neurons to further transmit the message generated. Adequate message propagation is fundamental for reliable neuronal communication. Any disruption leads to malfunctioning brain.

In case of malfunctioning network on board, an interest is to develop either non-invasive or invasive methods that can control neuronal communication (excluding the 'traditional' pharmaceutical drug delivery approaches which effects are highly debatable). In terms of non-invasive methods, recent studies performed showed that electromagnetic radiation (EMR) exposure and magnetic fields (MF) can improve the cognitive abilities of the brain eliminating the possibility that the positive effects were caused by increased temperature [1]–[3]. In terms of invasive methods, an ideal futuristic approach is to develop nano-machines based on graphene or carbon nanotubes, that can deliver drugs, mimic neurotransmitters, or expose neurons to EMR/MF at very specific targets, minimizing risk of side effects [4], [5]. With such devices, one becomes less dependent on specific physical relationships since the possibilities of neural control ranges from field exposure to delivery of specific chemical agents and mimicking of certain neurotransmitter molecules.

The objective of this paper is to develop a possible graph-based strategy of driving the neuronal communication through controlled concentration of calcium ions,Ca^{2+}, since findings from neuroscience proved they play a crucial role in neuronal communication performance. It is believed that any disruption in the processes that regulates calcium concentration levels (referred as *calcium dysregulation*), will lead to dramatic changes in neuronal functioning. A possible treatment in brains with calcium dysregulation is, therefore, to develop techniques that re-establish proper calcium regulation. It will be shown in Section II that $[Ca^{2+}]_i$ can be altered through changes in membrane's potential evoked either by drugs, external fields, nano-machines, or other means such as optogenetics (a neuromodulation technique employed in neuroscience that uses a combination of techniques from optics and genetics to control and monitor the activities of individual neurons), or classic "current clamp" technique (used to stimulate a neuron

via electric current that enters a cell by means of glass micropipettes (electrodes)). This paper promotes a usage of specified graph theory concepts for selection of targeted neurons whose controlled $[Ca^{2+}]_i$ overcome calcium dysregulation issue and lead to adequate formation of strategy assisting in treatment and diagnosis of neuronal diseases. Since graph theory has recently started being utilized in neuroscience, the most appropriate graph theory measures for the analysis of brain have not been selected. Nonetheless, a set of those that might have certain relevance for neuronal nano-communication performance analysis and neuroscience applications are deployed here. Treated neurons shall further be able to affect the behavior of its postsynaptic neurons, as long as they are (somehow) treated long enough to make relevant processes reach the steady state.

II. Relation between Intracellular Calcium Concentration and Membrane's Potential

Having a strategy defined in this paper being confined to neuronal communication dependent on $[Ca^{2+}]_i$, it seems reasonable to state a relation between $[Ca^{2+}]_i$ and neuronal response characterized with its membrane potential fluctuations.

The $[Ca^{2+}]_i$ can be expressed by the following differential equation [6]

$$\frac{\mathrm{d}[Ca^{2+}]_i}{\mathrm{d}t} = -I_{Ca} - \frac{[Ca^{2+}]_i}{\tau_{Ca}}, \tag{1}$$

where I_{Ca} is the total calcium current through the membrane, and τ_{Ca} is the time constant for cytosolic Ca^{2+} removal. i_{Ca} is mainly flowing into the neuron since external concentration, denoted as $[Ca^{2+}]_e$, is much greater than internal, i.e. $[Ca^{2+}]_e >> [Ca^{2+}]_i$. Typical resting values are $[Ca^{2+}]_e \sim$ 1 mM scale and $[Ca^{2+}]_i \sim 100$ nM scale. Rather few ions are needed in order to raise $[Ca^{2+}]_i$ significantly. A concentration of around 10 μM require about 300-400 Ca^{2+}. The solution to (1) is

$$[Ca^{2+}]_i(t) = e^{-\frac{t}{\tau_{Ca}}} \int_0^t e^{\frac{t'}{\tau_{Ca}}} I_{Ca}(t')\mathrm{d}t'. \tag{2}$$

The Goldman, Hodgkin and Katz (GHK) equation [6] relates I_{Ca} to $[Ca^{2+}]_e$ and $[Ca^{2+}]_i$

$$I_{Ca} = \mathscr{P}_{Ca} 2vF \frac{[Ca^{2+}]_i - [Ca^{2+}]_e e^{-v}}{1 - e^{-v}}, \tag{3}$$

where $v = 2V_mF/(RT)$, V_m is the membrane potential, F is Faraday constant, R is the gas constant, and T the absolute temperature (in Kelvin). $\mathscr{P}_{Ca}$ is the cell membranes permeability to Ca^{2+} and works as a constant of proportionality between molar flux density over a unit membrane area and the concentration difference over the membrane, i.e. $S = -\mathscr{P}_{Ca}\left([Ca^{2+}]_i - [Ca^{2+}]_e\right)$. According to (3), I_{Ca} directly depends on the membrane potential. One possible way of controlling $[Ca^{2+}]_i$ is through change in membrane voltage V_m, since it also drives the $[Ca^{2+}]_i$ according to (2). Hence, to change the neuronal membrane potential, it is sufficient to inject or induce a convenient current.

In what follows, an answer to the following question will be given: *How to recruit the neurons whose $[Ca^{2+}]_i$ is to be controlled when large-scale malfunctioning network is to be treated?*

III. Graph Theoretical Modeling of Neuronal Connectivity

The anatomical configuration of brain networks can quantitatively been characterized by graph representing either inter-neuronal- or inter-regional connectivity. If inter-neuronal connectivity is analyzed, structural networks can be described as graphs which are composed of vertices (nodes) denoting neurons, that linked by edges representing their physical connections, i.e. synapses. Conversely, if inter-regional connectivity is analyzed, graph theory describes structural networks as graphs which are composed of nodes denoting brain regions, further linked by edges representing axonal projections. The former approach is subject in this paper. Thereby, graph G, representing neuronal network that consists of V set of vertices and E set of edges, denoted as $G(V,E)$, is analyzed here. Concerning the notation, both of the terms node and vertex denoting neuronal cells will be equally present.

Quantitative characterization of anatomical patterns can be mathematically described through matrices, such as the adjacency matrix (AM), also called the connection matrix. An AM is with non-zero entries a_{ij} if the connection, i.e. synapse, is present between neurons i and j; otherwise, a_{ij} is zero. For any given synapse, the synaptic weight (strength) will depend on the number of vesicle release sites n, the release probability p, and a measure on the postsynaptic effect q. Although we lack a formula describing release probability as a function of $[Ca^{2+}]_i$, and thereby w_{ij}, a formula for w_{ij} as a function of $[Ca^{2+}]_e$ has been experimentally found by Dodge and Rahminoff [7],

$$w_{ij} \sim \left(\frac{[Ca^{2+}]_e/K_c}{[Ca^{2+}]_e/K_c + [Mg^{2+}]_e/K_m + 1}\right)^a, \tag{4}$$

where K_c and K_m are equilibrium constants of calcium and magnesium, respectively. a is the number of independent sites Ca^{2+} must bind to in order to release neurotransmitters. Depending on the nature of the ion flow, the synapses representing connections can have either an excitatory, depolarizing, or inhibitory, hyperpolarizing, effect on the postsynaptic neuron. Hence, excitatory and inhibitory synaptic connections correspond to positive and negative a_{ij} values in AM, respectively.

Anatomically, many compartments (e.g. dendrites) of single presynaptic neuron i can be connected to postsynaptic neuron j, making a numerous synaptic connections. In presented analysis, it is assumed that weights of all the connections are superimposed resulting a_{ij}. Furthermore, each two neurons can both be the pre- and postsynaptic ones to each other, in which case they mutually share two connections represented by different values of a_{ij} and a_{ji} in AM. Additionally, the graph $G(V,E)$ representing neuronal network is directed due

to the one-way axonal transmission[1]. This ultimately makes the $G(V,E)$ being directed, weighted, and signed. Excluding the self-connections, all diagonal elements of AM are zero.

With the neuronal network mapped (usually by means of functional magnetic resonance imaging (fMRI) or electrophysiological techniques, such as electroencephalography (EEG) and magnetoencephalography (MEG)), one can reveal the connectivity patterns and structure formed of densely connected clusters linked together in a small-world network [8]. Small-worldness combines high levels of local clustering among neurons, associated with high efficiency of information transfer and robustness, and short paths, that link all neurons of the regional network. Small-world organization is intermediate between that of random networks, characterized with short overall path length and low level of local clustering, and that of regular networks, characterized with long path length accompanied by high level of clustering. Such a structure is essential not only for regions of specialized neurons, but also for the brain in general, since it combines two fundamental functioning aspects: *segregation*, where the similar specialized neuronal units are organized in densely connected groups that are interconnected, and the functional *integration*, which allows the collaboration of a large number of neurons to build cognitive states [9].

IV. Strategy of Diagnosis and Treatment of Neuronal Disorders

As shown, graph interpretation of neuronal connectivity provides researchers with connectional relationships between individual neurons and ability to computationally analyze the connectivity patterns. Although this research is mostly confined to formation of adequate strategy assisting in diagnosis and treatment of neurons struck by mental disorders, e.g. hippocampal neurons, two-dimensional spatial representations of a local subnetwork of 131 neurons with 764 unidirectional connections within *Caenorhabditis elegans* rostral ganglia (available in [10] and [11]), can be used for analysis without any loss of generality.

In terms of network quantification, the first parameter that can be easily computed is the average connection density, k. Value of k represents the number of all non-zero entries in AM, divided by the maximal possible number of connections, i.e. $n^2 - n$ (n stands for number of vertices in V set) for a directed graph, excluding self-connections. The sparser the $G(V,E)$, the lower its average connection density. The average connection densities vary depending on the particular anatomical structure. For instance, connected cells across the entire cerebral cortex typically produce k values ranging from 10^{-7} to 10^{-6} [12]. Conversely, local connection densities are significantly higher as shown for analyzed *Caenorhabditis elegans* network, where $k = 0.0449$ in particular. Although the result suggests a dense property of network, k value in addition to binary, unweighted, and, at this point still unsigned AM, does not provide a detailed insight into the functional organization of the region. Nonetheless, potential application of average connection density is in diagnosis: decreased k value of monitored brain subnetwork can mark disorder producing increased failures in neuronal links.

To keep the reasoning as realistic as possible, modification of 20% of the number of synapses in AM, is introduced giving them negative weights (considering inhibitory synapses) according to (4). Modified $G(V,E)$ is visualized as shown in Fig 1. Spatial two-dimensional positions represent the position of the soma of individual neurons. X-axis colormap ranging from -10 to 10 is used to differentiate between inhibitory and excitatory synapses. Note that distribution of inhibitory and excitatory cells is not temporal, unlike the time-variable synaptic weight values. Different patterns the synaptic weights form are beyond the long terms potentiation (LTP) and long term depression (LTD) processes. LTP refers to strengthening of the connection between neurons, and is usually initiated when the activity of connected neurons are synchronized (or highly correlated). LTD refers to a weakening of the connection between neurons, and is usually initiated when the activity of neurons are nonsynchronized (or have low correlation). Synchronized and nonsynchronized neuronal activity in the hippocampus is crucial for memory formation. During learning, the LTP and LTD processes will create memory as certain connection patterns in the (hippocampal) neuron network.

Presented spatial and temporal visualization can also be used in neuroscience as one possible way of diagnosis. For instance, significant reduction of clustering and loss of small-worldness might provide a *clinically useful diagnostic marker* indicating an Alzheimer's disease [8], [13]. Moreover, graph theory measures can be used to create a certain *visualization tool* that might lead to efficient and effective process of human regulation of $[Ca^{2+}]_i$. In what follows, concepts build upon the outdegree, clustering coefficient, and node and edge centrality criteria are presented.

A. Centrality Criterion

Neurotransmitter release process and modification of post-synaptic neuron's conductivities drive the neuronal abilities to communicate and build reliable connections. Potential cellular changes, however, should be designed to avoid small-scale alterations. To produce large-scale effects and effectively cure a brain region of interest, 'important' targets should be determined.

In graph theory and network analysis, centrality of a vertex measures its importance within a graph. Centrality concepts were first developed in social networks, but produce general answers to questions like "how influential a person is within a social network", "how important a room is within a building", and "how well-used a road is within an urban network". In this study, centrality is applied to examine *how involved a cell is in neuronal communication system* in order to identify those to be monitored, "guided", and controlled in terms of $[Ca^{2+}]_i$. There are variety of measures of centrality that

[1]Axon conducts the APs typically in a direction outward from the soma. This propagation is called orthodromic propagation. Antidromic propagation in the reverse direction is not excluded under a certain scenario.

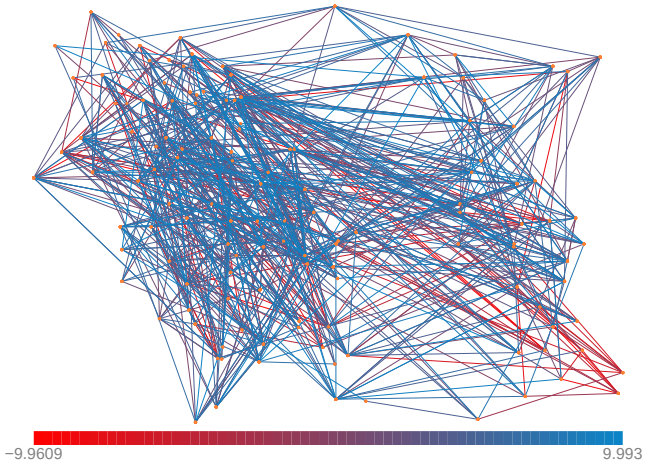


Fig. 1. Two-dimensional spatio-temporal pattern of weighted excitatory and inhibitory synapses. Red lines denote the inhibitory connections, whereas blue lines denote excitatory ones.

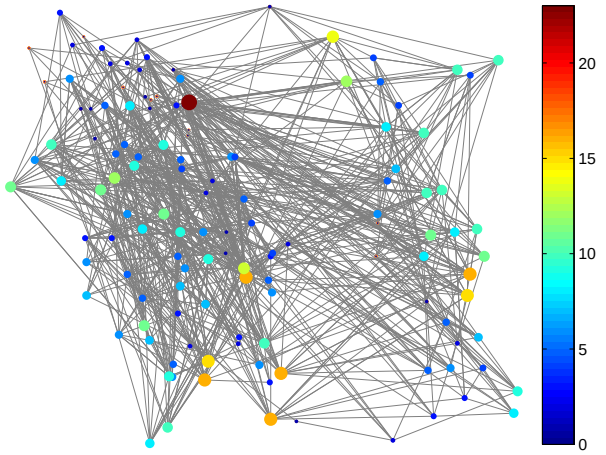


Fig. 2. Representation of neurons according to the outdegree criterion. Blue color indicates cells with no importance communication-wise. Important and potentially treated neurons are orange and red colored.

focus on different concepts and definitions, such as degree, betweenness, closeness, and eigenvector. As historically first, conceptually simplest, and very useful measure of centrality, degree is applied here to find neurons with the most numerous synaptic connections. Additionally, betweenness centrality is also deployed in what follows since that criterion is introduced as a measure for quantifying the control of a human on the communication between other humans in a social network [14]. Conversely, closeness and eigenvector centrality do not qualify themselves as usable criteria in this study due to the unmatched physical interpretations. For instance, closeness can be regarded as a measure of how long it will take to spread information from analyzed vertex to all other vertices sequentially, thereby being completely irrelevant in $[Ca^{2+}]_i$ regulation strategy.

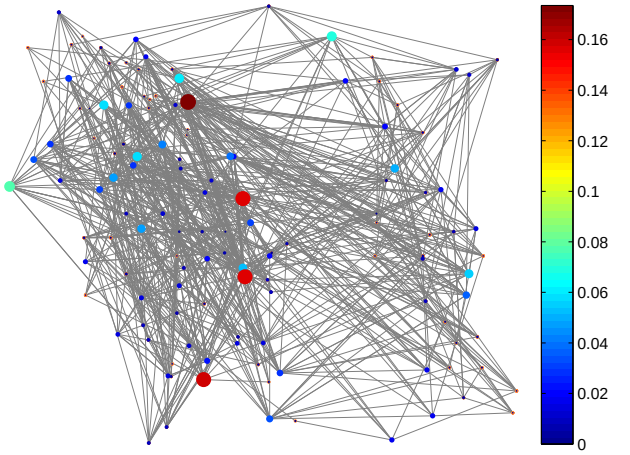
1) Outdegree centrality: The degree of a vertex in a network is the number of edges attached to it. Computationally, degree is the sum of vertex's indegree and outdegree values defined as the number of incoming (afferent) or outgoing (efferent) edges, respectively. In many cases, vertices with the highest degrees in a network play important roles in functioning of the system. Hence, degree can be a useful guide for focusing attention on the system's most crucial elements. In particular, outdegrees, $od(v)$, are of interest here due to the obvious functional interpretation: high outdegree of analyzed neuron indicates a large number of potential functional targets, i.e. receivers. Although the corresponding values are subject to constraints due to the growth, tissue volume, or metabolic limitations [12], the criterion satisfactorily qualifies an adequate neurons whose $[Ca^{2+}]_i$ is to be regulated. In such a way, any message sent from identified and treated neuron will be effectively spread throughout the network reaching the maximum number of recipients. Outdegree centrality criterion visualizes the $G(V,E)$ in a way shown in Fig. 2.

2) Betweenness centrality: Betweenness centrality also measures a vertex's or edge's importance in a network. If σ_{hj} is the total number of shortest paths from h to j, and $\sigma_{hj}(i)$ is the number of these shortest paths that pass through the vertex i, the betweenness of vertex i is defined as [9]

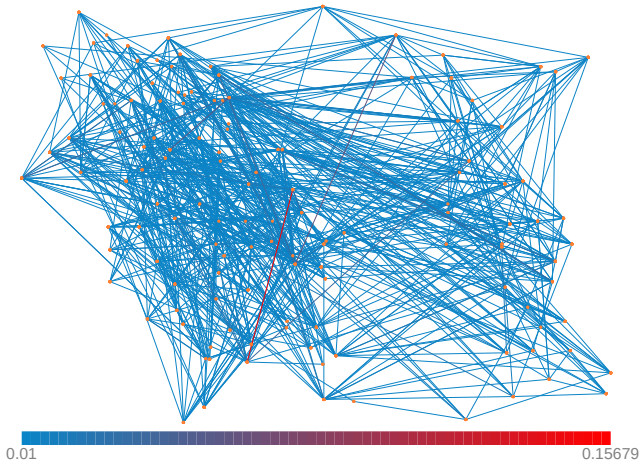
$$b_i = \sum_{h \neq j \neq i} \frac{\sigma_{hj}(i)}{\sigma_{hj}}. \tag{5}$$

The calculated value may be normalized by dividing through the number of pairs of vertices not including v, i.e. $(n-1)(n-2)$ for directed graphs. Analogously to the vertex betweenness, the betweenness centrality of edges is calculated as the number of shortest paths among all the possible vertex couples that pass through the given edge. Vertices and edges with high centrality values are assumed to be crucial for the graph connectivity and, thereby, shall be monitored in potentially malfunctioning brains. Neurodegenerative disease removing important neurons and synaptic connections leads to both anatomically and functionally disconnected clusters. The most important edges, called bridges, are particularly important for decreasing the average path length among neurons in a network, and for speeding up the information transmission.

Computationally, betweenness centrality of all vertices in $G(V,E)$ requires calculation of the shortest paths between all pairs of vertices. Inhibitory and excitatory cells, as well as synaptic weights that change versus time depending on LTP and LTD, should not be taken into account when calculating neurons' betweenness. Thus, for finding shortest paths of a binary directed graph with positive edge (but with no cycles), Floyd-Warshall algorithm (also known as Floyd's algorithm, Roy-Warshall algorithm, Roy-Floyd algorithm) is deployed. The algorithm compares all possible paths through the graph between each pair of vertices. An execution finds the lengths of the shortest paths between all pairs of vertices, though it does not return details of the paths themselves. When applied to local *Caenorhabditis elegans* subnetwork, Floyd-Warshall algorithm visualizes the network as shown in Fig. 3(a). Most



(a)



(b)

Fig. 3. (a) Representation of neurons according to the node betweenness centrality criterion. Red markers identify the most influential cells in the region. Death of these cells would produce two weakly connected- or even disconnected clusters which is further ruinous in terms of both anatomical and functional connectivity. (b) Representation of neurons according to the edge betweenness centrality criterion. Similarly, red lines denote crucial synaptic junctions which disunity produce harmful effects.

important neurons are again colored in red as ones that are either to be monitored in order to prevent disorders producing malfunctioning network, or whose $[Ca^{2+}]_i$ is to be regulated in case disorder has already took place. For analyzed network consisting of 131 frontal neurons, 4 cells are identified as essential in preserving the proper communication. In general, *the higher the number of important neurons, the lower the network robustness*.

Distance matrix that results the Floyd-Warshall algorithm execution provide calculation of average path length, l_{avg}, also one of the most important and basic graph theory measures. Short average path length is basically what makes the brain efficient in information transfer from one part to the other. It is also a function of degree (e.g. for a random network $l_{avg} \approx \log(n)/\log(< k >)$) ($< k >$ is the average degree). In brains with Alzheimer's disease, one can typically observe an increase of average path length, as described by Stam et. al in [15]. The average path length is calculated as

$$l_{avg} = \frac{1}{N(N-1)} \sum_{i \neq j} d(i,j),$$

where $d(i,j)$ denotes the shortest distance between i and j, and $d(i,j) = 0$ if j cannot be reached from i. For network analyzed here $l_{avg} = 3.1277$.

The Girvan-Newman algorithm extends the definition of node betweenness to the case of edges, defining the edge betweenness as the number of shortest paths between pairs of nodes that run along it. Edge betweenness may be interested in this study as a way of detecting important synapses. However, preserving those synapses is not trivial since execution of Girvan-Newman algorithm does not return details on the source and sink vertices, or equivalently the pre- and post-synaptic neurons. Therefore, information about the specified cell, whose $[Ca^{2+}]_i$ should be regularly kept, is missing. Nonetheless, it would not be superfluous if important synapses are visualized (see Fig. 3(b)).

B. Clustering criterion

The cluster coefficient of a vertex, $\gamma(v)$, is defined in [16]. As a measure of the degree to which vertices in a $G(V,E)$ tend to cluster together, it indicates how many connections are maintained between a vertex's neighbours. In particular scenario with directed graph on board, neighbours are all those neurons that are connected through an incoming or outgoing synaptic connections to the central neuron v. The ratio of actually existing connections between the neighboring neurons and the maximal number of such connections possible defines the neuron's cluster coefficient. If neuron does not have any

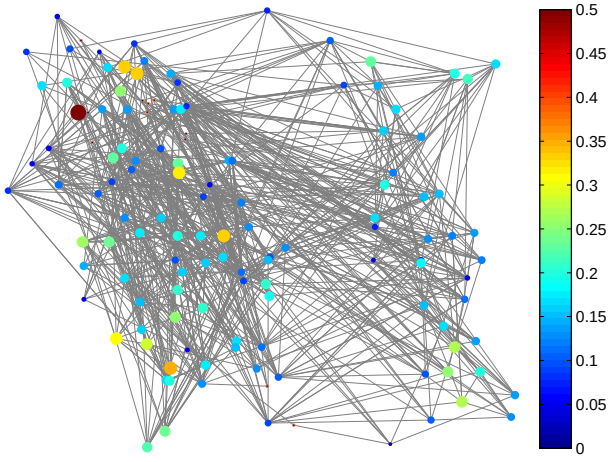


Fig. 4. Representation of neurons according to the clustering coefficient criterion. Red colored cell points to a region consisting of neurons that mutually share structural connections and function in a closely related manner.

neighbors, then $\gamma(v) = 0$, by convention. The average of the cluster coefficients determines the cluster coefficient of the $G(V,E)$. Albeit $\gamma(v)$ does not provide the information about the number or size of clustering groups, and only captures local connectivity patterns involving the direct neighbours of the central vertex [12], high $\gamma(v)$ points to a neuronal region consisting of groups of units that mutually share structural connections and function in a closely related manner. Accordingly, high $\gamma(v)$s set candidates whose controlled $[Ca^{2+}]_i$ properly regulates communication within the entire cluster (see Fig. 4). Note that cluster-based strategy might be destructive communication-wise, since $\gamma(v)$ can candidate an inhibitory cell which calcium is to be controlled.

V. Conclusion

This paper elaborates a significance of intracellular calcium concentration, $[Ca^{2+}]_i$, in performance of neuronal communication. High concentration levels usually lead to synchronized neuron firings, or equivalently, synaptic strengthening, whereas moderate levels lead to unsynchronized firings, or synaptic weakening. This paper also helps in development of an effective graph-based strategy of neurons selection whose electrical properties should be altered in order to evoke certain changes in $[Ca^{2+}]_i$ and communication abilities. Neuronal network is first described as graph and then quantitatively characterized through the adjacency matrix (AM). Specified graph theory concepts are utilized in formation of adequate strategy assisting in diagnosis and treatment of neuronal diseases. A set of graph theory measures that might have certain relevance for neuroscience applications is stated and deployed in this study to create a visualization tool that helps in selection of targeted neurons whose $[Ca^{2+}]_i$ is to be controlled. In such a way, problems in aging brains, strongly noticeable in brains with neurodegenerative diseases such as Alzheimer's disease, can be overcome. However, computationally demanding handling of huge AM describing large-scale networks, that are more likely to be recorded with functional magnetic resonance imaging, electroencephalography and magnetoencephalography, is not analyzed in this paper. Since an adult human brain has more than 100 billion neurons, recorded subnetworks can contain up to a billion cells forming an AM which presumably is to be stored in segments due to the limited memory resources. Segments of data should probably be further processed in a parallel or sequential computation fashion.

References

[1] G. W. Arendash, J. Sanchez-Ramos, T. Mori, M. Mamcarz, X. Lin, M. Runfeldt, L. Wang, G. Zhang, V. Sava, J. Tan, and C. Cao, "Electromagnetic field treatment protects against and reverses cognitive impairment in Alzheimer's disease mice," *Journal of Alzheimer's Disease*, vol. 19, pp. 191–210, 2010.

[2] N. Dragicevic, P. C. Bradshaw, M. Mamcarz, X. Lin, L. Wang, C. Cao, and G. W. Arendash, "Long-term electromagnetic field treatment enhances brain mitochondrial function of both Alzheimer's transgenic mice and normal mice: a mechanism for electromagnetic field-induced cognitive benefit," *Neuroscience*, vol. 185, pp. 135–149, 2011.

[3] J. Juutilainen, A. Hoyto, T. Kumlin, and J. Naarala, "Review of possible modulation-dependent biological effects of radiofrequency fields," *Bioelectromagnetics*, vol. 32, pp. 511–534, 2011.

[4] F. Mesiti and I. Balasingham, "Nanomachine-to-neuron communication interfaces for neuronal stimulation at nanoscale," *Journal on Selected Areas in Communication - Special Issue on Emerging Technologies in Communications*, In Press.

[5] M. Veletic, I. Balasingham, and Z. Babic, "On spectrum analysis for nanomachine-to-neuron communications," in *First International Black Sea Conference on Communications and Networking (BlackSeaCom)*, 2013, pp. 64–69.

[6] C. Koch, *Biophysics of Computation: Information Processing in Single Neurons*. Oxford University Press, Inc., 1999.

[7] F. A. Dodge and R. Rahamimoff, "Cooperative action of calcium ions in transmitter release at the neuromuscular junction," *J. Physiol.*, vol. 193, pp. 419–432, 1967.

[8] E. Bullmore and O. Sporns, "Complex brain networks: graph theoretical analysis of structural and functional systems," *Nature Reviews Neuroscience*, vol. 10, pp. 186–198, 2009.

[9] A. Barrat, M. Barthelemy, and A. Vespignani, *Dynamical processes on complex networks*. Cambridge University Press, 2008.

[10] M. Kaiser and C. C. Hilgetag, "Nonoptimal component placement, but short processing paths, due to long-distance projections in neural systems," *PLoS Computational Biology*, vol. 2, no. 7, 2006.

[11] (2013) Resources. [Online]. Available: http://www.biological-networks.org/?page_id=25

[12] O. Sporns, *Neuroscience database*. Springer US, 2003, ch. Graph Theory Methods for the Analysis of Neural Connectivity Patterns.

[13] (2013) Resources. [Online]. Available: http://www.alz.co.uk/

[14] L. C. Freeman, "A set of measures of centrality based upon betweenness," *Sociometry*, vol. 40, pp. 35–41, 1977.

[15] C. J. Stam, B. F. Jones, G. Nolte, M. Breakspear, and P. Scheltens, "Small-world networks and functional connectivity in alzheimer's disease," *Cerebral Cortex*, vol. 17, no. 1, pp. 92–99, 2007.

[16] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.

Use of 3D preoperative planning program in understanding of articular fracture treatment

M. Tomaževič*, A. Fischinger, A. Kristan*, D. Kreuh**, M. Cimerman*
* University Clinical Center Ljubljana, Department of Traumatology, Ljubljana, Slovenia
** Ekliptik l.t.d., Ljubljana, Slovenia

Abstract - Human joint is a complex 3D structure. When fractured, weight bearing articular surfaces of the joint have to be anatomically reduced and biomechanically stable coupled to the diaphysis with correct alignment of axis and rotation. To achieve this, soft tissue envelope should be traumatized as little as possible. We organized a course on preoperative planning of articular fractures with the use of computer 3D planning software, so surgeons could achieve the goal.

16 surgeons with different experiences of articular surgery, from residents to consultants, attended the course. First ordinary planning was done with 2D X-ray and CT images; later planning of the same cases was done with the use of 3D planning program. We compared chosen approach, reduction and use of implants in two fracture cases of knee joint. First was simple unicondylar and second complex bicondylar fracture of proximal tibia. Computer program used was EBS (Ekliptik, l.t.d.)

In the simple case plan remained the same in 61,5 % participants. In more complex case participants changed their plan in 100%.

3D preoperative planning is essential for understanding and treating articular fractures. 3D preoperative program is useful as an educational and as a working tool.

I. INTRODUCTION

Human joint is a complex 3D structure. It enables controlled movements in several planes. Due to movements even small incongruences can cause early posttraumatic arthrosis. In weight bearing joints, due to these incongruences, posttraumatic arthrosis can develop even sooner.[1], [2] Ligaments, joint capsule and muscles have an important role in stability of the joints. When there is a fracture in the joint we call these fractures articular fractures. Basic principle of treating articular fracture is to anatomically reduce the joint surface and to couple it in correct alignment to the diaphysis. In that way the length, axis and rotation of bony structure are assured and biomechanical bone preposition for the joint to work is done. [3] By achieving this, the stabilizing structures should not be additionally injured during the approach. To achieve all these goals the operation must be planned.[4] To plan correctly fracture should be understood, visualized and reduction planned. When reduction of the fracture fragments is planned, plan of fixation should be made. When all this steps are known the approach can be chosen and positioning of the patient can be planned. At the end closure of the wound and post-operative care can be planned. The key element to all this is to understand the fracture.

To be able to better understand the fracture, many computer programs for 3D visualization, segmentation and fracture reduction simulation were developed. [5]–[7]. But to our knowledge none of these programs could be completely operated by a surgeon. In cooperation between surgeons and computer engineers EBS software was developed called EBS (Ekliptik l.t.d.). EBS software uses DICOM images from CT and enables simulation of the classic X-ray, 3D reconstruction, fracture segmentation, fracture reduction and virtual fracture fixation. All the steps can be done on regular PC by the surgeon himself. Process from 3D reconstruction to virtual fracture fixation is shown with Fig 1-4. The procedure depends on the speed of computer, complexity of the fracture and on surgeon's skills from 10 – 60 minutes[8].

Preoperative planning course was organized, where 3D planning software was used. During the course we evaluated, whether software helps in understanding of the articular fractures. Our hypothesis was, that deeper understanding of the fracture will have influence on preoperative planning.

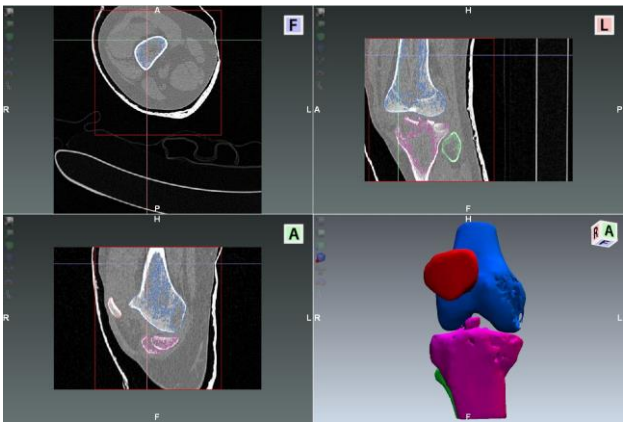


Figure 1: 3D model is built from DICOM images

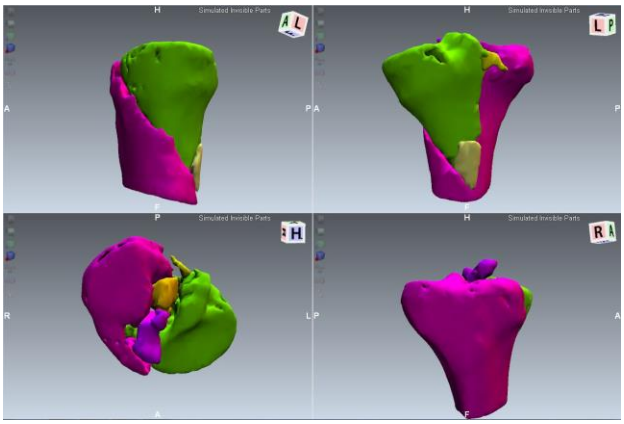


Figure 2: Fracture segmentation is done by the surgeon himself semiautomatic.

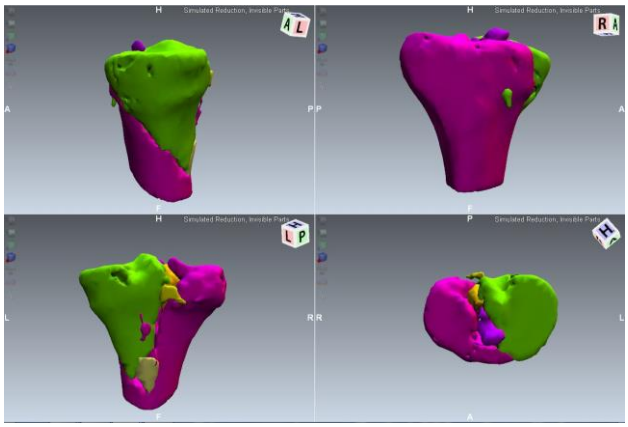


Figure 3: Fracture is virtually reduced.

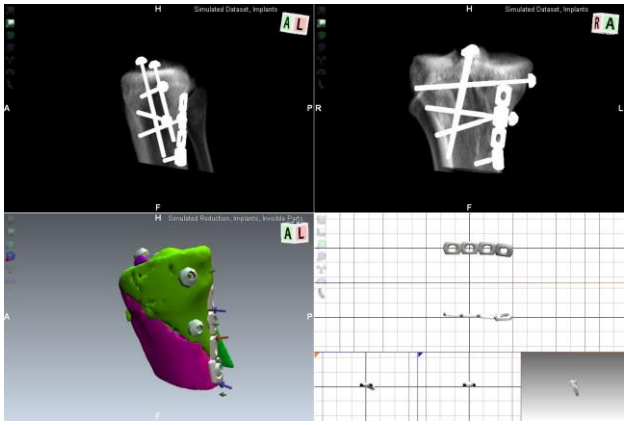


Figure 4: Virtual fixation of the fracture can be done. Contours of the plate can be seen and X-ray simulation can be done.

II. MATERIAL AND METHODS

Preoperative planning course with the use of EBS software was organized twice. 16 surgeons attended the courses. Each course was divided into two parts. First part was theoretical; it consisted of preoperative planning theory and instructions for the use of EBS software. Second part was practical; it consisted of real cases where preoperative planning was done. Planning for each case was done with the help of questionnaire. Attendants filled questionnaires with positioning of the patient, use of antibiotics, planned approach, reduction procedure, type of fixation, planned use of implants, and closing of the wound. Study of the fracture was first done with ordinary technique from classic X-ray images and 2D CT images. After this part first questionnaire was filled out. Later the study of the same cases was done with the help of EBS software. After EBS study second questionnaire was filled out. Evaluation of preoperative questionnaire was done whether there was a change in planned approach, position of the patient, planned reduction, type of fixation and choice of implants.

Due to unknown reason, three of the attendants did not return the questionnaire or only EBS questionnaire was filled out. For that reason we could evaluate only 13 questionnaires. Statistics was done only on approach, reduction type and implants where answers could be transformed in to nominal values; other answers were too long, subjective and descriptive and could not be transformed in to nominal values.

For comparison two cases of the same region were chosen; first one was simple unicondylar fracture of the proximal tibia and second was a more complex bicondylar fracture of the proximal tibia.

For 3D preoperative planning EBS software (Ekliptik l.t.d., Slovenia) was used.

Statistics was calculated with chi square test and due to small group likelihood ratio test was done. IBM SPSS Statistics 20.0 software was used.

III. RESULTS

Presented results show how preoperative planning with EBS software influences the understanding of the fracture. Better understanding of the fracture affects the choice of approach, choice of reduction technique, and choice of planned implants for fracture fixation. Results for the first and second case will be presented separately. First case was fracture of the proximal tibia according to Schatzker classification grade I, and second was fracture of the proximal tibia according to Schatzker classification grade IV.

A. Fracture of the proximal tibia Schatzker I

All the answers in first questionnaire were compared with the answers of the EBS questionnaire for each individual attendant. 5 attendants (38,5%) changed their plan at least at one point after planning with the use of EBS was done. With 7 attendants (61,5%) plan remained the same.

The planned operative approach remained the same with all the attendants. 12 decided for anterolateral approach and one for MIPO approach with ordinary and EBS preoperative planning. Statistics was done with chi square test, and the influence on approach was statistically insignificant at the level of $p = 0,05$ ($p = 1$).

There was a statistically important change in planned reduction type after the preoperative planning with EBS. Statistics was done with chi square test, and due to small number of attendants also with likelihood test and it was significant at the level of $p = 0,05$ ($p = 0,029$). The results are shown with graph in Fig. 5. Trend towards direct reduction is seen, arthrotomy (surgical opening of joint) is abandoned and indirect reduction remains the same.

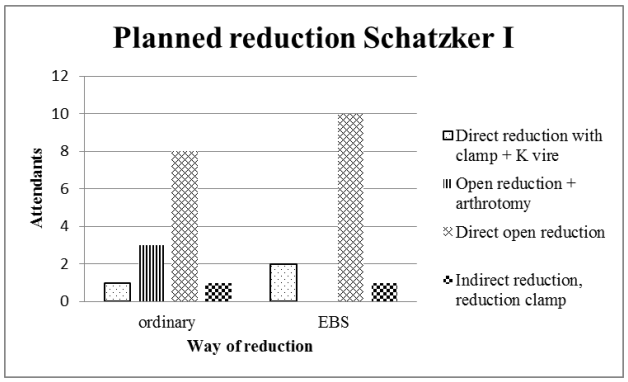


Figure 5: Planned way of reduction with ordinary and EBS preoperative planning in Schatzker I case

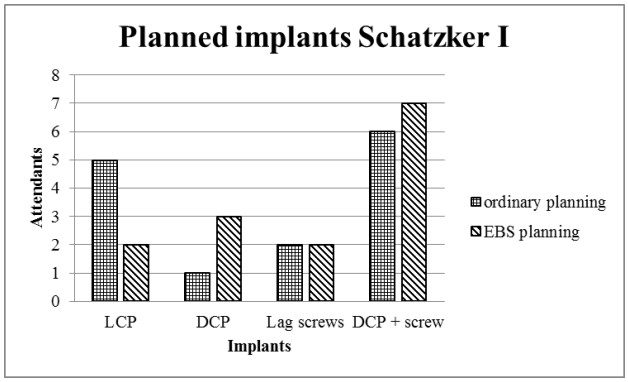


Figure 6: Planned use of implants with ordinary and EBS preoperative planning in Schatzker I case. (LCP – locking compression plate, DCP – dynamic compression plate)

Statistically important change in planned use of implants is also seen after the preoperative planning with EBS. Statistics was done with chi square test and likelihood test and it was significant at the level of p = 0 0,05 (p= 0,01). The results are shown in Fig. 6. After understanding of the fracture is better there is place for older not so expensive implants like dynamic compression plate (DCP).

B. Fracture of the proximal tibia Schatzker IV

All the answers in first questionnaire were compared with the answers of the EBS questionnaire for each individual attendant. 13 attendants (100%) changed their plan at least at one point after planning with the use of EBS was done.

There was a statistically important change in planned approach after the preoperative planning with EBS was done. Statistics was calculated with chi square test and likelihood ratio. It was significant at the level of p = 0,05 (p = 0,014). The results are shown in graph in Fig. 7. Planned medial approach is abandoned because the key for fracture reduction is at the posterior aspect of the tibia but the main fragment is on the medial side. All the attendants have decided for posteromedial approach with or without arthroscopy.

There was a statistically important change in planned reduction type after the preoperative planning with EBS. Statistics was calculated with chi square test and likelihood test. Change in planned way of reduction was significant at the level p = 0,05 (p = 0,000). As seen on Fig. 8 three ways of reduction that were planned with ordinary preoperative planning were completely abandoned after preoperative planning with the use of EBS software. There was a sliding mechanism of the fracture that needed a buttress plate to hold it in place and additional screws for stability were needed.

Statistically important change in planned use of implants is also seen after the preoperative planning with EBS software. Statistics was done with chi square test and likelihood test and it was significant at the level of p = 0 0,05 (p= 0,007). The results are shown in Fig. 5. It can be seen that after EBS preoperative planning choice of implants is a lot more uniform. A lot of implants that were planned with classic preoperative planning were completely abandoned after preoperative planning with the use of EBS software. There is less implants planned to be implanted and the implants are very low profile so that do not interfere with soft tissues.

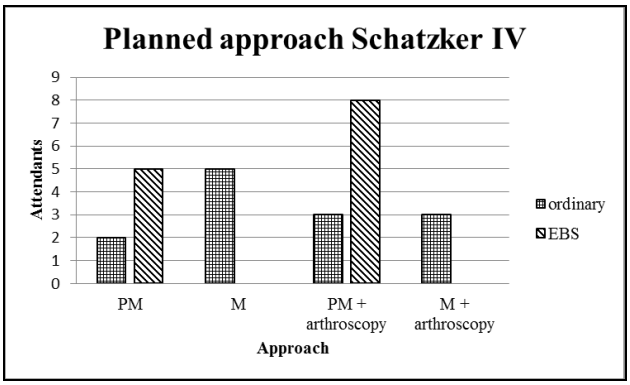


Figure 7: Planned approach with ordinary and EBS preoperative planning in Schatzker IV case (PM – posteromedial, M – medial)

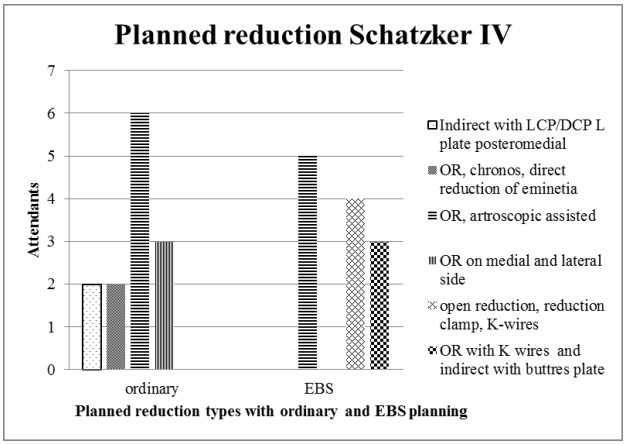


Figure 8: Planned way of reduction with ordinary and EBS preoperative planning in Schatzker IV case (OR – open reduction, Chronos – bone substitute)

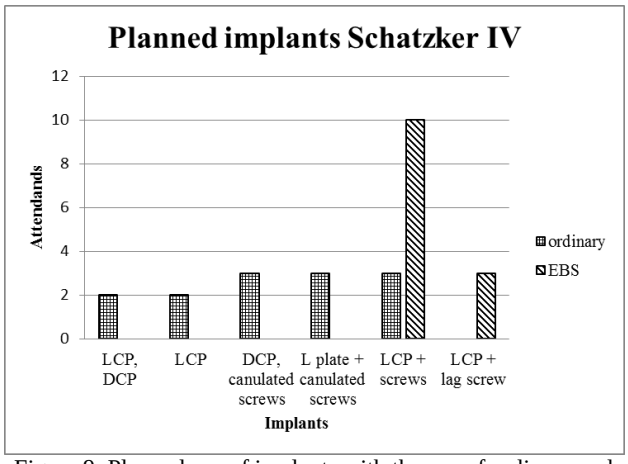


Figure 9: Planned use of implants with the use of ordinary and EBS preoperative planning (LCP – locking compression plate, DCP – dynamic compression plate, L plate – DCP in shape of an L)

C. *Comments on results*

It was expected that with easier Schatzker I case there will not be any statistically important differences. As it shows even in simple case the differences in preoperative planning exists. This means that even in simple case 3D image and fracture plane imaging helps to better understand the fracture.

It is seen that in more complex case the choice of treatment was a lot more uniformly spread among the attendants after the preoperative planning with EBS software. We assume that is because fracture understanding is easier when 3D model is seen and all the fracture planes are visible. Of course there is never just one solution to the problem, but usually when there are too many it is difficult to decide which one is good and which is not.

IV. CONLUSION

Fracture treatments are taught by classifying the fractures and according to fracture classification choose the appropriate approach and fixation. Unfortunately most of the classifications are based on 2D imaging. With 3D imaging understanding of the fractures is more profound. Until now 3D imaging was difficult to use in a preoperative planning. There was no appropriate 3D preoperative planning software. Making 3D solid models took too much time and it was too expensive.

With 3D preoperative planning software an important step toward understanding of the fracture patterns and fixation was made. By doing this study we proofed that 3D preoperative planning changes the thinking pattern about the fractures and improves understanding of the appropriate fracture fixation.

By significant changes in preoperative plan with the use of 3D preoperative planning program we proofed that 3D preoperative planning software changes understanding of the articular fractures in a better way.

3D preoperative program showed useful as an educational and as a working tool by making users to better understand the fracture.

REFERENCES

[1] T. D. Brown, D. D. Anderson, J. V. Nepola, R. J. Singerman, D. R. Pedersen, in R. A. Brand, „Contact stress aberrations following imprecise reduction of simple tibial plateau fractures“, *J. Orthop. Res. Off. Publ. Orthop. Res. Soc.*, let 6, št 6, str 851–862, 1988.

[2] T. Trumble, C. H. Allan, J. Miyano, J. M. Clark, S. Ott, D. E. Jones, P. Fernacola, M. Magnusson, in A. Tencer, „A preliminary study of joint surface changes after an intraarticular fracture: a sheep model of a tibia fracture with weight bearing after internal fixation“, *J. Orthop. Trauma*, let 15, št 5, str 326–332, jul 2001.

[3] R. W. Bucholz, J. D. Heckman, C. M. Court-Brown, in P. Tornetta, Ur, *Rockwood and Green's Fractures in Adults: Two Volumes Plus Integrated Content Website (Rockwood, Green, and Wilkins' Fractures) (Fractures in Adults*, Seventh. Lippincott Williams & Wilkins, 2009.

[4] J. Mast, R. Jakob, in R. Ganz, *Planning and reduction technique in fracture surgery*. Berlin; New York: Springer-Verlag, 1989.

[5] M. Cimerman in A. Kristan, „Preoperative planning in pelvic and acetabular surgery: the value of advanced computerised planning modules“, *Injury*, let 38, št 4, str 442–449, apr 2007.

[6] M. Citak, M. J. Gardner, D. Kendoff, S. Tarte, C. Krettek, L.-P. Nolte, in T. Hüfner, „Virtual 3D planning of acetabular fracture reduction“, *J. Orthop. Res. Off. Publ. Orthop. Res. Soc.*, let 26, št 4, str 547–552, apr 2008.

[7] C. Dahlen in H. Zwipp, „[Computer-assisted surgical planning. 3-D software for the PC]“, *Unfallchirurg*, let 104, št 6, str 466–479, jun 2001.

[8] N. Pallikarakis in P. D. Bamidis, *XII Mediterranean Conference on Medical and Biological Engineering and Computing 2010: MEDICON 2010, 27-30 May 2010, Chalkidiki, Greece*. Springer, 2010.

Education of cosmetic device users in Slovenia

Špela Elizabeta Soklič* and Jure Jelenc **
* Stane Žagar Primary School, Kranj, Slovenia
** Iskra Medical d.o.o., Ljubljana, Slovenia
jure@iskramedical.eu

Abstract - The education of cosmetic device users in Slovenia and across Europe is diverse. Although their level of knowledge varies, almost all users are allowed to perform the same cosmetic procedures. In Slovenia, all major stakeholders in the field of cosmetics are interested in increasing the level of knowledge of all participants, since knowledge and experience is the best assurance for safety and efficacy as well as customer satisfaction. This article reviews the education system and Slovenian organizations providing formal education in the discipline. The tremendous rate of biomedical innovation in cosmetic devices forces experts to get novel knowledge through education outside the government-regulated framework. Informal education provides a fast and efficient way to educate users in new procedures, but requires user's critical thinking since it is often commercially oriented. The rate of biomedical innovations implanted in the novel cosmetic procedures represents challenges to both formal and informal education. Lecturers face students' questions about novel procedures available on the market on a daily basis. For them, a big challenge is to be continuously acquainted with the new biomedical innovations implemented in the field so that they can provide evidence-based answers.

Keywords: cosmetic, beauty, aesthetic, device, user, school, faculty

I. INTRODUCTION

Cosmeticians practice a broad range of cosmetic procedures on the skin, face, nails and lips [1]. The English language refers to them with multiple expressions that are not necessarily synonyms: aestheticians, estheticians, cosmetic therapists, cosmetologists, beauticians, beauty practitioners, beauty therapists… The general scope of their activities is treating non-medical skin problems that are limited to the upper layer of skin (the epidermis) [2]. Even though in practice they employ therapeutic modalities that are known to penetrate through all layers of skin. Prime examples of such modalities are electrical current therapies, light therapies, ultrasound therapies, and application of creams optimized to enhance intradermal delivery. However, cosmeticians in Slovenia generally do not use technology that would allow them to physically penetrate the skin with the device itself.

The education of cosmeticians in Slovenia is diverse. Although their level of knowledge varies, the majority of professionals are allowed to perform the same cosmetic procedures. In the following sections, we will review the education system in Slovenia (Fig 1) and describe organizations that provide formal education in the discipline.

II. EDUCATION IN COSMETICS IN SLOVENIA

A. FORMAL EDUCATION IN COSMETICS IN SLOVENIA

Compulsory basic education consists of a nine-year primary school attended by pupils aged six to fifteen years [3]. After completing primary school, students may continue their education at any upper secondary school [3]. Education in cosmetics is possible at the second education level and above. Organizations that offer education in cosmetics are listed in Table 1. Post-compulsory education begins with upper secondary education. It is divided into general upper secondary education (4 years), technical upper secondary education (4-5 years) and vocational upper secondary education (2-4 years). General upper secondary (Gymnasium) schools provide general education and are aimed at pupils who plan to continue with tertiary education. Upper secondary vocational and technical education and training are divided into broad technical/vocational fields, within which several courses are available [3]. One of them is a 4-year cosmetic therapist course, which is also offered as a one to two year adult requalification program for adults that have already finished other secondary education programs.

Adults, who have acquired a secondary vocational qualification and have at least 3 years of work experience, may acquire a secondary technical education in an alternative way by taking a master craftsman, foreman or managerial examination with the competent association. In the field of cosmetics, the competent authority for the master craftsman exam is the Chamber of Craft and Small Business of Slovenia with the program for Master craftsman of cosmetics. Several competent associations offer managerial examinations for Massage therapists, Manicure specialists, Pedicure specialists and Make-up artists. The duration of the education needed to achieve these qualifications varies from 2 weeks to half a year.

Tertiary education in Slovenia comprises higher vocational college education and higher professional and academic education [4]. The short-cycle higher vocational college education has become a special part of the tertiary education system. It lasts two years and focuses on acquisition of practical knowledge and skills

in specific occupations. The two-year short-cycle higher vocational college education allows students to acquire a degree in Cosmetics. There are several types of higher education institutions in Slovenia: universities, art academies and independent higher education institutions. Most of these institutions are now introducing the Bologna process [4]. Currently, the Higher School of Applied Sciences (VIST) offers a 3-year professional education, which gives the student a bachelor's degree in Cosmetics, and a second two-year academic education, which gives students a master's degree in Cosmetics. The Faculty of Pharmacology at the University of Ljubljana offers bachelor's and master's degrees in Cosmetology, but is excluded from Table 1 due to differences in aims of the study. At the moment, there is no doctoral study of cosmetics in Slovenia.

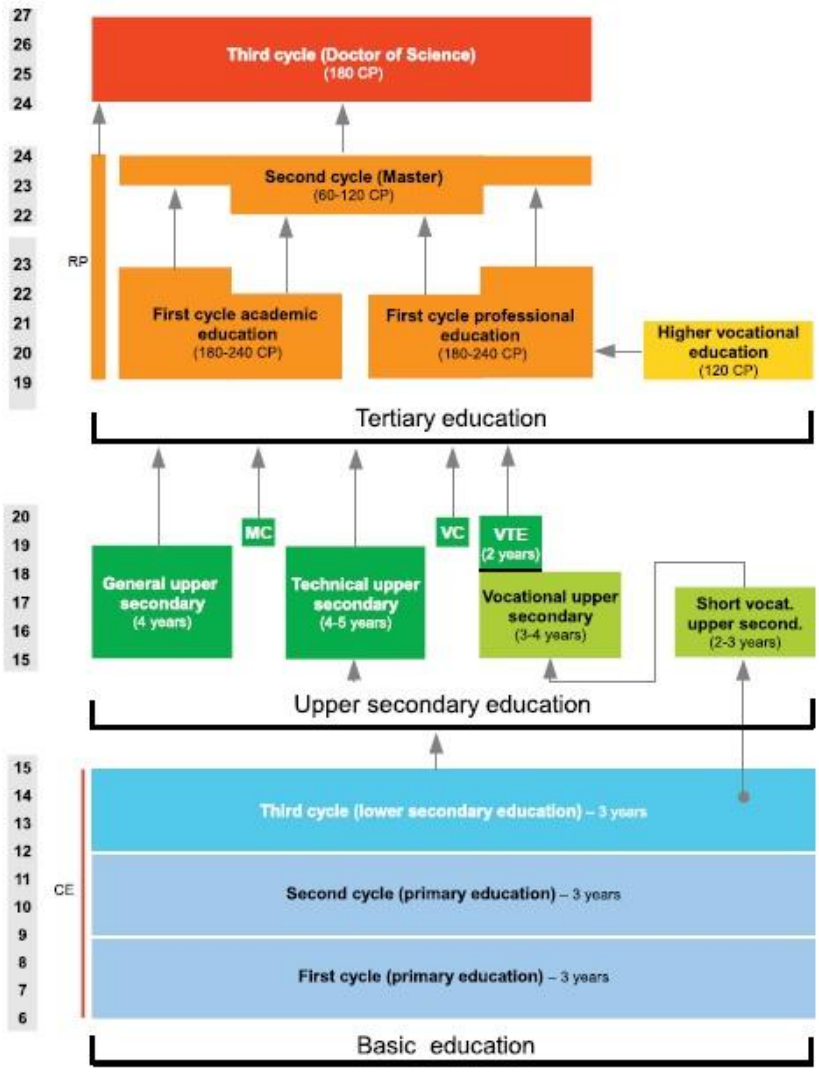


Fig 1: Formal education system in Slovenia. The figure has been modified from the reference [4].

Table 1: Organizations enrolling students in cosmetic education in the 2014/15 academic year combined with the professional/scientific titles earned by the students in the programs (unofficial translation).

UPPER SECONDARY EDUCATION	
Secondary Schools [5]	
Secondary School of Nursing and Chemistry Novo mesto	Beauty therapist
Secondary School Izola	Beauty therapist
Secondary School for Pharmacy, Cosmetics and Health Care Ljubljana	Beauty therapist
Secondary Medical School Maribor	Beauty therapist
Secondary Medical School Murska Sobota	Beauty therapist
Secondary Medical School Celje	Beauty therapist
Secondary School of Nursing and Chemistry Novo mesto	Beauty therapist
Secondary School Izola	Beauty therapist
Secondary School for Pharmacy, Cosmetics and Health Care Ljubljana	Beauty therapist
Secondary Medical School Maribor	Beauty therapist
Secondary Medical School Murska Sobota	Beauty therapist
Secondary Schools-Adult education [6]	
Zavod za izobraževanje in razvoj AI – Akademija za inovativnost beauty therapist	Beauty therapist
College DOBA beauty therapist	Beauty therapist
Secondary Medical School Celje beauty therapist	Beauty therapist
Secondary School Izola beauty therapist	Beauty therapist
Secondary Medical School Maribor beauty therapist	Beauty therapist
Cene Štupar Public Institute – Ljubljana, Center for Education beauty therapist	Beauty therapist

TERTIARY EDUCATION	
Higher vocational college education [7], [8]	
School Center of Novo mesto, Vocational College	Beauty technologist
Vocational College DOBA Maribor	Beauty technologist
Izobraževalno-razvojni zavod IZRAZ Vocational College	Beauty technologist
Vocational College for Cosmetics and Wellness Ljubljana	Beauty technologist
School Center of Novo mesto, Vocational College	Beauty technologist
Vocational College DOBA Maribor	Beauty technologist
First cycle professional education [9]	
Higher school of applied sciences (VIST) Ljubljana	Bachelor of Cosmetics
Second cycle education (Masters) [9]	
Higher school of applied sciences (VIST) Ljubljana	Master of Cosmetics

B. INFORMAL EDUCATION IN COSMETICS IN SLOVENIA

The tremendous rate of biomedical innovation in cosmetic products and cosmetic devices forces users and scholars to acquire novel knowledge. Like in every discipline, novel discoveries are first presented to experts in scientific journals, but are generally introduced to end users through informal education. A number of currently omnipresent devices in the business are the result of biomedical innovations over the last decades. Examples of these, now routinely used, technologies include light-based devices for hair removal and skin rejuvenation, low-frequency ultrasound and high-intensity ultrasound for local fat tissue reduction, radiofrequency devices for skin rejuvenation and wrinkle reduction, delivery vehicles and nanotechnology in cosmetics products. However, new methods pose new risks of complications during and after the procedures [10]. Most of these methods were invented and introduces to the cosmetic procedures only recently; therefore majority of cosmetic professionals gained no information about them during their formal education. Informal education provides a fast and efficient way to educate users in new procedures. Key stakeholders organizing informal education in Slovenia are the Chamber of craft and small business of Slovenia, education institutes, and companies that sell cosmetic products. Common forms of informal education include congresses, fairs, journals, invited lectures, and seminars with hands-on experiences. The majority of these events are at least partly sponsored by companies; therefore they require user's critical thinking since they can be commercially oriented. In Slovenia all major stakeholders in the field of cosmetics are interested in increasing the level of knowledge of all participants, since knowledge and experience is the best assurance for safety and efficacy as well as customer satisfaction.

III. DISCUSSION

Formal education in cosmetics in Slovenia offers a variety of levels in cosmetic education. The length of education varies from a few weeks to nine years for students that finished 4-year secondary education, 3-year bachelor education and 2-year master's degree. At the moment, the first generation of master students is finishing their education. But in the cosmetic saloons, private medical offices, spas and wellness centers there is a number of cosmetics professional with tertiary education

and an even greater number of cosmetic professionals with secondary school education. To the authors' knowledge, there is no published data on professional knowledge and expertise actually acquired by students in different levels of education or different educational organizations at the same level. At the secondary school level, the title of beauty therapist is granted after the student successfully passes a nation-wide exam. However, the standardized, practical exams are led and graded by teachers from the school that the student attended. Therefore, the possibility to compare results from different schools by these exams is severely limited.

Students wanting to enroll in cosmetic programs, organizations looking to hire new professionals and, most importantly, the end customers are limited in obtaining information about the quality of knowledge and expertise gained by students in formal cosmetic education programs.

To start their own private cosmetic practice, a cosmetician needs not only a formal education, but also a license (craft license) from the Chamber of Craft and Small Business of Slovenia [11]. However, the same license is issued for all formal education levels in cosmetics. Therefore, despite big differences in the level and duration of their education, all professionals are allowed to employ the same methods in their work.

The use of novel technologies is usually supported by informal education, which provides a significant amount of knowledge to both the lecturers in formal education and to the cosmetic professionals. The rate of biomedical innovations introduced into cosmetics is forcing lecturers and organizers of informal education to cooperate. The lecturer should be careful to accept only evidence-based claims as a valid answer to their questions. When in any doubt, all cosmetic professionals should gain the expert opinion of an experienced medical doctor to assure safety of their customers. Supplying cosmetic professionals with state-of-the-art, scientifically sound knowledge is crucial if cosmetic education is to gain the respect it deserves.

IV. Conlusion

Formal education for cosmetic professionals in Slovenia is diverse, ranging from secondary education level to a tertiary Master of Cosmetics program. The rate of biomedical innovations introduced into cosmetics forces lecturers and organizers of informal education to cooperate. All major stakeholders in the field of cosmetics are interested in increasing the level of knowledge of all participants, since knowledge and experience are the best assurance for safety and efficacy as well as customer satisfaction.

CONFLICT OF INTEREST

Jure Jelenc currently serves as a Head of research in Iskra Medical, a company that produces medical and cosmetic devices. As such, he is frequently involved in informal education of cosmetic device users. He also serves as a guest lecturer at organizations offering formal cosmetic education.

References

[1] A. Shai, H. I. Maibach, and R. Baran, *Handbook of cosmetic skin care*. London; Malden, MA, USA: Martin Dunitz ; Distributed in the U.S. by Blackwell Science, 2001.

[2] M. Lees and D. Young, "The Esthetician's Role in Skin Care," in *Evaluation and Treatment of the Aging Face*, M. L. Elson, Ed. New York, NY: Springer US, 1995, pp. 261–271.

[3] "Slovenia:Organisation of General Upper Secondary Education - Eurypedia." [Online]. Available: https://webgate.ec.europa.eu/fpfis/mwikis/eurydice/index.php/Slovenia:Organisation_of_General_Upper_Secondary_Education. [Accessed: 14-Feb-2014].

[4] N. Pajnič, P. Vilfan, and M. Širok, "Higher education in Slovenia." CMEPIUS, Centre of the Republic of Slovenia for Mobility and European Educational and Training Programmes, 2011.

[5] "Vpis v srednje šole | Ministrstvo za izobraževanje, znanost in šport." [Online]. Available: http://www.mizs.gov.si/si/delovna_podrocja/direktorat_za_srednje_in_visje_solstvo_ter_izobrazevanje_odraslih/srednjesolsko_izobrazevanje/vpis_v_srednje_sole/. [Accessed: 14-Feb-2014].

[6] "Srednješolski izobraževalni programi | Ministrstvo za izobraževanje, znanost in šport." [Online]. Available: http://www.mizs.gov.si/si/delovna_podrocja/direktorat_za_srednje_in_visje_solstvo_ter_izobrazevanje_odraslih/srednjesolsko_izobrazevanje/srednjesolski_izobrazevalni_programi/. [Accessed: 14-Feb-2014].

[7] "Vpis | Ministrstvo za izobraževanje, znanost in šport." [Online]. Available: http://www.mizs.gov.si/si/delovna_podrocja/direktorat_za_srednje_in_visje_solstvo_ter_izobrazevanje_odraslih/visjesolsko_izobrazevanje/vpis/. [Accessed: 14-Feb-2014].

[8] M. Kirsch and Y. Beernaert, *Short cycle higher education in Europe: level 5: the missing link*. EURASHE, 2011.

[9] "Razpis za vpis javni in koncesionirani | Ministrstvo za izobraževanje, znanost in šport." [Online]. Available: http://www.mizs.gov.si/si/delovna_podrocja/direktorat_za_visoko_solstvo/sektor_za_visoko_solstvo/razpis_za_vpis/razpis_za_vpis_javni_in_koncesionirani/#c20128. [Accessed: 14-Feb-2014].

[10] H. J. Brody, R. G. Geronemus, and P. K. Farris, "Beauty versus medicine: the nonphysician practice of dermatologic surgery," *Dermatol. Surg.*, vol. 29, no. 4, pp. 319–324, 2003.

[11] V. Franca, E. Zirenstein, Z. Vodnik, A. Trnavčevič, M. Ruzzier, and A. Arzenšek, *Deregulacija poklicev*. Planet GV d.o.o., 2012.

Telehealth using ECG Sensor and Accelerometer

Hristijan Gjoreski*, Aleksandra Rashkovska**, Simon Kozina*, Mitja Luštrek*, Matjaž Gams*,
* Department of Intelligent Systems, Jožef Stefan Institute, Ljubljana, Slovenia
** Department of Communication Systems, Jožef Stefan Institute, Ljubljana, Slovenia
{hristijan.gjoreski, aleksandra.rashkovska, simon.kozina, mitja.lustrek, matjaz.gams}@ijs.si

Abstract - The increasing size of the elderly population is driving the development of ambient assisted living systems and telehealth. The recognition of the user's everyday activities and detection of alarming situations are important components of such systems. Moreover, the monitoring of vital signs, like the ECG, has a key role in telecare and telemonitoring systems. Therefore, in this paper we propose a system that monitors the user by combining an ECG sensor and two accelerometers. Our system recognizes the user's activities and detects falls using the accelerometer data. The ECG data is analyzed in order to extract relevant physiological signals: heart rate, respiration rate, etc. In order to improve the reliability and robustness of the system, the measured accelerometer signals can be combined with the ECG signal in order to detect anomalies in the user's behavior and heart-related problems. The proposed proof-of-concept system could contribute significantly to the quality, unobtrusiveness and robustness of the health care and patient safety.

I. INTRODUCTION

The world's population is aging rapidly, threatening to overwhelm the society's capacity to take care of its elderly members. The percentage of persons aged 65 or over in developed countries is projected to rise from 7.5% in 2009 to 16% in 2050 [1]. This is driving the development of innovative ambient assisted living (AAL) technologies to help the elderly live independently for longer and with minimal support from the working-age population [2, 3, 4]. To provide timely and proper assistance, AAL systems must understand the user's situation and context, making activity recognition (AR) an essential component [5, 6, 7]. Fall detection (FD) is an important component of many AAL systems because almost half of the hospitalizations of the elderly are caused by falls [8]. Fear of falling is an important cause for nursing home admission [9], and "the long lie" (not being able to get up and call for help) is an important predictor of death within six months [10].

The introduction of modern information and communication technologies (ICT), as support to medical activities, can increase the efficacy of the health care system and decrease its costs. Numerous studies have confirmed the benefit of the development of telecare systems [11, 12]. The proposed mobile approaches to telecare rely on body sensor networks that collect and manage recorded vital data [13].

This work was partly supported by the Slovene Human Resources Development and Scholarship funds and partly by the CHIRON project - ARTEMIS Joint Undertaking, under grant agreement No. 2009-1-100228. The authors acknowledge the financial support from the state budget by the Slovenian Research Agency under Grant P2-0095.

This paper presents a system that combines two sub-systems that correspond to the type of the sensor-data they analyze: (i) electrocardiograph (ECG) sub-system, and (ii) accelerometer (ACC) sub-system.

The ECG sub-system is based on Wireless Bipolar Body Electrode (WBBE) [14] that captures and preprocesses body surface potential signals [15]. An additional feature is the ability to extract hidden information from the body surface potential, like the respiration rate [16], or combine synchronized data from three sensors for reconstruction of the standard 12-lead ECG [17, 18].

The ACC sub-system uses two 3-axial ACCs attached on the user's chest and right thigh. By using data analysis techniques it recognizes user's activities and detect falls in real-time. The architecture of the sub-system is based on: (i) rules, which detect falls, recognize postures (static activities) and ensure the behavior of the sub-system is predictable and robust, and (ii) classifiers trained with machine learning (ML) algorithms, to recognize dynamic activities, for which the rules are not sufficiently accurate. The ACC sub-system was evaluated at the EvAAL-2013 AR competition [19] and awarded the first place. The evaluation was performed in a living lab using several criteria: recognition performance, user-acceptance, recognition delay, system installation complexity, and interoperability with other systems.

The rest of this paper is organized as follows. First, the design of the prototype system is given with a brief functional description. Next, the sensors and the methods for analysis of the sensor data are described. The operation of the system is presented in the Experiments section. Finally, the conclusions are summarized and some directions for future work are given.

II. SYSTEM DESCRIPTION

The current prototype system (see Fig.1) consists of an ECG sensor and two wearable ACCs. The ECG sensor is placed on an appropriate position on the user's chest. The ACCs are attached on the user's chest and the right thigh using elastic Velcro straps. The data analysis is performed on a laptop using the raw sensors data acquired through Bluetooth. However, for the future implementation of the system, we are considering a smartphone implementation and combining the chest ACC and ECG sensor into a single device. Moreover, this device is planned to be multifunctional [20, 21], and beside the ECG and ACC data, should also measure: body temperature, body humidity, skin resistance, light, etc.

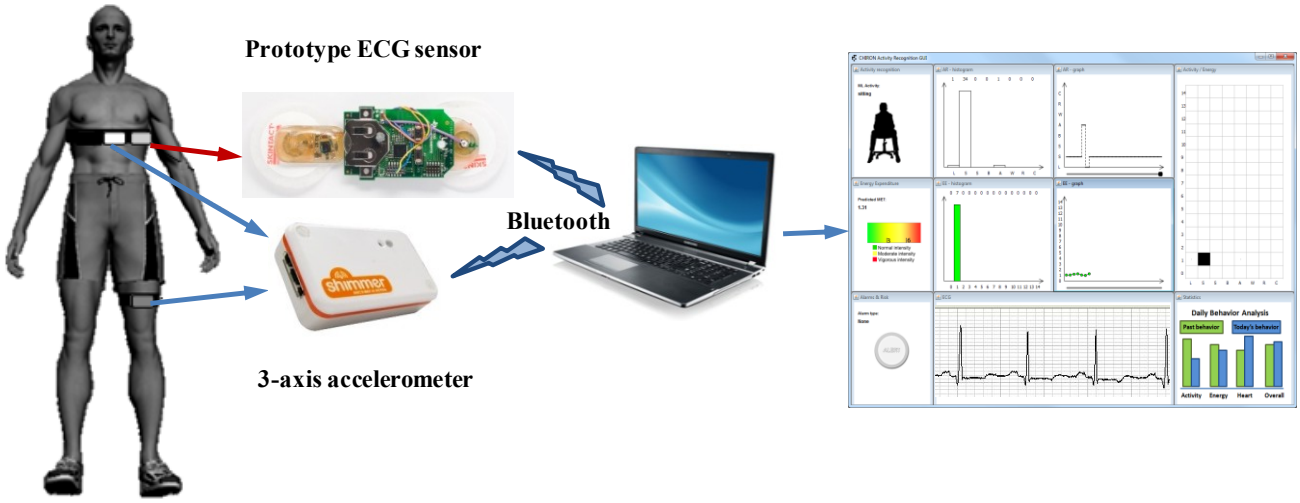


Figure 1. System design. Combination of the two solutions: ACC and ECG sub-systems.

A. *ECG sensor*

The ECG wireless body sensor designed from two self-adhesive electrodes positioned 5 cm apart, analog front-end, ultra-low-power (ULP) microcontroller and ULP Bluetooth V4.0 radio, ceramic chip antenna, and lithium coin battery, is shown in Fig.1 [22].

The analog front end is designed to suppress radio frequency interference. Signal amplification and filtering is made with a band-pass filter (0.15 Hz to 250 Hz, -3dB) before entering the microcontroller's 10-bit A/D converter with sampling frequency of 125 Hz for monitoring, and up to 1000 Hz for diagnostic and research cases. The 16-bit microcontroller (Texas Instruments, MSP430F2274) has 32 kB of FLASH, 1 kB of RAM and on-chip temperature sensor, all with an average power in active mode of 270 uA at 1 MHz. The microcontroller has enough peripheral lines to support other on-board sensors, like 3-axis accelerometer, skin temperature sensor and others. The Nordic Semiconductor's chip nRF8001 was used to support Bluetooth radio connectivity in peripheral (slave) mode. We have chosen lithium coin battery CR2032 as a power source for the body sensor. The nominal cell capacity is 220 mAh at discharge current of 200 uA, but due to pulse operation, a realistic capacity of 180 mAh can be expected.

B. *Accelerometer*

An accelerometer (ACC) is a sensor that measures the acceleration applied to the sensor and also the constant Earth's gravity. When the ACC is at rest, only Earth's gravity is measured. Using this information, the velocity of the sensor and also the sensor orientation can be estimated and therefore used for AR and FD. The platform used in this study has a 3-axis ACC, uses Bluetooth communication, and has 2 GB of storage, which is enough to store 3 months of sensor data for offline analysis. However, for the needs of this study, only real-time data acquisition and analysis was used. A potential, more in-depth, offline analysis is considered for future work.

III. METHODS

A. *ECG Data Analysis*

The measurement of ECG with sufficiently high amplitude and time resolution can support, besides standard heart rate analysis, also heart rate variability analysis [23], repolarization variability [24], ST-segment denivelation [25], detection of atrial fibrillation [26], characterization of arrhythmias, syncopes [27], sleep apneas, etc. Further, advanced approaches enable obtaining additional information from the recorded sensor signals. Two of them are listed below.

1) *Extraction of hidden information from WBBE*

It is known that body surface potentials include a lot of information that can be used for the estimation of personal medical status. As a successful case, we refer to the respiration rate extraction directly from WBBE [16]. We confirmed that several DLs on positions near the center of the chest provide adequate signals for ECG-derived respiration (EDR) algorithms that can reliably extract respiration rates from variations in the R-peak amplitudes. The EDR techniques, relevant to our work, are based on the observation that the positions of WBBEs on the chest surface move relative to the heart. Additionally, the transthoracic impedance varies as the lungs fill and empty. The proposed methodology is accurate enough for most practical cases and therefore useful for mobile health applications based on body sensors.

We have shown that a single WBBE is enough for the extraction of the respiration rate. However, if more electrodes are available, they can be applied to improve accuracy and robustness of the EDR results.

2) *Synchronized WBBEs on strategic positions*

The 12-lead ECG is the gold standard in cardiology and lies at the center of the decision pathway for the evaluation and management of patients. However, the conventional 12-lead ECG is obtained from ten electrodes connected with wires to a data collecting device. Its application therefore imposes obtrusion and has negative effects on patient's comfort. Moreover, due to the fact that the standard positions of the precordial electrodes are

often difficult to locate accurately, particularly in seriously damaged patients, the application of the conventional 12-lead ECG device can be impractical.

We have shown [17] that a small number of WBBEs − as few as three − can reliably reproduce a standard 12-lead ECG and are, therefore, suitable for real wireless applications. The proposed approach can be personalized in the sense that optimal positions of sensors and transformation matrices can be calculated for each individual [18].

B. *Accelerometer Data Analysis*

The ACC sub-system pipeline is shown in Fig.2. First, the sensors transmit the raw acceleration data over Bluetooth to the processing unit, i.e., laptop. The data from both sensors are then preprocessed: synchronized, filtered and segmented. Then the pipeline splits in two. On one side, the segmented data are transformed into feature vectors for the AR module, which recognizes the user's activity. On the other side, the FD module checks the acceleration for falls. If a fall pattern is recognized, the user's orientation is checked. If the orientation corresponds to lying, a fall is detected. Both the AR and FD modules are evaluating the user's situation every 250 milliseconds using the last 2 seconds of sensor data. For instance, if the current system time is denoted with t, the FD module evaluates fall events in the [t – 2 s, t – 1 s] interval, and the [t – 1 s, t s] interval is used to check if the user's orientation corresponds to lying. If the fall event is detected and the orientation is correct, the reported activity is falling, otherwise the reported activity is computed with the AR module in the [t – 2 s, t s] interval. The system thus reports the user's activity and detects falls with a two-second delay.

In the following sections, the AR and FD methods are briefly described. More details can be found in our previous work, [28, 29, 30] for AR, and [28, 31] for FD.

1) *Activity Recognition*

In the AR module, the activities are recognized by a three-level scheme [30]. The AR scheme was developed after empirical analysis of the data, which showed that some activities (such as cycling) are better recognized by a classifier trained only to distinguish that particular activity from the others. Therefore, on the first level, the feature vector is fed into a classifier trained by the Random Forest ML algorithm to distinguish cycling from the other activities. If the activity is not classified as cycling, the feature vector is passed to the second level, where the activities are recognized by rules. On this level, only the features that represent the sensor orientation the best are used (using component of the acceleration that corresponds to the gravity). The following activities are recognized at this level: sitting, lying, bending, and upright posture. If the recognized activity is the upright posture, the third level of AR is used to distinguish between standing and walking. The feature vector is fed into another Random Forest classifier, which is trained to separate these two activities.

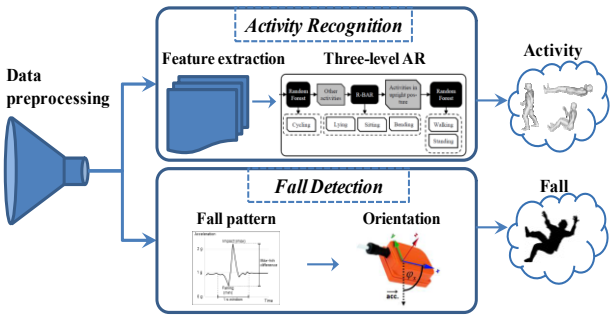


Figure 2. The data and recognition flow in the ACC sub-system.

2) *Fall Detection*

A typical acceleration pattern during a fall, measured by an ACC placed on the abdomen, is a decrease in acceleration followed by an increase [30]. This is because an ACC, when stationary, registers around 10 m/s^2 acceleration (denoted also as 1 g), and during free fall 0 m/s^2 acceleration. When a person starts falling, the acceleration decreases from 1 g to around 0.5 g (perfect free fall is never achieved). Upon the impact with the ground, a short strong increase in the acceleration is measured.

To detect fall patterns, we used the length of the acceleration vector to ignore the direction of the acceleration. The minimum and the maximum acceleration within a one-second window were measured. If the difference between them exceeded 1 g and the maximum came after the minimum, a fall pattern was found. We augmented the fall-pattern detection with the measurement of the user's orientation after a potential fall. We assumed that the orientation of the user's body after a fall couldn't be upright. Therefore, a fall was detected if a fall pattern was detected and the orientation in the next second was not upright.

IV. EXPERIMENTS

A. *ECG sensor performance results*

ECG signals from proximal electrodes have different appearance than the signals from standard ECG. An example is shown in Fig.3. The baseline is wandering and complete disappearance of the P, S and U waves can be observed. Note that the R and T waves could have opposite polarity, which is rarely seen in standard ECG leads. It is evident that the position of the electrodes has a significant impact on the measured ECG signal. Detailed analyses are possible only after significant preprocessing targeted to specific goals, e.g. heart rate determination, arrhythmia identification and characterization, ST segment denivelation, etc.

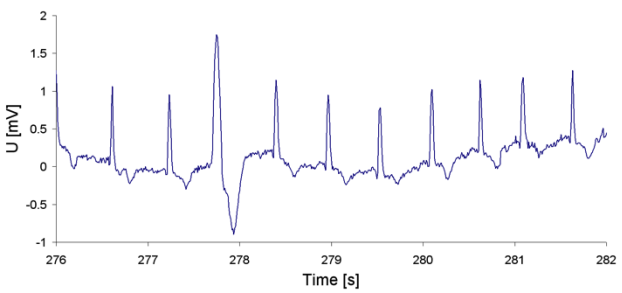


Figure 3. Raw ECG signal from electrodes at a distance of 5 cm. The person is walking around. The third heart beat is a ventricular extra systole with large amplitude and prolonged timing of the QRST complex.

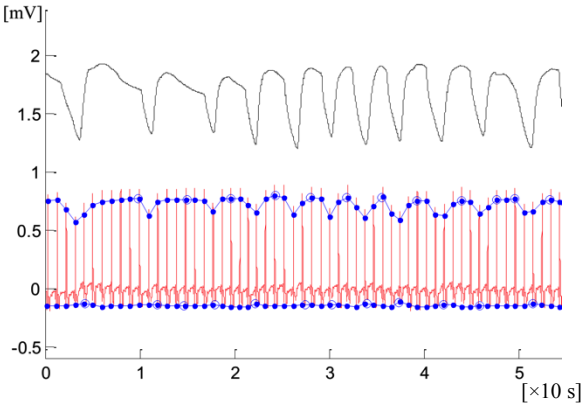


Figure 4. 53 seconds interval of an ECG (red) measured with a WBBE positioned in the chest center and 11 respiration intervals (black) measured by a thermistor near the front of the nose. 58 R-peaks (blue filled points) are detected and all 12 respiration interval are identified (blue circles).

For example, to determine the respiration rate, we found the best positions for the respiration body sensor [16] and proved that a single sensor of body surface potentials difference on proximal skin electrodes can be used for combined measurements of respiratory and cardiac activities (Fig.4).

B. Activity Recognition & Fall detection

The evaluation of the ACC sub-system was performed in order to check the recognition performance of the methods, using a pre-recorded dataset. The evaluation was performed on a complex, 90-minute scenario, recorded from 10 people. The scenario was designed to capture the real-life conditions of a person's behavior, although it was recorded in a laboratory.

Table I shows the performance of the ACC sub-system. The performance of the AR is high, achieving 99.04% F-measure score averaged over all activities. The performance of the FD shows that 93.3% of the falls were detected (recall value), and 66.7% of all the fall detections were actually falls (precision value), giving the final F-measure of 77.8%. The detailed FD results (Table II) show that the first event – tripping (quick uncontrolled fall) was detected each time (15 out of all 15 events). The next event, fainting, was detected 13 out of 15 times. The next two events were the non-fall events that are difficult to distinguish from the fast falls because of the high acceleration. Because the FD module also checks the user's orientation after a potential fall, it was able to distinguish quickly sitting on the chair from the falls. However, this was not the case for quickly lying in the bed (13 detections − which are false because the event is non-fall). For correct recognition of this event, additional information would be needed, e.g., user's location [31].

C. Concurrent ECG and Accelerometer Data

Fig.5 shows activity experiment with concurrent measurements from ECG and ACC sensors when the user is lying, sitting and walking.

ACC sensors give clear information for distinguishing walking from sitting or lying. One can also notice lower beat rates in lying and sitting compared to walking, which implies eventual lower energy expenditure. Detailed analysis of ECG indicates also that beat rate of lying is even lower that the one while sitting, which is supplementary information. With longer measurements, the long-term changes in the beat rate variability could give additional information about the time scale. Eventual extra systoles could point out a susceptibility to arrhythmia, etc.

TABLE II. ACC SUB-SYSTEM–AR AND FD RESULTS

Performance	*Activity Recognition*	*Fall Detection*
Recall	99.22%	93.33%
Precision	98.85%	66.67%
F-measure	99.04%	77.78%

TABLE I. FD DETAILED RESULTS

	Events	*Detected/All*
Fall events	Tripping	15/15
	Fainting	13/15
Non-fall events (false detections)	Quickly lying	13/15
	Quickly sitting	1/15
	Other	0

V. CONCLUSIONS

We showed that the fusion of ECG and ACC sensors data can provide supplementary information about the status of the monitored user. Based on this, it is possible to better understand the context of user's health state and activity and therefore better reason about his/her health and behavior status. The reasoning may include classification of the detected anomaly into several levels: (i) low-risk warning, e.g., higher heart-rate detected during sedentary activity, (ii) medium-risk warning, e.g., gait anomaly detected, (iii) high-risk warning e.g., very high heart-rate detected during longer period of time, and (iv) alarming situation detected that requires medical attention, e.g., the user has fallen and is lying without movement.

We are currently developing and testing a new multifunctional sensor node that consists of several sensors to obtain synchronized data about vital bio-signs and activities of the monitored users. For the final implementation of the system we plan three-phase

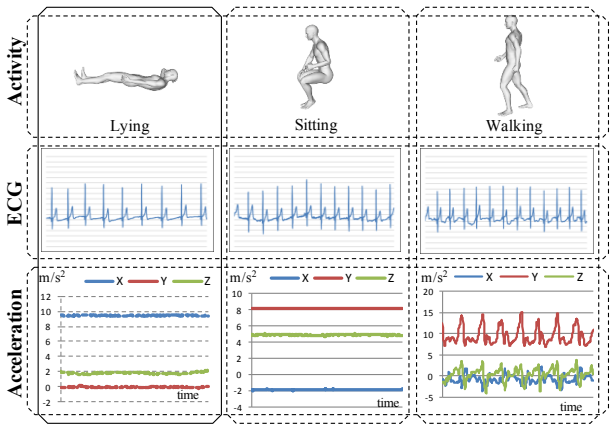


Figure 5. Concurrent ECG and ACC data.

monitoring and data analysis: (i) short-term behavior and health analysis - focusing on the last several minutes of data (alarming situations, falls, arrhythmias); (ii) medium-term behavior analysis - focusing on the past day (gait analysis); and (iii) long-term behavior analysis (daily/weekly anomalies) related to heart problems, less active days, etc.

REFERENCES

[1] United Nations 2009, World population ageing, Report

[2] A. Bourouis, M. Feham, and A. Bouchachia, "A new architecture of a ubiquitous health monitoring system: a prototype of cloud mobile health monitoring system," The Computing Research Repository, 2012.

[3] M. Luštrek, B. Kaluža, B. Cvetković, E. Dovgan, H. Gjoreski, V. Mirchevska, and M. Gams, "Confidence: ubiquitous care system to support independent living" DEMO at European Conference on Artificial Intelligence, 2012, pp. 1013-1014.

[4] B. Kaluža, B. Cvetković, E. Dovgan, H. Gjoreski, V. Mirchevska, M. Gams, and M. Luštrek, "A Multiagent care system to support independent living," International journal on artificial intelligence tools, 2014.

[5] D.A. Gregory, K. D. Anind, J. B. Peter, D. Nigel, S. Mark, and S. Pete, "Towards a better understanding of context and context-awareness," 1st International Symposium Handheld and Ubiquitous Computing, 1999, pp. 304-307.

[6] N. Vyas, J. Farringdon, D. Andre, and J. I. Stivoric, "Machine learning and sensor fusion for estimating continuous energy expenditure". Innovative Applications of Artificial Intelligence Conference, 2012, pp. 1613-1620.

[7] H. Gjoreski, B. Kaluža, M. Gams, R. Milić, and M. Luštrek. "Ensembles of multiple sensors for human energy expenditure estimation," Proceedings of the 2013 ACM international joint conference on Pervasive and Ubiquitous computing, Ubicomp, 2013, pp. 359-362.

[8] M. J. Hall, L. Fingerhut, and M. Heinen, "National Trend Data on Hospitalization of the Elderly for Injuries, 1979-2001. American Public Health Association (APHA), 2004.

[9] M. E. Tinetti and C. S. Williams, "Falls, Injuries Due to Falls, and the Risk of Admission to a Nursing Home," The New England Journal of Medicine, 1997, pp. 1279–1284.

[10] D. Wild, U. S. Nayak, and B. Isaacs, "How dangerous are falls in old people at home?," British Medical Journal (Clinical Research Edition), vol. 282, no. 6260, 1982, pp. 266–268.

[11] A.G. Ekeland, A. Bowes, and S. Flottorp, "Effectiveness of telemedicine: A systematic review of reviews," Int. J. Med. Inform., vol. 79, 2010, pp. 736-771.

[12] W. Chaboyer, L. Thalib, M. Foster, C. Ball, and B. Richards, "Predictors of Adverse Events in Patients After Discharge From the Intensive Care Unit," Am. J. Crit. Care, vol. 17, 2008, pp. 255-263.

[13] R. Trobec and U. Stanič, "Telehealth: A myth or reality?", Proceedings of MEET & GVS on the 34th International Convention MIPRO 2011, Opatija, Croatia, 2011, pp. 295-300.

[14] R. Trobec, M. Depolli, and V. Avbelj, "Wireless network of bipolar body electrodes," Proceedings of the 7th International Conference on Wireless On-demand Network Systems and Services, WONS 2010, pp. 145–149.

[15] R Trobec, "Computer analysis of multichannel ECG," Computers in Biology and Medicine, vol. 33 , pp. 215-226, 2003.

[16] R.Trobec, A. Rashkovska, and V. Avbelj, "Two proximal skin electrodes - a respiration rate body sensor," Sensors, vol. 12, 2012, pp. 13813-13828.

[17] R. Trobec and I. Tomašić, "Synthesis of the 12-lead electrocardiogram from differential leads," IEEE Trans. Inf. Technol. Biomed., vol. 15, 2011, pp. 615-621.

[18] I. Tomašić, S. Frljak, and R. Trobec, "Estimating the Universal Positions of Wireless Body Electrodes for Measuring Cardiac Electrical Activity", IEEE Transactions on Biomedical Engineering, vol. 60, 2013, pp. 3368-3374.

[19] EvAAL competition. http://evaal.aaloa.org/ [Accessed: November, 2013]

[20] R. Trobec, V. Avbelj, and A. Rashkovska, "Multi-functional Wireless Body Sensor", Proceedings of GLOBAL HEALTH 2012: The First Internacional Conference on Global Health Challenges, 2012, pp. 71-74.

[21] R. Trobec, V. Avbelj, and A. Rashkovska, "Multi-functionality of Wireless Body Sensors", Transactions on Internet Reserch, vol. 10 (1), 2014.

[22] K. Bregar and V. Avbelj, "Multi-Functional Wireless Body Sensor Analysis of Autonomy," MIPRO 2013, 35th International Convention, Opatija, Croatia, May 20-24, 2013, pp. 346-349.

[23] J.M. Kalisnik, V. Avbelj, R. Trobec, and B. Gersak, "Position-dependent changes in vagal modulation after coronary artery bypass grafting," Comput. biol. med., vol. 37, 2007, pp. 1404-1408.

[24] A. Rashkovska and V. Avbelj, "Signal processing methods for ST variability assessment in ECG," MIPRO 2013, 35th International Convention, Opatija, Croatia, May 20-24, 2013, pp. 355-358.

[25] A. Smrdel and F. Jager, "Automated detection of transient ST-segment episodes in 24h electrocardiograms," Med. biol. eng. comput., vol. 42, 2004, pp. 303-311.

[26] B. Geršak et al., "European experience of the convergent atrial fibrillation procedure: Multicenter outcomes in consecutive patients", J Thorac Cardiovasc Surg., Early publication Aug. 2013.

[27] V. Avbelj and R. Trobec, "A closer look at electrocardiographic P waves before and during spontaneous cardioinhibitory syncope," International journal of cardiology, vol. 166, 2013, pp. e59-e61.

[28] H. Gjoreski, M. Luštrek, and M. Gams, "Accelerometer Placement for Posture Recognition and Fall Detection," In The 7th International Conference on Intelligent Environments, 2011, pp. 47–54.

[29] S. Kozina, H. Gjoreski, M. Gams, and M. Luštrek,"Three-layer Activity Recognition Combining Domain Knowledge and Meta-classification," Journal of Medical and Biological Engineering, vol 33, no 4, 2013.

[30] S. Kozina, H. Gjoreski, M. Gams, and M. Luštrek, "Efficient Activity Recognition and Fall Detection Using Accelerometers," Evaluating AAL Systems Through Competitive Benchmarking Communications in Computer and Information Science, Volume 386, 2013, pp 13-23.

[31] H. Gjoreski, M. Luštrek, and M. Gams, "Context-Based Fall Detection using Inertial and Location Sensors," International Joint Conference on Ambient Intelligence, Lecture notes in computer science, 2012, pp. 1-16.

Children obesity treatment support with telemedicine

M. Kozlovszky*,**, K. Karóczkai**, B. Jókai*, B. Ruda*, D. Kacsuk*, Zs. Meixner*
* MTA SZTAKI/Laboratory of Parallel and Distributed Computing, Budapest, Hungary
** Óbuda University/John von Neumann Faculty of Informatics, Budapest, Hungary
{kozlovszky.miklos, jokai.balazs}@nik.uni-obuda.hu, {ruda.balazs, kacsuk.daniel, meixner.zsolt}@biotech.uni-obuda.hu, {karoczka}@sztaki.hu

Abstract - Obesity is considered an endemic within the whole European continent, just as in most part of the developed world. Obesity poses an increasing risk, childhood obesity has more than doubled in children and tripled in adolescents in the past 30 years. To assess the impact of obesity we need to take into account the severity of the problems that obesity might cause, as risk is usually interpreted as the product of probability of occurrence and severity. In the last few decades, increased occurrence of several morbidities was casually linked to obesity. Recently health problems targeting children have increasing trends in well developed countries. Our research is focusing on obesity related health problems in children, and our offered ICT solutions enables healthcare service providers to remotely monitor patient's (children's) activities and weight. The solution also helps to assign workout programs to the children according to the collected sensor data. The used workout programs base not only on widely used sport activities (etc. running or gymnastics), but cover non-conventional sport activities like playing with interactive game consoles at home. The system contains an Android based DAQ (Data Acquisition Device), sensors and a server side portal, which collects all measured patient information in the data center. The solution combines existing; self developed client and server side applications and frameworks, and commodity sensor solutions. The main goal of our work is to adapt the already existing monitoring solution to children. Kids behave and use such monitoring environments differently due to their different perception and knowledge level. They need different GUIs; they live different lifestyle as elderly patients. The communication lines of the system are compliant to the EU and national regulations. The developed system was tested and a prototype system was installed successfully in a real environment. Sensor data acquisition on the personal mobile device enables both end-users and care givers/professionals to provide better and more effective health monitoring and facilitate patient's treatment. The paper describes the internal architecture of the software solution and its main functionalities.

Keywords

Obesity treatment, health monitoring, telemedicine, sensor data acquisition

I. INTRODUCTION

Obesity is considered an endemic not only in Hungary but within whole European continent, just as in most part of the developed world. According to, 1.5 million Hungarian is definitely obese, whilst an additional 2.7 million can be considered overweight. Together, this is almost the half of the Hungarian population. To assess the risk of obesity we need to take into account the severity of the problems that obesity might cause, as risk is usually interpreted as the product of probability of occurrence and severity. In the last few decades, increased occurrence of several morbidities was casually linked to obesity, such as non-insulin dependent diabetes mellitus (NIDDM), stroke, ischemic heart disease (IHD), immunological and reproductive dysfunctions and certain neoplasms. From these, stroke and IHD is amongst the first three in the list of mortality rates in Hungary. These all confirm that obesity poses a significant risk, high enough to be interesting from the public health point of view as well. Early intervention is needed to prevent the onset of obesity and obesity-associated comorbidities, which intervention should be focused to be the most effective. Early, focused intervention presumes however an efficient screening method which can select those people who are prone to obesity even when they are healthy otherwise. This screening has many difficulties in practice, starting as early as how to measure obesity. In the frame of our national scale project a combined Diabetes & Obesity monitoring solution have been developed for children.

This paper shows how we have built up our remote patient monitoring environment using a client side software, gaming console to increase children's motivation to do exercises and the DrHealth portal.

II. SYSTEM OVERVIEW

Recently health problems targeting children have increasing trends in well developed countries. The platform is focusing on Diabetes and Obesity related health problems in children, enables healthcare service providers to remotely monitor patient's (children's) blood sugar level and weight. The solution is also help to assign workout programs to the children according to the

collected sensor data. Workout programs not only supporting widely used sport activities (etc. running or gymnastics), but cover non-conventional sport activities like playing with interactive game consoles at home. The system contains Android based DAQ (Data Acquisition Device) connected with external sensors, a gaming console and a server side portal, which collects all measured patient information in the central data center. The solution combines existing; self developed client and server side applications and frameworks and already developed sensor solutions. The main highlight of the project is to adapt the already existing monitoring solution to children. Kids behave and use such monitoring environments differently due to their different perception and knowledge level. They need different GUIs; they live different lifestyle as elderly patients. The communication lines of the system are compliant to the EU and national regulations. Security is a pivotal point of the system. Health records and patient sensor are transported in an encrypted way.

A. *DAQit - Mobile data acquisition framework*

DAQit Framework is a mobile data acquisition (DAQ) solution, which collects personalized health information of the end-user, store, analyze and visualize it on the smart device (DroidHUB) and optionally sends it towards to the datacenter (DrHealth) for further processing. The smart mobile device is capable to collect information from a large set of various wireless (Bluetooth, and WiFi) and wired (USB) sensors.

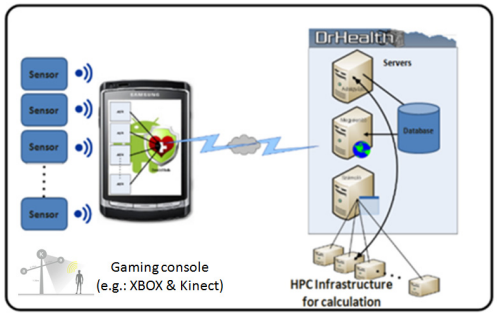


Figure 1. System overview

Embedded sensors of the mobile device provide additional useful status information (such as: user location, magnetic or noise level, acceleration, temperature, etc.). The user interface of our software solution is suitable for different skilled users, highly configurable and provides diary functionality to store information (can act as a diet log, or even can be used as a pain diary).

B. *DroidHub – DAQ on smart client devices*

DroidHub is an Android based mobile data acquisition solution implemented within the DAQit framework. DroidHub has the following basic software functionalities/features:

- Multi-language support (English, German, Italian, and Hungarian).
- Silent sensor DAQ mode (automatic data collection) via Bluetooth.
- Health status visualization (with statistics and data mining facility).
- Manual DAQ mode (optionally GUI initiated sensor data collection).
- Data compression and encryption during data transmission.
- Automatic sensor data pre-evaluation at the DroidHub.
- Full featured location aware mobile emergency alarm.
- Real-time sensor data forwarding to the central data collector server
- Store and forward mode to enable offline data collection

C. *DroidHUB - Sensor set*

Sensors can provide digital fingerprint of the patient's/person's psychophysical status and performance. A wide range of sensors is supported to collect detailed and accurate information about the patient's health status remotely:

• patient's blood glucose/sugar level

• patient's actual pulse, and oxygenization level

• patient's ECG signals

• patient's blood pressure

• patient movement within the house, with the usage of wall mounted sensors

• patient medication, with sensors of the medicament dispenser

• patient's eating habit, with sensors placed on the refrigerator

• patient activity (non-invasively rest/activity cycles, motor activity)

• patient's weight

• environmental information, about room temperature, humidity etc.



Figure 2. DroidHub sensor set and the GUI

We categorize sensors according to their data collection methods: active or passive type sensors. In our solution we are using a wide range of sensors (both passive and active) to collect detailed and accurate information about the patient remotely. Some of our

measurements are used for long-term monitoring (minutes-hours-days), others are just being able to provide snapshot measurements (e.g.: blood glucose level checking). Long term monitoring can become an invisible health data acquisition during patient activities.

Sensor name	Data collection type	Long-term monitoring	Snapshot like measurement
ECG holter	passive	yes	yes
Blood pressure holter	passive	yes	yes
Blood glucose meter	active	no	yes
patient's actual pulse	passive	yes	yes
Oxygenization level (SpO2)	passive	yes	yes
Simple ECG	active	no	yes
Blood pressure	active	no	yes
Movement (wall mounted sensors)	passive	yes	no
Medication (medicament dispenser)	passive/active	yes	no
Eating habit (sensors placed on the refrigerator)	passive	yes	no
Rest/activity cycle monitoring, gross motor activity	passive	yes	no
Weight	active	no	yes
Environmental information (room temperature, humidity etc.)	passive	yes	yes

Table 1. Sensor and measurement

1) Diary / reminder services

Besides pure data collection our DroidHub enables to do correlation analysis of stored data series. Reminders are intended to notify user in pre-defined appointments about data collection needs. This feature enables to track data in certain periods like measuring patient's blood glucose level on a regular basis.

D. Game console support - Training facility

Our solution is compatible with different type of game consoles such as the Wii[1] or XBOX 360 [2] equipped with the Kinect™ [3] sensor. Kinect is a motion sensing input device developed by Microsoft for the Xbox 360 video game console and Windows PC. It has a built-in RGB camera (640×480 pixels @ 30 Hz), an IR emitter, plus an IR depth sensor, a multi-array microphone, and a 3D accelerometer. It has an operational range between 40 cm and 3 m. Obesity and diabetes requires significant training activities. Motivation is a key factor, which is very hard to increase sometimes in obese children. We have created a metrics to evaluate available games on the software game market. We have checked gamer statistics of the last 5 years, and collected feedbacks about the major, well-known console games. We have identified interactive games which have good acceptance levels by children (kids like it, according to the game statistics). We are targeting special games which are using motion sensors to control the game. We have combined the game console and our DroidHub equipped with activity sensors. We are able to measure the full energy consumption of the gamer during their play.

E. DrHealth - Data management center

On the server side we have realized a generic Data Center and a portal based graphical user interface. The GUI is a Liferay [4] based open source, service-rich front-end with the following features:

- Data management and data sharing
- Data visualization (multi-sensor)
- Real-time monitoring
- Strong authentication
- Easy-to-use web interface
- Data analysis
- Data storage

It supports full training programs composed of training exercises. The portal can be used to develop training programs. The portal based GUI is accessible via the Internet after the required user authentication. Instead of real-time monitoring, the practitioners can access the uploaded and archived training data. The defined training programs and the collected sensor data about each exercise are archived to the user's data space. The GUI is capable to let practitioners search and retrieve their patient's training data.



Figure 3. Multiuser (a) and single user (b) activity/health monitoring

DrHealth is collecting all the activity information and calculate the burned calories of the kids during the training activities.

III. SUMMARY AND FUTURE WORK

We have built up a combined obesity and diabetes monitoring solution with our DAQit framework, which is using external sensors, and game console systems as training devices to stimulate and monitor patient's training activities. Sensor data acquisition on the personal mobile

device enables both end-users and care givers to provide better and more effective health monitoring and facilitate prevention. We have integrated the DAQit framework with various indoor game console systems (Wii, and Microsoft XBOX). The patient can play with the game console and health parameter measurements tracking the effectiveness of the activity. The server side system enables both patient and practitioners to do comparative analysis between the trainings, and later reuse the data also for trend analysis. The developed solutions provide important feedbacks about training efficiency and health status both to the patient' parents and to the medical experts.

IV. ACKNOWLEDGMENTS

The projects have been supported by the European Union. The authors would like to thank GOP-1.1.1-11-2012-0055- Dialogic „DIALOGIC – Mathematical model based decision support system for diabetes monitoring" for its financial support. This work makes use some of the results produced by the AALAMSRK project (Research & development of assisted living patterns and their integration into decision support system to promote the quality of life at home), Hungarian National Technology Programme, A1, Life sciences project. Authors would like to thank for their financial support hereby.

REFERENCES

[1] http://hu.wikipedia.org/wiki/Wii [acc. 09.02.2014]

[2] http://hu.wikipedia.org/wiki/Xbox_360 [acc. 10.01.2014]

[3] http://www.microsoft.com/en-us/kinectforwindowsdev [acc. 07.01.2014]

[4] http://www.liferay.com/documentation/liferay-portal/6.2/user-guide [acc. 08.01.2014]

Analysis of sEMG Signals Associated with Isometric Contraction of Triceps Brachii Muscles Using Multifractals

Kiran Marri and Swaminathan Ramakrishnan
Non-Invasive Imaging and Diagnostics Laboratory
Biomedical Engineering Group, Department of Applied Mechanics
Indian Institute of Technology Madras, India.
Email ID: kirankmr@gmail.com and sramki@iitm.ac.in

***Abstract*—In this work, isometric contraction of triceps brachii muscles are analyzed using surface electromyography (sEMG) signals and multifractals. The signals are recorded from 20 healthy adult volunteers at three different angles of elbow flexions namely 30º, 60º and 90º using standard experimental protocol. The recorded signals are preprocessed and subjected to Multifractal Detrended Fluctuation Analysis (MF-DFA). The three extracted MF-DFA features namely, peak exponent event, rare exponent event and smooth exponent event, are used for further analysis. The results show that multifractal spectrum response are distinct for all the three angles of flexion. It is observed that multifractal spectrum is shifting from higher order exponents to lower order with the increase in the angle of flexion. The features are statistically highly significant between 30º and 90º flexion. The smooth exponent is found to be statistically highly significant for both 60º-90º and 30º-90º flexion. It appears that this method of multifractal technique is an useful approach to understand muscle contractions in fatigue and in varied clinical conditions.**

Keywords—Surface EMG; Isometric Contraction; Triceps Brachii; Multifractal; Detrended Fluctuation Analysis;

I. INTRODUCTION

Muscle tissues are categorized into smooth, cardiac and skeletal muscles based on their structure and functionality. The skeletal muscles are striated and attached to skeleton to facilitate movement, force generation and interactive action through the process of excitation, extension and contraction. Muscle tissues are made up of muscle fibers and innervated by alpha-motoneuron, and together they form the basic functional unit of neuromuscular system known as motor unit [1-2]. Based on the motor unit ability to contract and resistant to fatigue, they are classified into slow-twitch and fast-twitch muscle fibers [3]. The slow-twitch muscle fibers are fatigue resistant and their properties are useful for endurance-based activities like repeated contraction with lesser force [4]. The fast-twitch muscle fibers are prone to fatigue but provide good contraction and force in rapid time [5]. The percentage of slow-twitch and fast-twitch fibers vary based on skeletal muscle function and purpose. The triceps brachii is normally a three-head muscle in the upper arm and running along the humerus bone connected to the shoulders and elbow [6]. The triceps brachii is antagonist muscle and biceps brachii is agonist muscle providing stability to arm for different interactive actions and flexions. The maximum force and power output produced by muscle are influenced by numerous factors such as muscle fiber size, muscle length, angle and physical properties of the fiber-tendon attachment, fiber to muscle length ratio, fiber type, number of cross-bridges in parallel, rate of force development, and force-frequency relationship [7].

The study of muscles during elbow flexion is gaining field of sports medicine [8], rehabilitation [9], robotics, myoelectric and prosthetics control [10-12]. The muscle activities of the biceps brachii and brachioradialis showed angle-related changes in the order of 55° (largest value) followed by 70° and 90° [13]. The muscle length is dependent on angle of joint and is found to influence the muscle force [8]. A study conducted by Kim indicated that the angular variations due to flexion has an influence on the muscle and activation behavior [14].

The Surface Electromyography (sEMG) is a non-invasive technique of recording the electrical activity of muscle contractions [15]. The varying intense contraction increases the complexity of analyzing the sEMG signal [16]. The presence of high frequency (small time scales) and low frequency components (large time scales) can complicate the analysis of sEMG signals in analysis [17].

There is an increasing evidence that physiological signals are generated by complex self-regulating system and may have fractal structure [18-19]. The fractals are referred to a signal that can be split into smaller parts that is approximately similar to a reduced-size of whole signal. This property is called self-similarity and proposed by Mandelbrot [20]. The multifractal analysis based on detrended fluctuation analysis (MF-DFA) is introduced to determine the scaling properties and for detecting the long-range correlation for non-stationary series [21]. This MFDFA method is successfully implemented in heart rate variability [18], human gait analysis [22] and EEG [23]. The patterns of sEMG exhibit self-similarity

property [24]. The area of singularity spectrum based on fractal technique is computed in isometric contraction for biceps brachii [25]. Distinct features are reported for biceps brachii muscles during isometric contraction [26-27].

The aim of this work is to analyze of sEMG signals recorded from triceps brachii muscle during isometric contraction using multifractal detrended fluctuation technique at various angles of flexion. Three different features are derived from the multifractal spectral approach and are used for further analysis.

II. METHODOLOGY

A. Experimental Protocol and Signal Acquisition

In this study, 20 healthy volunteers without any history of neuromuscular disorder, participated in performing an isometric contraction of elbow extension by moving the forearm. A dumbbell measuring 6-kilogram is given to the participant for exerting the load on the muscles. The experimentation protocol is explained to each participant before the start of the study, and written informed consent form are collected. The subject is requested to maintain upright body position and hold the dumbbell at 30º, 60º and 90º flexion (dumbbell is parallel to the ground). Each participant is advised to perform this exercise using his or her dominant hand for a duration of at least 3 seconds and signals are recorded.

B. Signal Acquisition

The skin around the triceps brachii area is cleaned and abraded before placing the surface electrodes (Ag/AgCl electrodes: 1cm diameter; 3 cm inter-electrode distance). The reference electrode is placed on the elbow. The sEMG signals are recorded from the triceps brachii muscle in a bipolar configuration. The signals are amplified (Gain=1000) and sampled at 10000 Hz with Biopac MP36 system (24-bit data acquisition; CMMR 110 db) [16]. The signals are preprocessed offline using a butterworth band-pass filter (10Hz – 400 Hz), filtered using a notch filter (50 Hz). The signals are down sampled at 1000 samples per second for computational reasons.

C. Multifractals and Detrended Fluctuation Analysis

The term fractal is related to statistical self-similarity. The class of signal with varying singularity exponent and regularities is called as multifractal. The detrended fluctuation analysis (DFA) method is proven to be useful in revealing scaling properties of monofractal signals and in identifying the extent of long-range correlations in non-stationary time series [22] . The multifractal generalization of DFA (MF-DFA) is based on the identification of scaling of higher order moment related to power-law dependent signal length. The MF-DFA technique consists of below steps:

Step 1: For the sEMG sequence, x_k, of length N, the integrated signal profile $Y(i)$ is computed as

$$Y(i) = \sum_{k=1}^{i} [x_k - \mu \quad] \tag{1}$$

$i = 1 \ldots N;\ \mu\ is\ mean$

The profile is then divided into N_s segments of equal length s without any overlap ($N_s = N/s$). The length N may not be multiple of time scale s and may lead to a profile tail remainder. Hence the same procedure is repeated from the other side of the series to get $2N_s$ segments.

Step 2: For each segment $\boldsymbol{v}$, the local trend is calculated by fitting a polynomial $\boldsymbol{y_v}(\boldsymbol{i})$. The variance is then determined as

$$F^2(s, v) = \frac{1}{s}\sum_{i=1}^{s}\{Y[N - (v - N_s)s + 1] - y_v(i)\}^2 \tag{2}$$

Step 3: The qth order mean fluctuation function is obtained by averaging over all possible segment length s,

$$F_q(s) = \left\{\frac{1}{2N_s}\sum_{v=1}^{2N_s}[F^2(s, v)]^{q/2}\right\}^{1/q} \tag{3}$$

Step 4: The above steps are repeated for various scales to establish a relationship between detrended fluctuation function, $\boldsymbol{F_q}$and the size scale s as

$$F_q(s) \sim s^{Hq} \tag{4}$$

where Hq is called the generalized Hurst exponent. The power-law scaling is represented as a linear function on the double log plot for different range of s by the fluctuation function. This relationship is represented as

$$H_q(q) = [\tau(q) + 1]/q \tag{5}$$

The generalized Hurst exponent and the multifractal spectrum is related by Legendre transform [21] and deriving the multifractal parameters as

$$Dh\,(Hq) = q[Hq - Hq(q)] + 1 \tag{6}$$

where Dh is the Hausdorff dimension and Hq is the generalized Hurst exponent.

D. Features of Multifractals

The organization of multifractal series is characterized by evaluating number and size of fractal sets that influences the time scale. Based on the multifractal spectrum, it is possible to derive the quantity and quality of irregularities in the time series [21]. The shape of the spectrum is dependent on the periodic pattern. Higher value of Hurst exponent in the time series is an indication of more periodicity and less complexity. The sEMG signal are subjected to MF-DFA with 1000 samples (1 second data) and then continued with a 75% overlap for the duration of the recorded signal. This procedure is repeated for all the three angles of isometric contraction. The multifractal spectrum is averaged for each angle for analysis. The features used in this study are explained below [21],[23]

Peak Exponent Event (PET): The value of generalized Hurst exponent in multifractal spectrum is taken as peak exponent when the Hausdorff dimension equals one. If there are more than one Hausdorff dimension having a value of one, then the least exponent is considered as peak exponent.

Rare Exponent Event (RET): The left most value of generalized Hurst exponent in the multifractal spectrum is defined as rare exponent event. This feature is a representation of non-stationary characteristics of the signal.

Smooth Exponent Event (SET): The right most value of generalized Hurst exponent in the multifractal spectrum is defined as smooth exponent event. This feature is a representation of stationary characteristics of the signal.

E. Statistical Analysis

Analysis of Variance (ANOVA) is an inferential method that is used to test the equality of three of more population means. In this study, one way ANOVA test is performed to investigate the effects of mean value of peak exponent, smooth exponent event and rare exponent event for each angle of isometric contraction. Post-hoc procedure ($p<0.05$) is used with Tukey to investigate differences between features for different angles.

III. RESULTS AND DISCUSSIONS

A. Raw sEMG signals and demographic data

The demographic data for 20 participants are represented in Table I. The representative sEMG signal from triceps brachii muscle at 30º, 60º and 90º are shown in Figure 1. The magnitude, frequency response, amplitude and pattern of sEMG signals are found to be complex during the isometric contraction at various angle of flexion. At 90º flexion, the amplitude of sEMG signals appears to be increasing steadily. The duration of the experiment performed by different participants varied from 3 seconds to 8 seconds. For computation purpose, the entire duration is considered and normalized.

B. Scaling Exponent function

The value of multifractal scaling exponent $\tau(q)$ plot for various values of q (ranging from -10 to 10) is represented for isometric contraction in Figure 5 for the three angles of flexion. The scaling behavior is varying for both positive and negative values of q and with angle of flexion. This is an indication that the sEMG signals are exhibiting multifractal behavior during isometric contraction. The scaling function at 30º flexion is having highest slope for q indicating influence of both small and large fluctuations. This may be due to lower firing rate of the motor units and relatively lesser force needed to sustain the contraction in isometric position at 30º as compared to higher angles. At 30º, the scaling exponent varied from 4.5 and -14 during isometric contraction. The larger variation of scaling exponent are indication of higher multifractal behavior during 30º than in 60º and 90º. This may be due to increase of self-similar behavior and higher multifractal property for 30º as compared to higher angles of flexion.

TABLE I. DEMOGRAPHIC DATA FOR 20 PARTICIPANTS

Parameter	Mean (± SD)
Age (years)	21.05 (0.911)
Weight (kg)	68.08 (13.37)
Height (cm)	170.6 (5.79)
Gender	18 male and 2 female Participants

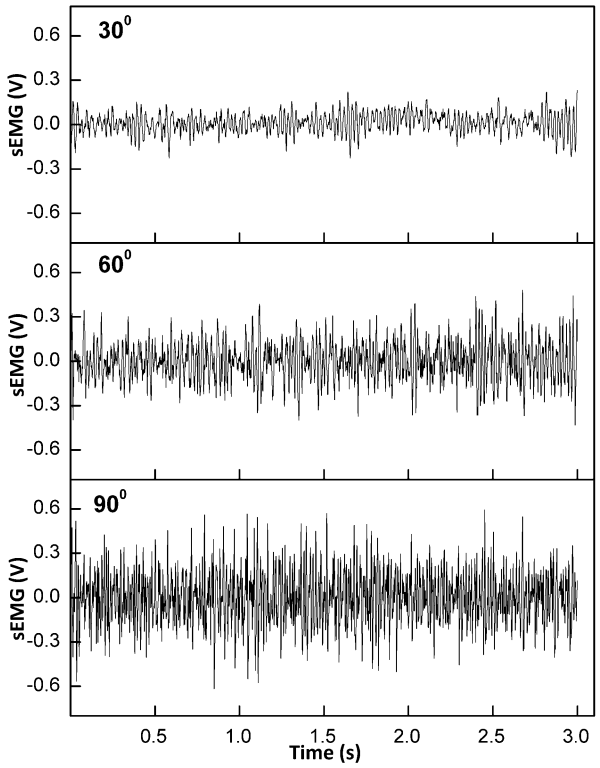


Figure 1. Representative 3 second sEMG signal during isometric contraction at 30º, 60º and 90º flexion

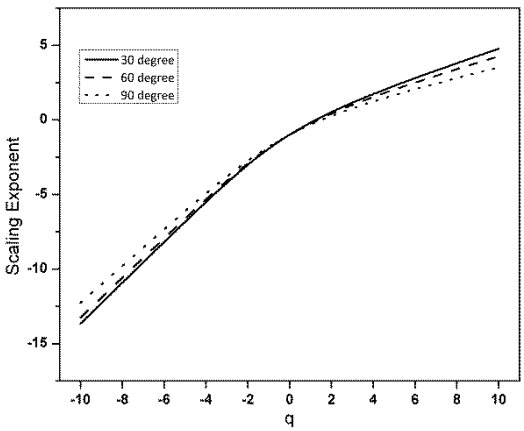


Figure 2. Average Scaling Exponent Function in Isometric Contraction

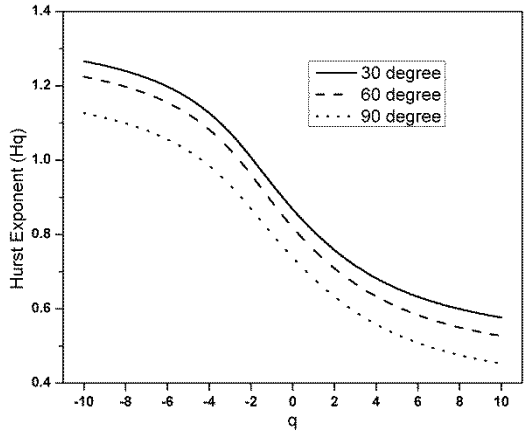


Figure 3. Average generalized Hurst Exponent for Isometric Contraction

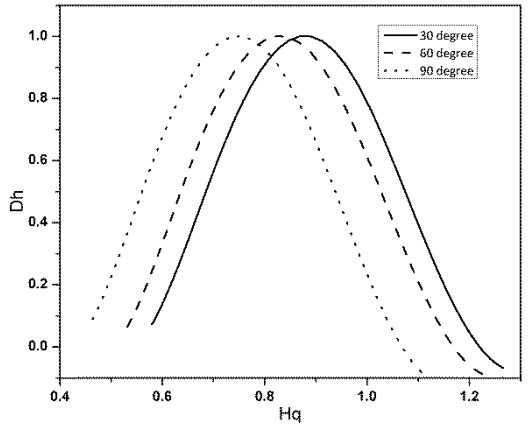


Figure 4. Average Multifractal Spectrum for Isometric Contraction

C. Genralized Hurst Exponent function

Figure 3 is the representation of averaged generalized Hurst exponent of sEMG signals for three different angle of flexion during isometric contraction. The decreasing exponent function with increasing moment is an indication that sEMG signals are exhibiting multifractal characteristics. This is observed for all the three angles of flexion. The values of Hurst exponent is found to be highest in 30º angle of flexion. Since the values of Hq are higher when q is lesser than 0, this is an indication that exponent function is dependent on smaller fluctuations in the sEMG signal than with the larger fluctuations. The highest impact of small fluctuation is observed at 30º flexion ($q < 0$) and highest impact of large fluctuations are also observed at 30º flexion ($q > 0$). The motor unit synchronization is higher when the muscle is flexed at 30º than higher degree of flexion. This is also seen in the Figure 1 where the amplitude of sEMG signals are relatively lower for 30º flexion and maximum at 90º flexion. The trend also indicates that sEMG signals are more stationary and lesser complex when the flexion is minimum (at 30º flexion) and increases with angle with respect to vertical axis.

D. Multifractal Spectrum

Figure 4 is the representation of average multifractal spectrum with various degree of elbow flexion in isometric contraction. The peak exponent is highest for 30º (0.89) and lowest at 90º (0.75). This is an indication that the sEMG signal in multifractal analysis experiences a left shift with the increase in elbow flexion. There is a steady decrease in self-similarity of the sEMG at 90º as compared to 30º flexion. This may be due to increase in motor unit activation for sustaining greater force in 90º flexion. The muscle has to exert maximum force as compared to 30º flexion. There is also a visible increase of large amplitude of sEMG at 90º. The trend indicates that with increasing angle of flexion, complexity of sEMG signals increases and thereby shifting the spectrum to lower values. The smooth exponent event and rare exponent event are also reducing with increasing angle due to the left shift of the spectrum.

The values of these three features are found to be statistically significant in all the three angles of flexion (Table II and III). This may be due higher activation at 90º flexion. The elbow at 90º position has greatest influence on triceps brachii muscle and hence higher firing rate leading to increased recruitment and de-recruitment of motor units. However in fatigue conditions, there is also a higher degree of synchronization due to these changes and thereby pushing the spectrum to higher regions. The smooth exponent event is showing statistically significance (> 0.05) for both 60º-90º and 30º-90º.

TABLE II. MF-DFA FEATURES FOR ISOMETRIC CONTRACTION WITH VARYING ANGLE OF FLEXION

Features	**30º**	**60º**	**90º**
	Mean (± SD)	**Mean (± SD)**	**Mean (± SD)**
Peak Exponent Event	0.89 (0.10)	0.84 (0.11)	0.75 (0.11)
Smooth Exponent Event	1.26 (0.10)	1.22 (0.13)	1.12 (0.13)
Rare Exponent Event	0.58 (0.11)	0.53 (0.11)	0.45 (0.09)

TABLE III. ANOVA ANLYSIS FOR THREE FEATURES AT THREE ANGLE OF FLEXION

Features	**Sigma for 30º- 60º**	**Sigma for 60º- 90º**	**Sigma for 30º- 90º**
Peak Exponent Event*	0	0	1
Smooth Exponent Event*	0	1	1
Rare Exponent Event*	0	0	1

{* At the 0.05 level, the population means are significantly different
Sigma as 1 indicates that the mean difference is significant at 0.05 level
Sigma as 0 indicates that the mean difference is not significant at 0.05 level}

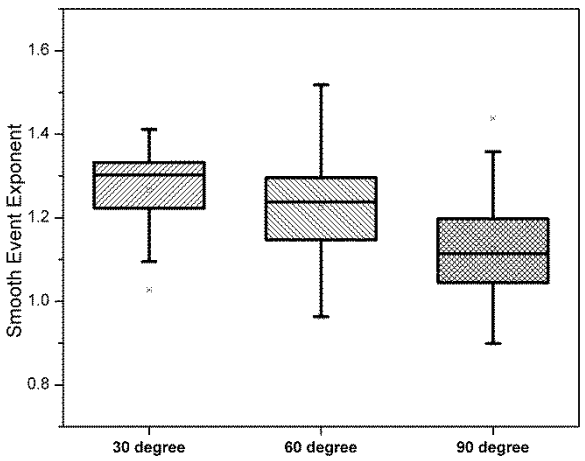


Figure 5. Smooth exponent event for three angles of isometric contraction

IV. CONCLUSION

In this work, isometric contraction of triceps brachii is analyzed using multifractal detrended fluctuation analysis. Three different features are extracted and are used for further analysis. Three different elbow flexions are considered for this study. The features from scaling exponent and multifractal spectrum are exhibiting multifractal characteristics sEMG signals during isometric contraction. The multifractal characteristics behavior is also observed in three angles of flexion.

The results show that the triceps brachii muscles are activated differently during isometric contraction at various angles of flexion. The signals are having highest multifractal characteristics when the participants are holding the dumbbell at 30 degree and the complexity of sEMG signals increases with increasing flexion.

The features computed such as peak exponent, rare exponent and smooth exponent are found to be statistically significant. The smooth exponent event appears to be useful feature for studying the isometric contraction. The study could further be extended to compare the variations of other muscles in the upper arm during isometric and dynamic contraction, and in studying various neuromuscular conditions.

REFERENCES

[1] Staron RS. Human skeletal muscle fiber types: delineation, development, and distribution. Can J Appl Physiol. ; 22:307–327, 1997

[2] Burke RE. Revisiting the notion of motor unit types.". Prog Brain Res.; 123:167–175, 1999

[3] McComas AJ. Skeletal Muscle: Form and Function. Champaign, Ill: Human Kinetics, 1996

[4] Burke RE. Motor unit types of cat triceps surae muscle. J Physiol.; 193:141–160, 1967

[5] Sieck GC, Prakash YS. Morphological adaptations of neuromuscular junctions depend on fiber type. Can J Appl Physiol. ; 22:197–230, 1997

[6] J. Chris, The Concise Book of Muscles, Second Edition, 2008

[7] Fitts R H, McDonald K S, Schluter J M, The determinants of skeletal muscle force and power: their adaptability with changes in activity pattern, .J Biomech.; 24 suppl 1:111-22, 1991

[8] Taewook Kang, Youngjoon Seo, Jaehoon Park, Eunseok Dong, Byungdo Seo M S, Dongwook Han,The Effects of Elbow Joint Angle Change on the Elbow Flexor Muscle Activation in Pulley with Weight Exercise ,J Phys Ther Sci.; 25(9): 1133–1136, 2013

[9] Thijs Krabben, Birgit I Molier, Annemieke Houwink, Johan S Rietman, Jaap H Buurke and Gerdienke B Prange,Circle drawing as evaluative movement task in stroke rehabilitation: an explorative study, Journal of NeuroEngineering and Rehabilitation, 8:15, 2011

[10] Artemiadis, Panagiotis K., and Kostas J. Kyriakopoulos. "A switching regime model for the EMG-based control of a robot arm." Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Trans : 53-63, 2011

[11] Cooper G, Sheret L, McMillan L, Siliverdis L, Shan N, Hodgin D, Kenny L, and Howard D, Inertial sensor-based knee flexion/extension angle estimation. Journal of Biomechanics, 42(16):2678–2685, December 2009.

[12] Ritik Looned, Jacob Webb, Zheng Gang Xiao and Carlo Menon, Assisting drinking with an affordable BCI-controlled wearable robot and electrical stimulation: a preliminary investigation, Journal of Neuro Engineering and Rehabilitation, 11:51, 2014

[13] Kapandji IA.: The physiology of the joints, 5th ed. Edinburgh: Churchill Livingstone, pp 138–142, 1983

[14] Kim JJ, Lee MH, Kim YJ: Comparison of the maximum EMG levels recorded in maximum effort isometric contractions at five different knee flexion angles.J Sport Biomech, 15: 197–206, 2005

[15] G. Venugopal, M. Navaneethakrishna, S. Ramakrishnan, Extraction and analysis of multiple time window features associated with muscle fatigue conditions using sEMG signals. Expert Systems with Applications; 41:2652– 2659, 2014 (Article in Press)

[16] D. Farina and R. Merletti, Comparison of algorithms for estimation of EMG variables during voluntary isometric contractions. Journal of Electromyography and Kinesiology ; 10: 337-349, 2000

[17] Sheean, Geoffrey L. "Quantification of motor unit action potential energy."Clinical Neurophysiology 123.3 : 621-625, 2012

[18] Wang G., Huang H., Xie H., Wang Z., Hu X. (2007). Multifractal analysis of ventricular fibrillation and ventricular tachycardia. Med. Eng. Phys. 29, 375, 2006

[19] Eke A., Herman P., Bassingthwaighte J. B., Raymond G. M., Percival D. B., Cannon M., Balla I., Ikényi C. Physiological time series: distinguish fractal noises from motions. Eur. J. Physiol. 439, 403–414, 2000

[20] Mandelbrot, B. B. The Fractal Geometry of Nature. Updated and augmented. New York: W. H. Freeman and Company, 1977

[21] J.W. Kantelhardt, S.A. Zschiegner, E. Koscielny-Bunde, A. Bunde, S. Havlin, and H.E. Stanley, Multifractal detrended fluctuation analysis of nonstationary time series, Physica A, 316(1) ,87-114, 2002

[22] Dutta S, Ghosh D and Chatterjee S, Multifractal detrended fluctuation analysis of human gait diseases. Front. Physiol. 4:274, 2013.

[23] Vladimir Kulish, Alexei Sourin and Olga Sourina, Human EEG seen as fractal time series: Mathematical analysis and visualization, Computers in Bio and Medicine, 36, 291-302, 2006

[24] C.J. Anmuth, G. Goldberg, N.H. Mayer, "Fractal dimension of EMG signals recorded with surface electrodes during isometric contractions is linearly correlated with muscle activation", Muscle Nerve 17, 953-954, 1994.

[25] Gang W, Xiao-mei R, Lei L, Zhi-zhong W. Multifractal analysis of surface EMG signals for assessing muscle fatigue during static contractions. J Zhejiang Univ Sci 8(6):910–5, 2007

[26] Talebinejad M, Adrian D.C. Chan, Ali Miri, Fatigue estimation using a novel multi-fractal detrended fluctuation analysis-based approach, J Electromyogr Kinesiol 20; 433–439, 2010

[27] Talebinejad M, Chan A, Miri A, Dansereau A. Fractal analysis of surface electromyography signals: a novel power spectrum-based method. J Electromyogr Kinesiol;2:1, 2008 M.S. Taqqu, V. Teverovsky, W. Willinger, Fractals 785, 1995

DISTRIBUTED COMPUTING

Improving data-parallel construction of δ_N-Nets with Maximum Dispersion

P. Zinterhof
* University of Salzburg/Department of Computer Science, Salzburg, Austria
peter.zinterhof3@sbg.ac.at

Abstract - Linear nearest-neighbor search in high-dimensional data exposes high computational complexity. In order to minimize search complexity we employ optimal δ_N-nets of rank N, which consist of a sub set of N vectors out of an initial code book E, yet approximate all En vectors of E by the least error (denoted by the number theoretical concept of dispersion) of all possible selections of N vectors.

In this paper we demonstrate a substantially improved algorithm, which results in a reduction of computational complexity from previously O (En N^2- N^3) to O(En^2). We demonstrate computational feasibility both on CPU- and GPGPU-based systems for any desired accuracy in the approximation of codebook E by some δ_N-net for real-world problem sizes in the range of En = [10^7..10^8]. In addition to the high speedup that can be achieved by the new algorithm, the GPGPU implementation typically shows speedups of one order of magnitude over a CPU implementation.
We also give a practical example of image segmentation within computed tomography scans in order to show one potential application of this new approach.

I. INTRODUCTION

Machine learning is now a rapidly growing field with myriads of already existing and potential future applications in science, technology, and economy. In this paper we want to focus on the well-known machine learning method of principal component analysis (PCA), which transforms potentially correlated data into linearly separable, uncorrelated 'principal components' or feature vectors. One of the most important characteristics of PCA is the fact, that the number of output features is usually much smaller than the number of data items in the original input data. Thus, PCA essentially acts as a dimensionality reduction scheme, which often forms the basis for building robust pattern recognition systems. Correspondingly, our motivation to employ PCA has been the task of image segmentation in computed tomography (CT)-scans, in which we found PCA to be valuable in that it does not suffer from over-fitting as many neural network based approaches often do.

Our application takes small image patches (64x64 pixels) from CT-scans, that comprise our labeled input data. In the next step, all principal components are computed, but only the 32 (out of 4096) most eminent components are being put into a code book data structure for later use in the recall process. In the recall (or segmentation) phase, again small image patches are taken and transformed into short vectors of 32 PCA-values. In the final classification step, new vectors are being compared to the vectors in the code book, based on their Euclidean distance. The unknown vector is then assigned the same image class as the 'winning' vector in the code book, which is the vector with minimal Euclidean distance to the previously unknown vector. Unfortunately, our impulse of refining segmentation quality by increasing the number of samples in the training phase also led to an increase in computational complexity. As has been described in [1], any attempt of optimizing this code book lookup, will fail on high dimensional data and all sorts of clustering-, and kD-trees-algorithms break down in this task, due to the 'curse of dimensionality'.

This leaves us with linear search within a large array of short vectors of length 32. Obviously, in addition to employing parallel computation, one of the most efficient ways to improve recall speed is to actually decrease the size of the code book itself. In a way, we would have needed some form of PCA on the PCA-based code book vectors themselves. But as the single components already are linearly separable, we do not consider this approach feasible. Instead, in the following chapters we will shortly recall the concepts of Metric Dispersion and δ_N-Nets(see [2] for the full definitions) and then show how the construction of δ_N-Nets can be vastly improved in terms of computationally complexity. δ_N-nets can then be used to approximate the original code book (or any set of vectors, for that matter) of feature vectors which greatly speeds up the process of segmenting of CT-scans.

II. METRIC DISPERSION

In the following section we recapitulate the main definitions from [6]. Four properties are defined for metric d(x,y)

$$d(x,y) \geq 0, x,y \in E \quad (1)$$

$$d[x,y] = 0 \Leftrightarrow x=y \quad (2)$$

$$d[x,y] = d[y,x], \forall\ x,y \quad (3)$$

$$d[x,z] \leq d[x,y] + d[y,z], \forall\ x,y,z \quad (4)$$

with $d(x,y) = \sqrt{\sum_{k=1}^{s}}$ defined in a finite vector space $E \neq 0$ over $\mathbb{R}^s$.

The dispersion δ_N for N points in E is defined as

$$\delta_N = \delta(x_1, x_2, \ldots, x_N) = max_{x \in E}[min(d(x, x_k))] \quad (5)$$

As E is a finite vector space, $max[min(d(x,x_k), k = 1,2,\ldots,N)]$ is finite, too. Furthermore, it is $\delta_N \geq \delta_{N+1} \geq \delta_{N+2} \ldots$

Now, let

$$min(d(x,x_k), k = 1,\ldots,N) \geq min(d(x,x_k), k = 1,\ldots,N,N+1) \quad (6)$$

for all $x \in E$. Then for all $y \in E$

$$max_{x \in E} min\, d(x,x_k), k = 1,\ldots,N) \geq min(d(x,x_k), k = 1,\ldots,N,N+1) \Rightarrow max_{x \in E} min\, d(x,x_k), k = 1,\ldots,N) \geq max_{y \in E} min(d(y,x_k), k = 1,\ldots,N,N+1) \quad (7)$$

and hence $\delta_N \geq \delta_{N+1}$. In other words, metric dispersion δ_N is equal to the radius of the largest hyper-sphere S that can be put around any point in E without getting into contact or incorporating any other point of E. More details on metric spaces and dispersion can be found in [2] and [5], respectively.

III. Related work

As pointed out in [6], the mathematical concept of dispersion has initially been developed by E. Hlawka. Dispersion deals with discrepancy measures of uniform distributions of N points, which may also be the case for our algorithm. But the proposed algorithm itself is not limited to uniform distributions. Another overview on the mathematical concepts and foundations can be found in [1]. As our work on δ_N-nets starts with [6], one can find some of the main ideas described there, too. The proposed algorithm shares motivation of approximating a code book of feature vectors as in [6], but we do not consider it a simple adaptation of the previous algorithm. Instead, it is a rather different approach, which also might be seen when comparing the extremely improved computational performance of the new algorithm.

IV. Improved Construction of Δ_N-nets

A. Algorithm

The construction of δ_N nets is based on an iterative algorithm, in which two seed points x_1, x_2 are chosen, such that (5) is fulfilled and dispersion δ_N net with N=2 is obtained.

Next, based on the seed points, dispersion δ_3 is computed by finding point x_3, which is given by $max[min(d(x_1,x_k), d(x_2,x_k)])$. Obviously, In order to find point x_k, we have to determine Euclidean distances between all pairs x_1, x_k and x_2, x_k. The crucial point here is the fact, that we do not only compute $max[min(d(x_1,x_k), d(x_2,x_k)])$, but also retain the minimum distance of each point in E to x_1, x_2.

At this stage we already have the next 'winning vector' x_k, with k=3 along with an array (D) of distances $[min(d(x_1,x_k), d(x_2,x_k)])$. Furthermore, we can update array D repeatedly by computing d(y, x_k) for all remaining points y in E and store the local minimum D[y] =min(D[y], d(y,x_k)). Also, the new maximum max(D[y]) is being computed 'on the fly' during the same iteration, which will mark the next 'winner' x_{k+1}. So we have an iterative computation of the $(k+1)^{th}$ vector during updates of the minimum distances of vector x_k in array D.

The most important aspect of this new algorithm, compared to (6) happens to be its linear O(N) complexity of the update process of array D. At no point during construction of a δ_N-net we have to compute more than N Euclidean distances for each vector of Dn, that is, one computation for each of the N iterations. This is in stark contrast to (6), where complexity linearly grows with each new vector M that we add to the δ_N-net, giving total complexity O((N-M)*(M)) for the mth vector. In other words, adding vectors to some δ_N-net is of constant complexity now, were it has been of exponential complexity in the previous algorithm.

For typical applications (see results section), we frequently encounter vector sets E with N = 10^6 to 10^7. These numbers proved to be of intractable computational complexity, even for decently sized clusters of GPGPU-accelerated systems. We therefore had to limit the number of vectors in E and especially limit the size of the resulting δ_N-net to modest levels of 10^4 - 10^5. Still, the observed run times frequently had been in the range of several weeks!

The described algorithm makes it feasible to compute very large δ_N-nets including all vectors of E, not just small sub sets of E, without having to resort to a supercomputing system.

```
find seed points x1, x2
update distance array D, declare follow-up winner X

for k=3: N
  xk = UpdateD (X)
  X=xk
next k
```

Figure 1: pseudo-code for construction of δ_N-nets

B. Data-parallel Implementation

We employ two functions that are based on CUDA for C for finding both seed points x_0, x_1 and for finding points x_k, respectively (Fig. 1). As distances between any two points are symmetric (3), the search for maximum distances in E are of triangular form. The CPU based driver routine loops over all N points of E (En), repeatedly calling a CUDA function 'GPU_MaxDistance (Xk)' for each point $0 \leq x_k \square En$. The number of threads in each CUDA block is given by 32 x BLO, with 'BLO' being an arbitrary integer constant, such that the product of 32 by BLO is smaller than the maximum number of supported threads per CUDA block on a given NVIDIA GPGPU. The rationale of this particular number of threads is that we aim to achieve good memory coalescing (hence multiples of 32 threads) and also good utilization of shared memory. By moving BLO copies of the vector x_k (line 11) into local shared memory, we avoid potential conflicts in memory access later on in the main loop (line 23 ff.). The same principle is enacted in line 26, where a block of BLO vectors is copied from the code book into local shared memory. We are then able to perform Euclidean distance computations strictly off-line, based on local shared memory only. This frees up the memory bus of the accelerator for access by additional CUDA blocks. Euclidean distance computations for a given vector are performed in parallel by all 32 threads in a WARP (line 32 ff.), but only one thread is scheduled to determine the nearest neighbor of the vector within our code book array 'base_d'. After having sifted through the entire code book, there do exist BLO local minima (nearest neighbors), which have to be reduced to a single global maximum (line 62 ff.). Again, only one thread of each WARP operated this procedure, but as this is a global operation that potentially interferes with CUDA threads of different CUDA blocks, a binary semaphore is used for arbitration. As no explicit semaphore constructs are offered in CUDA, the desired functionality of the semaphore is implemented by an atomic compare-and-swap function. By busy-waiting for value 0 on a global variable 'sema' and atomically setting it to the block index of the calling thread, access of one and only one thread is ensured. Global atomic operations are notoriously expensive on GPUs and some further optimization step could involve a slightly different scheme, which would restrict thread-synchronization to each CUDA block. Thus a number of gridDim.x maxima would arise instead of one truly global maximum as in the code shown (fig.2, lines 67 ff.). In preliminary tests we have moved the final computation of the maximum back to the CPU, which showed a slight improvement in execution times, but also increased code complexity to some extent.

Construction of all subsequent vectors of the δ_N-nets works analogously, with the algorithmic addition of storing local minima in an array and by masking out those vectors from E that already have become part of the δ_N-net. Masked vectors need not be taken into account any more and thereby reduce the computational load.

```
1   __global__ void GPU_ MaxDistance (
2               const float *  __restrict__ base_d,
3               int Xk,
4               int Vectors,
5               volatile float * __restrict__ maximum,
6               int * __restrict__ index)
7   __shared__                  float Winner[BLO][32];
8   __shared__  volatile   float dist[BLO][32];
9
10  if (threadIdx.y==0) {
11      Winner[0][threadIdx.x] = base_d[((unsigned  long )
12                              Xk*bestevs)+threadIdx.x];
13      }
14  __syncthreads();
15
16    if (threadIdx.y>0)
17      Winner[threadIdx.y][threadIdx.x] =
18                Winner[0][threadIdx.x];
19
20     max_temp=0.0;
21     temp=SENTINEL;
22
23  for (i=Xk+1+(BLO*blockIdx.x+threadIdx.y);
24       i < Vectors; i+=(BLO*gridDim.x))
25    {
26     dist[threadIdx.y][threadIdx.x]=base_h[((unsigned
27      long)i*bestevs)+threadIdx.x];
28      __syncthreads();
29
30  /*  ===== EUCLIDEAN DISTANCE ===== */
31
32    diff          =  Winner[threadIdx.y][threadIdx.x] -
33                          dist[threadIdx.y][threadIdx.x];
34          dist[threadIdx.y][threadIdx.x] =  diff*diff;
35   __syncthreads();
36
37  if (threadIdx.x <16) dist[threadIdx.y][threadIdx.x] +=
38  dist[threadIdx.y][threadIdx.x+16];
39     __syncthreads();  // can be omitted on Fermi or later
40
41  if (threadIdx.x <8 ) dist[threadIdx.y][threadIdx.x] +=
42  dist[threadIdx.y][threadIdx.x+8];
43     __syncthreads();  // can be omitted on Fermi or later
44
45  if (threadIdx.x <4 ) dist[threadIdx.y][threadIdx.x] +=
46  dist[threadIdx.y][threadIdx.x+4];
47     __syncthreads();  // can be omitted on Fermi or later
48
49     if (threadIdx.x <2 ) dist[threadIdx.y][threadIdx.x] +=
50  dist[threadIdx.y][threadIdx.x+2];
51     __syncthreads();  // can be omitted on Fermi or later
52
```

```
53   if (threadIdx.x==0)  {
54       dist[threadIdx.y][0] += dist[threadIdx.y][1];
55       if (dist[threadIdx.y][0] < temp)    {
56             temp = dist[threadIdx.y][0];
57             bestindex=i;
58             }
59         }  // endif  threadidx.x==0
60     }  // next i
61
62   if (threadIdx.x==0)   {
63     do {  // increment semaphore
64       old =atomicCAS (sema, 0, (blockIdx.x+1));
65         }    while (sema != (blockIdx.x+1));
66
67       if (temp <maximum[0]) {
68           maximum[0]=temp;
69           index[0] = bestindex;
70           __threadfence();
71           }
72       sema = 0; // release semaphore
73     }
74   }
```

Figure 2: CUDA code for finding maximum Euclidean distance to a given vector in an array of vectors

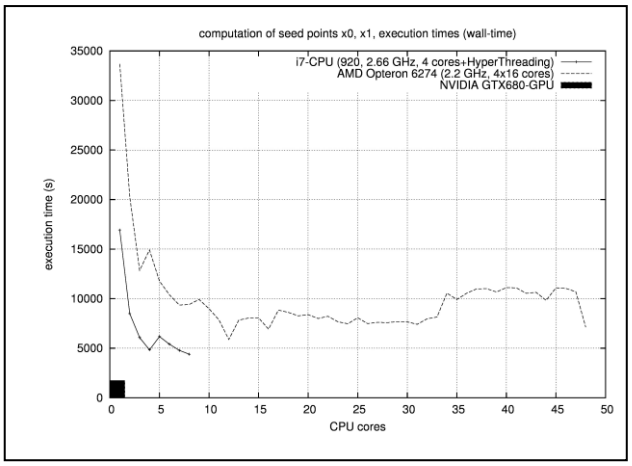


Figure 3: Scalability of seed point generation

V. RESULTS

We have tested the algorithm by implementing it both on standard CPU-multiprocessors and NVIDIA GPGPUs, respectively.
Figure 3 displays both scalability and performance of the generation of seed points for a δ_N-net with E=1 million for an INTEL i7-desktop CPU, a quad-socket system based on AMD Opteron server CPUs (SuperMicro), and a consumer GPGPU from NVIDIA. Parallelization has been based on GCC 4.4.3 with OpenMP and CUDA 5.5, respectively. The GPGPU has been operated in the i7-desktop system for all GPGPU-related tests.

On the i7-CPU the algorithm scales very well for up to 4 processing cores. By employing Intel HyperThreading we get only marginal improvements at the maximum of 8 cores. Given the ratio of memory per floating point operation of 4:3 we are inclined to attribute nearly constant execution times for core counts higher than 4 to a saturation of the memory system ('hitting the memory wall').
In the case of the quad-socket Opteron System very similar effects can be observed. Scalability is good for up to 12 cores, where peak performance is reached. total Execution times do not benefit from employing higher numbers of CPU cores.

CPU-code has been compiled with optimization (-O5 -funroll-loops), but without explicit use of SIMD instructions.

GPGPU-code outperformed any number of CPU-cores by a factor of at least 3.

Table 1 reflects a comparison of the i7-system and two different NVIDIA GPGPUS in the main task of constructing δ_N-nets. The latter two examples (rows 3 and 4 of table 1) can directly be compared to previous results given in (6). They emphasize the tremendous increase in performance which is a direct result of the new algorithm, far less of potential advances in computing machinery during the last two years.
In the case of a δ_N-net of 15.000 vectors out of a code book of 6.3 million entries, one single i7-core (2.66 GHz) running the new algorithm clearly outperforms two NVIDIA GPGPUs (GTX480, Fermi) running the previous algorithm (6) by a factor of 1.6, whereas a single GTX680 (Kepler) card excels by a factor of 69.85.

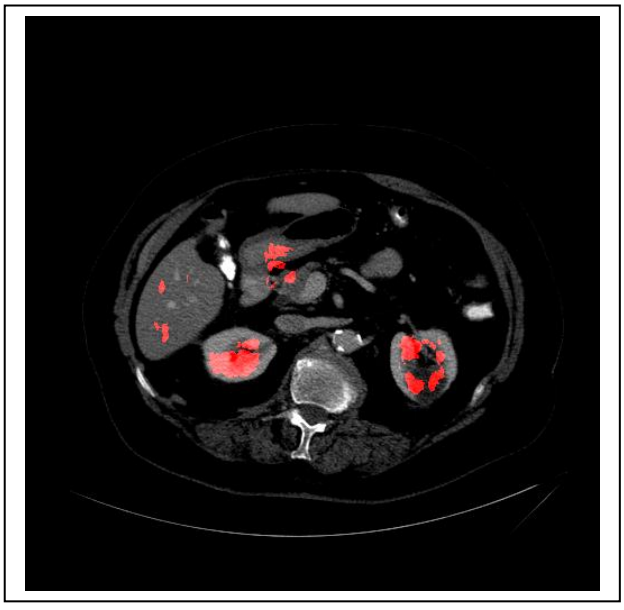


Figure 4: Segmentation based on random selection of 500k entries

Table 1: Computation of δ_N-nets, (wall-) times in s, factors in line 2 denote the speedups over the CPU versions.

	i7-CPU, 2.66 GHz	**NVIDIA GTX680**	**NVIDIA GTX 480**
Δ_N-net:50k vectors, E=100k vectors	543.94 1.0x	32.70 16.6x	51.75 10.5x
Δ_N-net :500k vectors, E=1m vectors	54433.2 1.0x	2642.84 20.5x	2856.46 19.0x
Δ_N-net:5500 vectors, E=6.3 mio. vectors	7852.6 1.0x	185.67 42.2x	243.97 32.1x
Δ_N-net:15000 vectors, E=6.3 mio. vectors	21392.8 1.0x	499.56 42.8x	652.86 32.7x

In a joint project with Salzburger Landeskrankenhaus (SALK) we developed a principal component (PCA) algorithm for segmenting computed tomography scans (CT). Data taken from some 80 patients have been labeled manually for training and by applying PCA a code book consisting of several million entries has been built from scratch. Each entry of this code book can be described as a vector of 32 float values (32 bit), along with some label data, e.g. 'kidney', 'cyst', etc.

By transforming previously unseen and hence unlabeled data of some new patient into vectors of 32 principal components, we can sift through the code book and find that entry 'W' (winner) with least Euclidean distance to the unclassified vector. Classification of that vector is then simply a matter of reading the label data accompanying the winning code book vector.

In reality, the process of segmenting an entire CT scan involves several million lookup calls in a code book of – again – several million entries. Obviously, this can be very time consuming and therefore we resorted to clusters of GPUs.

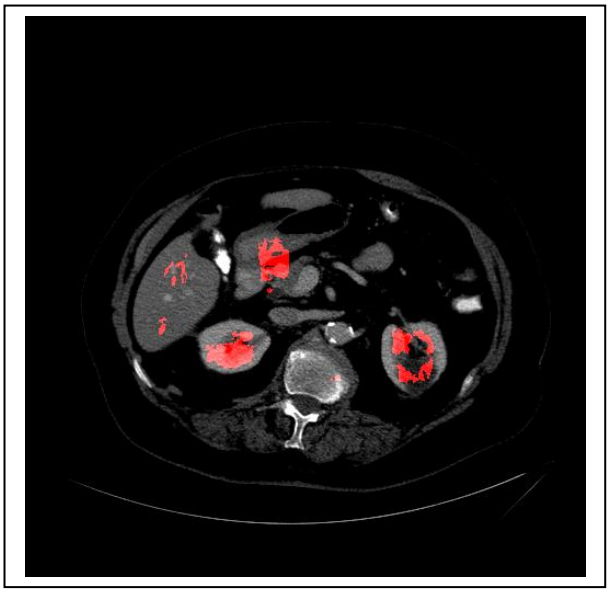


Figure 5: Segmentation based on a δ_N-net, N=500k entries

But even with the help of powerful compute accelerators run times had to be brought down further and we finally applied δ_N-nets for the approximation of very large code books.

Fig. 4 and Fig. 5 give an impression of the output of the segmentation process based on 500.000 vectors. Concerning Fig. 4, vectors have been chosen randomly, whereas Fig. 5 has been based on the corresponding δ_N-net.

False negatives could be reduced substantially over randomly chosen vectors, which we used the base line. Also, the number of correctly labeled pixels has been up by 25 % compared to the base line.

We therefore consider δ_N-nets a suitable approach for accelerating certain lookup operations in PCA-vectors.

VI. CONLUSION

We have given an overview of the mathematical concept of metric dispersion, on which we have based our improved algorithm for the construction of δ_N-nets. The improvement makes it computationally feasible to build very large δ_N-nets out of vast code books even on CPU-based hardware. We also demonstrate, that the new algorithm neatly fits into the data-parallel programming paradigm that recent GPGPUs rely on. Thus, we are able to effectively use GPGPUs during all algorithmic steps and hence improve execution times by at least one order of magnitude.

So, our contribution is two-fold. Firstly, we describe a new, much improved algorithm for the construction of δ_N-nets. Secondly, we show that efficient use of GPGPUs is possible in order to gain further speedups and construct δ_N-nets in reasonable time.

Further work will involve optimizations of the GPGPU implementation by employing multiple GPGPUs.

REFERENCES

[1] M. Drmota and R.F. Tichy, Sequences, Discrepancies and Applications, Springer, 1997.

[2] O. Forster, Analysis Band 2 Differential- und Integralgleichungen im R n, Gewöhnliche Differentialgleichungen, Vieweg-Studium 31: Grundkurs Mathematik, Vieweg, Braunschweig, 5. edition, 1999.

[3] E. Hlawka, Abschätzung von trigonometrischen Summen mittels diophantischer Approximationen, Österreichische Akademie d. Wissenschaften, SB II, pp. 43-50, 1976.

[4] R. Weber, Similarity search in High-Dimensional data spaces, 1988.

[5] A. Zhigljavsky and A. Zilinskas, Stochastic Global Optimization, Springer Optimization and Its Applications, Springer, New York, N.Y., 1998.

[6] P. Zinterhof, Data-Parallel Construction of Delta-N-Nets With Maximum Dispersion, Parallel Numerics 11 Conference Proceedings, Leibnitz, Austria, pp: 106-114, 2011

High Productivity Data Processing Analytics Methods with Applications

M. Riedel*+, A.S. Memon * and M.S. Memon*
* Forschungszentrum Juelich, Juelich Supercomputing Centre, Juelich, Germany
+ University of Iceland/School of Natural Sciences and Engineering, Reykjavik, Iceland
m.riedel@fz-juelich.de

Abstract - The term 'big data analytics' emerged in order to engage in the ever increasing amount of scientific and engineering data with general analytics techniques that support the often more domain-specific data analysis process. It is recognized that the big data challenge can only be adequately addressed when knowledge of various different fields such as data mining, machine learning algorithms, parallel processing, and data management practices are effectively combined. This paper thus describes some of the 'smart data analytics methods' that enable a high productivity data processing of large quantities of scientific data in order to enhance the data analysis efficiency. The paper thus aims to provide new insights how various fields can be successfully combined. Contributions of this paper include the concretization of the cross-industry standard process for data mining (CRISP-DM) process model in scientific environments using concrete machine learning algorithms (e.g. support vector machines that enable data classification) or data mining mechanisms (e.g. outlier detection in measurements). Serial and parallel approaches to specific data analysis challenges are discussed in the context of concrete earth science application data sets. Solutions also include various data visualizations that enable a better insight in the corresponding data analytics and analysis process.

I. INTRODUCTION

'Understanding climate change, finding alternative energy sources, and preserving the health of an ageing population are all cross-disciplinary problems that require high-performance data storage, smart analytics, transmission and mining to solve' [1]. Scientific disciplines face huge challenges in data processing with the ever increasing amount of scientific data [1,2,3]. This raises the demand for *'smart data analytics methods'*, which are required to provide a very concrete set of tools, techniques, frameworks, and infrastructures in order to perform *'big data analytics'* with real data sets. This is one of the important elements of data-intensive science [3] among others such as preserving and managing large-scale data sets for their *'active re-use'* (e.g. in-situ validations with real measurement data sets during a computational simulation on large-scale computing resources).

In order to effectively perform data-intensive science, it is important (i) to understand its relationship to other important and closely related concepts (e.g. machine learning algorithms, data mining practices, etc.), and (ii) to understand what are the fundamental techniques for efficiently using available resources (e.g. high performance computing, high throughput computing, big data sharing and access, etc.). Hence, by considering (i) and (ii), we want to achieve a *'high productivity data processing'* using a systematic approach and thus refine with this term the often generally used term 'big data analytics' with concrete methods. A wide variety of data analysis and analytics were arbitrary but we require a principled and systematic set of methods that guides the *'smart data analytics processes'*. This enters the field of process models such as the Cross Industry Standard Process for Data Mining (CRISP-DM) [4]. We require *'innovative perspectives on data'* as described by Provost et al. [5] who mentions that a *'data-science perspective provides practitioners with structure and principles, which give the data scientist a framework to systematically treat problems of extracting useful knowledge from data'*. This is particularly useful when the general data analytics step (performed by data scientists) precedes the often scientific domain-specific data analysis.

In earlier work [6], we already gave insights into several concrete *'big data processing approaches'* (e.g. iterative map-reduce, scientific workflows, databases, massively-parallel processing algorithms, visualization and steering, etc.) that we leverage in the context of large-scale computing (e.g. PRACE [7]) and collaborative data infrastructures (e.g. EUDAT [8]). This paper goes a step further in disentangling the set of closely interrelated *'big data analytics concepts'* described above in order to understand which techniques can be applied on which compute and storage resources to solve one specific data science problem. We provide methods in the context of two concrete data-intensive science case studies from the earth science domain we work with in a systematic fashion using CRISP-DM and selected algorithms and resources: (a) outlier detection [21] in measurement devices data, and (b) data classification [21] of multi-spectral images.

This paper is structured as follows. After the introduction to the problem domain, Section II provides a background to the concrete earth science data sets in question and briefly surveys systematic ways to perform *'data science'* with selected required concrete techniques. Section III offers the reader then methods in context of our two case studies well embedded in the overall systematic way of performing a high productivity processing of data. The paper ends with some concluding remarks of several findings and outlines future work elements.

II. Background and Related Work

A wide variety of process models exist that bear the potential to support 'smart data analytics processes' with a concrete structure that often refer to 'phases' with certain 'steps'. CRISP-DM [4] (cf. Figure 1) is the most suitable one that we refined for our (parallel) smart data analytics since it clearly outlines *'six concrete phases'* that are typical for a data mining project: *(a) problem/business understanding; (b) data understanding; (c) data preparation; (d) modeling; (e) evaluation; (6) deployment*. Given the page restriction, we summarize them only briefly based on the CRISP-DM user guide [9] including refinements towards *'smart data analytics'*, but would like to emphasize that we perform our work on the scientific data case studies using the CRISP-DM as a useful backbone process. To the best of our knowledge, a generally accepted *'big data analytics process model'* for science and engineering problems is not existing except disconnected approaches from several business domains (e.g. predictive/customer analytics processes, etc.).

The *'(a) problem/business understanding phase'* includes general steps (aka tasks) such as determining the objectives and goals (e.g. data analytics success criteria), performing a situation assessment (e.g. inventory of computing and storage resources). In phase *'(b) data understanding'*, the tasks includes the creation of a couple of reports about the initial data (collection and data sets), including data (collection) description, exploration options, and data quality statements. The third phase *'(c) data preparation'* refers to tasks that accurately described the data sets in question, their rational for inclusion/exclusion as well as required work on cleaning data or reformatting data if needed. The phase *'(d) modeling'* includes the selection of modeling techniques, test designs and related parameter settings to build the model. This phase depends highly on the scientific question that drives the 'smart data analytics' process as well as the next phase *'(e) evaluation'*, which stands for several tasks related to assess the analytics results w.r.t. the initial success criteria described in phase (a). The final phase *'(f) deployment'* enables a systematic way of performing analytics on a regular basis including a monitoring and maintenance plan. In our given context of *'smart data analytics'* we apply the CRISP-DM process not as an iterative process, but rather as a cyclic one refining information and models of phases as required.

According to Marbán et al. [10], CRISP-DM is the *'de facto standard for developing data mining and knowledge discovery projects'* that was one of the facts influencing our decision to use the CRISP-DM as a backbone process for our *'smart data analytics'* case studies. Another process model often mentioned in context and named after its phases is 'Sample, Explore, Modify, Model, and Assess (SEMMA)' [11] mostly driven by the SAS Institute. Although used in the business domain alongside a dedicated commercial tool, it rather concentrates on the modeling tasks in data mining compared to CRISP-DM that includes a more comprehensive 'perspective' on the overall data analytics process.

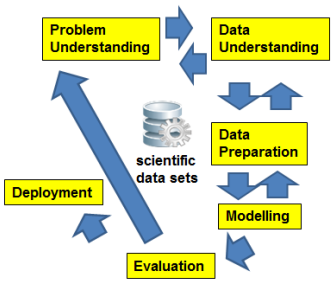


Figure 1. CRISP-DM process model (*modified from [9]*)

After the introduction of the major structure of performing *'smart data analytics'* in a systematic way, we briefly introduce our two data-intensive case studies in order to have a focus on relevant and concrete techniques required. The field (and overlaps) of data mining, machine learning, knowledge discovery, and (parallel) data analysis & analytics is very large and thus we need to concentrate on two very specific problems given real world data sets. At the same time we face the challenge that many algorithms in the field of machine learning, artificial intelligence, or data mining have not a long history in taking advantage of parallel and distributed computing techniques in the field of *'high performance computing (HPC)'* and *'high throughput computing (HTC)'*. In this contribution, it is the rate of interconnection of computing nodes that differentiates HPC (i.e. fast interconnect within clusters, e.g. Infiniband) from HTC (i.e. moderate interconnect, e.g. usual Ethernet as found in Grids and Cloud infrastructures today).

The first case study raises the demand to tackle the problem of *'automatic outlier/anomaly detection'* in measurement data sets that are part of the PANGAEA data collection [12]. With 7.9 billion data items (2013), better sensor equipment during the years, and its constant growing factor contribute to name this data source as *'big data'*. Hodge et al. [13] present a good survey of principles in outlier detection referred to as different types: *'Type 1 – Determine the outliers with no prior knowledge of the data (unsupervised clustering)'*; *'Type 2 – Model both normality and abnormality (supervised classification)'*; *'Type 3 – Model only normality or in a very few cases model abnormality (novelty detection and recognition) '*. We re-use these principles, but require a more fine-grained perspective when tackling a concrete data set with more elaborate elements for parallel approaches than mentioned in [13].

The second case study tackles the problem of *'classification of multi-spectral images'* in data sets that are part of data collections acquired by the Quickbird satellite [14]. This is just one *'sensor'* of many similar data sources that enable the study of changes in land usage, agricultural and forest climates. Experts refer to case studies like this as *'remote sensing science'* that aims to interpret information using sensors that are not in direct physical contact with the object being observed. Classification algorithms that work with these datasets have to treat both spectral and spatial information in order to obtain a satisfactory level of detection accuracy. A survey of related work in classification principles is nicely summarized by Moguerza et al. in [15] as follows: *'support vector machines (SVMs) have increasingly turned into a standard methodology in the computer science and engineering communities'*.

III. METHODS FOR SMART DATA ANALYTICS

Innovative methods for smart data analytics require a solid foundational basis in order to enable a high productivity processing of *'big data'* (i.e. towards petabytes) rather than being only usable with data sets of sizes that we often used in the past (i.e. gigabytes, terabytes). We therefore suggest that the methods are well embedded in a *'collaborative data infrastructure (CDI)'* design with additional services that optimize data access and provide added value with services that enable data sharing, persistence, and safe replication. In earlier work, Riedel et al. [16] proposed a *'data infrastructure reference model'* that follows a CDI approach and that consists of computing and storage resources relevant for our proposed concrete analytics methods. The design approach and some of its services (e.g. the B2SAFE safe data replication service) are currently implemented as part of the EUDAT data infrastructure.

In contrast to the rather elaborated reference architecture top-down work of the National Institute for Standard (NIST) big data working group [18], we aim to provide in this contribution a bottom-up approach based on concrete scientific case studies and their required reference architecture elements. In [6] we therefore suggest a derived concrete architecture design tailored for *'big data analytics'* in a context of a CDI while this section aims to refine this architectural setup based on concrete techniques and tools that are used in the context of two specific case studies. Scientific results of them are expected to be published in domain-specific publications and thus this paper focusses on the pure technical findings of combining a wide variety of available tools and frameworks for *'smart data analytics'*. Selected elements of the CRISP-DM process model tailored to our problem domain are given in context of the two case studies while full reports are already available or expected to become available as part of the EUDAT data infrastructure.

A. Automatic outlier detection in measurement datasets

By following the tailored CRISP-DM model phase *'(a) problem/business understanding'* we define the goal to automatically detect outliers in large quantities of measurement data, because manual *'quality control'* is slow, error-prone, and becomes increasingly impossible as the amount of data set grow significantly in type and volume. The architectural setup in Figure 2 illustrates selected parts of the inventory of computing and storage resources such as the PANGAEA collection archive (i.e. community storage data sets) and HPC resources (i.e. computing nodes available at the Juelich Supercomputing Centre and UoIceland respectively).

Selected elements of phase *'(b) data understanding'* include the task to collect initial data out of the large PANGAEA data collection that in our case refers to initial measurement datasets acquired by a cabled underwater observatory sensor in the open Swedish Fjord Koljoe Fjord (approximately 100 KM north of Gothenburg) installed as part of the HYPOX2 project [17]. The data sets are provided by PANGAEA data scientists that preserve and manage data sets in their own domain-specific data infrastructure.

The third phase *'(c) data preparation'* includes tasks of describing the structure of the data. In our case study we work with datasets that consists of a time series of a month/file (i.e. tab separated data files) with underwater measurements. We furthermore had to remove some metadata elements in the plain data files that created problems when using it with the R Statistical Computing tool [20].

The phase *'(d) modeling'* includes the selection of modeling techniques and in our case also the related appropriate outlier detection algorithms. There are currently two parallel approaches we explore as part of our 'smart data analytics' methods: (i) using the statistical computing tool R with the RMPI and outlier package (cf. Figure 2); (ii) using a new implementation by G. Fox et al. of the DOoR algorithm described by Bhaduri et al. in [22]. The automation of the approach is provided by using a HPC middleware (i.e. UNICORE [23]) that provided a seamless access to large-scale supercomputers. R scripts and the execution of outlier algorithms are performed with this system using workbench tools (cf. Figure 2).

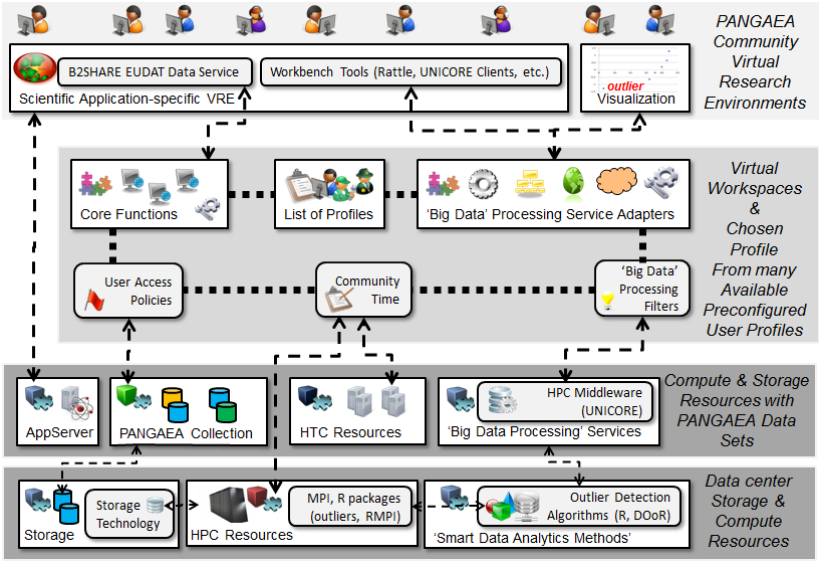


Figure 2. Architectural deployment for the automatic outlier detection in PANGAEA measurement datasets.

The *(e) evaluation'* phase, which stands for several tasks related to assess the analytics results w.r.t. the initial success criteria described in phase (a) is twofold: Firstly on a technical basis using data mining techniques; secondly, results of the automatic outlier detection are regularly shared with scientific experts from PANGAEA by using the *'B2SHARE EUDAT Data Service'* (cf. Figure 2). They evaluate which *'smart data analytics'* elements make sense and what approaches are not helpful. In addition, often used outlier data visualizations are available (e.g. normal probability plots, boxplots, histograms) as needed to further support this process (cf. Figure 2 top right).

The final phase *'(f) deployment'* enables a systematic way of performing automated outlier detection on PANGAEA datasets once the tests with initial datasets have been shown successful over time. This concrete 'smart data analytics method' is well embedded in a sustainable data infrastructure offering storage resources and services mostly independent of project funding, because of a network of trusted data centers in EUDAT and PRACE.

B. *Classification of multi-spectral images*

We start again with our tailored CRISP-DM process model, but perform a quicker walk-through as the ideas of each phase are the same, but the given data analytics problem is usually different. In phase *'(a) problem/business understanding'* we define the goal of performing a data classification of buildings in street map with high accuracy using multi-spectral image datasets. These data sets constantly grow in volume and thus require a more powerful classification approach and thus we define a *'speed-up of data classification'* as a clear objective in this particular case study. The architectural setup in Figure 3 illustrates selected parts of the inventory of computing and storage resources such as parts of the Quickbird satellite data collection (i.e. community storage data sets) and HTC resources (i.e. computing nodes available at the UoIceland and as part of FutureGrid [24] respectively).

Phase *'(b) data understanding'* include the task to collect initial multi-spectral image data sets out of the large commercial Quickbird data collection that in our case is done by the domain scientists that work with us.

In the phase *'(c) data preparation'* we perform tasks to describe the structure of the data. In our case study we work with datasets that are in fact different types of images acquired by the QuickBird satellite: (i) a high-resolution panchromatic image or resolution 0.6m and a low-resolution (2.4m) multi-spectral image.

The phase *'(d) modeling'* includes the selection of modeling techniques and in our case also the related appropriate parallel data classification algorithm. As described above, one of the most commonly applied principle technique in data classification are support vector machines (SVMs) [21]. There is currently one parallel approach we explore as part of our 'smart data analytics' method: using Twister [25,26] as shown in Figure 3. Twister improves the original map-reduce [27] approach (e.g. used in Apache Hadoop [28]) with iterative capabilities, which can be used to train an SVM classifier as described by Zhang and Fox in [29]. We need to mention here that the implementation in turn is based on the widely known libSVM library [30].

Also in this case study, the *(e) evaluation'* phase is twofold: Firstly on a technical basis using machine learning techniques separating the labelled data in three distinct sets: training data, test data and validation data; secondly, results of the data classification results are regularly shared with scientific experts by using the *'B2SHARE EUDAT Data Service'* (cf. Figure 3). They evaluate which *'smart data analytics'* elements make sense and what data classifiers have not enough accuracy. In addition, visualizations are available (e.g. speed-up diagrams compared to serial approaches) as needed to further indicate that the training time of the SVM classifier was significantly faster. This is especially the case when considering the observed used manual approaches by using Matlab SVMs [31].

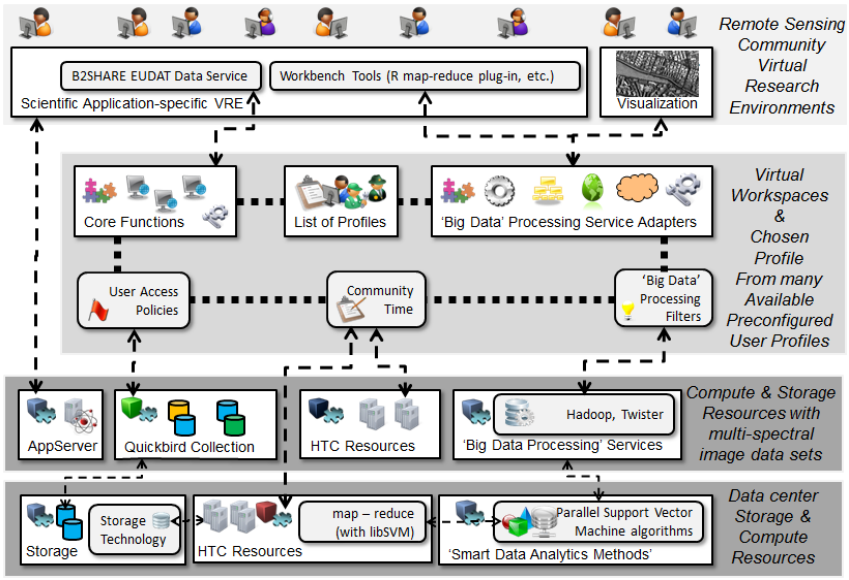


Figure 3. Architectural deployment for the data classification using parallel support vector machines for Quickbird masurement datasets.

The final phase *'(f) deployment'* enables a systematic way of performing data classification of multi-spectral image datasets once the tests with initial Quickbird image datasets have shown good accuracy of the trained parallel SVM classifiers. This concrete *'smart data analytics method'* is well embedded in a sustainable data infrastructure offering computing and storage resources mostly independent of project funding, because of a network of trusted data centers in EUDAT with computational resources .

IV. CONCLUSIONS

This section gives the reader some concluding remarks although we are in the early phases of our *'smart data analytics process'* in the described case studies – but that is a process that might be never finished given the ever increasing amounts of big data sets. We are nevertheless able to derive some initial conclusions about structuring the process and thus performing *'big data analytics'* in a systematic way in the concrete area of *'outlier detection'* and *'image data classification'* of scientific datasets. Detailed findings that relate to the clever and innovative combinations of technology required for the 'smart data analytics' described in this paper are actively contributed and discussed as part of the Big Data Analytics (BDA) [19] interest group of the Research Data Alliance (RDA). This enables the guidance of non-experts for *'big data analytics'* and sharing of findings with the broader community (e.g. with members of the US NIST big data activities).

We firstly conclude that the general reference model design of a CDI offers 'added value services' (e.g. data sharing, replication, access optimization, data transfers) to the *'smart data analytics process'*. It offers a solid foundation to perform *'higher level analytics'* based on concrete algorithms and tools without the need to tackle simple problems such as deploying data transfer protocols. We observe that services like B2SHARE represent an important element in a seamless cooperation between general data analytics experts and scientific domain-specific experts for rather temporary scientific data results whereby only a fraction qualifies for long-term data preservation (e.g. using the B2SAFE EUDAT data service).

Secondly we are able to conclude that the CRISP-DM is well suited as a backbone for a process we referred to as *'smart data analytics'* although being originally designed for traditional data mining and knowledge discovery. Tailoring this model by describing large-scale compute and storage resources, clarifying the concrete machine learning algorithms and data mining techniques, tools, and service deployments make the approach feasible to guide the *'high productivity data processing'* of large quantities of scientific data. These findings are considered to be part of a report compiled by the RDA BDA group.

Thirdly, we can conclude that the architectural setup of the method for automatic outlier detection is plausible, however, we also have to acknowledge that the work on tailoring the algorithm to the specific problem at hand requires many (parameter) configurations (e.g. within the HPC system UNICORE). Nevertheless, we identified all the required elements and implementations in order to use the method in practice today. Tailoring the CRISP-DM process model, when performing outlier detection, is very useful too, because of many reasons one of which is the clarity of describing all relevant aspects of the concrete proposed method (resources, data, algorithms, tools, etc.). This clarity is an important fact during the communication of the general data analytics expert and the PANGAEA scientists. Using an easy and open approach for sharing the data for validation purposes by using the B2SHARE EUDAT data service was generally viewed as an important element of the method.

We can further conclude that the architectural setup of using parallel SVMs for classification of multi-spectral image data is possible to implement today. Nevertheless, the parameter tuning and creating the classifiers is not straightforward and needs time to achieve reasonable results. These findings are important, but further raise the demand for several software packages that need to be maintained and installed on relevant underlying hardware resources (i.e. Twister and libSVM). Hence, achieving a speed-up required for 'big data' is possible, but also requires a distributed system (e.g. with map-reduce). Using CRISP-DM and the B2SHARE EUDAT data service was equally useful in this case study as described in the concluding remarks of the previous case study.

ACKNOWLEDGMENT

The work is partially funded by the EUDAT project that is co-funded by the EC 7th Framework Programme (Grant Agreement no: 283304). We thank Robert Huber (MARUM Bremen, Germany) for his close cooperation in the context of the PANGAEA outlier detection case study. We are also grateful to Gabriele Cavallaro (University of Iceland) for his close cooperation in the context of the spectral image analysis. We further thank Geoffrey Fox and Zhanquan Sun for their consultation and sharing of implementations in context of parallel data analytics algorithms. Finally, thanks to Tomas Philip Runarsson for his advice in mathematical aspects of statistical data mining algorithms.

REFERENCES

[1] Wood J, et al.*,'Riding the wave - How Europe can gain from the rising tide of scientific data'*, European Union, Report of the High Level Expert Group of Scientific Data, 2010.

[2] KE Partners, *'A Surfboard for Riding the Wave - Towards a four country action programme on research data',* 2011

[3] DOE ASCAC Data Subcommittee Report, *'Synergistic Challenges in Data-Intensive Science and Exascale Computing'*, 2013

[4] Shearer C., '*The CRISP-DM model: the new blueprint for data mining'*, J Data Warehousing (2000); 5:13—22.

[5] Provost F. and Fawcett T., *'Data Science and its relationship to Big Data and Data-Driven Decision Making'*, doi: 10.1089/big.2013.1508, Big Data, Vol.1(1), 2013

[6] Riedel M., Memon M., Memon A., Fiameni G., Cacciari C., and Lippert Th.**, '***High Productivity Processing - Engaging in Big Data around Distributed Computing***'**, Proceedings of the 36th MIPRO Convention, 2013

[7] Lippert Th., Eickermann Th., and Erwin D., 'PRACE: Europe's Supercomputing Research Infrastructure', Advances in Parallel Computing, Vol. 22, doi:10.3233/978-1-61499-041-3-7, 2012

[8] Lecarpentier, D; Wittenburg, P; Elbers, W; Michelini, A; Kanso, R; Coveney, P; Baxter, R; *'EUDAT: A New Cross-Disciplinary Data Infrastructure for Science'*, International Journal of Digital Curation, Vol. 8 (1) 279 – 287, 2013

[9] Pete Chapman, *'CRISP-DM User Guide'*, 1999, Online: http://lyle.smu.edu/~mhd/8331f03/crisp.pdf

[10] Marbán Ó., Mariscal G., Segovia J., *'A Data Mining & Knowledge Discovery Process Model'*, Book Data Mining and Knowledge Discovery in Real Life Applications, ISBN 978-3-902613-53-0, pp. 438-453, 2009

[11] Azevedo A. and Santos M.F., *'KDD, SEMMA and CRISP-DM: a parallel overview'*, in Proceedings of the IADIS European Conference on Data Mining 2008, pp 182-185.

[12] PANGAEA – Data Publisher for Earth & Environmental Science, 2014, Online: http://www.pangaea.de/

[13] Hodge V.J. and Austin J., *'A survey of Outlier Detection Methodologies'*, Artificial Intelligence Review 22: 85–126, 2004

[14] Quickbird Satellite Sensor, Satelllite Imaging Corporation, Online: http://www.satimagingcorp.com/satellite-sensors/quickbird.html

[15] Moguerza J.M. and Munoz A., 'Support Vector Machines with Applications', Statistical Science, Vol. 21, No. 3, pp. 332 – 336, doi:10.1214/088342306000000493

[16] M. Riedel, P. Wittenburg, J. Reetz, M. van de Sanden, J. Rybicki, B. von St. Vieth, G. Fiameni, G. Mariani, A. Michelini, C. Cacciari, W. Elbers, D. Broeder, R. Verkerk, E. Erastova, M. Lautenschlaeger, R. Budig, H. Thielmann, P. Coveney, S. Zasada, A. Haidar, O. Buechner, C. Manzano, S. Memon, S. Memon, H. Helin, J. Suhonen, D. Lecarpentier, K. Koski and Th. Lippert *'A Data Infrastructure Reference Model with Applications: Towards Realization of a ScienceTube Vision with a Data Replication Service',* Journal of Internet Services and Applications, Volume 4, Issue 1

[17] HYPOX2 Project Web Page, *'In situ monitoring of oxygen depletion in hypoxic ecosystems of coastal and open-seas, and land-locked water bodies'*, Online: http://www.hypox.net/

[18] NIST Big Data Public Working Group, Online: http://bigdatawg.nist.gov/home.php

[19] Research Data Alliance (RDA), Big Data Analytics (BDA) interest group, Online: https://rd-alliance.org/internal-groups/big-data-analytics-ig.html

[20] Ross I., and Gentleman R., *'R: a language for data analysis and graphics'*, Journal of computational and graphical statistics, Vol 5.3, pp. 299-314, 1996

[21] Tan P., Steinbach M., and Kumar V., *'Introduction to Data Mining'*, ISBN-10: 0321321367, Addison Wesley, 2006

[22] Bhaduri K., Matthews B., Giannella C., *'Algorithms for Speeding up Distance-Based Outlier Detection'*, in proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, pp. 859-867, 2011

[23] Streit A., Erwin D., Lippert Th., Mallmann D., Menday R., Rambadt M., Riedel M., Romberg M., Schuller B., Wieder Ph., *'UNICORE – From project results to production grids'*, Advances in Parallel Computing, Volume 14, pp. 357–376, 2005

[24] von Laszewski G., et al., *'Design of the futuregrid experiment management framework'*, Proceedings of the Gateway Computing Environments Workshop (GCE), 2010.

[25] Ekanayake J., Li H., Zhang B., Gunarathne Th., Bae S. Qiu, Fox G., *'Twister: a runtime for iterative mapreduce'*, in proceedings of the 19th ACM International Symposium on High Performance Distributed Computing. ACM, 2010.

[26] Qiu J., Ruan Y., Wu T., B. Zhang, A. Hughes, G. Fox., *'Hybrid cloud and cluster computing paradigms for life science applications'*, BMC bioinformatics 11.Suppl 12, 2010

[27] Dean, J. and Ghemawat S., *'MapReduce: simplified data processing on large clusters'*, *Communications of the ACM* 51.1, pp. 107-113, 2008

[28] White T., 'Hadoop: The Definitive Guide', 2nd Edition, O'Reilly Media, ISBN 978-1-4493-8973-4 2010

[29] Sun Z., and Fox G., 'Study on Parallel SVM Based on MapReduce', In Proceedings of the international conference on parallel and distributed processing techniques and applications, 2012.

[30] Chang C., and Lin C., *'LIBSVM: a library for support vector machines'*, in ACM Transactions on Intelligent Systems and Technology (TIST) 2.3, 27, 2011

[31] Pelckmans K., et al., *'LS-SVMlab: a matlab/c toolbox for least squares support vector machines'*, Tutorial, KU Leuven-ESAT, Leuven, Belgium, 2002

Optimal Block Size for Matrix Multiplication Using Blocking

Sasko Ristov, Marjan Gusev
University Ss Cyril and Methodius, FCSE,
Rugjer Boshkovic 16,
Skopje, Macedonia
Email: {sashko.ristov,marjan.gushev}@finki.ukim.mk

Goran Velkoski
Innovation LLC,
Belasica 2,
Skopje, Macedonia
Email: goran.velkoski@innovation.com.mk

***Abstract*—Matrix multiplication is a widely used algorithm in today's computing. Speeding up the multiplication of huge matrices is imperative for scientists and they are trying to discover the fastest algorithm. Blocking the matrices reduces the cache misses since both blocks can be stored in L1 cache and thus only the first access of an element will result in cache miss and all other access will be a cache hit. However, choosing the block size is not the only optimization. In this paper we analyze the impact of various block size ($M \cdot K$ and $K \cdot N$) on the performance. The realized experiments use different parameter values for K and predefined values of the parameters M and N, to test algorithm behavior in different cache regions.**

The results of the experiments show three phenomena. The first phenomenon is that if M is greater than N, then choosing the first block $M \cdot N$ will achieve a significantly greater speed. The second states that, if the second parameter N is increased for constant M has no significant influence on the performance. Finally, the third phenomenon is that the speed decreases significantly if N is increasing for constant M.

***Index Terms*—Cache Memory, Matrix Multiplication, Processor, Performance.**

I. Introduction

Today's world is unthinkable without efficient matrix multiplication. Apart of being an essential linear algebra algorithm for solving linear systems of equations, it is widely used for linear transformations in image processing. That is, moving some points (translation), changing left- and right-handed ordering (reflection), and a combination of both are all operations that require matrix multiplication. The increased image size and resolution require faster operations, and if possible, in real time.

Matrix multiplication is a compute intensive algorithm with $O(N^3)$ complexity, memory demanding with $O(N^2)$ complexity. It is also defined as a cache intensive algorithm [1] of order matrix size N because each element is accessed N times. Due to these characteristics, the matrix multiplication requires huge computing (faster CPU with huge cache and big RAM memory) and storage resources for multiplying greater matrices. Big matrices cannot be placed in the cache memory, which generates a lot of the cache misses and reduces the algorithm performance.

Sscientists are enforced to optimize the algorithm for a specific hardware parameters. Dividing the matrices to blocks and multiplying each block among each other enforces the multiplications to work only with a small amount of data that can be stored in the L1 cache, which significantly improves the performance, especially when multiplying greater matrices [2].

However, the cache size is not the only important cache parameter. Cache replacement policy and cache type (inclusive or exclusive) are important cache parameters, but they impact more the naive matrix multiplication algorithm, rather than the blocking algorithm, since the number of cache misses is reduced. However, cache line (and the matrix elements' size), cache levels, n-way set associativity are more important, especially the latter [3]. In our previous work [4], we improved the blocking algorithm by choosing the blocks as rectangles, rather than squares. This was more emphasized for AMD processors due to their small associativity. In this paper, we extend this research to determine the optimal block size of the blocking algorithm, that is, we analyze the impact of the selected sizes ($M \cdot K$ and $K \cdot N$) when the same number of operations is executed. Several default values for the parameters M and N are chosen, and the value of parameter K varies within a given range for the behavior in different cache regions to be analyzed [1].

The rest of the paper is organized as follows. Section II presents the related research that we found in the literature, which speed up the matrix multiplication. We present the testing methodology used for the experiments in order to provide reliable results in Section III. The results of the experiments are presented in Section IV. The results are discussed in Section V. Finally, we conclude the paper and present the plans for future work in Section VI.

II. Related Work

This section presents the various techniques to speedup the matrix multiplication algorithm. We have grouped them in several classes: Reducing the number of operations, reducing the average memory access time, using many core processors and parallelization, and other various techniques will speedup the matrix multiplication. Anchev et al. [5] overviews several different techniques. An obvious speeding up is to run the multiplication on a faster CPU, but the CPU clock frequency is reaching a limit [6], also known as power wall.

A. Speeding up by reducing the number of operations

The goal of the researchers is to reduce the asymptotical power of N (matrix size) for matrix multiplication algorithm, i.e., the number of floating point operations. The first revolutionary reduction was proposed by the well known Strassen's algorithm [7], which reduced the power from 3 to 2.8, replacing some multiplication operations with add. Several other authors decreased this value to 2.3727 [8].

However, these theoretical improvements are efficient only for huge matrices that cannot be stored in the memory, and the cache size does not affect the average behavior a lot.

B. Using parallelization

Greater speedup is achieved if parallelization is used. Matrix multiplication algorithm is a school example of good algorithm for parallelization because it does not have data dependency. There are many examples of parallelization on CPU by using OpenMP, MPI, and MPI with OpenMP. Another example is using CUDA, or even MPI with CUDA if GPGPU cluster is used, as well as using the data flow and systolic arrays [9].

C. Speeding up by reducing the average memory access time

The total execution time consists of two parts: the processing time and the memory access latency. Therefore, the idea of speeding up the average value of memory access can provide a sufficient speedup, usually higher than the approach based on reducing the number multiplication operations.

Matrix multiplication is a cache intensive algorithm and its performance depends on the cache memory behaviour. Ristov and Gusev showed why and when the performance drawback will appear for a particular cache memory parameters [3]. A possible solution to avoid the associativity problem is padding the first element of the matrix [10].

Another very important algorithm is blocking the matrices. This algorithm even increases the number of operations, but significantly reduces the average memory access time. Our improved 2D/1D algorithm [4] is faster than the blocking algorithm with squared blocks for AMD processors due to their small associativity.

In this paper we focus on reducing the memory access time since this speeds up the cache intensive algorithms.

D. Other techniques

Drevet et al. [11] analyzed the optimization techniques for small matrix multiplication. Loop unrolling is another technique that uses the processor pipelining.

III. Testing Methodology

In this section we present the testing methodology used for the experiments as we can achieve reliable results.

A. Testing Algorithm

We use a multiplication of two matrix blocks. Denote the blocks of the first matrix A by $BA_{M \cdot K}$ and the blocks of the matrix B by $BB_{K \cdot N}$. The matrix elements are stored as double precision numbers, each with 8 bytes.

TABLE I
Test cases

Test Case	Notation	$BA_{M \cdot K}$	$BB_{K \cdot N}$
1	56x56	$56 \cdot K$	$K \cdot 56$
2	56x224	$56 \cdot K$	$K \cdot 224$
3	56x896	$56 \cdot K$	$K \cdot 896$
4	224x56	$224 \cdot K$	$K \cdot 56$
5	224x224	$224 \cdot K$	$K \cdot 224$
6	224x896	$224 \cdot K$	$K \cdot 896$
7	896x56	$896 \cdot K$	$K \cdot 56$
8	896x224	$896 \cdot K$	$K \cdot 224$
9	896x896	$896 \cdot K$	$K \cdot 896$

B. Testing Environment

The testing environment consists of Intel(R) Xeon(R) CPU X5647 @ 2.93GHz with 8GB RAM memory. The CPU is quad core, and each core has its own private L1 (32KB) and L2 (256KB) caches. All 4 cores share L3 cache of 12MB.

The server is configured with Linux Ubuntu 12.04 operating system and C++ is compiled with gcc compiler.

C. The Test Cases

Several test cases are conducted with predefined values of parameters M and M, and by varying all integer values of the parameter K in a given region to analyze the performance for fitting the blocks in all cache regions. The region $1 \leq K \leq 600$ was used for smaller predefined values of M and N, and the region $1 \leq K \leq 300$ for greater values.

The three values of M and N, i.e., 56, 224 and 896 are chosen to avoid the associativity problem [3]. That is, 56 matrix elements can be placed in 7 cache blocks, rather than problematic 8 blocks. The other two values are a product of 56 by 4 and 16, again avoiding associativity problem. Each combination of pairs (M, N) covers fitting of blocks in corresponding cache regions L_1 to L_3. L_4 is used as a level 4 cache or main memory if missing.

D. Test Data

We measure the execution time $T(M, N, K)$ for a particular value of the parameter K in each test case. Next, we use $T(M, N, K)$ to calculate the *Speed* $V(M, N, K)$ in each test case by using the definition for effective speed [12]. That is, the speed $V(M, N, K)$ is calculated by dividing the number of operations and measured execution time as defined in (1), for each test case for particular values of the parameters M and N.

$$V(M, N, K) = \frac{2 \cdot M \cdot N \cdot K}{T(M, N, K)} \tag{1}$$

E. Test Goals

The test cases have three goals. The first goal is to determine which pair of block sizes provides better performance if the blocks have the same number of elements, but exchanged parameters. That is, to compare the performance of test case pairs (2, 4), (3, 7) and (6, 8). For example, we multiply the blocks of test case 2 where $M_2 = 56 < N_2 = 224$. Figure 1

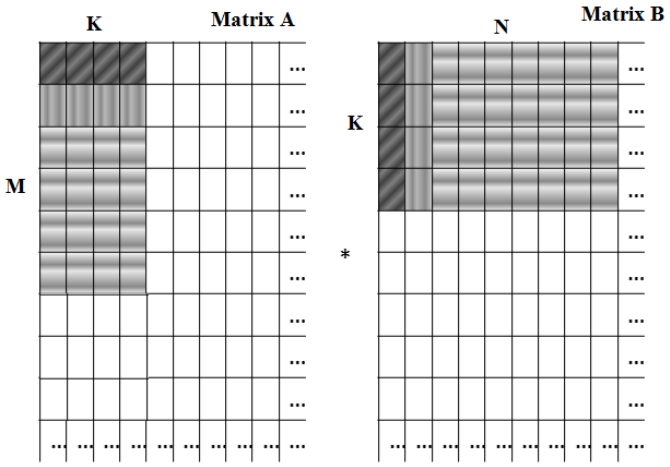


Fig. 1. Example of test cases 2, 3 and 6, where $M < N$ (Horizontal Rectangle)

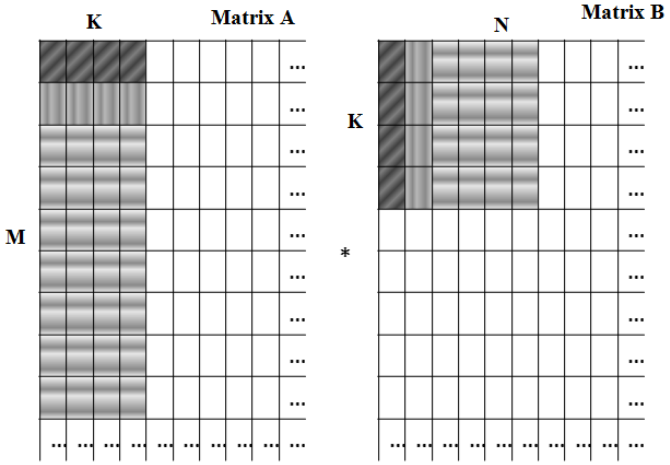


Fig. 2. Example of test cases 4, 7 and 8, where $M > N$ (Vertical Rectangle)

depicts an example of test case 2 with $M = 6 < N = 9$. The parameter K is variable.

Further on, we compare the achieved speed with test case 4, whose values for the parameters M and N are exchanged, that is, $M_4 = N_2 = 224 > N_4 = M_2 = 56$. Figure 2 depicts an example of test case 4 with exchanged values of test case 2, i.e., $M = 9 < N = 6$. The parameter K also varies.

The second goal is to determine the impact of the parameter M and the third goal is to determine the impact of the parameter N, varying the parameter K as explained in Section III-C.

IV. The Results

This section presents the results of the experiments. We are analyzing which block matrix rectangle multiplication provides the best performance for the same number of operations. Further on, we analyze the performance impact of the parameters M and N, while one of them is constant.

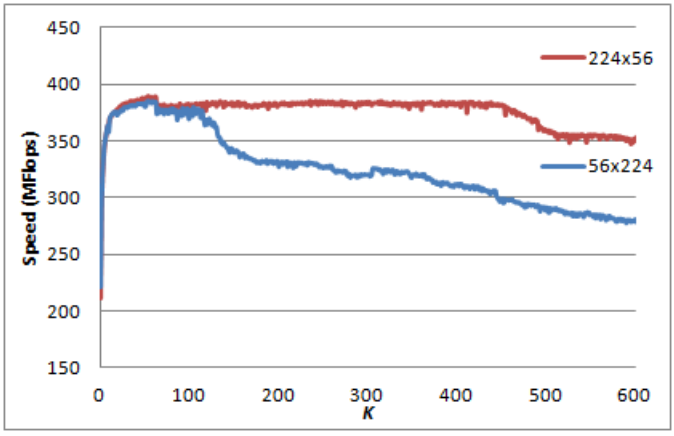


Fig. 3. The speed as a function of parameter K for test cases 56x224 vs 224x56

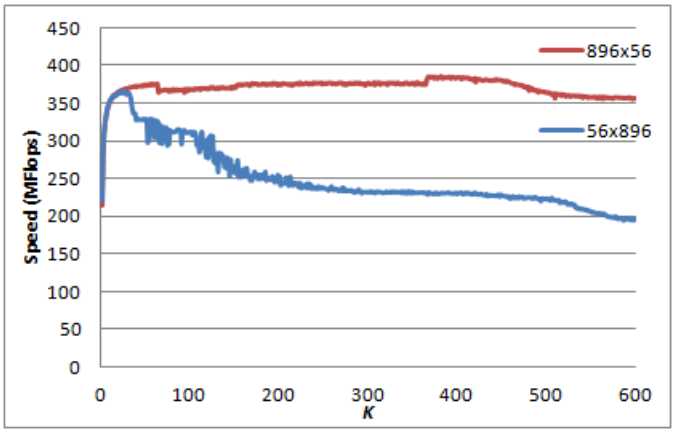


Fig. 4. The speed as a function of parameter K for test cases 56x896 vs 896x56

A. Horizontal or vertical rectangle

Figure 3 compares the results of test cases 56x224 vs 224x56. Both curves are identical until $K = 116$ when the speed for test case 56x224 decreases when K is increased. This is the point when both matrices together exceed L2 cache of 256KB.

The comparison of the results of the test cases 56x896 vs 896x56 is depicted in Figure 4. Very similar results are observed as the previous comparison. That is, both curves are identical very shortly until $K = 35$, which is also the point when both matrices together exceed the L2 cache size.

Very similar results are achieved of the comparison of the test cases 224x896 vs 896x224, which are depicted in Figure 4. The threshold is $K = 30$ depending of the L2 cache size.

B. Performance impact of the parameter M

This section elaborates the results of the comparison of the test cases to determine the impact of a particular rectangle height, that is, M, if the other parameter N is constant.

Figure 6 compares the results of test cases 56x56, 224x56 and 896x56, that is, different values of M for constant value

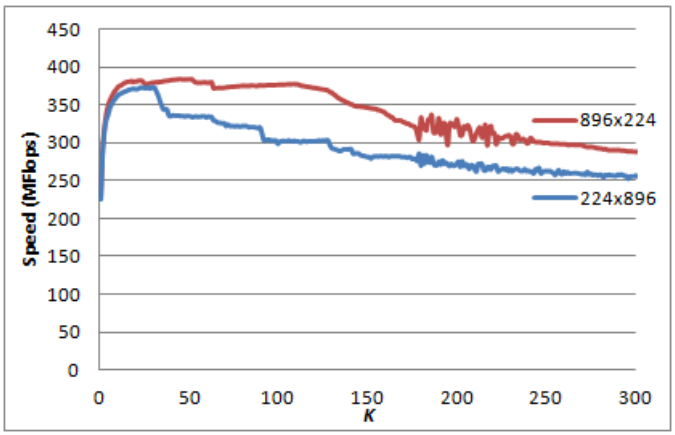


Fig. 5. The speed as a function of parameter K for test cases 224x896 vs 896x224

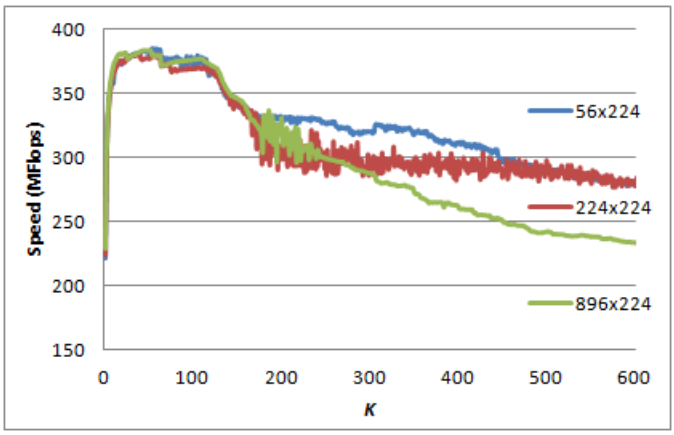


Fig. 7. The speed as a function of parameter K for test cases 56x224, 224x224 and 896x224

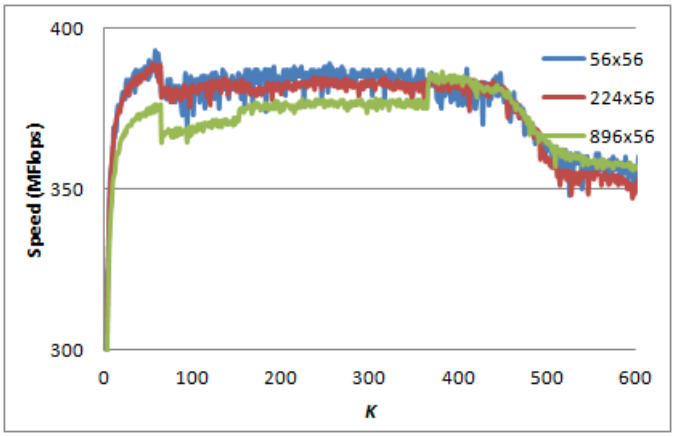


Fig. 6. The speed as a function of parameter K for test cases 56x56, 224x56 and 896x56

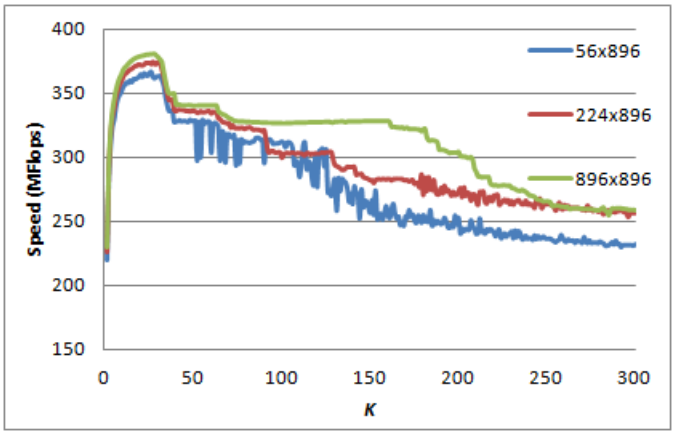


Fig. 8. The speed as a function of parameter K for test cases 56x896, 224x896 and 896x896

$N = 56$. All three curves are similar regardless of the parameter K. This means that increasing the parameter M does not impact on the performance.

Two regions are observed with a threshold $K = 450$, i.e., the left region with greater performance, and the right region with lower performance. Also, there is a small discrepancy for the test case 896x56. It has worse performance than others in the left region, but better performance in the right region.

Similar results, but moved to the left, can be observed in Figure 7, which compares the results of the test cases 56x224, 224x224 and 896x224. These matrices are greater than those with $N = 56$, and therefore, the performance of the test case 896x224 suffers for greater K. But still, all three curves have very similar performance until $K = 160$.

Figure 8 depicts different results for test cases 56x896, 224x896 and 896x896 compared to both previous comparisons for huge K. However, for small K, the curves are also similar.

C. Performance impact of the parameter N

This section elaborates the results of the comparison of the test cases to determine the impact of particular rectangle width, that is, N, if the other parameter M is constant.

Figure 9 compares the results of test cases 56x56, 56x224 and 56x896, that is, different values of N for constant value $M = 56$. The results show that increasing the parameter N negatively impacts the algorithm performance, which is emphasized for greater values for parameter K.

The same conclusion can be derived when comparing the test cases 224x56, 224x224 and 224x896, whose results are depicted in Figure 10. Also, the results of the comparison of the test cases 896x56, 896x224 and 896x896 are similar, as Figure 11 depicts.

V. Discussion

A general conclusion for the three comparisons of the test cases to determine the impact of parameters M and N with the same number of operations (greater height or width of the block rectangle) is that choosing a rectangle depends only when both matrices cannot be placed in L2 cache, favoring the rectangle with greater M, that is, the vertical rectangle. Another important result is that the speed is increased until L2 for horizontal rectangle, while the speed keeps its value for

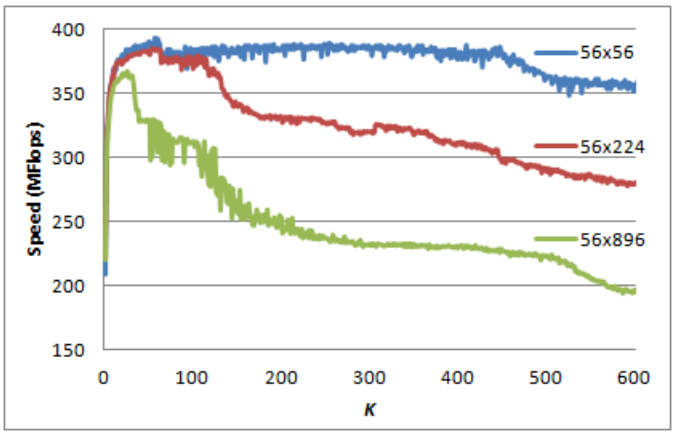


Fig. 9. The speed as a function of parameter K for test cases 56x56, 56x224 and 56x896

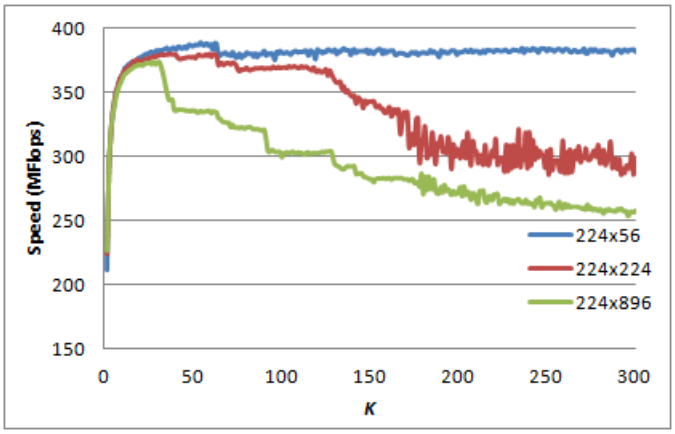


Fig. 10. The speed as a function of parameter K for test cases 224x56, 2224x224 and 224x896

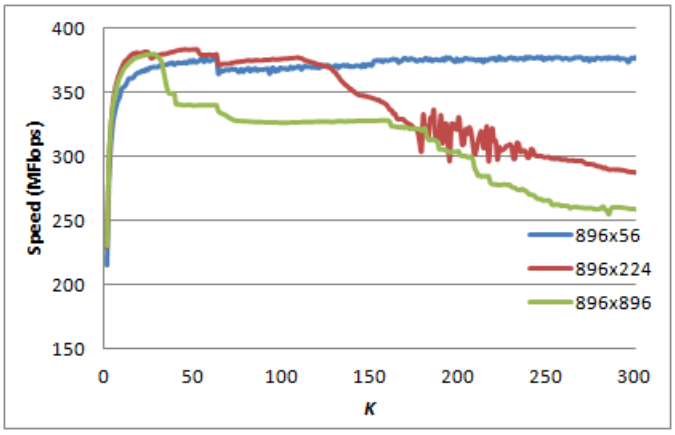


Fig. 11. The speed as a function of parameter K for test cases 896x56, 896x224 and 896x896

vertical rectangle. This leads to conclusion that the blocking can be choosed with greater blocks since the number of operations will be smaller, and thus the overall execution will be faster than traditional blocking.

The impact of the separated parameters M and N is also analyzed. Their impact on the algorithm performance is totally different, but the common feature is that the impact is increased for greater values of parameter K.

Comparing the figures 6, 7 and 8, we can conclude that the parameter M has a small impact to the performance, especially for smaller values of the parameter N. Again, this also leads to the similar conclusion, that blocks can be even greater than L1 cache.

The conclusion for the parameter N is totally opposite. Figures 9, 10 and 11 concludes the same, that is, the increasing of the parameter N significantly reduce the algorithm performance.

VI. Conclusion and Future Work

Dividing matrices into blocks that can be placed in L1 cache speeds up the matrix-matrix execution because it reduces the cache misses. That is, once the block matrix elements are placed in the L1 cache, they are used several times by their common multiplication.

The results of the experiments conducted in this research show three phenomena:

- Choosing as much greater height M of the first matrix provides better performance for the same product $M \cdot N$;
- The value of the parameter M can be chosen arbitrary, and therefore, it should be selected according to the value of already selected value of parameter N; and
- Parameter N should be as small as possible.

Apart of these phenomenons, we determine that greater speeds can be achieved if both blocks (by one of each matrix) are greater than L1 cache size, but smaller than the L2 cache size. These results could speed up the matrix multiplication with blocking the matrices even more.

Although these results are compliant with our previous results for AMD CPU [4], we will conduct the same experiments on AMD CPUs, as well as increase the number of the experiments for other values of parameters M and N. Another research will be towards parallelization of these experiments using multi-core CPUs, as well as GPUs, which also have set associative caches.

References

[1] M. Gusev and S. Ristov, "Matrix multiplication performance analysis in virtualized shared memory multiprocessor," in *MIPRO, 2012 Proceedings of the 35th International Convention*. IEEE, 2012, pp. 251–256.

[2] J. L. Hennessy and D. A. Patterson, "Computer Architecture, Fifth Edition: A Quantitative Approach," MA, USA, 2012.

[3] S. Ristov and M. Gusev, "Achieving maximum performance for matrix multiplication using set associative cache," in *The 8th International Conference on, Computing Technology and Information Management (ICCM2012), IEEE Conference Publications*, ser. ICNIT '12, vol. 2, 2012, pp. 542–547.

[4] M. Gusev, S. Ristov, and G. Velkoski, "Hybrid 2d/1d blocking as optimal matrix-matrix multiplication," in *ICT Innovations 2012*, ser. Advances in Intelligent Systems and Computing, S. Markovski and M. Gusev, Eds. Springer Berlin Heidelberg, 2013, vol. 207, pp. 13–22. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-37169-1_2

[5] N. Anchev, M. Gusev, S. Ristov, and B. Atanasovski, "Some optimization techniques of the matrix multiplication algorithm," in *Information Technology Interfaces, Proceedings of the ITI 2013 35th International Conference on, IEEE Conference Proceedings*, Cavtat, Croatia, Jun. 2013, pp. 71–76.

[6] C. Mack, "Fifty years of moore's law," *Semiconductor Manufacturing, IEEE Transactions on*, vol. 24, no. 2, pp. 202–207, May 2011.

[7] V. Strassen, "Gaussian elimination is not optimal," *Numerische Mathematik*, vol. 13, no. 4, pp. 354–356, 1969. [Online]. Available: http://dx.doi.org/10.1007/BF02165411

[8] V. V. Williams, "Multiplying matrices faster than coppersmith-winograd," in *Proceedings of the 44th Symposium on Theory of Computing*, ser. STOC '12. New York, NY, USA: ACM, 2012, pp. 887–898. [Online]. Available: http://doi.acm.org/10.1145/2213977.2214056

[9] I. Milovanovic, T. Tokic, M. Stojcev, E. Milovanovic, and N. Novakovic, "Mapping matrix multiplication algorithm onto optimal fault-tolerant systolic array," in *Microelectronics, 2000. Proceedings. 2000 22nd International Conference on*, vol. 2, 2000, pp. 711–714 vol.2.

[10] S. Williams, L. Oliker, R. Vuduc, J. Shalf, K. Yelick, and J. Demmel, "Optimization of sparse matrix-vector multiplication on emerging multicore platforms," *Parallel Comput.*, vol. 35, no. 3, pp. 178–194, Mar. 2009.

[11] C.-E. Drevet, M. N. Islam, and E. Schost, "Optimization techniques for small matrix multiplication," *Theoretical Computer Science*, vol. 412, no. 22, pp. 2219 – 2236, 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0304397510007036

[12] B. Lipshitz, G. Ballard, J. Demmel, and O. Schwartz, "Communication-avoiding parallel strassen: implementation and performance," in *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*, ser. SC '12. IEEE Computer Society Press, 2012, pp. 101:1–101:11.

Extensible Framework for Execution of Distributed Genetic Algorithms on Grid Clusters

D. Georgiev* and E. Atanassov*
* Institute of Information and Communication Technologies, BAS, Sofia, Bulgaria
dobromir@parallel.bas.bg, emanouil@parallel.bas.bg

Abstract - Genetic algorithms are effective metaheuristic optimization methods, based on the principles of natural selection and genetics. Although they are able to find adequate solutions in acceptable time for small problems, their execution time increases rapidly for problems with large search space and complex fitness landscape. Computational grids provide an appropriate platform for the execution of parallel genetic algorithms. However, when using multiple grid clusters for a single computational task, one has to comply with a number of technical and administrative limitations.

In this paper we describe a new implementation of GA, written in C++11, which provides a mechanism for parallel execution and an extensible programming interface. Our framework utilizes MPI, ZeroMQ and the newest generic programming facilities of C++, enabling the execution of a distributed GA on multiple grid clusters. Various tests were performed on clusters from the European Grid Infrastructure, which prove the efficiency of our approach.

I. INTRODUCTION

Genetic algorithms (GA) are metaheuristic optimization techniques, based on mechanisms analogous to those in biological evolution. Using the concept that environmental pressure enforces natural selection, these algorithms perform a heuristic search by iteratively evaluating a group of points in the search space, called a population, and replacing some of the individuals with newly generated ones. Due to their black-box nature and their ability to find adequate solutions in acceptable time, they have been successfully applied in a variety of disciplines [5], [4]. However, with the increase of the problem size and the complexity of the function to be optimized, the necessary time to find good solutions increases greatly [1]. This has motivated a lot of efforts for improving the runtime of genetic algorithms and one of the most promising approach thus far is to use parallel implementations of Gas [6], [7], [8].

Currently, there exist a few open source implementations of GAs written in C or C++. However, most of them are not portable between platforms, have not been updated for the past few years and are developed for compilers which at the time of development did not support important new features from the C and C++ standards. In some C++ implementations this is obvious when we take into consideration the sparing use of templates and standard containers. This combination of programming practices and old target platforms have led to code, which is hard to modify and extend for parallel execution. On the other hand, there have been several efforts for creating a new GA library, written in idiomatic C++ with heavy usage of templates and the generic programming approach. However, most of these projects have not yet been able to build a strong community of developers and are either still in design phase or are in stagnation with unresolved bugs, which renders them practically unusable. This motivated our work in developing our own general purpose genetic algorithm library.

These issues with the available GA implementations impede the development of parallel implementations.

There are three main types of parallel genetic algorithms (PGA) [5]:

- Global or master-slave GAs
- Fine-grained parallel GAs
- Distributed GAs

The master-slave GA uses a single population, like the normal GA, but evaluation of the fitness function is distributed among several slave processors, while the master processor executes all other steps in the GA. This approach is simple to implement as it does not require a lot of changes to the original algorithm and it is widely used when the evaluation of the fitness function dominates the overall runtime. This type of PGAs are suited for both shared-memory and distributed-memory platforms.

The fine-grained parallel GA is suited for massively parallel hardware like GPUs. It uses a single population, which is spatially distributed among the different processors. However, selection and reproduction are limited to small overlapping neighborhoods.

The distributed genetic algorithms (DGA) also known as multi-deme GAs, are one of the most widely used GAs, despite their inherent complexity [10], [3]. In DGAs, the individuals are distributed among several populations and periodically populations exchange individuals. This exchange is called migration. Despite their popularity, they are one of the hardest to comprehend as the effects of the migration are not fully understood. The multi-deme GAs are also called distributed GAs because they are usually implemented on machines with distributed memory - supercomputers or clusters [11]. Since the computation to communication ratio is high, they are usually classified as coarse-grained, which makes them

suitable for execution on geographically distributed computing resources, for example grids, as well as volunteer computing and desktop computing initiatives.

Numerous variations of these algorithms are possible [9]. The DGA with a combination of the global GA is particularly suitable for clusters of multi-core servers, where each node executes a master-slave GA with migration between nodes [12]. On the other hand the combination of DGA with fine-grained GA can be effective on clusters of GPU-equipped nodes.

In this paper we present an implementation of a distributed genetic algorithm which executes by launching multiple grid jobs on different grid sites. Each job can be either a serial job, which executes a GA with migration and runs on a single processor or an MPI job, which executes a distributed genetic algorithm with migration both between the processes of the job and with other jobs.

Naturally, when considering parallel implementations of GAs and parallel execution, there are several possible target platforms. One of our main concerns when developing our framework for executing DGAs was to make it platform independent in the sense that it could be ran efficiently on grid, cloud and high-performance computing resources. Naturally this required a communication model and an implementation which are easy to tune for each type of resources and for an arbitrary combination of them. The developed model uses ZeroMQ for communication between each computational task and in the case of parallel tasks it utilizes MPI. Balancing between these two approaches enables effective execution on these three types of computational resources.

In this paper we describe our experience in executing the framework in grid [2] environment, particularly using the European Grid Infrastructure.

In the next section we describe our general purpose implementation of GAs, using C++ generic programming via templates. In Section 3 we lay out the technical features and limitations of the grid environment, which govern our communication patterns. In Section 4 we describe our parallelization approach and communication models as well as the functionality provided to end users. Experiments and benchmarks are shown in Section 5.

II. General purpose genetic algorithm library

For the purposes of this framework, we developed a general purpose GA library, called NAGA. Addressing the problems that we discovered in most currently open source GA implementations, we designed our library with extensibility and ease of use in mind. It is written in C++, using the latest edition of the C++ standard - C++11, taking into consideration which parts of the standard are supported by the major compilers to make the code as portable between platforms as possible. NAGA has been successfully tested with GNU Compiler Collection 4.4.x, Intel C++ Compiler 12.1.x and Microsoft Visual C++ 2010 and 2012. The core library provides several classes implementing GAs, selectors, crossover and mutation operators, etc. Each GA provides an interface which can be easily extended for developing a DGA.

III. Technical and administrative limitations of the grid

A core limitation of the computing environment provided by EGI is the inability for direct communication between grid jobs. In practice this limits the maximal number of processors a computational task may use to the maximal number of processors provided by a single grid site. Submitting a large grid job however, implies greater time the job stays in the execution queue until resources are freed. Also, jobs requiring more resources are usually given lower priority by the Local Resource Management System operating the grid site. Using multiple grid jobs, which communicate is impossible most of the times due to the usage of firewalls which filter inbound connections or due to compute nodes running behind a NAT. Additionally, most grid sites consider opening ports for inbound connections a violation of their acceptable use policy.

Another technical specificity of the grid is its high heterogeneity. Each provider may use different versions of several operating systems, different hardware and different versions of libraries. This usually results in different runtime of the same applications on different grid sites.

IV. Parallel implementation

A. Inter-job communication

There are several technical and administrative limitations of the European Grid Infrastructure. As noted earlier, the usage of firewalls, which filter inbound connections, or of private IPs on the compute nodes, requires any connection with the computational task to be initiated by the task itself. Naturally, this makes direct communication between grid jobs impossible. In our work we adopted an approach which resolves this problem, by introducing a new component, called a central server, which acts as a message broker for the inter-job communication. This central server is accessible through public Internet, i.e. computational tasks can directly connect to it.

It should be noted that using this message broker adds an extra layer of indirection, which implies an increase of the latency for each communication exchange. This necessitates using a high-performance communication library in order to mitigate the negative effects of the retransmission of messages. Another necessary feature that should be supported by the library is the capability to transmit messages asynchronously, i.e. the call should be non-blocking in normal circumstances. Additionally, the library should be portable between different platforms. When considering grid usage, the library should at least be able to run on the Linux distributions used in EGI - Scientific Linux, CentOS and other derivatives of Red Hat Enterprise Linux, as well as on Debian. The library should work efficiently even with a large amount of grid jobs, i.e., a large amount of messages.

We based our implementation on ØMQ (also ZeroMQ, 0MQ or ZMQ). ZeroMQ is a high-performance library for asynchronous messaging. It provides facilities for working with message queues, but does not require an

explicit message broker. It is written in C++ and provides an API for C, Python and over 40 other programming languages. It is compatible with POSIX Sockets and with Windows Sockets API (Winsock). ZeroMQ provides an API centered around sockets, which generalizes IP and inter-process communication sockets used in POSIX.

ZeroMQ provides three main types of network topologies and the appropriate types of sockets:

- Service bus - implements the standard request-reply pattern through the usage of REQ and REP sockets
- Data distribution tree - implements the publish-subscribe pattern, with subscribers being able to subscribe to specific messages through filters. It uses the PUB and SUB sockets.
- Parallelized pipeline - implements the push-pull pattern with support for multiple providers and workers. It uses the PUSH and PULL sockets.

Through the usage of these three pairs of sockets we define three distinct communication channels between the central server and each grid job - REQ-REP, PUB-SUB and PUSH-PULL. Each grid job has three sockets REQ, SUB and PUSH, while the central server has REP, PUB and PULL sockets.

The REQ-REP channel is used for initialization of the grid job. In several communication exchanges the grid job requests and receives a unique identifier, possibly seed values that other computational tasks may have saved and a permission to begin running the GA.

The PUSH-PULL channel provides a method for passing messages from the grid job to the central server. As, in contrast to REQ-REP, PUSH-PULL does not require a confirmation that the message is received, the grid job can continue working on the GA immediately after the ZeroMQ API call. Each of the messages is prefixed with the grid job ID.

In its event loop the central server polls the REP and PULL sockets and processes any messages that may have arrived. In the case of an incoming request on the REP socket, if the message states the grid job doesn't yet have an ID, the server generates a new ID and creates a new entry which describes the new task. The server sends a reply consisting of the new ID. Subsequent requests from the grid job may require individuals to be inserted in the otherwise completely random population and a request to begin work. Messages incoming on the PULL socket consist of individuals and are always prefixed with the grid job ID. Depending on the message, these individuals are either stored as the best ones found by the grid job thus far or are inserted in the migration queue. The migration queue is a list of queues, with different queues for each of the grid jobs. When the individuals are sent to the central server as part of a migration, they are inserted in the queue of the grid job and a response form a randomly chosen queue is returned. The probability of choosing a queue is proportional to its size.

The third communication channel uses the PUB-SUB sockets. After an identifier is granted to each grid job, it subscribes itself to all messages beginning with its ID and beginning with the ID dedicated to the central server - 0. The PUB-SUB channel serves two major purposes. After the migration group is chosen from the migration queue as described above, it is prefixed with the ID of the grid job and sent through the PUB-SUB channel. As the PUB-SUB channel does not require a confirmation that the message is received, the central server continues work immediately after the ZeroMQ call. The second major use of PUB-SUB channel is to send control messages to the grid jobs. Each control message is prefixed with the central server ID and is treated as a signal and all grid job receive it as they are subscribed to messages with said prefix. This provides a mechanism to control the formed distributed system and is the main way for enforcing balancing, proper termination and monitoring of the grid jobs.

This topology allows for non-blocking messaging between the central server and the grid jobs and most notably all communication channels are initiated by the grid jobs.

B. Using parallel jobs

The specified system is independent of the manner of execution of the grid jobs, i.e. it can be applied to jobs consisting of independent processes and to parallel jobs using MPI. However, when considering MPI jobs, using a single communication endpoint is more efficient for several reasons:

Using a single job ID for all the processes in the grid job will ensure that a message between two of the processes cannot be sent through the central server. Such messages could only be sent through a communication channel specific to that job - MPI.

- Although the quantity of the messages will remain unaltered, the necessary data structures for each job, thus the total memory used by the central server will be an order of magnitude less.
- This approach enables transferring all communication related tasks to a single process of the grid job.
- MPI implementations are optimized for communication between tightly coupled servers. That is why we give MPI precedence over ZeroMQ when considering communication within a single grid job.
- Using a single process to control MPI and ZeroMQ communications within a single tasks enables the other processes to work asynchronously and resolves the problem of implicit synchronization. When considering fitness functions whose runtime depends on the exact input, this is a quality that the distributed system should possess.

Given these reasons, we used a modification of the master-slave communication pattern typically used with MPI. While most implementations of the master-slave model require the master to take an equal part in the computations and do additional tasks like preprocessing,

postprocessing and overall management of the slave processes, we have chosen to completely separate the master of all computations related to GA itself. In our implementation the master slave is in charge of all communication with the central server - global communication, and also manages all communication within the grid job - local communication. While this means that one of the processes of the grid job will not do any actual computations, this enables all slave processes to work asynchronously and to perform minimal work in regard to communication [13].

A single iteration of the main loop of a slave process can be roughly described as a sequence of a GA iterations, local communication and global communication. During both local and global communication, the slave process sends a message to the master process, through non-blocking MPI calls, and probes if it has incoming MPI messages. It should be noted that local and global communications do not occur on each iteration of the GA.

The master uses two queues to manage the local and global communication. The local queue stores individuals sent from one slave process to another as part of a migration, while the global queue stores individuals sent by other grid jobs. In a single iteration of its event loop the master probes for incoming MPI messages and polls the ZeroMQ SUB socket. Incoming migration from the SUB socket is stored in the global migration queue, while signals are immediately sent to all slave processes using MPI. There are several types of MPI communication exchanges between the master and the slaves, each marked with a special tag:

- Global communication exchange - the master receives a message containing a number of individuals to be sent to a different job. If the global migration queue is not empty it returns the first subpopulation to the slave. After that, the master prepends the grid job ID to the message and sends it upstream to the central server using the PUSH-PULL channel.
- Local communication exchange - in this case the master receives a message containing a number of individuals to be sent to a different process in the same job. If the local migration queue is not empty, it replies to the slave process with a group of individuals. After that it inserts the message in the local migration queue.
- Upstream results - in this case the slave process sends to the master the best results found so far. The master sends them upstream to the central server.
- Slave-to-master signal - in this case a slave process sends a signal to the master process. Signals may be sent to signify that the GA job should terminate due to reaching some computational limit, reaching a plateau of the fitness function, or finding a solution satisfying some criteria.
- Master-to-slave signal - in this case the master has received a signal on the SUB socket and sends it to each of the slaves.
- Slave-to-master signal acknowledge - used to signify to the master that the slave process has acted upon the signal appropriately.

C. Heterogeneity

The described system is heterogeneous in two aspects. Firstly, each grid job may use a different number of processes and carry out global communication at different rate. Additionally the different grid jobs may execute on different grid sites with different hardware configuration - processors, memory, interconnect. While the first aspect can be easily controlled by the user, the second is not subject to control. The heterogeneity of the computational grid may lead to unbalanced computations and give precedence to some populations over others. Periodically using the PUB-SUB channel, the central server sends control signals to each grid job containing the ratio between the quantity of the migration groups originating from each of its slave processes and the average for all processes in the system. Using these numbers, each grid job can improve the global migration rate by increasing the number of iterations per global communication. This improves the balance between the populations of different grid jobs, without lowering the rate of local communications within each grid job.

D. End user functionality

Naturally, such a dynamic and complex system could easily become hard to manage even for experienced users and requires easy to use tools for management. This is why we developed a web-based interface to monitor and manage grid jobs. The interface is a stand-alone process which runs the Python web framework Tornado. User requests are sent through the REQ-REP channel to the central server. Using the web interface the user can extract the best results so far as well as various statistics about the jobs. The web interface also provides means for terminating any or all of the jobs.

Results of the execution of the GA can be stored on the central server and flushed to a file upon terminating the grid jobs and the server. Additionally, the user may use the result stores provided by the framework. They are both stand-alone Python processes with a specific ID that utilize the PUB-SUB channel to store the best results in a SQLite database or in a file in a log-like manner.

It should be noted that the result stores and the web-based interface are not necessarily running on the same machine as the central server and they conform to the

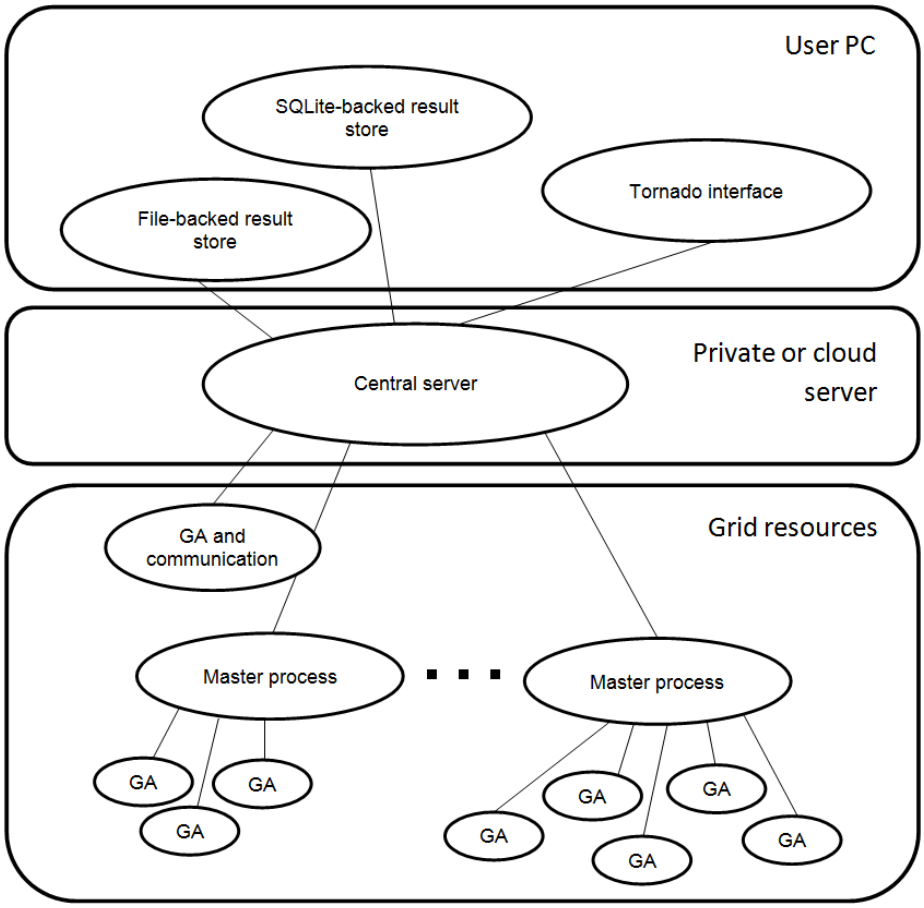


Figure 1. A schematic representation of the fully deployed system

same limitations as the grid jobs. This ultimately enables the user to run the central server on a machine accessible through public Internet, whether private or cloud resources, while grid jobs are ran on various grid sites and the result stores and the web interface run on his or her personal computer as presented in Figure 1.

We have also developed a code generation tool, which generates C++ code based on an XML file describing the GA. This enables users with limited knowledge of C++, who need only the core features of the library to actively use it.

V. Experiments and Benchmarks

The presented benchmarks were all performed using grid sites from EGI. Efficiency is presented both in terms of parallel efficiency and in processed genomes until a solution is found. Due to the randomness involved in the GAs results are averaged among several experiments. The chosen problem for the benchmarks is 3-SAT, which is a special case of the boolean satisfiability problem and is an NP-complete problem.

In our experiments the genome is a binary vector representing the boolean variables. The fitness function is the number of satisfied clauses of the problem. We used the roulette wheel selection (RWS). Migration is based on RWS when determining which individuals to be sent to other GA instances and RWS with inversed weights for determining which individuals should be replaced. The population size is 500. The crossover operator implements two-point crossover, while the mutation operator flips a bit with probability 0.9, or tries all possible flips of a bit and chooses the best one with probability 0.1. Table I shows the number of needed iterations for different configurations.

Table II represents the parallel efficiency in terms of throughput. The last two columns show the parallel efficiency.

TABLE I. Needed iterations for different configurations. Grid jobs are defined with the notation number_of_communication_processes + number_of_copmute_processes

Configuration	*Processes*	*Compute processes*	*Iterations*
1	1	1	52387
1+4	5	4	78253
1+4,1+2	8	6	83452
1+4, 2*(1+2)	11	8	92488
16*1	16	16	170046
2*(1+15)	32	30	231017
2*(1+15), 1+7	40	37	262459
8*(1+7)	64	56	471559
4*(1+15)	64	60	516154
8*(1+15)	128	120	1278933
6*(1+15), 8*(1+7), 16*(1+3)	224	194	1859213

TABLE II. PARALLEL EFFICIENCY WITH DIFFERENT CONFIGURATIONS. GRID JOBS ARE DEFINED WITH THE NOTATION NUMBER_OF_COMMUNICATION_PROCESSES + NUMBER_OF_COPMUTE_PROCESSES

Configuration	*Processes*	*Compute processes*	*Populations per second*	*Parallel efficiency*	*Parallel efficiency of compute nodes only*
1	1	1	237	100.00%	100.00%
1+4	5	4	936	78.99%	98.73%
1+4,1+2	8	6	1401	73.89%	98.52%
1+4, 2*(1+2)	11	8	1851	71.00%	97.63%
16*1	16	16	3011	79.40%	79.40%
2*(1+15)	32	30	7032	92.72%	98.90%
2*(1+15), 1+7	40	37	8654	91.29%	98.69%
8*(1+7)	64	56	13115	86.46%	98.82%
4*(1+15)	64	60	13671	90.13%	96.14%
8*(1+15)	128	120	27408	90.35%	96.37%
6*(1+15), 8*(1+7), 8*(1+3)	224	194	45089	84.84%	97.96%

VI. CONCLUSION AND FUTURE WORK

The presented framework demonstrated excellent parallel efficiency in terms of throughput of individuals. The communication model used in the framework overcomes the technical and administrative limitations of the grid environment and allow for usage of both serial and parallel MPI jobs. The model is designed to be portable and does not enforce any particular limitations on the execution environment, making it suitable for execution on cloud and HPC resources as well. The usage of both ZeroMQ and MPI provide support for complex topologies. This allows the user not only to comply with any resource limitations, but to experiment and analyze the effects of different migration patterns.

The developed general purpose genetic algorithm library NAGA uses the newest edition of the C++ standard and provides an easy to use interface, which enables easy extension of the provided core functionalities and it can be used for implementing other evolutionary algorithms as well. The heavy use of templates allows for significant code reuse, while preserving the speed of C++.

Many evolutionary algorithms which require no common knowledge among all the individuals can be easily implemented through our framework e.g. differential evolution and memetic algorithms. However, parallel versions of algorithms like particle swarm optimization, which are based on common knowledge, require carefully chosen communication strategies in order to preserve the speed of execution.

ACKNOWLEDGEMENT

This work is supported by the National Science Fund of Bulgaria under Grants DCVP 02/1 (SuperCA++) and the European Commission under EU FP7 project EGI-InSPIRE (contract number RI-261323).

REFERENCES

[1] J.H. Holland, "Adaptation in Natural and Artificial Systems." MIT Press Cambridge, MA, USA, 1992

[2] Ian Foster, Carl Kesselman, Steve Tuecke, "The Anatomy of the Grid", International J. Supercomputer Applications, 15 (3), 2001.

[3] Woodrow Wilson Bledsoe. "An analysis of genetic populations." Technical Report, Panoramic Research Inc., Palo Alto, California, USA, 1962.

[4] Braun H. C. "On Solving Travelling salesman Problems by Genetic Algorithms." In Schwefel H.-P., Manner R., Eds., Parallel Problem Solving from Nature, p. 129–133, Springer-Verlag (Berlin), 1990.

[5] Cantu-Paz, E. (1998). "A survey of parallel genetic algorithms." Calculateurs paralleles, reseaux et systems repartis, 10(2), 141-171.

[6] Eroftiev, A. A.; Timofeeva, N. E.; Savin, A. N., "Parallel computing in application to global optimization problem solving," MIPRO, 2011 Proceedings of the 34th International Convention , vol., no., pp.185,190, 23-27 May 2011

[7] Golub, Marin, and Leo Budin. "An asynchronous model of global parallel genetic algorithms." In Second ICSC Symposium on Engineering of Intelligent Systems EIS2000, University of Paisley, Scotland, UK, pp. 353-359. 2000.

[8] Jakobovic, Domagoj, Marin Golub, and Marko Cupic. "Asynchronous and implicitly parallel evolutionary computation models." Soft Computing (2013): 1-12.

[9] Gordon V. S., Whitley D., "Serial and Parallel Genetic Algorithms as Function Optimizers." In Forrest S., Ed., Proceedings of 5th ICGA, p. 177–183, Morgan Kaufmann (San Mateo, CA), 1993.

[10] Tanese R., "Distributed genetic algorithms." In Schaffer J. D., Ed., Proceedings of the 3rd ICGA, p. 434–439, Morgan Kaufmann (San Mateo, CA), 1989.

[11] Lin S.-C., Punch W., Goodman E., "Coarse-Grain Parallel Genetic Algorithms: Categorization and New Approach." In 6th IEEE Symposium on Parallel and Distributed Processing, IEEE Computer Society Press (Los Alamitos, CA), October 1994.

[12] Lim, Dudy, et al. "Efficient hierarchical parallel genetic algorithms using grid computing." Future Generation Computer Systems 23.4 (2007): 658-670.

[13] Atanassov, E., Georgiev, D., & Manev, N. (2013). "Number Theory Algorithms on GPU Clusters." In High-Performance Computing Infrastructure for South East Europe's Research Communities (pp. 131-138). Springer International Publishing.

The Performance Impact Analysis of Loop Unrolling

Goran Velkoski
Innovation LLC,
Vostanichka 118, Skopje, Macedonia
Email: goran.velkoski@innovation.com.mk

Marjan Gusev,
Univ. Ss Cyril and Methodius, FINKI
Rugjer Boshkovic 16, Skopje, Macedonia
Email: marjan.gushev@finki.ukim.mk

Sasko Ristov
Univ. Ss Cyril and Methodius, FINKI
Rugjer Boshkovic 16, Skopje, Macedonia
Email: sashko.ristov@finki.ukim.mk

***Abstract*—Loop unrolling is a well known technique, which usually results with speedup of a program that contains loops. The effect is obtained by reducing the operations that require counter increases and branch jumps at the end of the loops.**

This paper analyzes the impact of loop unrolling on various processor types and memory patterns. The experiments show a high correlation between the cache and the problem size. The loop unrolling results with a higher speedup for the execution of a smaller size problem, while it does not have impact for a problem whose size is greater than the capacity of the last level cache size, due to the huge number of cache misses.

Another important result is that the loop unrolling achieves greater speedup on Intel, rather than AMD CPU. In this paper we analyze and discuss the various behaviors of loop unrolling.

***Index Terms*—AMD, Intel, Performance**

I. Introduction

Recent trends in programming tend to build heavy applications. Yet, many communities work tirelessly in the field of code optimization since we live in a time where the computing power is becoming a utility offered as a service. Optimizing algorithms can be achieved in many ways. Some of them try to optimize code based on processing architectures, while others find different and better algorithms that do the same job more efficiently [1], [2].

Compilers, on the other hand, which are build to be smart, compile based on certain architectures in order to optimize the translated assembly code and to provide the best performance [3]. Compilers are the Queen of Computing Science and Technology and they have long been the bridge from applications to systems [4]. Nowadays, they determine which architectural features should be implemented in the new hardware, as well as which new language features will be effective for software developers. There are a lot of optimization methods that the compilers implement.

Loop unrolling (also called loop unwinding) is one of the most common methods that the compilers implement to optimize the loops. It utilizes the CPU pipeline in the best possible way. Other methods for loop optimization include: peephole optimizations, local optimizations, global optimizations, machine code optimization, etc.

The goal of the loop unrolling is to increase the program performance by reducing branching penalties and pointer arithmetics and increasing the efficiency of pipelining [5]. The process of loop unwinding is often achieved by dispersing the innermost loops to several independent instructions. Therefore the level of usage of this method was limited because it had to be specified by the user for maximum performance. In order to eliminate this limitation, loop quantization was introduced as a general technique for automatically unwinding multiply nested loops [6].

Loop unrolling directly impacts on program speedup by achieving ILP (Instruction-level parallelism) and has its slight impact on power consumption and energy usage [7].

The loop unrolling optimization method can be used on the popular Graphics Processing Units (GPUs), as massively parallel processors with tremendous computational power and high memory bandwidth. Optimizing programs on this kind of platforms run up to 70 percent faster when compared to the unoptimized versions [8].

In this paper we target the loop unrolling optimization technique and try to find the impact of cache memory on the performance when it comes to computation intensive algorithms. 2 hypothesis are target in this research: *H1* The cache bunch size proportionally impacts the speedup, and *H2* The speedup does not depend or lightly depends on the hardware architecture.

Some of the most used and most optimized algorithms are the linear algebra algorithms, therefore we used the matrix-matrix multiplication (MMM) algorithm to find the required correlation of software optimization that are build based on hardware properties. This algorithm contains three nested loops, it is cache intensive and generates large number of cache misses, making it a good testbed to find a correlation and test the hypothesis.

The rest of the paper is organized as follows. The related work is presented in Section II. Section III describes the used methodology for the experiments. The results of the experiments and a comparison of different levels of loop unrolling are presented in Section IV. We discuss these results in Section V. The conclusion and future work are specified in Section VI.

II. Related work

This section presents the related work of the loop unrolling technique correlating with the hardware and the MMM algorithm performance and optimization.

A. MMM optimization methods

Anchev et al. presents a nice overview of several optimization techniques for the MMM algorithm [9]. Similar research that provides an overview of cache optimization techniques is presented by Kowarchik and Weis [10]. They all conclude that the optimizations according to the cache memory can yield a significant execution speedup, especially when they are applied to numerically intensive codes. Furthermore, future computer architecture trends motivate research efforts to focus on memory hierarchy optimizations.

The optimization of a linear algebra algorithms can be also oriented towards reducing the number of operations that are executed. Furthermore, the MMM algorithm may be accelerated asymptotically. Coppersmith and Winograd [11] made a novel use of the Salem-Spencer Theorem, which gave a fairly dense set of integers with no three-term arithmetic progression. The first revolutionary reduction for the MMM algorithm was proposed by well known Strassen's algorithm [12], which reduced the power to 2.8. Several other authors decreased it to 2.3727 [13].

Gusev et al. [14] provide an optimization of the MMM algorithm using hybrid blocking algorithm that preserves the number of operations and yet provides speedup compared with the original algorithm. The CPU affinity may also be used in order to provide better performance. Velkoski et al. [15], [16] conclude that using loosely coupled cores with private L3 cache is better than using tightly coupled cores with shared last level L3 cache when the problem size fits L3 cache of loosely coupled CPU cores, but in the same time cannot be placed in only one L3 cache of tightly coupled CPU cores. Gusev and Ristov derived an interesting conclusion about the MMM algorithm in a virtual environment, that is, it is better environment when the problem size fits in the private per core cache memory, but the performance suffers in the shared memory [17].

B. The loop unrolling method usage

The loop unrolling optimization method is widely used while compiling programs for high performance solutions. It is not applicable to FOR-loops only. Huang and Leng [18] generalize the loop unrolling method for any loop construct. Kisuki et. al [19] discuss how to simultaneously select the tile sizes and unroll factors in order to optimize the programs via iterative compilation.

Automatic source-to-source transformation is presented by Cronsioe et al. [20], who prove that the loop unrolling approach can achieve up to 67% performance gain on high performance and low energy consumption machines.

Stephenson and Amarasinghe [21] predict the unroll factors using supervised classification. Using this method, they managed to correctly predict the enroll factor for 65% of the loops in their dataset and provide 5% overall performance improvement for the SPEC 2000 benchmark suite.

We found contradictory results observed in different research papers about the energy consumption. Ayala et. al [22] present how loop unrolling techniques can decrease the energy consumption of the Multiprocessor System-on-Chip architectures up to 60%. Nevertheless, Seng and Tullsen [23] find that loop unrolling provides a very minor benefit in terms of reducing energy.

C. Loop unrolling in hardware solutions

The loop unrolling method is often used when optimizing the programs for hardware based solutions. Furthermore, the loop unrolling method hides the memory latency [24].

The loop unrolling method can be implemented when using variety of devices. Kurra et al. [25] study the effect of the loop unrolling factor on the controllers generated during High Level Synthesis (HLS).

Dong et al. [26] study the performance impact of the loop unrolling method when it is executed on reconfigurable hardware boards, i.e., FPGAs. Loop division on multiple smaller loops is proved to be appropriate when using performance aware devices such as FPGAs [27]. Since the loop unrolling method is not the only one, Dragomir et al. [28] propose a combined technique (loop unrolling and loop shifting) in order to relocate the function calls for kernel based reconfigurable FPGAs.

Optimizing loop performance is also required on clustered VLIW architectures. Yi et al. [29] describe how loop unrolling method alongside with unroll and jam can enhance the parallelism of the loop in order to limit the effects of inter-cluster data transfers.

The loop unrolling method is highly effective for optimization when high performance architectures are considered. Lokuciejewski and Marwedel [30] present loop unrolling method tailored towards real-time systems and achieve an average reduction of up to 13.7%.

We have not found any research paper that analyzes the performance of loop unrolling on various hardware architecture for different cache size for cache intensive algorithms.

III. Testing methodology

To provide reliable results we define a relevant testing methodology, where each test case is tested upon its correctness.

A. Testing algorithm

The MMM algorithm defined in (1) is used to create test data.

$$C_{N \cdot N} = A_{N \cdot N} \cdot B_{N \cdot N} \tag{1}$$

Each element c_i of result matrix $C_{N \cdot N}$ is calculated as an inner product of row i of matrix $A_{N \cdot N}$ and the column j of matrix $B_{N \cdot N}$, for each $i, j = 0, 1, \ldots, N-1$. For simplification, we use squared matrices $A_{N \cdot N}$ and $B_{N \cdot N}$. Each matrix element is stored as double precision number (8 bytes).

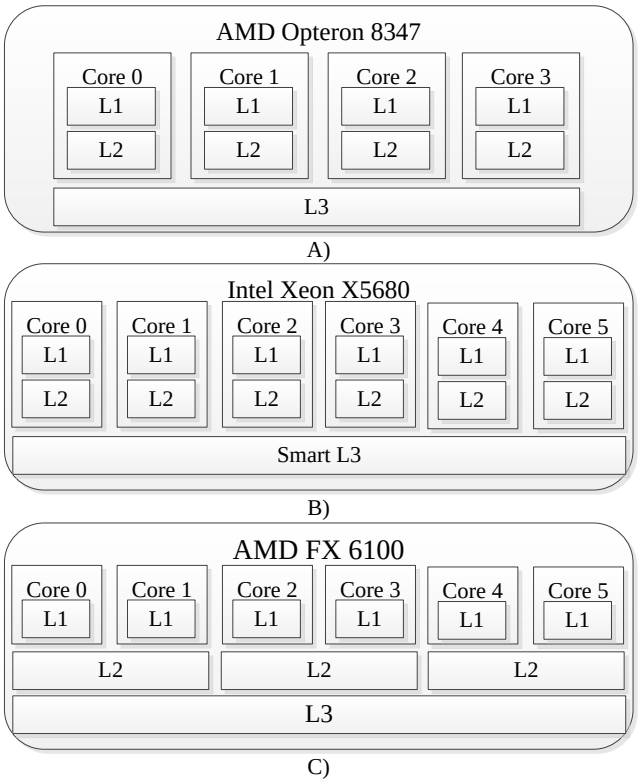


Fig. 1. CPU architectures

B. Unrolling the MMM algorithm

Five implementations of the MMM algorithm are used, such that each implementation is sequential execution with different bunch of loops unrolled. Furthermore, the code is compiled using the GCC 4.5 compiler with no optimization in order to evade its loop unrolling and other optimizations. The implementations are:

- *Implementation 1:* 1 only loop, (no loop unrolling);
- *Implementation 2:* 2 loops are unrolled;
- *Implementation 3:* 4 loops are unrolled;
- *Implementation 4:* 8 loops are unrolled; and
- *Implementation 5:* 16 loops are unrolled.

C. Testing environments

Three different architectures are used as testing environments: AMD(tm) Opteron(tm) 8347, Intel(tm) Xeon(tm) CPU X5680 and AMD(tm) FX 6100. Since the AMD Opteron 8347 is a representative of the older Barcelona architecture, we used the newer AMD processor with the Zambezi architecture. The following sections briefly describe the architecture of the three CPUs.

1) The AMD(tm) Opteron(tm) 8347 architecture: It has 4 cores each with 64KB private L1 (both data and instruction) cache and 512KB private L2 cache. Additionally, all four cores share 2MB L3 cache (Figure 1-A).

2) The Intel(tm) Xeon(tm) CPU X5680 architecture: This CPU has 6 cores, each with 32 KB 8-way set associative L1 data cache dedicated per core and 256 KB 8-way set associative L2 cache dedicated per core. All 6 cores share 12 MB 16-way set associative Smart L3 cache (Figure 1-B).

3) The AMD(tm) FX 6100 architecture: The AMD(tm) FX 6100 CPU has 6 cores, each with 64 KB 4-way set associative L1 data cache private per core. There is 3×64 KB of 2-way associative instruction L1 cache. Each instruction cache is shared between 2 cores. There are 3×2 MB of 16-way associative L2 shared exclusive caches. All cores share the L3 cache that is sized 8 MB and has up to 64-way associativity (Figure 1-C).

D. Experiments

Five different experiments are executed, such that each experiment corresponds to particular implementation of the algorithm defined in Section III-B. In each experiment we use the same single core of a CPU using $GOMP_CPU_AFFINITY = 0$. Many test cases are conducted in a single experiment, such that the MMM algorithm is executed varying the matrix size N, where $N = 16, 32, \ldots, 2000$.

E. Test data

The execution time $T(U)$ is measured for each test case in each experiment. Based on this parameter, we calculate the *Speed* $VU(U)$ using (2) in order to compare the performance impact of loop unrolling optimization method on multiple architectures with different core and cache architectures. The parameter U denotes the bunch size of the unrolling and it has several values $U \in \{1, 2, 4, 8, 16\}$.

$$VU(U) = \frac{2 \cdot N^3}{T(U) \cdot 10^9} \tag{2}$$

We introduce the *Relative Speedup* $R(U)$ as defined in (3) to compare the impact of unrolling bunch size on a particular environment and test the hypothesis H1.

$$R(U) = \frac{VU(U)}{VU(1)} \tag{3}$$

IV. Evaluation of results

This section presents the achieved speed of all test cases in realized experiments defined in Section III-B. Evaluation determines the correlation between the CPU cache size, the loop unrolling bunch size and the problem size.

A. The results on the AMD Opteron 8347 CPU

Figure 2 presents the speed for different levels of loop unrolling on AMD Opteron 8347 processor. We observe the existence of 2 different regions. In the first region ($0 < N < 512$) a greater speed is achieved and the differences in speed are observed for different loop unrolling levels. In the second region ($512 \geq N < 2000$), there is no visible difference based on the loop unrolling level and the speed converges.

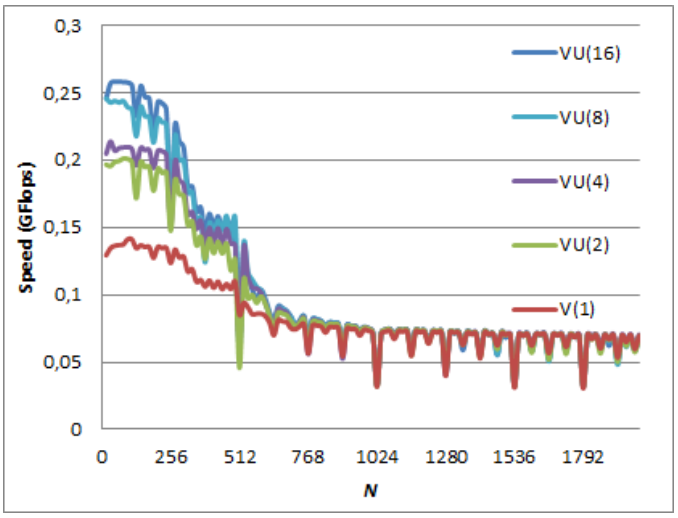


Fig. 2. Speed $VU(U)$ for the experiments executed on AMD Opteron 8347

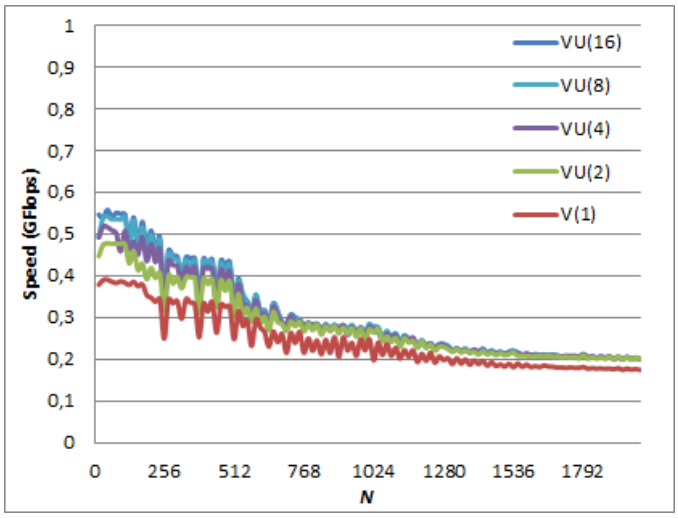


Fig. 3. Speed $VU(U)$ for the experiments executed on Intel Xeon X5680

For smaller matrices the speed is proportional to the loops that are unrolled. The greatest speed is measured for the MMM algorithm with loop bunch size 16, while the lowest speed is measured for the implementation without loop unrolling.

The regular drawbacks at $N = 512, N = 640, N = 768, \ldots, \ldots$ appear due to cache associativity problem. Ristov and Gusev [31] present detailed theoretical analysis and proof for its appearance.

B. The results on Intel Xeon X5680 CPU

The speeds for different levels of loop unrolling executed on Intel Xeon X5680 processor are depicted in Figure 3. Compared to the previous architecture, we can not distinguish the same regions since the experiments with no loop unrolling never converge with the rest of the implementations' results. The implementations with loop unrolling create the same two regions as the previous CPU, with the only difference, i.e., the performance is similar when the loops are unrolled with 8 or 16 bunches. Yet, we can observe that the speed is proportionally dependent for the smaller matrices ($N < 512$).

C. The results on AMD(tm) FX 6100 CPU

Figure 4 depicts the achieved speed for different levels of loop unrolling on the new Zambezi core architecture by AMD, i.e., the AMD(tm) FX 6100 CPU.

The loop unrolling in this case proves to be essential since in each section the number of unrolled loops proportionally increases the speed of the algorithm.

Nevertheless, for very small matrices ($N < 128$) we observe the greatest speed. Except the loop unrolling with bunch size 8 and 16, all other implementations are registered with speed degradation for greater matrices. Another degradation of performance is also observed for the other scenarios on different stages.

Another important conclusion is that the implementations of the loop unrolling with bunch size 8 and 16 tend to have the

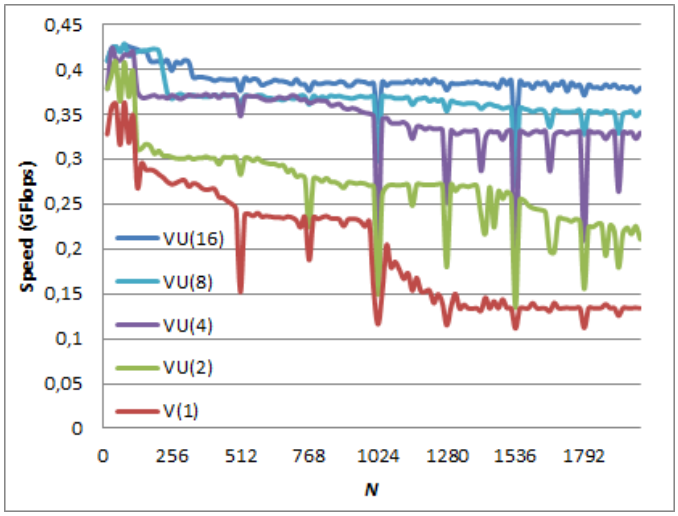


Fig. 4. Speed $VU(U)$ for the experiments executed on AMD FX 6100

stable speed for each test case. Moreover, the scenario with bunch size 8 provides better performance than the implementation with bunch size 16 for matrix size $N < 256$.

V. Discussion

In this section we will discuss the results gathered from the experiments and plot the conclusions from this paper. In Section IV we observed how the loop unrolling optimization method impacts the speed for different matrix sizes. We will expand our discussion by making a comparison for different levels of loop unrolling, as well as whether a relevant conclusion can be derived for the hardware impact on the loop unrolling.

A. Loop unrolling bunch size impact

This section analyzes the loop unrolling bunch size impact on a particular CPU through the relative speedup.

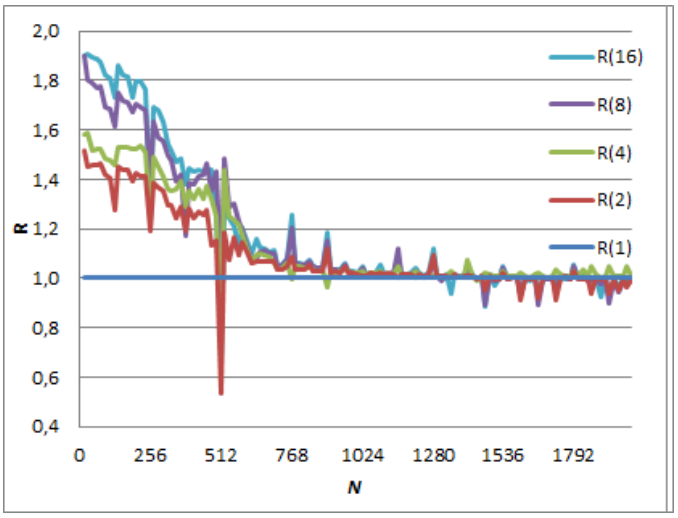


Fig. 5. The relative speedup $R(U)$ for the experiments executed on AMD Opteron 8347

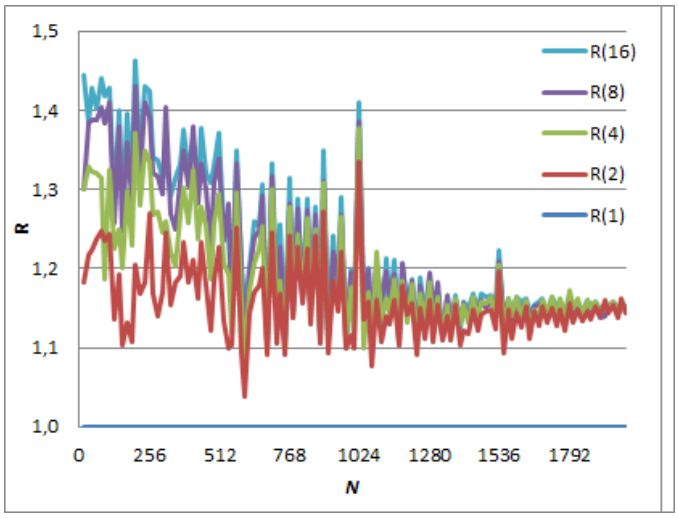


Fig. 6. The relative speedup $R(U)$ for the experiments executed on Intel Xeon X5680

1) Bunch size impact on AMD Opteron 8347: The speedup provided by the method of loop unrolling on AMD Opteron 8347 is presented in Figure 5. For larger matrices, that is, when they cannot be stored completely in the last level cache, the loop unrolling method on this architecture does not make any difference. The speedup proportionally increases with the number of unrolled loops. The loop unrolling is without effect when the number of cache misses increases, because the response mostly depends on the memory access, rather than the operations.

2) Bunch size impact on Intel Xeon X5680 CPU: Figure 6 presents that regardless of the matrix size and/or number of unrolled loops, the loop unrolling optimization method speed ups the response time on the Intel Xeon architecture. Therefore the loop unrolling optimization must be highly regarded in these architectures. The speedup in this case is proportional to the number of unrolled loops for smaller matrices and it diverges to the value speedup when using bunch size 2 for greater matrices that cannot be placed in the last level L3 cache.

3) Bunch size impact on AMD(tm) FX 6100 CPU: In Figure 7 again we observe the greater relative speedup for all matrix sizes and levels of loop unrolling. Additionally, we observe totally opposite results on the Zambezi AMD architecture. The greater relative speedup is recorded for larger matrices than for the smaller matrices. This confirms the previous observation from Section IV where we state that the loop unrolling provides sustainable performance of the MMM algorithm.

The larger number of unrolled loops results with increased speedup. This is most emphasized for larger matrices. Therefore, when compared with previous CPU architectures, a small paradox occurs, since the loop unrolling optimization method provides better performance when the number of memory accesses and cache misses increase.

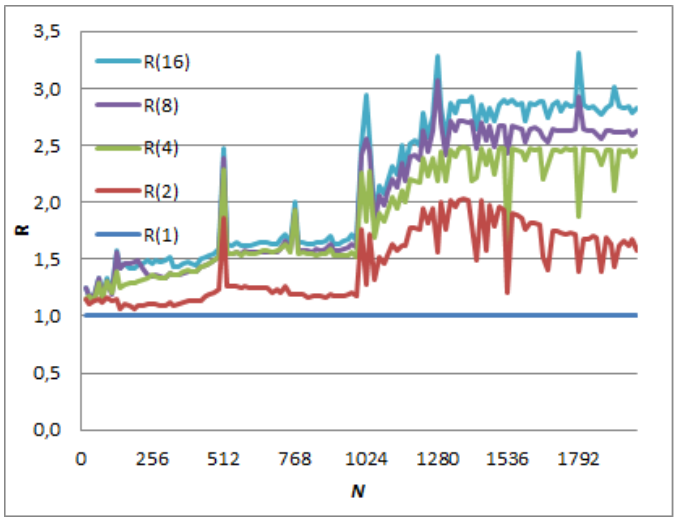


Fig. 7. The relative speedup $R(U)$ for the experiments executed on AMD FX 6100

VI. Conclusion and Future Work

In this paper we compared the loop unrolling technique for algorithm optimization using the dense MMM algorithm. We measured the performance for three processors manufactured by two different manufacturers AMD and Intel. The experiments were conducted varying the bunch size of loop unrolling and the matrix sizes.

Two different conclusions can be derived. The first conclusion is that the loop unrolling speeds up the cache intensive algorithms for the greater bunch size. However, the speed up depends on the CPU architecture. This is more emphasized for smaller matrices that can be placed in private per core cache memory on AMD Opteron and Intel Xeon, while it is significantly emphasized for greater matrices on the AMD Zambezi architecture.

Although the AMD Opteron and Intel Xeon are different vendor CPUs, the MMM algorithm with and without loop unrolling is very similar, compared to the AMD Zambezi architecture. The difference is in the cache organization and its associativity. This proves our hypothesis that there is a correlation between the CPU cache architecture and organization.

We will continue our research towards mathematically modeling the dependency of loop unrolling optimization technique and its correlation with the cache memory architectures and its capacity. Afterwards, we will try to generalize this theory with other most used algorithms that have loops, such as dot product with one loop, or matrix-vector multiplication with two loops. All these algorithms will be executed on more different CPU architectures, both with sequential and parallel implementation.

REFERENCES

[1] J. A. Stratton, C. Rodrigues, I. Sung, L.-W. Chang, N. Anssari, W.-M. W. HWU, N. Obeid *et al.*, "Algorithm and data optimization techniques for scaling to massively threaded systems," *Computer*, vol. 45, no. 8, pp. 26–32, 2012.

[2] G. D. Liu, W. H. Wen-mei, and N. Obeid, "Algorithm and data optimization techniques for scaling to massively threaded systems."

[3] A. J. Offutt and W. M. Craft, "Using compiler optimization techniques to detect equivalent mutants," *Software Testing, Verification and Reliability*, vol. 4, no. 3, pp. 131–154, 1994.

[4] R. Allen and K. Kennedy, *Optimizing compilers for modern architectures*. Morgan Kaufmann San Francisco, 2002.

[5] A. Aiken and A. Nicolau, "Perfect pipelining: A new loop parallelization technique," in *ESOP'88*. Springer, 1988, pp. 221–235.

[6] A. Nicolau, "Loop quantization: A generalized loop unwinding technique," *Journal of Parallel and Distributed Computing*, vol. 5, no. 5, pp. 568–586, 1988.

[7] M. Booshehri, A. Malekpour, and P. Luksch, "An improving method for loop unrolling," *arXiv preprint arXiv:1308.0698*, 2013.

[8] G. S. Murthy, M. Ravishankar, M. M. Baskaran, and P. Sadayappan, "Optimal loop unrolling for gpgpu programs," in *Parallel & Distributed Processing (IPDPS), 2010 IEEE International Symposium on*. IEEE, 2010, pp. 1–11.

[9] N. Anchev, M. Gusev, S. Ristov, and B. Atanasovski, "Some optimization techniques of the matrix multiplication algorithm," in *Information Technology Interfaces, Proceedings of the ITI 2013 35th International Conference on, IEEE Conference Proceedings*, Cavtat, Croatia, Jun. 2013, pp. 71–76.

[10] M. Kowarschik and C. Weiß, "An overview of cache optimization techniques and cache-aware numerical algorithms," in *Algorithms for Memory Hierarchies*. Springer, 2003, pp. 213–232.

[11] D. Coppersmith and S. Winograd, "Matrix multiplication via arithmetic progressions," *Journal of symbolic computation*, vol. 9, no. 3, pp. 251–280, 1990.

[12] V. Strassen, "Gaussian elimination is not optimal," *Numerische Mathematik*, vol. 13, no. 4, pp. 354–356, 1969. [Online]. Available: http://dx.doi.org/10.1007/BF02165411

[13] V. V. Williams, "Multiplying matrices faster than coppersmith-winograd," in *Proceedings of the 44th Symposium on Theory of Computing*, ser. STOC '12. New York, NY, USA: ACM, 2012, pp. 887–898. [Online]. Available: http://doi.acm.org/10.1145/2213977.2214056

[14] M. Gusev, S. Ristov, and G. Velkoski, "Hybrid 2d/1d blocking as optimal matrix-matrix multiplication," in *ICT Innovations 2012*, ser. Advances in Intelligent Systems and Computing, S. Markovski and M. Gusev, Eds. Springer Berlin Heidelberg, 2013, vol. 207, pp. 13–22. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-37169-1_2

[15] G. Velkoski, S. Ristov, and M. Gusev, "Affnity-aware hpc applications in multichip and multicore multiprocessor," in *Information Technology Interfaces, Proceedings of the ITI 2013 35th International Conference on, IEEE Conference Proceedings*, Cavtat, Croatia, Jun. 2013, pp. 95–100.

[16] ——, "Loosely or tightly coupled affinity for matrix - vector multiplication," in *Information Communication Technology Electronics Microelectronics (MIPRO), 2013 36th International Convention on*. Opatija, Croatia: IEEE, May 2013, pp. 228–233.

[17] M. Gusev and S. Ristov, "Matrix multiplication performance analysis in virtualized shared memory multiprocessor," in *MIPRO, 2012 Proceedings of the 35th International Convention*. IEEE, 2012, pp. 251–256.

[18] J. Huang and T. Leng, "Generalized loop-unrolling: a method for program speedup," *Proceedings 1999 IEEE Symposium on Application-Specific Systems and Software Engineering and Technology. ASSET'99 (Cat. No.PR00122)*, pp. 244–248, 1999. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=756775

[19] T. Kisuki, P. Knijnenburg, and M. O'Boyle, "Combined selection of tile sizes and unroll factors using iterative compilation," *Proceedings 2000 International Conference on Parallel Architectures and Compilation Techniques (Cat. No.PR00622)*, pp. 237–246, 2000. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=888348

[20] J. Cronsioe, B. Videau, and V. Marangozova-Martin, "BOAST: Bringing Optimization through Automatic Source-to-Source Transformations," *2013 IEEE 7th International Symposium on Embedded Multicore Socs*, pp. 129–134, Sep. 2013. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6657917

[21] M. Stephenson and S. Amarasinghe, "Predicting Unroll Factors Using Supervised Classification," in *International Symposium on Code Generation and Optimization (CGO'05)*. IEEE Comput. Soc, 2000.

[22] L. Ayala, D. Atienza, and L. Marisa, "Optimal Loop-Unrolling Mechanisms and Architectural Extensions for an Energy-Efficient Design of Shared Register Files in MPSoCs," in *Innovative Architectures for Future Generation High-Performance Processors and Systems (IWIA'05)*. IEEE Comput. Soc, 2005.

[23] J. Seng and D. Tullsen, "The effect of compiler optimizations on Pentium 4 power consumption," in *Seventh Workshop on Interaction Between Compilers and Computer Architectures, 2003. INTERACT-7 2003. Proceedings.* IEEE Comput. Soc., 2003, pp. 51–56. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1192355

[24] Y. T. Sato Hiroyuki, "Characteristics of Loop Unrolling Effect: Software Pipelining and Memory Latency Hiding," in *Inovative Architecture for Future Generation High-Performance Processors and Systems (IWIA '01)*. IEEE Comput. Soc, 2001.

[25] S. Kurra, N. K. Singh, and P. R. Panda, "The Impact of Loop Unrolling on Controller Delay in High Level Synthesis," *2007 Design, Automation & Test in Europe Conference & Exhibition*, vol. 1, pp. 1–6, Apr. 2007. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4211828

[26] Y. Dong, J. Zhou, Y. Dou, L. Deng, and J. Zhao, "Impact of Loop Unrolling on Area, Throughput and Clock Frequency for Window Operations Based on a Data Schedule Method," *2008 Congress on Image and Signal Processing*, pp. 641–645, 2008. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4566234

[27] O. S. Dragomir and K. Bertels, "Loop distribution for K-loops on Reconfigurable Architectures," *2011 Design, Automation & Test in Europe*, pp. 1–6, Mar. 2011. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5763245

[28] O. S. Dragomir, T. Stefanov, and K. Bertels, "Optimal Loop Unrolling and Shifting for Reconfigurable Architectures," *ACM Transactions on Reconfigurable Technology and Systems*, vol. 2, no. 4, pp. 25:1 – 25:24, 2009.

[29] Q. Yi, S. Carr, and P. Sweany, "Optimizing Loop Performance for Clustered VLIW Architectures," in *International Conference on Parallel Architetures and Compilation Techniques (PACT'02)*. IEEE Comput. Soc, 2002.

[30] P. Lokuciejewski and P. Marwedel, "Combining Worst-Case Timing Models, Loop Unrolling, and Static Loop Analysis for WCET Minimization," *2009 21st Euromicro Conference on Real-Time Systems*, pp. 35–44, Jul. 2009. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5161500

[31] S. Ristov and M. Gusev, "Achieving maximum performance for matrix multiplication using set associative cache," in *The 8th International Conference on, Computing Technology and Information Management (ICCM2012), IEEE Conference Publications*, ser. ICNIT '12, vol. 2, 2012, pp. 542–547.

Developing a Ticket Management SaaS solution

Marjan Gusev
Ss. Cyril and Methodius University,
1000 Skopje, Macedonia
marjan.gushev@finki.ukim.mk

Sasko Ristov
Ss. Cyril and Methodius University,
1000 Skopje, Macedonia
sashko.ristov@finki.ukim.mk

Pano Gushev
Innovation,
1000 Skopje, Macedonia
pano.gushev@innovation.com.mk

***Abstract*—Development of a Software as a Service cloud solution is not a trivial task. Apart from traditional software life-cycle, the developers should implement elasticity and scalability, which are probably the biggest challenges from software organizational and architecture aspect. In this paper we present the way we addressed several challenges concerning development of such a solution by designing several versions that reflect various implementations. The final organization was selected to be the one that offers the best performance among various candidates.**

***Keywords-bug reporting, issue tracking, software as a service, cloud computing, elasticity*.**

I. Introduction

Companies around the world develop various software products every day. Their relationship with the customers depends on the quality of software they deliver. Modern software engineering principles and concepts introduce customers in testing processes early in the development process in order to design more reliable applications. This is especially important in terms of increasing the customer satisfaction and efficient development, such as the iterative spiral design or agile software development [1].

Almost every company uses at least one software application. A lot of small and medium companies cannot afford to buy expensive systems, but, still they need to improve some processes in the company or relationships with the customers. Customer relation software solutions usually use an approach where a software solution keeps track and manages with issues that happen between the company and customers. Such an example is the issue tracking system, or bug reporting software where issues (bugs) are reported by customers, a ticket is associated to the given issue (or bug) and the company provides a service that maintains the list of tickets, (issues or bugs) until they are approved to be finalized.

In many software projects, ticket issuing (bug tracking systems) plays a central role: they allow users to communicate with developers to let them know about any problem and to request new features. In addition, developers can keep track of any unresolved issues or bugs, and request more information from users [2].

A lot of software applications were developed and offered on the market, even as free solutions, but non IT companies still have the problem of installation, management, maintenance and upgrades of these applications. This kind of solutions are usually self-hosting solutions that require appropriate hardware and software architecture, which is an extra investment for non IT companies.

Cloud solutions can solve the problem for both IT and non IT companies, offering issue tracking (or bug reporting systems) as a service [3]. Bug reporting delivered as a service has attracted research recently. For example, TaaS (Testing-as-a-Service) [4], is proposed as another prominent service, that enables enterprise and consumers to save the cost of complicated maintenance and upgrade effort. It provides a very high range of permutations and combinations to test the product and examine its performance, including a bug reporting system.

Several approaches use conventional ticket issuing (or bug reporting) systems, hosted on a PaaS (Platform as a Service) to obtain a cloud solution. For example, it is conceivable to automate the process of creating the development and test environments on demand using Amazon EC2 and expanding a bug tracking system [5]. However, these approaches do not have the flexibility that a SaaS (Software as a Service) solution provides, and do not offer a high level of scalability and elasticity. In this paper we present a software development process to design and develop a SaaS solution for a ticket management system (TMS).

The rest of the paper is organized as follows. Section II presents the essential concepts about TMSs and cloud computing, including several development phases from traditional bug reporting systems to a modern one. The challenges of a modern TMS and its architecture are presented in Section III, while the cloud solution software engineering is presented in Section IV. We discuss the benefits of this cloud solution and compare it with several existing architectures and organizations in Section V. Finally, Section VI is devoted to conclusion and future work.

II. Background

Cloud computing is essentially the next phase of innovation and adoption of a platform for computing, networking, and storage technologies designed to provide rapid time to market and drastic cost reductions [6].

Computation, storage, networking infrastructure, as well as development and deployment platforms are made available on-demand within minutes. Sophisticated futuristic applications are made possible by the abstracted, auto-scaling computation platform provided by cloud computing [7].

Cloud computing offers flexibility and financial benefits. It also lowers the entry point to high performance computing, allowing organizations to leverage computing power that they neither have the capital budget, nor the operational expertise for. Halpert [8] explains that cloud computing hasn't drastically altered existing technologies, but rather it has succeeded as a result of the collaboration of several existing technologies.

TMSs and their other implementations like bug reporting systems or issue tracking systems, are those systems, where a ticket is associated to a given object, such as reported bug, and the software monitors the processes around the ticket's life-cycle.

Some of the existing products are open source, while the others are commercial. Although a lot of companies have developed their own systems, these solutions usually take a lot of time and resources. After developing them, companies must take care for its maintenance and continue developing the upgrades. Also there is an extensive list of open source and free solutions, or commercial solutions which can provide.

An example of analysis of existing bug reporting solutions is given by Gushev et al. [3]. Hosting such a solution can cause a lot of problems to companies, including additional resources for management and administration. Therefore, a cloud solution can save a lot of resources and time.

These systems can be used both for quality assurance and task management. They allow companies to collect, track and assign tasks related to user satisfaction, for example, software development, when a new product or a service launches during beta testing and marketing research. A web based solution can benefit from a worldwide access by any user.

The system provides a centralized overview of tickets and their state. The list of pending items provides valuable input for quality assurance.

III. Design of a Cloud TMS

A. System Concept and User Requirements

The overall scenario is presented in Figure 1. Users click on menu options or active buttons and links sending a request to the system. The system starts the generation of web pages and sends the response to the user.

TMS application defines five different types of user groups: Application administrator (AM), Manager (PM), Resolver (RE), Customer manager (CM) and Customer (CU). The basic design of the Ticket Management (Bug Reporting) application consists of 3 segments. The first segment is the main application management (AM), the second segment is the provider (PM, RE) and the third segment is the customer (CM, CU). AM can create managers to manage the subscriptions. PM can create users of all groups except AM. RE user changes the ticket (bug) status after resolving. CM user can create other CM or CU users. CU creates a new ticket (reports a new bug) and answers to changed statuses. PM decides who will respond to tickets or associate workers on different modules of a given project (application) and CM decides who will report bugs for a given module.

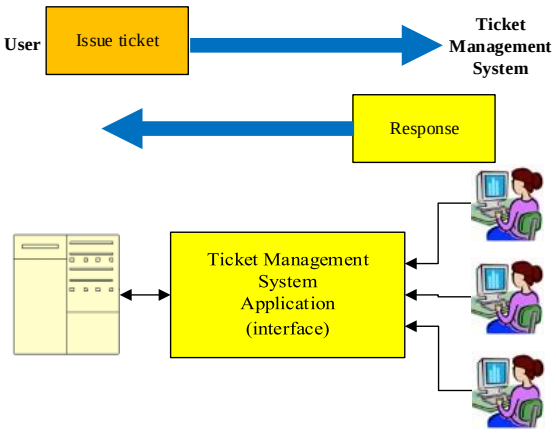


Fig. 1: Overall Scenario

TABLE I: User Access Rights

ID	Functions	AM	PM	RE	CM	CU
F1	Registration of company	Yes	No	No	No	No
F2	Customer administration	Yes	No	No	Yes	No
F3	Bug reporting	No	No	No	Yes	Yes
F4	Bug status administration	No	Yes	Yes	No	No
F5	Approving of corrections	No	No	No	Yes	Yes
F7	Create test cases	Yes	Yes	Yes	Yes	Yes
F8	Execute test cases	Yes	Yes	Yes	Yes	Yes
F9	Summary reports	No	Yes	Yes	Yes	Yes
F10	Detailed reports	No	Yes	No	Yes	No
F11	Site analytics	Yes	Yes	No	Yes	No
F12	Email notifications	No	Yes	Yes	Yes	Yes

Table I presents user characteristics by identifying basic and powered functions to each user group.

B. Ticket life-cycle

Functioning of the TMS includes tracking of various ticket states. An initiator creates a ticket (or issue), and then the system starts to process the ticket (or issue) until the initiator closes the issue (or ticket). Several states are defined in the ticket life-cycle, as depicted in Figure 2.

A *new* state is a result of the initiation activity that creates a ticket (an issue or a bug). Acknowledged or *pending* or open state is a kind of flag for the customer indicating that the supplier has opened the ticket or is aware of it. It will certainly mean that the supplier reads the ticket issue (reported bug) and will soon associate it to a resolver. Assigned or *in progress* state is an indicator that shows that the issue (bug) was assigned to a resolver to solve it. After finalizing the resolving activity and providing all corrective measures for a given bug or fixing the opened issue, the state changes to *resolved* or fixed state. *Rejected*, feedback or reopened state is a result of an activity by a customer for rejecting the provided answers for the ticket or solved bugs. *Closed* state is the final state in the lifecycle that appears as a result of verified confirmation that the issue is closed or bug was solved.

Three core functions correspond to the ticket management: ticket creation, resolving and approving. CU, CM and RE user

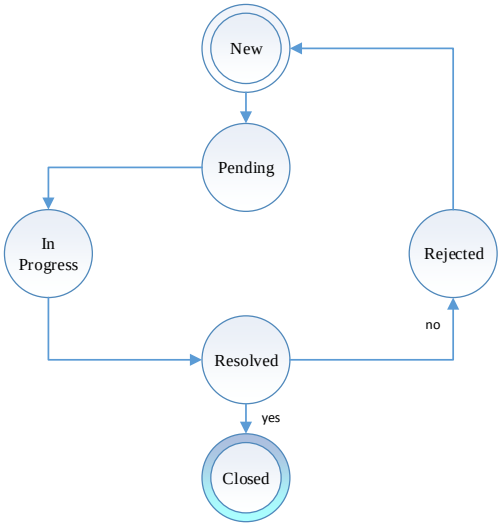


Fig. 2: Life-cycle of a ticket

groups can create a ticket (report a bug). Usually the system is intended for users from the CU group to create a ticket, but a more general system (as defined here) makes a possibility for users from the RE group to create a ticket, reporting an issue.

Issuing a ticket starts its lifecycle. The next step is identification of a moment when somebody has opened a ticket and is aware of its content, as associated with the pending state. Then the ticket state cannot be changed by customers (users from a CU or CM group), and only those responsible to resolve it (users from a RE or CM group) can change its state to resolved. CM users can allocate another users from the RE group to respond if necessary to realize the core resolving function and move across three progress states from pending, via in progress to resolved.

When a ticket issue is resolved, once again the ticket initiator can change its status by approving it. A careful inspection of the provided answer or the solution can lead to approval or rejection. If the ticket initiator sends an approval, then the ticket life cycle terminates with the closed state. If there is a rejection, the ticket status is once again associated to the pending state starting the life-cycle once again. It means that all the work on resolving the ticket issue (bug) is not satisfying the customer and it has to be analyzed once again seeing it in the to-do list of the responsible resolver (or users from the RE group).

C. Functional Description

The main TMS functions are explained as follows:

- ***Registration of a company*** - AM fills all necessary details about the company and sends them to the database. TMS saves those details in the database and returns a message to AM.
- ***Customer administration*** - A user sends request to TMS with user details. TMS receives those details and updates the database. Then it returns a response message.
- ***Administration of resolvers*** - A user sends a request to TMS with user details. TMS receives those details and updates the database. Then it returns a response message.
- ***Ticket issuing*** - CU sends a request to TMS with details. TMS receives those details and updates the database. Then it returns a response message.
- ***Resolving*** - RE responds to the ticket. TMS updates the database and returns the response message to CU.
- ***Approving*** - CU approves or rejects the response. TMS receives those details and updates the database.
- ***Creating test cases*** - CU creates test cases for review of functionalities. TMS stores those details and available to all interested parties.
- ***Summary reports*** - A user sends a request to TMS and receives a specific type of report with requested information.
- ***Detailed reports*** - A user sends a request to TMS with a specification of details and receives a detailed report.
- ***Site analytics*** - A user sends a request to TMS for site analytics and receives appropriate content as response.
- ***Email notification*** - A user sends a request to TMS for a change in a ticket state as an e-mail identification.

TABLE II: Client server scenario of TMS transactions

Role	Activity	Description
User	Create and submit a request	A user creates and submits a request to TMS
TMS	Receive and process a request	TMS receives and processes the request

D. Activity description and process rules

Table II describes a typical scenario in the client server model and the simplified activity diagram for core ticket management functions is presented in Figure 3. A customer can create a new ticket and the TMS saves the ticket details in the database. Each request is completed and does not inherit from the referenced requests. TMS returns a message to the users when the ticket associated to them changes its state as defined in Figure 2. RE can update the ticket status and CU can approve or reject this response. Additionally, the user can specify various reports, including summary, detailed reports and site analytics for all actions performed on the TMS. Finally, TMS responds by providing a relative message on the web interface or can send an e-mail or other personal message to the users. Information constraints relevant to this collaboration are captured in the referenced request data model(s).

E. Data model

Figure 4 shows the UML diagram of the models and their relations. The Database Provider level abstracts the logic for access to the database and enables an efficient programming interface for its use. In this level all the semantic for physical access to the database is encapsulated through collection of Database Provider objects. The internal structure of the database consists of several models. Each model of the internal

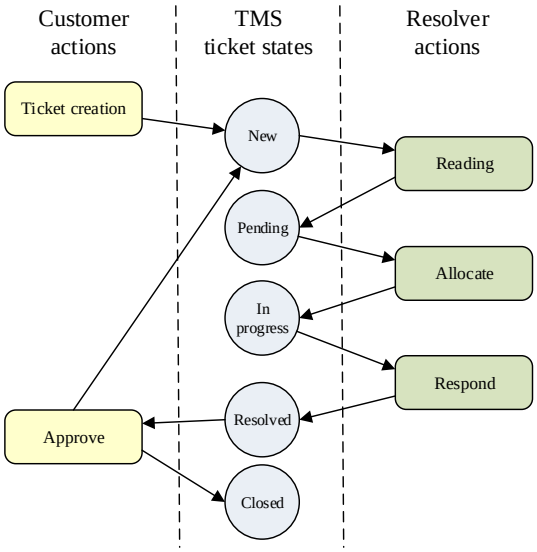


Fig. 3: Simplified activity diagram

TABLE III: TMS basic architecture modules

SMM	CMM	CFM
General settings	Company management	Create a ticket
Subscription admin	Company defaults	Change a ticket status
Finance management	Company subscriptions	Create test-cases
Users admin	Projects and modules	Execute test-cases
System reporting	Company users admin	Summary reports
System analytics	Customer management	Detailed reports
Security	Reporting	Site analytics
E-mail settings	E-mail settings	E-mail notification

structure of the portal describes one business area. Each model can consist of several sub-models.

IV. SaaS solution

This section presents basic software modules and organization principles to exploit elasticity and scalability.

A. Basic architecture modules

TMS consists of three core modules: System management module (SMM), Company management module (CMM) and Core functions module (CFM). The Additional functional module (AFM) is optional. Functionalities of all modules are described in Table III.

SMM is the core of TMS cloud solution. This module is general for all companies. It manages all resource provisioning in the cloud. This module is always active to provide user management, company subscription management, authentication, authorization etc.

CMM is specific for a company that uses the TMS as a service for its customers. This module is responsible for all general configurations within a company.

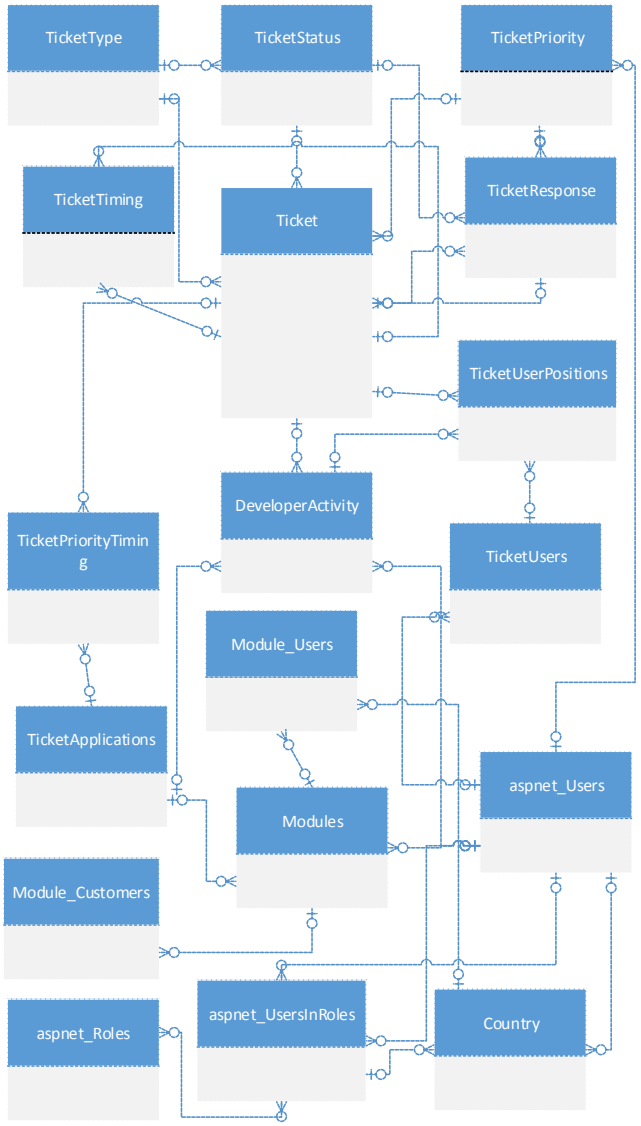


Fig. 4: Simplified version of data model

Apart from defining relations with the customers, it is always active by providing user management, customer management, project and module management, etc.

The basic CFM is also a company specific module. This module is responsible for all essential process to realize ticket management (bug reporting) within a company. It enables a connection between companies (providers) and their customers. This model provides functionalities such as: ticket creation, responding by a resolver, customer approving of the response, creation and execution of test-cases etc.

The Additional functional module (AFM) is dedicated for add-on functions, such as enabling a possibility to create and execute test cases, especially useful for bug reporting systems. There are also other types of add-ons that are possible and are beyond the limits of this paper.

TABLE IV: Classification and clustering into modules

Functionality	Static / Dynamic	Module
registration	Static	SMM
customer admin	Static	CMM
resolver admin	Static	CMM
ticket issuing	Dynamic	CFM
ticket reading	Dynamic	CFM
allocating resolvers	Dynamic	CFM
resolving	Dynamic	CFM
approving	Dynamic	CFM
test-case creation	Static	AFM
test-case executing	Dynamic	AFM
reporting	Static	CMM
notification	Dynamic	CFM

B. Design concepts

Probably the biggest challenge in developing a SaaS solutions is enabling elasticity and scalability. Most users do not think about this when developing a solution, although these characteristics define the main difference between the cloud solutions and the conventional web applications. A conventional approach to solve this issue is enabling an automated resource broker, which analyzes the processor utilization and if a processor where a virtual machine (VM) is allocated gets close to the saturation, then a new VM is activated. This approach works with allocation of threads, and whenever a new transaction is started, the new thread is scheduled to corresponding thread. This extension with resource broker is designed mainly in the operating system or VM scheduler.

However, our approach is not based on this idea, since the tests have shown that it lacks flexibility and often the scheduling requires a lot of resources and takes longer processing time. Web services provide better performance when the resources are orchestrated in many smaller VMs, rather than one big VM [9].

Our approach is based on determination of static and dynamic software parts in the execution [10]. Then the dynamic parts are scheduled on separate smaller VMs, which require less resources and offer higher performance. The static processing is realized via the conventional approach of allocating threads on appropriate VM via resource broker of the VM broker scheduler. This approach offers higher throughput and easier management, and also increased performances. Also, this will improve the performance since each VM will work with smaller database, which will improve the query response time. The design process concerns a two step procedure: the first step is classification of static and dynamic parts, and the second is developing a new architecture to support elasticity and scalability. The result of the classification is presented in Table IV. The next section presents the organization of the new cloud-based TMS.

C. Optimized cloud organization

The new proposed TMS is presented in Figure 5. The static VM (always active) contains the SMM module, while the other

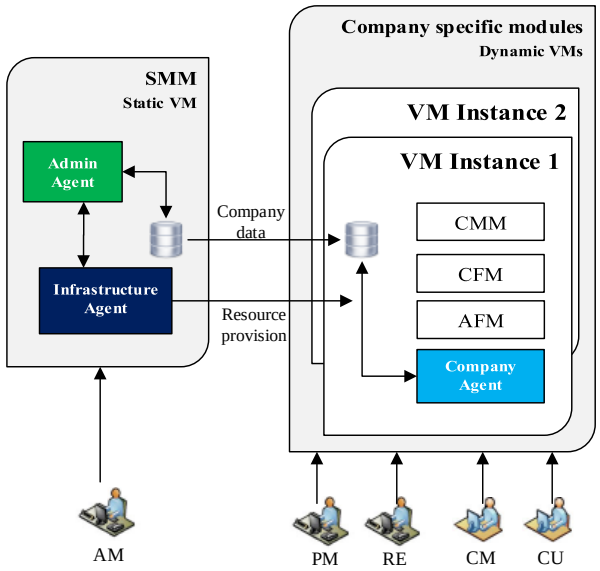


Fig. 5: TMS cloud model

VMs are dynamic and can work in two different modes, the Active mode when there is at last one active company and Inactive mode when there isn't any active company. They contain three modules, i.e., CMM, CFM and AFM.

This system has the very important Broker module in order to communicate with different company specific modules and acts as a role of system service orchestrator. Figure 5 presents three agents that are a part of the Broker module. So let us explain how this system organization works.

SMM is the main module of this system. This module can be hosted either on a bare metal server, or on a fixed static VM with constant resources. It is always able to provide authentication and authorization on the system, company registration, subscriptions and contracts, finance management, user management etc. This module uses two agents:

- *Admin agent*, which provides authentication and authorization, user management, company registration etc.
- *Infrastructure agent*, which communicates with authentication and authorization module of Admin agent to create and shut down the instances in the Company specific module providing the optimal resource utilization and reducing the overall cost for renting the hardware resources.

All other VMs are dynamically associated for each company subscribed to the system. They can be activated when a user of a company authenticates to the system, providing a ticket management function.

Each dynamic VM hosts two core software modules: the CMM module, which provides general management of the company and the CFM module, which realizes the specific ticket management, independently for each company. Additionally dynamic part of the AFM module is hosted on each dynamic VM.

The *Company agent* is hosted on the same VM, to enable a communication with SMM and data synchronization between the SMM and VM.

Note that the static parts of the AFM i CMM will be part of the admin agent and be present in the SMM.

V. Discussion

Open source solutions usually offer complete TMS (Self-Hosted) that can be downloaded and installed by a company. However, this initiates a situation, where companies must have their own hardware and software architecture or to upgrade the existing to fulfill the requirements of the TMS they choose. The installation process is not an easy one, which enforces the companies to employ a person who knows how the system works. Each upgrade for the system must be installed manually. Upgrades usually come with compatibility issues. All these issues can cost the company a lot of time and resources. Also, a company should provide test environment where each patch or new version will be tested prior deploying in the product environment.

A typical IaaS (Infrastructure as a Service) hosting implementation actually requires a self-hosting solution, and the customer is released of worries to buy the server and maintain it. The PaaS solution requires also a self-hosting solution, but in this case the customer uses a platform with an operating system and runtime environment, and even a database offered by the cloud provider. A real cloud implementation is the development of a SaaS solution, where the cloud provider gives a complete service to its customers.

The purpose of developing this type of software or introducing application is to realize web service for bug reporting. This service will be offered as a cloud computing solution to other companies in the facilitation process with customers when delivering software or other products.

Customers often asks themselves what happens when the technology goes forward and the existing applications need to be upgraded to work properly. Sooner or later, something will go wrong [11]. Usually that happens after the testing phase, when the customers start to use your system. For that reason the companies need a good support software for bug and issue tracking.

A lot of developed TMS systems are self-hosting solutions, and recently some have developed cloud solutions. The cloud solutions are more flexible, scalable and most importantly cost effective than self-hosting solutions for companies [3].

Some cloud solutions try to solve the bug reporting duplication problem and the negative impact of it in the development and maintenance process [12]. But on the other hand, bug duplicates are not considered as harmful, because they often provide additional information [11]. Our solution solves this problem by categorization of bugs. Every developer and customer is responsible for a specific module.

Cloud solutions must serve a lot of concurrent customers, so the performance is very important. In this paper we intend to describe only the cloud architecture that offers efficiency, scalability and the best performance, while in [13] we focus on the performance analysis.

VI. Conclusion and Future Work

We have developed a new cloud based system that works in optimized mode. The main functions were classified into static and dynamic components and then clustered into specific modules SMM, CMM, CFM and AFM. Three more agents were designed to allow optimized resource scheduling: Admin Agent, Infrastructure Agent and Company Agent. Our design uses a static VM hosting the SMM with Admin and Infrastructure Agents. The dynamic VMs instances are invoked whenever the admin module triggers a need. Each dynamic VM consists of three modules CMM, CFM and AFM and the Company agent.

The purpose of developing this type of software or introducing application is to realize a web service for ticket management (bug reporting). This service can be offered as cloud computing solution to other companies in the facilitation process with customers for customer satisfaction, quality assurance or when delivering software or other products.

In the future we plan to develop a SaaS solution that deals with these issues and open questions, and realize series of experiments to analyze the performance and the cost of the new cloud based TMS concept.

References

[1] I. Sommerville, "Software engineering. (9th edition)," 2010.

[2] S. Just, R. Premraj, and T. Zimmermann, "Towards the next generation of bug tracking systems," in *Visual Languages and Human-Centric Computing, 2008. VL/HCC 2008. IEEE Symposium on*. IEEE, 2008, pp. 82–85.

[3] P. Gushev, A. Guseva, S. Ristov, and M. Gusev, "Cloud solutions for bug reporting," in *XLVIII Int. Scientific Conference on Information, Communication and Energy Systems and Technologies (ICEST)*, 2013, pp. 227–230.

[4] A. K. Srivastava, D. K. Yadav, and S. K. Pandey, "Taas: An evolution of testing services using cloud computing," *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, vol. 1, no. 10, pp. pp–42, 2012.

[5] R. Paryani, "Nimbo cloud computing," Ph.D. dissertation, San Diego State University, 2011.

[6] B. Williams and G. Reese, *The Economics of Cloud Computing*. Cisco Systems, 2012.

[7] D. Sitaram and G. Manjunath, *Moving to the Cloud: Developing Apps in the New World of Cloud Computing*. Syngress, 2011.

[8] B. Halpert, *Auditing cloud computing: A security and privacy guide*. Wiley, 2011, vol. 21.

[9] M. Gusev, S. Ristov, G. Velkoski, and M. Simjanoska, "Optimal resource allocation to host web services in cloud," in *Proceedings of the 2013 IEEE Sixth International Conference on Cloud Computing*, ser. CLOUD '13. CA, USA: IEEE Computer Society, June 2013, pp. 948–949.

[10] S. Ristov, M. Gusev, G. Armenski, K. Bozinoski, and G. Velkoski, "Architecture and organization of e-assessment cloud solution," in *IEEE Global Engineering Education Conference (EDUCON), 2013 IEEE*. Berlin, Germany: IEEE, Mar. 2013, pp. 736–743, best Paper Award.

[11] S. Just, R. Premraj, and T. Zimmermann, "Towards the next generation of bug tracking systems," in *Visual Languages and Human-Centric Computing, 2008. VL/HCC 2008. IEEE Symposium on*, 2008, pp. 82–85.

[12] Y. Cavalcanti, P. Da Mota Silveira Neto, E. de Almeida, D. Lucredio, C. da Cunha, and S. de Lemos Meira, "One step more to understand the bug report duplication problem," in *Software Engineering (SBES), 2010 Brazilian Symposium on*, 2010, pp. 148–157.

[13] P. Gushev, S. Ristov, and M. Gusev, "Performance Analysis of a SaaS Bug Reporting System," *Tech Report UKIM FINKI*, no. IIT:03-01, pp. 1–6, Jan. 2014.

Alert Notification as a Service

Marjan Gusev, Sasko Ristov
University Ss Cyril and Methodius, FCSE,
Rugjer Boshkovic 16, Skopje, Macedonia
Email: {marjan.gushev, sashko.ristov}@finki.ukim.mk

Goran Velkoski, Pano Gushev
Innovation LLC,
Belasica 2, 1000 Skopje, Macedonia
Email: {goran.velkoski, pano.gushev}@innovation.com.mk

***Abstract*—Digital revolution offers instant information on millions of web sites. In plenty of available data, customers are interested to find information in context of appearance of an object or a certain property as soon as it is published by a given site. Some sites offer push notification to customers if they are subscribed to such a service. However, not all sites implement push service, and in these cases customers frequently visit the web site and wait for an occurrence, such as sport result, rank list, etc. In this paper we present how an intelligent software agent can realise this task instead of users that are manually visiting the web site and send information as if the web site has implemented a push service.**

This novel idea is good for both customers and web site providers. Customers would subscribe to get the information as soon as it is available, and web site providers would like to enable push service without changes in the web site software, that might be very costly. Provision of this kind of notification can be realised via sophisticated mobile phone alerts, text messages, twitter or Facebook notification, or even supported by new operating systems like Maveriks OS or web sites such as Google notifications. In this paper we present details and benefits of a such system.

***Index Terms*—push service, notification, alert service, cloud computing, intelligent agent**

I. Introduction

Remote access to web content is of utmost importance. Searching for information on the Internet can sometimes be the most straightforward job, but on other occasions a laborious and frustrating task for both experienced and casual users [1].

The fact that being well informed is key to almost everyone can surely be counted amongst the reasons for the expansion of social networking.

Borgatti et al. [2] indicate that people are gathering a considerable amount of information from shared informations by their friends and/or acquaintances.

Additionally, Lerman [3] expresses the need of aggregated news on a single website or software makes the news aggregators very popular and a successful business possibility. Powered by additional social information this process may provide an effective social recommendation system.

Furthermore, people are creatures of habit when it comes to web activity, which is very similar to the actual ”autopilot” navigation that is present in nature [4]. We all have our daily routine when it comes to starting the day on the web. Some do it when they wake up, others at their workplace. What is in common is that we all do it, except that we just have different habits such as opening the mail inbox, replying emails, checking the daily news and friends' activities, etc.

Not all sites provide push services and enable notifications. In fact, there are only few sites, such as sport news or similar information sites that send push information to customers. The new digital era initiates situations where, a lot of people are waiting for a notification to appear on certain web pages. For example, one would like to find out if there is a change in a given property associated to a certain key phrase. In this case the user is frequently visiting all those web sites looking for a given text or keyword.

Yet, overwhelmed by more important daily activities, the web occupation is often limited. This is one of the reasons for the success of feed gathering systems that provide data from multiple sites on a single web site or software based on the user required feed. Understandably, this kind of systems are expected from the popular web search engine providers such as Google, Microsoft, Yahoo etc. Clearly, the reason for this is their engine that is able to provide quick news indexing and availability to the users.

Indeed, Google Reader [5] was one of the most popular solutions that provided this functionalities to the users by using the popular RSS [6] (Rich Site Summary) news feed to gather information across the Web.

The problem occurs because not all sites provide RSS feed for their content. Our motivation is to build a cloud based system that provides an engine for site indexing and automatic feed creation for all of the sites. The overall idea is to realise this as a Service.

The expected affliction here is the performance issue since we are not just gathering RSS feeds but also crawl sites that do not offer this possibility. Therefore, in order to provide a scalable and elastic service, we find cloud computing to be a valid solution to do so.

We call the new service Alert Notification as a Service (ANS) because of its notification nature, and place it in the Cloud as a novel cloud service. Additionally, the service provides a possibility of a variety of notification services such as: Email, Facebook Notification, Skype Notification, Google Cloud Messaging (GCM) [7], SMS, Twiter, LinkedIn etc.

This novel cloud service can be offered to the end user, but also to the cloud service providers in order to provide better indexing for their hosted sites, and they can also offer this as a Service. By doing this, the sites will not be burdened with the RSS feed requirement and in the same time the users may get to the much needed results quicker and in the form they require.

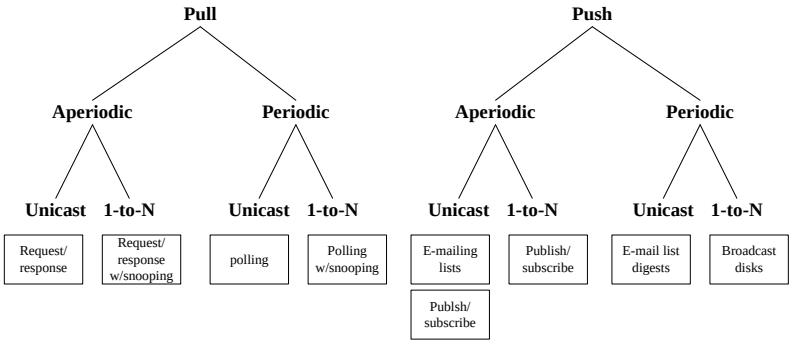


Fig. 1. Data delivery options for notification [8]

This paper concentrates on the development of a web service that will automatically visit all these web sites and alert if there is a change in the existing text or appearance of a certain keyword phrase. It will act as a kind of an intelligent agent and will periodically visit the pages instead of users to find a particular occurrence of their interest.

The rest of the paper is organised as follows. Related work in the area of push notification services and news feed services is presented in Section II. Section III describes the concept of ANS. The details of the system requirements and initial system architecture are discussed in Section IV. In Section V, we present the proposed system architecture intended for push notifications delivery. Section VI discusses the notification as a service system. The conclusion is specified in Section VII.

II. Related Work

In this section we will discuss some similar currently popular services and software solutions and compare it to our approach. The huge variety in offer of similar solutions confirms the popularity of this kind of systems.

We already started discussing similar software possibilities such as Google Reader in Section I. This is just an honourable mention to the software that Google released. Yet, the solution was discontinued and as of July 15, 2013 all subscription data is permanently and irrevocably deleted [5]. Google reader was an RSS feed aggregator and since it was deprecated a lot, for other solutions emerged to replace it.

Just by using the alternative to list [9] we can find more than 200 applications present that will offer similar services [10]. Some notable mentions are: Feedly [11], NewsBlur [12], Feedspot [13], etc. Almost all of these solutions offer a way to organise and read web content that is automatically aggregated on a single website or a software. Additionally, some of the solutions even offer statistics and synchronisation among multiple devices or even between friends when integrated with social networks.

Recently, Google released a new popular solution called Google Alerts that allows users to monitor the web for interesting new content based on search queries [14]. Besides the initial popularity of the system, one cannot miss out the more than few issues that the system has [15].

Therefore a lot of professional level and entry level applications are available on the market as both open source and proprietary solutions. Dembak [16] identified popular alternatives for this system. We will just mention some of them and summarise their functionalities. Talkwalker, Social Mention, Mention, Trendrr, Bloomberg Professional, Meltwater News, etc. are some of the popular solutions that provide functionalities for easy tracking information on specific pages or on the Web. Functionalities such as alerts in real time from web via email or in a RSS feed reader, social engine tracking, news importance ranking, analytics are available in these solutions.

There are various technologies that can be used to develop a notification service. The basic technology used was the push concept. Push technology stems from a very simple idea with ”in-your-face” nature [8]. Rather than requiring to explicitly request (i.e., ”pull”) the information that they need, data can be sent to users without having them specifically ask for it.

Push technology has evolved in last 15 years, and even the WWW consortium is trying to establish a description of Push API [17]. It defines how a web application can use an API via the user agent that will communicate to the push server and application server. HTML5 also supports the push technology [18] by defining Server-Sent Events (SSE) and standardising Comet for all standards-compliant web browsers. The Server-Sent Events specification ”defines an API for opening an HTTP connection for receiving push notifications from a server.” [19].

Push and Pull technologies use different data delivery options [8], as presented in Figure 1. Aperiodic pull is using the request response communication, while the periodic one is using polling. Push technology pushes the information without request or polling. In this case the user has to subscribe for a service. If a site would like to realise a push service, then the corresponding technology has to be used and this requires more resources.

Push notifications may be delivered via various methods which include standardised protocols (e.g. Server-Sent Events [SSE], the GSM Short Message Service [GSM-SMS], SIP MESSAGE [RFC3428], or OMA Push [OMA-PUSH]), or via

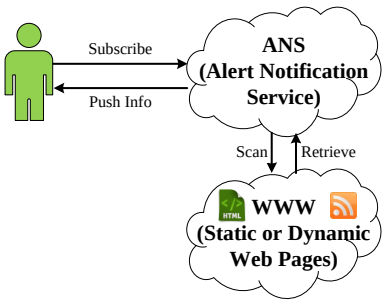


Fig. 2. Concept of ANS

browser-specific methods.

RSS is a realisation of push technology or a web feed, usually used for syndicated news service when one would like to read all the content that is pushed by a given server. Restful services are those Resource oriented architecture [20] and Representational State Transfer (REST) [21] can be used to realise push by concentrating on the roles of the components, the constraints of their interaction with other components and their interpretation of significant data elements. One of the alternatives is using SOAP while building web services. The main idea is to change the standard client server communication with a stateless communication by sending a representation of each application state with appropriate links.

Pushlets are another example of a push notification technology [22], where the server never terminates the response and the client is automatically notified about new events.

III. Concept of ANS

Our new solution helps all those sites that would not like to invest a lot of resources to implement various push services. They will still remain the same static or dynamic web sites. The concept of the ANS is presented in Figure 2. The user subscribes to the service and chooses the notification delivery method. Then the ANS is performing all the tasks, by frequent scanning the existing web site and finding a specific condition for an alert to occur.

In our solution we build all technology resources, similar to those described in W3C specification of a Push API [17]. However, we design a creation of a full service and do not require that the user has to change their concept of realisation the web application. Our service is realised as a web application where a user can register and subscribe using any web browser and can specify the way the notification can be received by choosing any conventional methods like e-mail, text message, social network or operating system notification.

Main functionalities can be summarised as in Table I.

Upon registration one can login to the main admin page. The main admin page can allow to define the alerts by selecting web pages and corresponding text to appear or to be changed from the last occurrence. Two main functionalities are possible within this activity:

TABLE I
Main functionalities of the Alert Notification Service

Functionality	Description
Definition	Specification of a web page, search key phrase, a property or occurrence of a keyword as an alert
Scanning	Crawling, visiting the web page and retrieving the required web content
Triggering	Controlling the crawling frequency, comparing the content with alert constraints and triggering the alert
Delivery	Activating the delivery channel and transmitting the alert message

- Select a web page where there is a change in text occurrence. For example, one would be interested to know when a new published paper will be indexed in a particular database, such as ieeexplore, and which paper, if more papers are waiting for publishing; and
- Select a web page with occurrence of a given text phrase. For example, my name has appeared in a blog or news.

The next functionality is selection of delivery method. The user can select among various delivery channels, such as personal message delivery via electronic mails or text messaging on a mobile device, personal messaging on social networks, or recently as a functionality of the operating system, such as Maveriks OS, or functionality of the web browser, such as Google Chrome.

The main functionality is realisation of scanning the existing web sites. Actually, the system should be able to do what a typical user is doing when visiting a web page, looking for an information. It means not just visiting a certain web page and finding an occurrence of a text within a paragraph, but also clicking on certain buttons or entering a value in certain fields. It will simulate a human entering data on an interactive web page, with a possibility to select a certain radio button or an item in a drop down list, to enter a value in a certain text field, or to click an activation button. Then the system will be able to trace the file and find a predefined keyword phrase, by detecting if there is a change in the context or in the occurrence of the keyword phrase.

This new system will enable the web sites that do not have push notification, alerting system or use corresponding API to realize alert notification. It will actually build push notification for any web page.

The alerting system can be realized conventionally by e-mail, skype, twitter or facebook messaging, or can be realized as a web page where one can see notifications, or as an add-on tool for the web browser. An example of occurrence is the alert bell on Google chrome.

It can be thought of as a system for users who would like to be notified for a certain change, or for cloud service providers who would like to build push functionality for classical web sites that do not support this technology.

IV. System Requirements

In this section we will discuss details of the system requirements and define the initial system architecture.

The system can be a service for users who would like to be notified about occurrence of a key phrase on a certain web page. The user interface should enable definition of the alerts and browsing of the results. Additionally, the ANS system may be offered to cloud service providers who would like to build push functionality for classical web sites that do not support this technology.

The main system functionalities are specified in Table I. Here we elaborate details of each functionality.

A. Definition

The system should allow user friendly definition of alerts. An alert can be defined by providing the following information:

- **key phrase** The user is looking for a certain key phrase in the web content. It may be one word or a sequence of several words. For example, one might look for a certain ranking to find if a name of a given person appeared in the list;
- **keyword as property of the key phrase** The user is looking for a certain property of the key phrase or just defining an occurrence of the keyword. For example, one might likes to find out the ranking of the person in the list and retrieve the position number;
- **alert constraint** The user should be able to define when to trigger an alert. For example, it may be an occurrence of a key phrase, change in its property, such as position number, combination with other keywords, or when the occurrence is changed etc.;
- **web page** The user is providing the url of the required web page;
- **defining a macro** The user should be able to define a macro of provided keyboard or mouse actions that leads to a occurrence of a specific web page content. For example, one might click on a certain radio button, select an option in a drop-down list, enter a specific text, click on action button in the given web page to filter the web content and dynamically choose the web content;
- **scanning frequency** The user must be able to select how often the site should be crawled for updates; and
- **delivery channels** The user should be able to select delivery channels for alert notification. For example, one might select e-mail and SMS text message on a mobile device as notification delivery channel.

B. Scanning

Scanning is realised as a crawling engine that will frequently visit a certain web page and execute various selection or activation actions. The final output is retrieving the web content on the visited web page, avoiding as much junk information as possible.

The prerequisite for this requirement is different from all existing software solutions that mostly use RSS feed from a specific site to retrieve a corresponding web content. This service must be able to extract information from any web site by crawling through its content. The crawling frequency is a parameter that can be defined by the user.

Additionally a user can select a web site that is offering RSS feeds and use this service as a kind of filter that selects a key phrase occurrence as defined alert. In this case the crawling is not started, and this functionality is actually a typical news aggregation functionality.

C. Triggering

Triggering refers to implementation of the main core control in the web service. Once the web content is retrieved by the scanning service, it is analysed and cleared from junk information. The selected information is then compared to the definition of the alert constraint. A positive outcome of the alert constraint condition triggers an alert message which is then scheduled for transmission via delivery channels.

This functionality is also novel compared to other present similar systems. The innovative approach is required in order to extract just a predefined selected part of the web content that the users need. Let's say that we want to follow the stock information on various stock exchanges. We can just simply follow the web page RSS feed with plenty of junk information and then extract the required information. The system must provide the user means to select a specific information. Definition of alert constraint is the realisation of an agent that checks if a certain condition is fulfilled triggering an appropriate alert. Then this triggers a push notification via selected delivery channels.

Additionally the user must be able to record the user interface input to filter the web content. It consists of several mouse and keyboard actions for a given web page that result in presentation of a specific content. For example, the user may be a scientist who is interested when a new journal paper is referenced on IEEEXplore database. The user should be able to navigate to the IEEEXplore search enter the search query he desires and check if in the list of the papers there is some change. The whole navigation and input on the page should be recorded in order to enable repeated query of the same request.

The whole functionality must have a web interface i.e. all the activities should be recorded using just web browser. This provides the means to use the same functionality across multiple platforms and additionally across multiple devices.

D. Delivery

When the alert is triggered the corresponding delivery channel is activated.

The system can activate a push notification to the end users as a kind of a real time notification using various messaging channels. The users can define some of the following delivery channels to get alert notifications:

- personal messaging via Email;
- personal messaging via text message (SMS) on a mobile device;
- personal messaging via social networks, like Facebook, Twiter, LinkedIn, etc.;

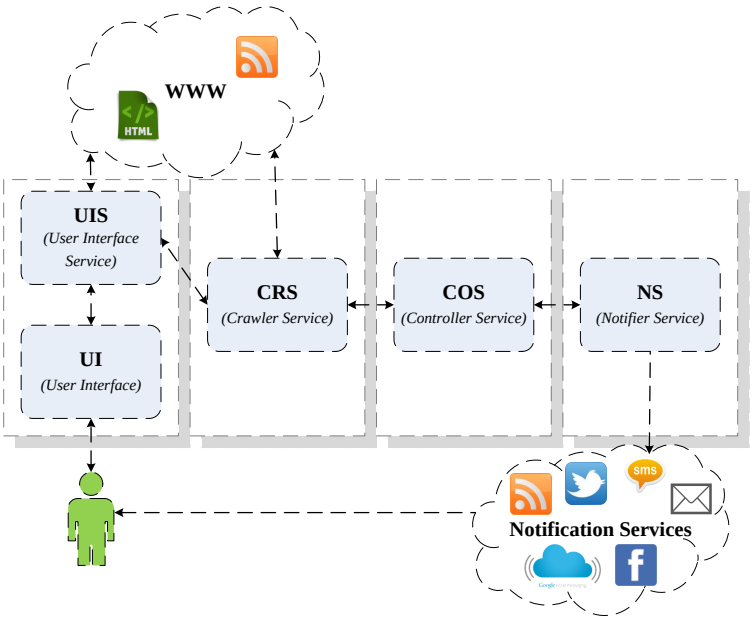


Fig. 3. Architecture of ANS

- cloud messaging, like Google Cloud Messaging (GCM);
- operating system notification, such as Maveriks OS, IOS, etc.;
- web browser notification, such as Google Chrome; and
- news feed delivery RSS.

This is also a functionality, which is not provided by the other similar solutions.

E. Other functionalities

Next we present other functionalities that may define the user interface of the complete ANS service. We start with content organisation and grouping. This functionality enables better user experience and is performance bound. The system is expected to be used by multiple users and often some will request similar if not even the same queries. Therefore, the system must organise and group the queries based on a required web page, or a specific query. This way the system should visit one single page and index it for information ones to cover as much as possible user queries.

The next functionality concerns the statistics of usage and alert generations. It may be important for future use to analyse typical behaviour and schedule corresponding resources for cloud service providers and site owners. The final goal is to realise an intelligent agent that will predict user activities and provide an information faster, and with greater accuracy.

The overall system should be compatible with different notification methods. It should support all existing push notification technologies and provide an API that can communicate with other web pages, push notification systems, RSS feeds, or similar scheduling systems, such as calendars. Realisation of an API that communicates with other systems and can integrate with operating systems and web browsers from one side, and web and push servers from other side is another feature of our service compared to the other similar systems.

V. ANS System Architecture

The main architecture of the new defined ANS system is presented in Figure 3. Core modules are realising the main functionalities. The main User Interface Service (UIS) brought by a web browser realises the functionality to define the alerts, including recorder and simple web interface components.

The Crawler Service (CRS) realises the scanning functionality and outputs appropriate content to the Controller Service (COS), which realises the triggering functionality. COS activates the CRS on a specific timing to realise frequent web page scanning. It also analyses the provided content, filters the relevant information from the junk and stores the relevant data. If the condition defined by the alert constraint is fulfilled it triggers the Notifier Service (NS) which realises the delivery functionality. NS activates corresponding notification channels to the users.

Additionally the system requirements must be performance and power aware. Therefore the system must be scalable and flexible in order to provide virtually unlimited scalability.

VI. Discussion

The advantages of push technologies are straightforward [8]. The traditional pull approach requires that users know a priori when and where to look for data or that they spend an inordinate amount of time polling known sites for updates and/or hunting on the network for relevant sites. Push relieves the user of these burdens.

There are several problems with the implementation of the push technology. The web technology changes a lot and the crawler system might need constant update to enable quality

TABLE II
AUTONOMOUS SOFTWARE AGENT PROPERTIES THAT OUR SYSTEM HAS

Property	Description
reactive	responds to a change in required information for a given content
autonomous	controls the process of triggering without direct human interaction
continuous	runs as continuous process over time
trustworthiness	adheres to learned behaviour and is truthful

scan of the web page and allow transfer of the related content from the web site to the system. Thus the system can make a good decision to find out if a condition for triggering occurs.

This system acts as an autonomous software intelligent agent according to the definition given in [23]. It scans the web page instead of the users and finds if a certain condition is fulfilled. The environment is the web, sensing correlates to scanning, and the action that produces change of state is comparison. The algorithm of realisation of the intelligent agent in this case follows a simple principle to learn by an example given by the user and sense the relevant information from the collected content.

The described online learning system with adaptive testing has several important agent's properties as presented in Table II. Our intention was not to present all characteristics and agent features, as given by Badica et al. [24] but to cover the most essential and specific properties analysed by our system.

Another problem is the realisation of the decision making algorithm that triggers an action. It is not a simple comparison of two strings, but it is realised by a rather complex actions. This happens since various technologies can hide the relevant information. For example, the information can be positioned it in tables, frames, in-line text, or various modules, so the realisation of the web scanner needs a careful algorithms to include all relevant cases. Various browsers present the relevant information with different algorithms, so when the user learns the crawler, has to be careful to use appropriate browser, which is supported by the crawler.

VII. CONCLUSION

In this paper, we have introduced a new system that actually builds on all those classical web sites, where the user is pulling information and realises the push of required information. In this case the user specifies what can trigger and push the information to the user. The system is built as a software as a service cloud solution.

One might think that this system is just an agent that provides push functionality to the conventional web sites. Instead of recompiling the software on enormous dynamic web sites and realise push technology for them, this system enables adding just a service with the same functionality that one can subscribe. In reality it repeats the actions, one frequently performs when looking for a certain information, and acts if a human reads the site and finds out a change in information he/she is looking for. This is represented as an alert and the user will be informed as soon as there is a change in information he/she is looking for.

Modern operating systems, like Maveriks, IOS or others include alert notification features if there is a change in some system information, like calendar events, received mail, finished actions, etc. Some modern web browsers also build notification alerts, such as Google Chrome, which is looking for a news on different web sites which correlates to the one the user is interested. This is a part of the conventional Google sprawling action, which compares users keywords with new scanned web pages. However, this is quite different from our service, which looks for a change in information in a predefined web page.

Adding functionalities of our newly introduced system may be part of all new releases of modern operating systems and web browsers. They are beneficial for both the users looking for a change in information or by providing push technology to conventional web site, beneficial for enormous number of cloud and web site providers.

REFERENCES

[1] C. Hölscher and G. Strube, "Web search behavior of internet experts and newbies," *Computer networks*, vol. 33, no. 1, pp. 337–346, 2000.
[2] S. P. Borgatti and R. Cross, "A relational view of information seeking and learning in social networks," *Management science*, vol. 49, no. 4, pp. 432–445, 2003.
[3] K. Lerman, "Social information processing in news aggregation," *Internet Computing, IEEE*, vol. 11, no. 6, pp. 16–28, 2007.
[4] J. Prendergrast, B. Foley, V. Menne, and A. K. Isaac, "Creatures of habit?" *The art of behaviour change, London*, 2008.
[5] "About Google Reader." [Online]. Available: http://www.google.com/reader/about/
[6] "RSS 2.0 Specification (version 2.0.11)." [Online]. Available: http://www.rssboard.org/rss-specification
[7] Google. [Online]. Available: http://developer.android.com/google/gcm/index.html
[8] M. Franklin and S. Zdonik, "data in your face: push technology in perspective," in *ACM SIGMOD Record*, vol. 27, no. 2. ACM, 1998, pp. 516–519.
[9] "AlternativeTo.net." [Online]. Available: http://alternativeto.net/
[10] "Google Reader Alternatives on AlternativesTo.net." [Online]. Available: http://alternativeto.net/software/google-reader/
[11] "Feedly, Homepage." [Online]. Available: http://feedly.com/#welcome
[12] "NewsBlur, Homepage." [Online]. Available: http://www.newsblur.com/
[13] "Feedspot, Homepage." [Online]. Available: http://www.feedspot.com/
[14] "Google Alerts." [Online]. Available: http://www.google.com/alerts
[15] R. Grant, "Google Alerts 'broken,' 'useless,' and slowed to a 'trickle'," 2013. [Online]. Available: http://venturebeat.com/2013/03/21/google-alerts-broken-useless-and-slowed-to-a-trickle/
[16] Y. Dembak, "7 apps to help you replace Google Alerts," 2013. [Online]. Available: http://venturebeat.com/2013/05/13/google-alerts-replacements/
[17] W3PUSH. (2013, Aug.). [Online]. Available: http://www.w3.org/TR/push-api/#dfn-appserver
[18] W3HTML5. (2013, Aug.). [Online]. Available: http://www.w3.org/TR/html5
[19] TodayJava. (2010, Mar.). [Online]. Available: https://today.java.net/article/2010/03/31/html5-server-push-technologies-part-1#still
[20] L. Richardson and S. Ruby, *RESTful web services*. O'Reilly, 2008.
[21] R. T. Fielding, "Architectural styles and the design of network-based software architectures," Ph.D. dissertation, University of California, 2000.
[22] J. van den Broecke, "Pushlets white paper," *Just Objects BV August*, vol. 6, 2002.
[23] S. Franklin and A. Graesser, "Is it an agent, or just a program?: A taxonomy for autonomous agents," *Intelligent Agents III Agent Theories, Architectures, and Languages*, pp. 21–35, 1997.
[24] C. Bădică, Z. Budimac, H.-D. Burkhard, and M. Ivanović, "Software agents: languages, tools, platforms," *Computer Science and Information Systems/ComSIS*, vol. 8, no. 2, pp. 255–298, 2011.

Crowd computing to support EGI scientists

Jozsef Kovacs, Robert Lovas
MTA SZTAKI, Budapest, Hungary
{jozsef.kovacs, robert.lovas}@sztaki.mta.hu

Abstract – The International Desktop Grid Federation (IDGF) aims to collect scientists, IT operators, developers, etc. who are interested in Desktop Grid computing and related technologies. In the on-going EU Framework Programme 7 "International Desktop Grid Federation - Support Project" (IDGF-SP) there is a special focus on spreading the technology among institutes and universities by providing support in different areas like desktop grid installation, maintenance, application porting, dissemination and other topics. The members of IDGF-SP are therefore continuously maintaining and improving a network of volunteer and campus-wide BOINC projects to form a reference, production level infrastructure. The IDGF-SP infrastructure consists of volunteer and institutional BOINC projects, bridges to/from EGI (European Grid Infrastructure) type grids (e.g. gLite), cloud resources, science gateways, monitoring and accounting systems, and virtualization support for the volunteer resources. The infrastructure is currently open for EGI scientists through bridges. Several EGI communities are already utilizing volunteer resources with their own applications while the aim of the newly established EGI Virtual Team is to increase the number of these communities and success stories. The paper introduces the IDGF-SP infrastructure, its usage by the supported communities and some of the most interesting technical challenges of the integration to the EGI ecosystem. We will also show how the overall usage is moving towards crowd computing where human intelligence and other resources from citizens might be also utilized in scientific experiments.

I. INTRODUCTION

There are over a billion PCs in the world, and most of these PCs can be located in citizens' homes and, to a lesser extent, in universities. Huge portion of these computers remain idle most of the time but a few millions of them are already active in supporting science in a volunteer computing grid and use their idle time to run scientific applications. The potential growth of this computing capacity is obviously enormous. Many Desktop Grids have therefore decided to found the International Desktop Grid Federation (IDGF) [1] to help each other improving their e-Infrastructures.

The IDGF-Support Project [2] is working on giving IDGF a boost in two important areas. First, it helps considerably with increasing the number of citizens that donate computing time to e-Science. Targeted communication activities and setting-up a network of "ambassadors" is one of the possibilities. On the other hand, it also helps universities' e-infrastructures to include otherwise idle PCs from their class rooms and offices. In addition IDGF-SP collects and analyses data that will help deploying idle PCs in an effective and energy efficient way. It has been shown that Desktop Grids can contribute to Green IT if used in the correct way. IDGF-SP collects data to underpin and advocate best practices. As a result of IDGF-SP, the number of citizen volunteers donating computing to e-Science increases. By employing unused PCs in private Desktop Grid, universities and other research organizations, they saves on their costs on providing computer capacity for their scientists. The federation helps strengthening the co-operation amongst Desktop Grid e-Infrastructure operators. IDGF-SP encourages and helps the participating Desktop Grid providers to integrate their infrastructures into the main e-Science environment.

In the IDGF-SP consortium, the members are operating desktop grid sites with different setup an operational policies. SZTAKI is operating two umbrella BOINC [3] projects called EDGeS@home and SZTAKI Desktop Grid that hosts applications from different scientific areas. The first devoted to European scientists while the second is serving Hungarian scientists. Another example is AlmereGrid, the first CityGrid in the world. It runs a large Volunteer Grid in production and several Grid and Cloud test environments. The main goal of the AlmereGrid volunteer Desktop Grid is to involve citizens, companies and organizations in the city of Almere into helping science. The focus on the city means AlmereGrid can participate in events in the city, organizes exhibitions, goes to schools, etc. "Mathematics is fun, and everybody can help with finding new mathematical results." That is the main message of ABC@Home. They operate one of the largest Volunteer Desktop Grids in the world, with the goal of working on the last Theorem of Fermat. However, just as important is to get younger generation involved in this. Hence ABC@Home works together with several organizations to provide a lecture programme and practical mathematical exercises for kids. They can test the mathematics themselves and then see how it works on the Volunteer Desktop Grid. Charity Engine from the UK takes a complete different approach than traditional volunteer grids. With Charity Engine the calculations done are paid for by the organizations that do the calculations. The money earned by Charity Engines is donated to a number of charities. To even encourage people more to participate, part of the money collected by Charity Engine is used in a prize draw amongst people that donate computing time. Another directions is the University of Westminster Local Desktop Grid (WLDG) that connects laboratory PCs of the University of Westminster (London, UK) into a BOINC based Desktop Grid infrastructure. The university is set over four main campuses and some additional smaller locations in Central- and North-West-London each of them offering a variable number of Windows based dual core PCs for teaching purposes. Over 1600 of these ma-

chines are connected to the WLDG serving the various research areas in the University.

The second chapter of this paper gives an overview about the IDGF-SP infrastructure (and its underlying technologies) that includes among others the above described Desktop Grids. The third chapter outlines its integration to the most relevant European representative of e-Infrastructures (EGI). The fourth chapter shifts the focus onto the scientific user communities, and the last chapter shows the ways how the different key players of Desktop Grids, i.e. the volunteers and the scientific users can converge to each other through citizen science and crowd-computing.

II. THE IDGF-SP INFRASTRUCTURE

The IDGF-SP infrastructure (see Figure 1) contains a core set of volunteer and campus desktop grid sites (introduced previously). Currently, seven sites are associated as core sites, meanwhile new sites can join to the infrastructure. Each site has its own operational policy; however there are some common points. Most of the sites are providing 3G Bridge [4] service in order to provide interface for incoming jobs. Through 3G Bridge, jobs can typically arrive through gLite [5] virtual organizations (upper left corner) or Scientific Gateways (bottom left corner).

EGI user communities can submit their own applications to this infrastructure. To support non-validated and non-preregistered BOINC applications, we are providing the GBAC [6] virtualization mechanism. Whenever a job for a non-registered application arrives, it is encapsulated inside a virtual machine (set up by GBAC) on the volunteer resources (upper right corner) based on the VirtualBox virtualization software. BOINC has a support mechanism to handle virtualization on the volunteer resources. The other source of jobs can arrive through web-based scientific gateways like the publicly accessible Autodock portal [7] providing several execution scenarios based on the well-known Autodock molecule docking software.

There are a big number of applications ported for BOINC, they can be accessible from different repositories (middle left) like the one called EDGI [8] where applications have a deployable version for BOINC. In case of user request, administrators can download and install the selected applications. Most of the applications are also uploaded and documented in the EGI repository.

BOINC projects are operated/maintained and monitored continuously by admins. Automatic monitoring system also detects situations when long delay occurs for some jobs. Due to voluntarism, the computing resources are untrusted and unreliable. BOINC can handle unreliability however, to handle jobs having to much delay, a mechanism is automatically initiates: new cloud resources with SLA (bottom right corner) are instantiated on-demand and delayed jobs are redirected to those reliable resources. The redirection mechanism contains a job selection, cloud resources management and job reassignment functionalities.

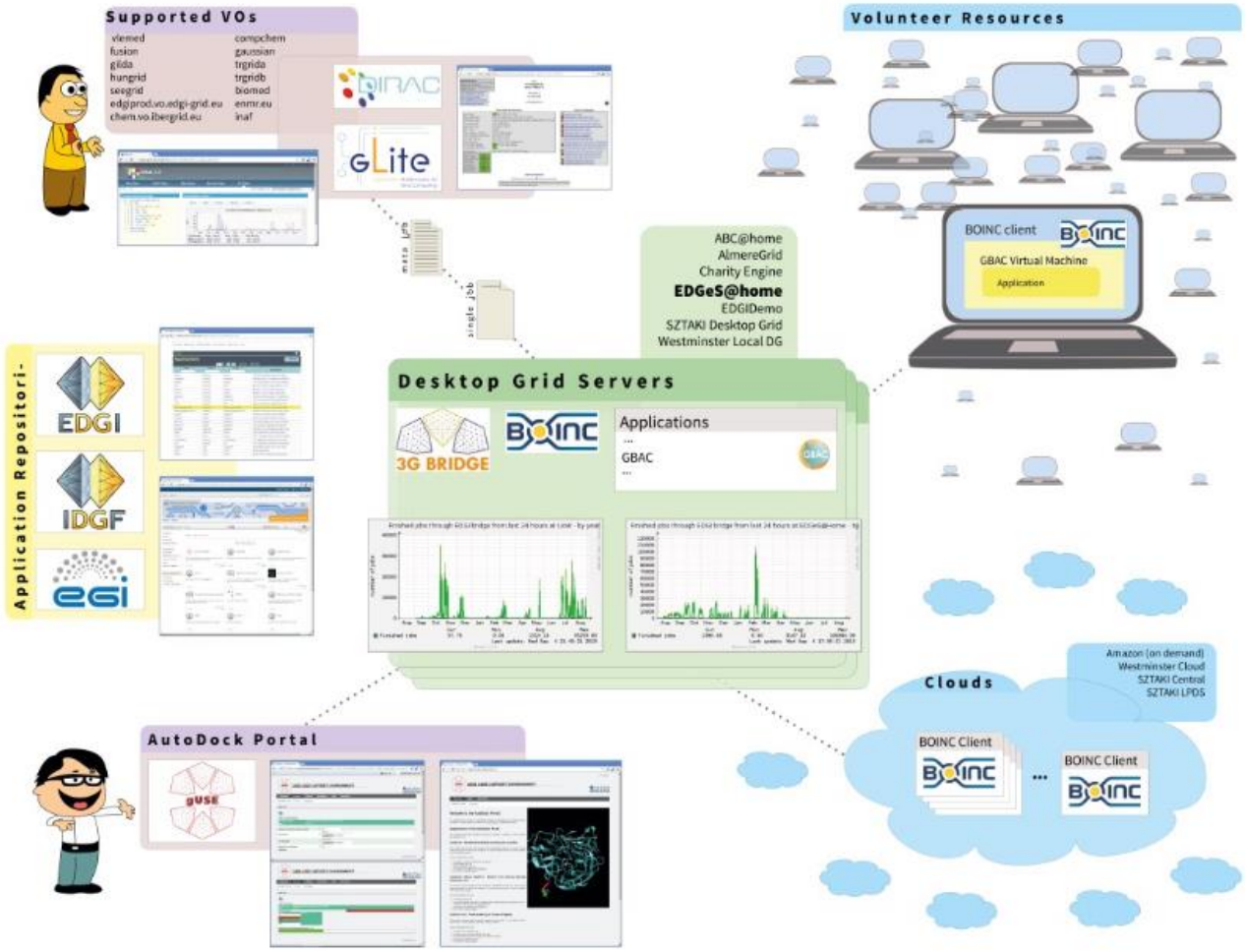


Figure 1 Overview of the IDGF-SP infrastructure

The overall infrastructure is operated and managed by the consortium members providing user, volunteer and infrastructure support, too.

III. INTEGRATION WITH EGI

The IDGF-SP project has a strong focus on the EGI related research communities [9]. At the EGI Community Forum (May 2013), IDGF-SP has started organizing a new Virtual Team in the frame of EGI-InSPIRE [10] project in order to improve its presence/acceptance further and to promote Desktop Grids among the EGI scientific communities. After the negotiations and feedbacks received from the potential invited partners and from the EGI.eu experts, the VT proposal has been officially approved by EGI and the Desktop Grid VT [11] has been launched (with open call for participation) in October 2013.

The project provides support for EGI communities to help them executing their application on the IDGF-SP infrastructure. To execute the applications in a seamless way integration of the IDGF-SP infrastructure into the EGI ecosystem is an important goal [12]. There are activities targeting technology tracking to keep the software components up to date with the recent requirements while there are also activities to increase the level of integration in different areas like job execution, accounting, user support, etc. Based on the bridging technology - the EDGI project developed - it is already possible to submit jobs from gLite to Desktop Grids, however there are other aspects of integration that are missing or not fully implemented. Recent results in terms of integration are as follows:

1) launched the IDGF Support Unit [14]

In cooperation with the EGI GGUS team, the IDGF-SP project has set up a support unit in order to provide the EGI compliant support service for the EGI scientists and resource providers. The support unit collects Desktop Grid experts from various areas (application porting, infrastructure maintenance, etc.), is registered and accessible through the GGUS ticketing system [15] operated by EGI.

2) completed the accounting related integration [12]

The EGI accounting system is based on the APEL accounting tool that collects accounting data from sites participating in the EGI infrastructure.

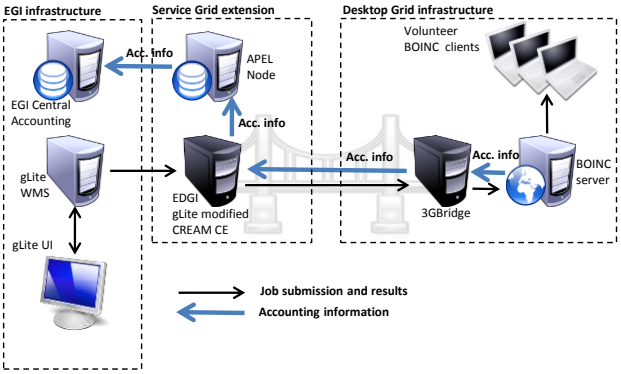


Figure 2 Overview of the combined EGI – DG accouting architecture

The accounting information is gathered from different sensors into a central accounting database where it is processed to generate statistical summaries. These are available through the EGI Accounting Portal [16] for view in different details by Users, VO Managers, Site Administrators, etc. In order to attach the Desktop Grid site as accounting information provider, it must 1) generate accounting information for each job, and 2) must deliver it to the appropriate service. The Desktop Grid site represented by a BOINC server by default keeps track of the workunits and gathers accounting information from the clients about the finished workunits in its database. This information is collected by the 3G Bridge tool for each individual job once a job is finished. The modified CREAM Computing Element queries this information for each job and synchronizes with the site level Apel [17] node. This Apel node acts as a sensor in the Apel network and regularly sends the accounting information of the site to the central EGI accounting server. Finally, the EGI users can browse accounting information about the attached DG sites through the EGI Accounting Portal [16].

3) offered EDGeS@home capacities into the EGI Federated Resource Pool [18]

One of the largest IDGF-SP volunteer project offered for the EGI scientists as a distributed computing infrastructure is the EDGeS@home BOINC-based project. EDGeS@home has already provided more than 300 Million Cobblestones in BOINC terminology computing power for the scientists since October 2009. In a nutshell, BOINC's unit of credit (the Cobblestone) is granted based on the computations performed by the volunteer's computer, and equals to 1/200 day of CPU time on a reference computer that does 1,000 MFLOPS based on the Whetstone benchmark. Using EGI metrics, it would mean also millions of CPU hours. Figure 3 shows the recently provided CPU hours in the EGI accounting portal for executing the workunits (jobs) originated from WeNMR.

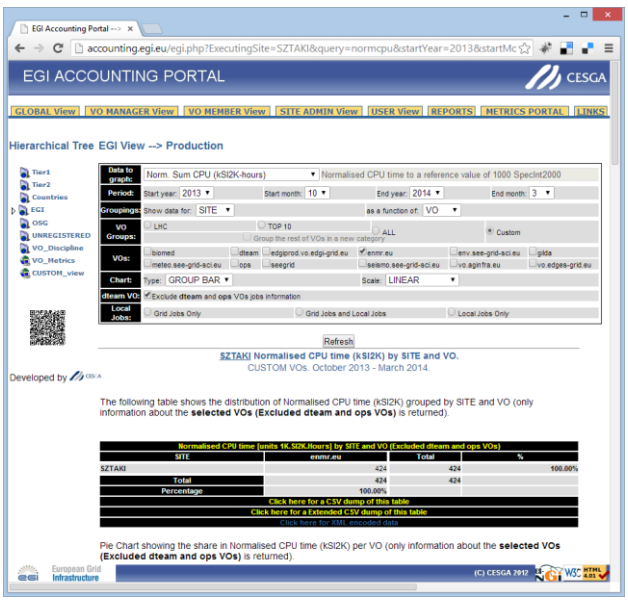


Figure 3 Accounting info presented by the EGI accounting portal

IV. SUPPORTING USER COMMUNITIES

In this section, the most representative usage scenarios are introduced through several user communities already utilizing volunteer resources with the help of IDGF.

WeNMR [19] is a Virtual Research Community supported by EGI, the largest one within the life science area. According to their introduction; WeNMR aims at bringing together complementary research teams in the structural biology and life science area into a virtual research community at a worldwide level and provide them with a platform integrating and streamlining the computational approaches necessary for NMR and SAXS data analysis and structural modelling.

Thanks to the virtualization (by GBAC) and to the bridging technology IDGF maintains, the WeNMR grid is able to forward jobs submitted by their scientific users to volunteer resources behind the dedicated computing element (cr2.edgi-grid.eu). Currently, jobs for the CNS application are regularly submitted to EDGeS@home. The Crystallography & NMR System (CNS) has been designed to provide a flexible multi-level hierarchical approach for the most commonly used algorithms in macromolecular structure determination. However, this software required some dependencies (deployed binaries, environment variables, etc.) to be resolved which was easily applied by the virtual machine GBAC sets up on the volunteer resources.

The DIRAC [20] (Distributed Infrastructure with Remote Agent Control) project is a complete Grid solution for a community of users needing access to distributed computing resources. DIRAC forms a layer between a particular community and various compute resources to allow optimized, transparent and reliable usage. The Workload Management System with Pilot Jobs introduced by the DIRAC project is now widely used in various grid infrastructures. This concept allows aggregating in a single system computing resources of different nature, such as computational grids, clouds or clusters, transparently for the users.

In this scenario, the DIRAC automatic pilot submission mechanism makes sure that as many pilots are running as possible. It submits pilots to a dedicated computing element (cr2.edgi-grid.eu) which forwards them as GBAC jobs to EDGeS@home. Once a volunteer starts running a GBAC workunit with the DIRAC pilot, the pilot comes to live and starts downloading and executing DIRAC jobs on the resource. In contrast to the previous scenario, the job submitted to the IDGF infrastructure is not performing the computation by itself, but rather attaches the resource to another grid to receive computational jobs. This is a non-regular, but interesting use case.

BioVeL [21] is a virtual e-laboratory that supports research on biodiversity issues using large amounts of data from cross-disciplinary sources. BioVeL offers the possibility to use computerized "workflows" (series of data analysis steps) to process data, be that from one's own research and/or from existing sources.

In this third scenario, two BioVel applications (BBGC and BBGC-MuSo) has been ported to BOINC as native applications. The ported applications have been deployed on EDGeS@home and jobs can be submitted through the 3GBridge interface. The BIOVEL portal exposes different workflow executions as services for the users. By executing a given workflow, jobs (for the aforementioned two applications) are generated and submitted. Similarly to the first two scenarios, a high-level portal hides the details of handling the underlying grid.

The fourth scenario is implementing a different approach in terms of job generation and submission. Mathematicians from the Department of Computer Algebra of the Eotvos Lorand University of Sciences is working on discovering the distribution of prime numbers based on the Riemann-Siegel [22] function. The main objective of this research is to generate lots of large values of the zeta function, in order to reveal new patterns related to prime numbers. For this purpose, the zeta-search application has been developed which is able to predict the value of the zeta function for a given number.

This application is an ideal parameter sweep one, where the parameter is a single number (increasing towards infinite) and output is the predicted value of zeta-function. Currently, SZTAKI Desktop Grid is hosting the application, thousands of jobs are continuously generated for the volunteer resources. In this scenario job generation has been fully automated and performed locally on the BOINC server, no need for user interface in this case.

V. TOWARDS CROWD COMPUTING

Citizen science can be considered as a new and innovative approach in order to involve the general public in scientific processes in several ways. This special approach helps people understand what scientists are dealing with in their 'ivory towers' by letting the 'outsiders' (i.e. citizens) participate and engaged in various scientific research and experiments.

One of the main aims of the FP7 SOCIENTIZE project [23] is to "promote the usage of science infrastructures (dedicated and external resources), including professional and amateur scientists, and also sets up a network where infrastructure providers and researchers recruit volunteers from a general public to perform science at home".



Figure 4 Screenshot on on-going Mind Paths experiment using pyBossa by SOCIENTIZE (source: [24])

Socientize has been implementing a set of applications that enable volunteers to become scientists, encouraging active public participation in scientific research processes and showing the capabilities of open community based resources, such as

- Citizens scientists can help scientists collect data that will be analysed by professional researchers.
- Performing tasks that require human cognition, knowledge or intelligence such as image classification, image analysis from scientific infrastructure, transcription, and more e.g. within a web browser.
- Experiments with citizen sensor networks.
- Mental skill games related to science experiments (see Figure 4).

Some of the experiments from SOCIENTIZE are based on the PyBossa [25] platform. PyBossa is an open source platform for so-called 'crowdsourcing' online (volunteer) assistance to perform tasks that require the above mentioned human cognition, knowledge or intelligence. PyBossa was inspired by the BOSSA crowdsourcing engine but is has been implemented in Python language.

Concerning the background of a related application area; there is a long-term and active research activity at MTA SZTAKI / DSD in the field of plagiarism search, some results are already provided as service to the public:

1) Monolingual plagiarism search

The KOPI Online Plagiarism Search Portal [26] is an open service that enables the users to check for identical or similar contents between their own documents and the files uploaded by other authors. This check function works in any European language, due to the language-independent algorithm.

2) Check for translated plagiarism

KOPI is is able to identify quotations - and also their translations - taken from the English Wikipedia. KOPI can run searches in Hungarian, English and German (and according to the plans it is being continuously improved to support other languages as well).

3) Measuring the amount of citations

KOPI is able to measure the total amount of citations in a document in case the source is available; the overlap between two documents; and also the originality of a written work in a set of documents.

4) Find the original author

With its constantly growing database KOPI is able to find the original author of a work, or search for works which cite from a given document/author.

The KOPI users and the related user scenarios can be classified in 3 categories:

1) Higher education

For teachers KOPI gives the opportunity to compare theses and home assignments to all documents uploaded before, or to the KOPI database and to the full content of Wikipedia.

Students are also able to check their own written documents to see e.g. if the amount of citations has exceeded the limit set by their home institution. They can also protect their theses by uploading them under their names.

2) Research and development

Authors can upload their written materials to the KOPI Portal so they are free to be published anywhere. KOPI protects copyright by showing who used the document and the author can see if the other party cited his work accordingly or not. Reviewers can use KOPI to compare the submitted work with other works of the same author, with the sources cited, so they are able to screen out improper citations and self-plagiarism cases.

KOPI helps scientific conference organizers make the submitted papers as unique as possible; screen out repeated submissions; and search for interesting papers and authors of the same or similar topics.

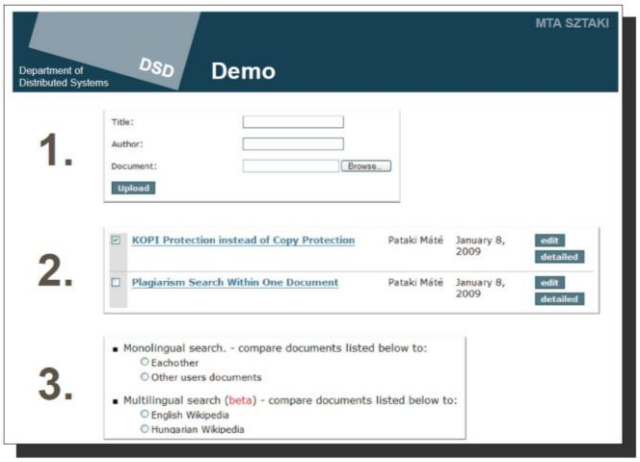


Figure 5 Screenshot from demo/presentation given by the representative of KOPI service at the 'Joint agINFRA and SCI-BUS workshop' supported by IDGF-SP (source: [27])

3) Companies, publishers

Contents of publications or corporate websites are often subject to plagiarism, so it is of great importance for press- and book-publishers, blog authors and for the online and offline media as a whole to protect their property from copyright infringement. They can easily do so by uploading their contents to the KOPI Portal, which will protect them against content theft by easily identifying the original author.

In general, the KOPI Plagiarism Search Portal enables all users to upload documents and start a plagiarism search using these documents. Uploaded written works are comparable with each other, and with the full database of the system. The users can check for plagiarism not only in tens of thousands of documents uploaded by the registered users, but can also compare documents to several digital data collections and the Wikipedia containing millions of entries and articles.

KOPI uses the SZTAKI Desktop GRID to pre-process the immense database of Wikipedia. In this way, the KOPI users can help improve the KOPI service by running the

Desktop Grid client on their machine. More details can be found in [26].

Joint research between IDGF-SP and SOCIENTIZE to involve more citizen scientists

The detected level of similarity (given in percentage) by the KOPI service sometimes leaves the final decision to the above listed users; whether the checked document has been the result of any kind of plagiarism or not. In order to enhance the KOPI service and provide more inputs for the KOPI related research activities at the same time, some joint activities have been initiated and partly negotiated between IDGF-SP and SOCIENTIZE projects. The collaboration is to cover the following topics:

1) IDGF-SP (KOPI) modifies the terms of use of the KOPI service that will optionally allow the KOPI service to send some pieces of the uploaded text to third-party contributors (to volunteers from SOCIENTIZE) if the KOPI user explicitly requests this new enhanced service.
2) IDGF-SP collects pieces of texts from the KOPI users (who explicitly requested this service) where the automatically detected level of plagiarism indicates that more consideration/attention would be required (i.e. the plagiarism cannot be detected with enough high certainty).
3) IDGF-SP provides these pieces of texts to the SOCIENTIZE project (in the given protocol and format, accepted by pyBossa) that must be compared and checked by the volunteers.
4) SOCIENTIZE project provides and operates the above described pyBossa framework as the required platform, and sends back the results from the volunteers to the KOPI service.
5) SOCIENTIZE and IDGF-SP advertise jointly the experiment to attract volunteers.
6) IDGF-SP contributes to the green paper of the SOCIENTIZE project.

The planned collaboration is to be formalized in a Memorandum of Understanding between the two projects.

According to the draft version of the green paper issued by the SOCIENTIZE project: "Citizen Science refers to the general public engagement in scientific research activities when citizens actively contribute to science either with their intellectual effort or surrounding knowledge or with their tools and resources." As the result of the above outlined joint activity KOPI with its enhanced service will form a special citizen science experiment where the volunteers (citizen scientist) might be classified in two ways:

Beneficiaries

- directly through the provided plagiarism search service (created by the researchers)
- indirectly through the better education of the citizens' children since the teachers will be able to use the enhanced KOPI service in their everyday work more efficiently

Active contributors

- with uploaded documents that serve input for plagiarism search,
- with their spare computer capacities to do preprocessing of all WIKIPEDIA articles time-to-time by the IDGF supported SZTAKI Desktop Grid
- with their human intellectual efforts to check documents where the plagiarism cannot be detected automatically with enough high certainty

Moreover, the entire experiment serves a solid base for multidisciplinary research activities in the fields of computer- and social sciences, and also linguistics. Some results have been already published in scientific journals and more publications are to be expected after the successful experiment.

VI. CONLUSION

Crowd computing has been gaining more and more attention. There are several benefits of the model; the scientists are able to harness the power of the general public, and the citizens will get more information and insight view about the science. As the paper illustrated, the number of application areas have been increasing, and IDGF is close to reach the level and the critical mass to be a fully integrated and vital part of the e-Infrastructure related ecosystem.

ACKNOWLEDGMENT

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no 261556 (EDGI) and under grant agreement no 312297 (IDGF-SP).

REFERENCES

[1] IDGF, http://desktopgridfederation.org
[2] IDGF-SP project, http://idgf-sp.eu
[3] D. Anderson, BOINC: A system for public resource computing and storage. Proceedings of the 5th IEEE/ACM International GRID Workshop, pp. 1 7, 2004.
[4] Z. Farkas, P. Kacsuk, Z. Balaton, and G. Gombas, Interoperability of BOINC and EGEE, Future Generation Computer Systems, vol. 26, no. 8, pp. 1092 1103, 2010. [Online]. Available: http://www.sciencedirect.com/science/article/B6V06-5046M26-8/2/b33ce4a0dc54204e34cdff1c90c577eb
[5] gLite, http://glite.web.cern.ch/glite
[6] Attila Marosi, József Kovács, Peter Kacsuk, Towards a volunteer cloud system, Future Generation Computer Systems, Volume 29, Issue 6, Pages: 1442-1451, 2013, ISSN 0167-739X, http://www.sciencedirect.com/science/article/pii/S0167739X12000660, 10.1016/j.future.2012.03.013
[7] Public Autodock portal, https://autodock-portal.sztaki.hu
[8] EDGI repository, http://edgi-repo.cpc.wmin.ac.uk/repository
[9] EGI Communities, http://www.egi.eu/community
[10] EGI-Inspire project, https://www.egi.eu/about/egi-inspire
[11] EGI Desktop Grid VT, http://go.egi.eu/vtdg
[12] EGI-DG integration, https://wiki.egi.eu/wiki/Desktop_Grids_integration
[13] EDGI project, http://edgi-project.eu
[14] IDGF Support Unit, https://wiki.egi.eu/wiki/GGUS:IDGF_FAQ

[15] GGUS, https://ggus.eu/pages/home.php
[16] EGI Accounting Portal, http://accounting.egi.eu/egi.php
[17] Apel, https://wiki.egi.eu/wiki/APEL
[18] EGI resource pool, https://wiki.egi.eu/wiki/Resource_Pool
[19] WeNMR community, https://www.wenmr.eu
[20] DIRAC community, http://diracgrid.org
[21] BIOVEL community, http://www.biovel.eu
[22] Riemann zeta search project, http://www.riemann-siegel.com
[23] The FP7 SOCIENTIZE project, http://www.socientize.eu
[24] SOCIENTIZE with PyBossa, http://pybossa.socientize.eu/pybossa/app/Semantics
[25] PyBossa, http://pybossa.com
[26] Máté Pataki, Attila Marosi: „Searching for Translated Plagiarism with the Help of Desktop Grids", Journal of Grid Computing, 11(1), pp. 149-166, ISSN 1570-7873, 10.1007/s10723-012-9224-5, http://dx.doi.org/10.1007/s10723-012-9224-5
[27] Joint agINFRA and SCI-BUS workshop, http://www.lpds.sztaki.hu/aginfra-osd

Obtaining Massive Data Sets for Contextual Experiments in Quantum Information

Mladen Pavičić
University of Zagreb, Faculty of Civil Engineering, Physics Chair, Zagreb, Croatia

***Abstract*—Finding even the simplest Kochen-Specker sets for contextual experiments—a hot subject in the field of quantum information theory, theoretically and experimentally—by brute force would take all the clusters on Earth more than the Age of Universe. Human ingenuity found about a dozen such sets in the last 40 years. I will present how our group reduced this finding of contextual needles in a haystack to a manageable task, by first visualising them via hypergraphs, human-derived from the space symmetries, and then running jobs that generate the sets—billions of them—in parallel.**

I. Introduction

A measurement of any observable of a classical system should yield the same value independently of whether we carry out the measurement of the observable simultaneously with another measurement of another observable of the same system or not. In other words, classical theories do not depend on arrangements in which measurements are carried out, i.e., on their "context," and we say that classical theories are *non-contextual* and that all their observables can be ascribed predetermined values.

In contrast, a measurement of an observable of a quantum system can yield different values depending on other measurements of another observables that we measure simultaneously with the considered observable. So, a quantum theory does depend on arrangements in which measurements are carried out and we say that quantum theories are *contextual* and that their observables cannot be ascribed predetermined values.

Not all experimental arrangements for a quantum system are contextual, though. Actually, until a few years ago only about a dozen such sets of quantum states to which no predetermined values can be ascribed were known. The sets are called Kochen-Specker (KS) sets and they are needed not only for a better understanding of quantum theories and measurements [1]–[3], but also for constructing of quantum gates in the field of quantum information and quantum computation and for handling quantum sets in general [4], [5]. So, it is of interest to know how often we can encounter such sets, i.e., how many KS sets we can generate.

II. Formalism

Let us first define KS sets more precisely.

Definition 1. Every KS set is a set of vectors in a Hilbert space $\mathcal{H}^n$, $n \geq 3$ to which it is impossible to assign 1's and 0's in such a way that:

1. No two orthogonal vectors are both assigned the value 1;

2. Not all of any mutually orthogonal vectors are assigned the value 0.

In this presentation we shall limit ourselves to $n = 4$, i.e., a 4-dim Hilbert space. Systems can be either a pair of qubits (qubit is a 2-dim quantum system) or a spin-3/2 particles. Sets of mutually orthogonal vectors in a 4-dim space we call tetrads. A KS set is a union of such tetrads.

Mutual orthogonality of 4 vectors in a 4-dim Hilbert space is represented by the following 6 nonlinear equations:

$$\begin{aligned}
\mathbf{a}_A \cdot \mathbf{a}_B &= a_{A1}a_{B1} + a_{A2}a_{B2} + a_{A3}a_{B3} + a_{A4}a_{B4} = 0,\\
\mathbf{a}_A \cdot \mathbf{a}_C &= a_{A1}a_{C1} + a_{A2}a_{C2} + a_{A3}a_{C3} + a_{A4}a_{C4} = 0,\\
\mathbf{a}_A \cdot \mathbf{a}_D &= a_{A1}a_{D1} + a_{A2}a_{D2} + a_{A3}a_{D3} + a_{A4}a_{D4} = 0,\\
\mathbf{a}_B \cdot \mathbf{a}_C &= a_{B1}a_{C1} + a_{B2}a_{C2} + a_{B3}a_{C3} + a_{B4}a_{C4} = 0,\\
\mathbf{a}_B \cdot \mathbf{a}_D &= a_{B1}a_{D1} + a_{B2}a_{D2} + a_{B3}a_{D3} + a_{B4}a_{D4} = 0,\\
\mathbf{a}_C \cdot \mathbf{a}_D &= a_{C1}a_{D1} + a_{C2}a_{D2} + a_{C3}a_{D3} + a_{C4}a_{D4} = 0.
\end{aligned}$$

Now, it might seem that the problem can be approached by a brute computational force. We ascribe values 0 and 1 to various sets of connected tetrads of vectors and as soon as we find a set for which we cannot do that it is a KS set. However, this is easier said than done because there are billions of such sets even if we limit ourselves only to unit components along each of four axes for all vectors. Besides, the equations are nonlinear and it is a hard problem with only a few equations, let alone billions.

Fortunately, in 2000 I realised that these equations can be reduced to a generation and then filtering of hypergraphs, in particular McKay-Megill-Pavičić (MMP) hypergraphs, which Brendan D. McKay, Norman D. Megill, and I defined previously for another purpose.

Definition 2. We define MMP hypergraphs as follows [6]

(i) Every vertex belongs to at least one edge;

(ii) Every edge contains at least 3 vertices;

(iii) Edges that intersect each other in $n-2$ vertices contain at least n vertices.

This definition enables us to formulate algorithms for exhaustive generation of MMP hypergraphs. In this work we shall work with subsets of a 60-75 (60 vertices - 75 edges) master set obtained by P. K. Aravind who derived it from another even bigger set using geometric symmetries [7]. From this master set we generate smaller hypergraphs that correspond to subsets of the 60-75 set with a specified number of edges deleted.

For any experimental application it is not viable to consider all KS (sub)sets but only those that can be experimentally

distinguished. Hence, we extract *critical* non-redundant non-isomorphic KS sets. "Critical" means that they are minimal in the sense that no orthogonal tetrads can be removed without causing the KS contradiction to disappear.

The "only" difficulty we face is the sheer size of these generated subsets—we are dealing with a haystack of 2^{75} or 38 sextillion subsets, in which we wish to find certain "needles," i.e., critical KS sets. In Sec. III we will present the algorithms and programs which enable us to deal with the haystack.

The hypergraphs we obtain reflect only the orthogonal structure of KS sets and do not in any way refer to the vector components of the original 60-75 KS set. This is yet another aspect in which the present method differs from the parity-proof method we used in [8], which relies on the vector components of the vectors in each KS set that were inherited from the original 60-75 set. For each hypergraph we can, however, find appropriate vector components with our program `vectorfind` or by interval analysis we developed in [6].

We encode MMP hypergraphs by means of alphanumeric and other printable ASCII characters. Each vertex is represented by one of the following characters: 1 2 3 4 5 6 7 8 9 A B C D E F G H I J K L M N O P Q R S T U V W X Y Z a b c d e f g h i j k l m n o p q r s t u v w x y z ! " # $ % & ' ( ) * - / : ; < = > ? @ [\ ] ^ _ ` { | } ~ , and then again all these characters prefixed by '+', then prefixed by '++', etc.

Each edge is represented by a string of characters that represent vertices within a single line. Edges are separated by commas. The line must end with a full stop. Skipping of characters is allowed. A line forms a representation of a hypergraph. The order of the edges is irrelevant. The numbers of vertices and edges are unlimited. We often present MMP hypergraphs starting with edges forming the biggest loop to facilitate their possible drawing.

In Fig. 1 we show a graphical representation of the minimal (26-13) critical KS set we found and which we shall now use to show a correspondence between the vector and the MMP hypergraph representation of any KS set.

Each vertex represents a vector in a 4-dim space. For instance, G= $\{1,0,0,0\}$, F= $\{0,1,0,0\}$, E= $\{0,0,1,0\}$, D= $\{0,0,0,1\}$. They are mutually orthogonal and that means they form an edge—GFED. Our program `vectorfind` can assign all vectors, that correspond to edges from the 26-13 set, component values from the set $\{0,\pm1,\pm(\sqrt{5}+1)/2,\pm(\sqrt{5}-1)/2\}$ and that means that the system of equations that define all orthogonalities for the 26-13 does have a solution. Now our program `states01` (which exhaustively verifies all possible assignments) checks whether all the vertices can be ascribed 0 and 1 according to the KS rules 1 and 2 above and verifies that it is not possible. The main point here is that we can always go from MMP hypergraphs to vectors and back and that `states01` works with MMP hypergraphs. MMP hypergraphs are linear while the system of equations describing mutual orthogonality of vectors are nonlinear. Therefore the evaluation of MMP hypergraphs by means of `states01` is exponentially faster than solving nonlinear equations and this is what makes our generation of KS sets feasible. While the algorithm used in `states01` is comparatively fast, the verification of KS sets for MMP hypergraphs with an odd number of edges can be even faster using our "parity proofs" algorithm described in [8].

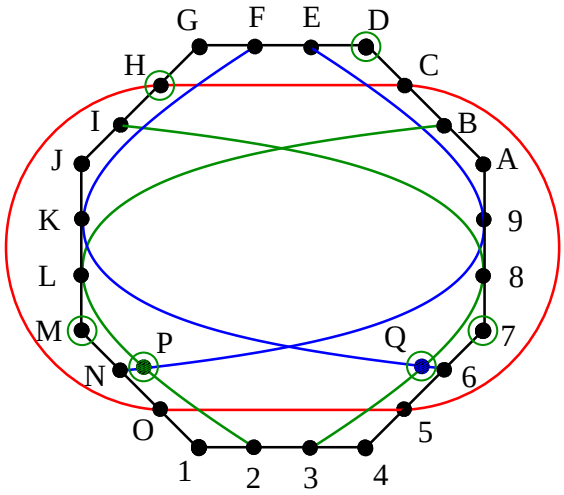


Fig. 1. The smallest KS set 26-13 (with 26 vertices and 13 edges). Possible assignments of 1s are indicated by circles (7,D,H,M,P,Q)—the remaining vertices can only be assigned 0s. Hence, all vertices in edge 1234 can only be assigned 0s and Def. 1 is satisfied. MMP encoding of the shown 26-13 reads: 1234,4567,789A,ABCD,DEFG,GHIJ,JKLM,MNO1,O5CH,2PLB,NP9E,6FQK,3Q8I..

We stress here that once we visualised, i.e., drew, MMP hypergraphs, we can drop the ASCII signs and compare and classify them graphically referring only to, e.g., the biggest loops they allow, the highest and lowest numbers of edges vertices allow, whether they allow parity proofs, the highest number of 1s vertices can be assigned, etc., all of which properties have their own role in experimental implementation of critical KS sets. Of course, we can always go back to vector representation so as to assign arbitrary ASCII signs to vertices and then run `vectorfind` on them.

III. Results

To generate critical KS sets we applied an algorithm and a program (`mmpstrip`) which strip edges from the master 60-75 MMP, then subsequently from the so obtained MMPs, and so on. The number of MMPs we can obtain in this way grows exponentially and in less then 10 steps it would make busy all CPUs and would overflow all storage capacities on the Globe.

Fortunately the greatest portion of the generated MMPs are isomorphic to each other and by using B. D. McKay's algorithm and program `shortd` which he derived from his program `nauty`, we are able to greatly reduce the number of MMPs we have to generate to eventually obtain the required critical KS sets. However, even their number is far too big for an exhaustive generation and even clusters with hundreds of CPUs cannot make more than 13 steps in an acceptable time.

Therefore the only option left is a probabilistic generation combined with 0-1 and isomorphic filtering. That excludes parallel processing on a cluster because the programs for stripping, isomorphic reduction, and 0-1 filtering are all probabilistic so far as the sequential steps are concerned (although,

of course, the end result is deterministic). Thus the times of parallel running greatly differ and it is much more efficient to put a large number of jobs in queue and rerun those that come out first. The details are as follows.

As shown in the flowchart 1, we first have to generate—by stripping edges—a sufficient number of MMPs, in several steps, to compensate for an inherent homogeneous distribution of critical MMPs in the generation "down-tree." After each step, i.e., with a sufficient number of MMPs within each step, we reiterate the procedure until we reach about " 10^{10} MMPs. Then we carry out a "random reduction" by means of `states01` which strips random edges from the input MMPs until they are critical.

brute force	10^{25} nonlinear equations	Age of the Universe
	⇓ bulk generation ⇓	
bare MMP isomorphic generation	10^{10} MMP hypergraphs	1 month cluster (100 CPUs)
	⇓ random reduction ⇓	
random stripping	10^{8} MMPs	$<$ 1 day cluster (100 CPUs)
	⇓ selective generation ⇓	
$0-1$ and isomorphic filtering	10^{4} MMPs	$<$ 1 hour cluster (100 CPUs)
	⇓	
isomorphic filtering	one MMP ($26-13$ KS set)	$<$ 20 min (1 CPU)

(1)

The majority of so obtained criticals are isomorphic and mixed. Therefore we have to rerun the random reduction and selective generation many times and each time we have to merge new MMPs with the already obtained ones so as to build a cumulative set of all possible KS. However, not even then can we be sure that we obtained samples of all kinds of MMPs. For example, after 3500 generated 60-41 critical MMPs (critical KS sets) among more than 1.5 billion critical KS sets we still have not obtained a single one with a parity proof although there is at least one because we obtained it via a parity algorithm from [8]. On the other hand a single 47-30 appeared only after 1 billion of generated KS criticals and it cannot be generated via the parity algorithm since it has an even number of edges. We present statistics of obtained critical KS sets in Fig. 2. Detailed presentation of all obtained results will appear in a subsequent publication.

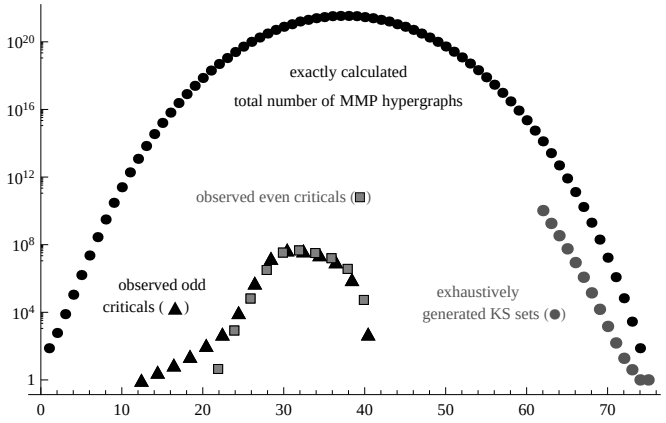


Fig. 2. Statistics calculated for subsets of 60-75 given on a logarithmic scale. There are more than 10^9 critical KS sets. Given numbers of critical KS sets with 13 to 27 edges (on the x-axis) are exhaustive. The number of criticals with 32 edges is the biggest; we estimate that they do not exceed 10^{10}. Given numbers of noncritical KS sets with more than 61 edges are also exhaustive.

Details of the programs `mmpstrip`, `states01`, `vectorfind`, and `shortd` and the algorithms behind them are given in [6]–[9].

IV. Conclusion

We have obtained a massive number of contextual sets, more specifically critical Kochen-Specker (KS) sets, by first translating a condition imposed on quantum states—orthogonality—into a condition imposed on correlated MMP hypergraphs and then handling the latter hypergraphs to generate the sets. In the standard Hilbert space the generation lead to huge sets of nonlinear equations and therefore to an exponentially complex problem. Generation of MMP hypergraphs is however a statistically polynomial problem and is therefore feasible, although lengthy as shown in the previous section. "Statistically polynomial" here means that although the algorithms behind the programs `mmpstrip` and `states01` might in general take an exponentially increasing time with larger sets, this on average does not happen because of the inner structure of the KS sets.

Obtaining large KS sets is important for better understanding of how quantum system behave and a better insight into their engineering within experiments. The most significant result we obtained is that efficiency of contextual measurements does not drop with large KS sets. Only the number of required measurement rises but the efficiency of each of them can be the same for some of the largest sets as for the smallest one. This is in contrast with most other quantum gates where the efficiency drops when the number of states increases, due to coherence problems. On the other hand, some of our previous conjectures based on a smaller number of sets [9] should be amended. For instance, the peak for the KS criticals turn out to be much lower, the distribution of them of a different shape, and the KS criticals with more than 41 edges nonexistent.

Also a very important contribution of a computational approach to contextual sets and similar problems is contained in our development of algorithms that enable solving huge systems of nonlinear equations via handling MMP hypergraphs on clusters. This is important not only because we already arrived at new results with new KS sets in 3-, 4-, and 6-dim Hilbert spaces, but also because it opens a new approach to solving systems of equations in an analogous way in other quantum and classical problems—whenever we can translate a condition imposed on a system of equations into a condition imposed on hypergraphs, we should be able to solve the system via handling and filtering the hypergraphs.

ACKNOWLEDGMENT

Computational support was provided by the cluster Isabella of the Zagreb University Computing Centre and by the Croatian National Grid Infrastructure. My special thanks go to Emir Imamagić for his generous help with problems I encountered while carrying out calculations on these clusters.

REFERENCES

[1] V. D'Ambrosio, I. Herbauts, E. Amselem, E. Nagali, M. Bourennane, F. Sciarrino, and A. Cabello, Experimental Implementation of a Kochen-Specker Set of Quantum Tests, *Phys. Rev. X* **3**, 011012–1–10 (2013).

[2] H. Bartosik, J. Klep, C. Schmitzer, S. Sponar, A. Cabello, H. Rauch, and Y. Hasegawa, Experimental Test of Quantum Contextuality in Neutron Interferometry, *Phys. Rev. Lett.* **103**, 040403–1–4 (2009).

[3] G. Kirchmair, F. Zähringer, R. Gerritsma, M. Kleinmann, O. Gühne, A. Cabello, R. Blatt, and C. F. Roos, State-Independent Experimental Test of Quantum Contextuality, *Nature* **460**, 494–497 (2009).

[4] M. Pavičić, *Quantum Computation and Quantum Communication: Theory and Experiments*, Springer, New York, 2005.

[5] M. Pavičić, *Companion to Quantum Computation and Communication*, Wiley-VCH, Weinheim, 2013.

[6] M. Pavičić, J.-P. Merlet, B. D. McKay, and N. D. Megill, Kochen-Specker Vectors, *J. Phys. A* **38**, 1577–1592 (2005), and **38**, 3709 (2005) (corrigendum).

[7] M. Pavičić, N. D. Megill, P. K. Aravind, and M. Waegell, New Class of 4-Dim Kochen-Specker Sets, *J. Math. Phys.* **52**, 022104–1–9 (2011).

[8] M. Waegell, P. K. Aravind, N. D. Megill, and M. Pavičić, Parity Proofs of the Bell-Kochen-Specker Theorem Based on the 600-cell, *Found. Phys.* **41**, 883–904 (2010).

[9] N. D. Megill, K. Fresl, M. Waegell, P. K. Aravind, and M. Pavičić, Probabilistic Generation of Quantum Contextual Sets, *Phys. Lett. A* **375**, 3419–3424 (2011).

Parallel Implicit Time Integration for Particle-based Models on Graphics Clusters

Adrian Sabou, Dorian Gorgan
Computer Science Department
Technical University of Cluj-Napoca
Str. G. Baritiu 28, 400027, Cluj-Napoca, Romania
Email: {adrian.sabou, dorian.gorgan}@cs.utcluj.ro

***Abstract*—Particle-based models are widely spread in the field of computer graphics, and mainly used for real-time simulations of soft deformable bodies. One of the high-demanding computational components of such physically-based simulations is numerical time integration. As the size of the models increases, this component becomes a potential bottleneck in the simulation process, thus parallelism must be deployed in order to preserve the real-time attribute. This paper presents an approach to parallel implicit numeric time integration for graphics clusters. To deal with very large models, we employ a static domain decomposition approach and we fully utilize the massively data parallel capabilities of GPU clusters. We discuss the parallel conjugate gradient algorithm used for time integration and we present a parallel method for generating the very large sparse matrices and vectors involved.**

I. Introduction

Modeling and simulating three-dimensional dynamic surfaces represents one of the main research areas of Computer Graphics and probably one of the most fascinating. The most prominent techniques for such simulations are physically-based methods, which have become an important part of the entertainment industry and are regularly employed in computer games, feature films and various computer based animations, for visual effects involving interacting fluids, textile materials, deformable and rigid bodies. However, taking into account that most such simulations need to be interactive and considering the computational expensive numerical methods employed, coupled with the very large sized models that the ever growing strive for realism demands, efficient parallel techniques must be developed in order to fully utilize modern High-Performance Computing (HPC) architectures such as the Graphical Processing Unit (GPU) and GPU clusters for real-time particle-based simulations.

One of the most important subdomain of physical-based modeling is particle-based modeling, where surfaces are approximated through sets of discrete points having various physical properties such as mass, volume, speed, acceleration. The behavior of such surfaces along with their interaction with the environment is governed by the laws of physics, more specifically by the forces that act upon particles. Particle-based modeling is a natural choice since in the real world interactions concerning deformable surfaces occur at a molecular level and, in theory, given a sufficient number of particles, any such surface can be accurately modeled. The particle-based model for simulating cloth was firstly introduced by Provot [1] and is still utilized in simulations where interactivity is a key requirement.

When used to simulate rigid cloth, such models generate systems of stiff differential equations with respect to time. In order to solve such systems one must employ numerical time integration. Implicit integration has the advantage of being unconditionally stable for large time steps, but can quickly become a major bottleneck as the size of the models increases, mostly due to operations involving the large sparse matrices of these systems. In computer graphics, the popular Conjugate Gradient (CG) algorithm is used to solve these differential equations, allowing balancing accuracy against simulation speed.

In our previous work towards particle-based simulation of 3D dynamic surfaces on graphics clusters we developed basic parallelization strategies for the time integration stage [2]. We proposed a data parallel approach for the velocity Verlet explicit scheme [3] and, by using a hybrid CPU/GPU parallelization technique and a sort-last approach to centralized visualization, we obtained interactive frame rates for models exceeding one million particles. This paper expands our previous work and presents a GPU cluster oriented approach to parallel implicit numerical integration for particle-based models. It is based on a parallel Conjugate Gradient algorithm implemented in the Vienna Computing Library (ViennaCL) [4], a high level linear algebra library for GPUs and Multi-Core CPUs. We propose a parallel method for updating the large sparse matrices that serve as input to the CG algorithm, tailored specifically to the particularities of the model used for simulating cloth and the Compressed Sparse Row format used by most linear algebra libraries. The rest of the paper is organized as follows: Section II highlights important related work. Section III presents an overview of the particle-based model used for simulating cloth. Section IV describes our approach for parallel implicit integration on GPU cluster. We report on performance measurements in section V.

II. Related Works

There are numerous approaches to physically-based simulation and the ways in which internal forces are computed differ considerably. However, all of them lead to a set of ordinary differential equations with respect to time, which have to be solved through numerical time integration. The equations in question usually describe the dynamic evolution of the simulated system as an initial value problem. Numerical solutions for such problems is an intensively studied field, and, for a detailed and comprehensive overview, we refer the interested reader to the standard textbooks by Hairer et al. [5]

[6] [7]. An analysis of time integration schemes in the context of pysically-based modeling in computer graphics can be found in the works by Hauth et al. [8] and Volino et al. [9].

Implicit integration methods have caught the attention of the computer graphics community after Baraff and Witkin [10] demonstrated their performance and stability for complex physically-based simulation scenarios.They used the semi-implicit Euler scheme and solved the arising linear system with a modified conjugate gradients method. This approach was adopted in numerous subsequent work and is still widespread today, although later research identified considerable numerical damping as an inherent drawback.

Among the first to implement Conjugate Gradient solvers on GPU were Bolz et al. [11] and Krüger and Westermann [12]. Their early implementations were based on fragment shaders and had to deal with the lack of scatter operations, limited floating-point precision and slow texture switching.

Keckeisen and Blochinger [13] published work focusing on parallelizing the numerical integration phase of cloth simulations using commodity PC clusters. Their stated goals were to develop an efficient parallel system for irregular cloth meshes, to achieve good scalability on relevant problem sizes, and to utilize a cost effective computing platform. Thomaszewski and Blochinger [14] continued on this direction by improving the physically-based simulation of cloth through parallelization of the collision phase as well as the numerical integration. Parallelization of the implicit numerical time integration phase was done at sparse matrix-vector multiplication level and on preconditioning, since implicit methods assume a number of sequential iterations in order to converge to a solution.In their subsequent work, Thomaszewski et al. [15] also explored parallelization on shared memory architectures. For implicit time integration they used static problem decomposition and solved the linear system arising in this context using a data-parallel preconditioned conjugate gradient algorithm.

The work of Verschoor and Jalba [16] is the most related to the research presented in this paper, in terms of the proposed solution for updating the large sparse matrices that the Conjugate Gradient algorithm uses. While their solution closely resembles ours, they use it in the context of elastically deformable models based on the Finite Element Method (FEM) and restrict themselves to experimenting on a single GPU equipped node. We also employ a *scatter* approach (i.e. all values of a local matrix are added to their corresponding values in the global matrix) instead of the *gather* strategy, due to the particularities of the particle-based model used for cloth that allow us to avoid any race conditions.

III. Particle-Based Simulation for Cloth

The parallelization techniques described in this paper apply to any specific approach for particle-based simulations, providing it uses implicit time integration. However, for our initial performance measurements we chose a simple particle-based model for cloth, the mass-spring model.

A. Particle-based Model

Mass-spring systems represent the simplest and most direct way of modeling deformable objects and from this simplicity comes their power. The mass-spring model most frequently used for simulating cloth and fabric was introduced by Xavier Provot [1] and is based only on simple, linear springs, thus avoiding complex particle interactions. Provot's model assumes that the mass of the cloth is concentrated in discrete points rather than being distributed along the surface, thus the realism of the model is conditioned by the density of particles. The model contains point masses and three different types of springs arranged in a rectangular grid, as illustrated in Figure 1:

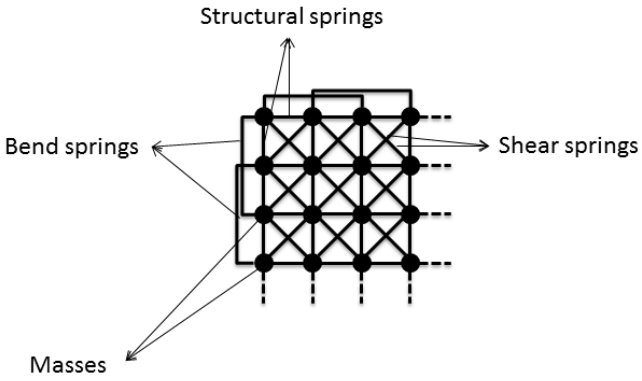


Fig. 1. Mass-Spring Mesh Model for Simulating Cloth

- Springs linking neighbor vertices horizontally and vertically are called "stretch" or "structural" springs;
- Springs linking neighbor vertices obliquely are called "shear" springs;
- Springs linking second degree neighbor vertices horizontally and vertically are called "bend" springs.

As their name indicates, each type of spring resists to stretching, shearing and bending respectively. Setting spring constants of these spring types independently enables us to mimic different types of cloth and rubbery objects.

The interaction potential between particles is given by internal and external forces acting on them. Internal forces are the elastic forces arising in springs and damping forces. To compute the elastic force between particle i and particle j we use Hooke's law of elasticity given by the formula:

$$\mathbf{F}_{i,j}^{el} = k \cdot \mathbf{x}_{i,j} \cdot (1 - \frac{l}{|\mathbf{x}_{i,j}|}) \tag{1}$$

where $\mathbf{x}_i$ and $\mathbf{x}_j$ are the positions of particle i and j respectively and $\mathbf{x}_{i,j} = \mathbf{x}_i - \mathbf{x}_j$, k is the spring's stiffness constant and l is the spring's rest length. The internal force acting on particle i due to elastic forces is computed by summing the forces between particle i and all its neighbors.

When considering the case of friction with air, complex drag effects may appear on real cloth surfaces. However, when considering the simulated case, if is often desirable to linearize the damping forces. House and Breen [17] defined the linearized version of mass-spring damping as related to the velocity of particles:

$$\mathbf{F}_{\mathbf{i,j}}^{\mathbf{damp}} = -\gamma \mathbf{v}_{i,j} \tag{2}$$

where γ is the damping factor and $\mathbf{v}_{i,j} = \mathbf{v}_i - \mathbf{v}_j$.

External forces can differ in nature, depending on the simulation. Forces most frequently included are gravity, wind and collision response.

In order to derive the behavior of such a system from its interaction potential, we must employ numeric time integration.

B. Numerical time integration

There are numerous approaches to particle-based simulation and the ways in which internal forces are computed differ considerably. However, all of them lead to a set of ordinary differential equations with respect to time, which have to be solved through numerical time integration. The equations in question usually describe the dynamic evolution of the simulated system as an initial value problem. The behavior of mass-springs systems is governed by Newton's second law of dynamics:

$$\mathbf{F} = m\ddot{\mathbf{x}} \tag{3}$$

where m is the mass and $\ddot{\mathbf{x}}$ is the second order derivative of displacement, namely the acceleration.

Implicit integration methods have caught the attention of the computer graphics community after Baraff and Witkin [10] demonstrated their performance and stability for complex physically-based simulation scenarios. They used the semi-implicit Euler scheme and solved the arising linear system with a modified conjugate gradients method. This approach was adopted in numerous subsequent work and is still widespread today, although later research identified considerable numerical damping as an inherent drawback. Unlike explicit integration, which compute the state of the next time step out of a direct extrapolation of previous steps, using derivative evaluations, implicit methods deduce the state of the next time step from an equation system expressing some "reversibility" from the extrapolated solution.

For our work we choose an implicit Euler integration approach, which uses the following discretization to compute the change in positions and velocities:

$$\begin{cases} \Delta\mathbf{x} = (\mathbf{v}_t + \Delta\mathbf{v}) \cdot \Delta t \\ \Delta\mathbf{v} = \mathbf{M}^{-1} \cdot \mathbf{F}(\mathbf{x}_t + \Delta\mathbf{x}, \mathbf{v}_t + \Delta\mathbf{v}) \cdot \Delta t \end{cases} \tag{4}$$

where $\Delta\mathbf{x}$ and $\Delta\mathbf{v}$ are the position and velocity increment respectively, $\mathbf{x}_t$ and $\mathbf{v}_t$ are the current position and velocity, $\mathbf{M}$ is a diagonal matrix of size $3n \times 3n$, with n being the number of particles in the model, containing 3×3 submatrices with each particle's mass on the main diagonal, $\mathbf{F}$ is the force acting upon a particle and Δt is the time interval over which we integrate.

In order to linearize this system of equations, $\mathbf{F}(\mathbf{x}_t + \Delta\mathbf{x}, \mathbf{v}_t + \Delta\mathbf{v})$ is approximated in the following way:

$$\mathbf{F}(\mathbf{x}_t + \Delta\mathbf{x}, \mathbf{v}_t + \Delta\mathbf{v}) = \mathbf{F}(\mathbf{x}_t, \mathbf{v}_t) + \frac{\partial \mathbf{F}}{\partial \mathbf{x}} \cdot \Delta\mathbf{x} + \frac{\partial \mathbf{F}}{\partial \mathbf{v}} \cdot \Delta\mathbf{v} \tag{5}$$

where $\frac{\partial \mathbf{F}}{\partial \mathbf{x}}$ and $\frac{\partial \mathbf{F}}{\partial \mathbf{v}}$ are Jacobians of $\mathbf{F}$. When substituting this, solving for $\Delta\mathbf{v}$, we obtain:

$$\begin{aligned} &\left(\mathbf{I} - \mathbf{M}^{-1} \cdot \frac{\partial \mathbf{F}}{\partial \mathbf{v}} \cdot \Delta t - \mathbf{M}^{-1} \cdot \frac{\partial \mathbf{F}}{\partial \mathbf{x}} \cdot \Delta t^2\right) \cdot \Delta\mathbf{v} = \\ &= \mathbf{M}^{-1} \cdot \left(\mathbf{F}(\mathbf{x}_t, \mathbf{v}_t) + \frac{\partial \mathbf{F}}{\partial \mathbf{x}} \cdot \mathbf{v}_t \cdot \Delta t\right) \cdot \Delta t \end{aligned} \tag{6}$$

with the unknown $\Delta\mathbf{v}$. The numerical solutions of large systems of linear equations of the form $\mathbf{A}\mathbf{x} = \mathbf{b}$, where $\mathbf{x}$ is an unknown vector, $\mathbf{b}$ is a known vector and $\mathbf{A}$ is a known, square ($n \times n$), symmetric (i.e. $\mathbf{A}^T = \mathbf{A}$), positive-definite (i.e. $\mathbf{x}^T\mathbf{A}\mathbf{x} > 0, \forall \mathbf{x} \in \mathbb{R}^n, \mathbf{x} \neq 0$) (or positive-indefinite) matrix, are obtained in real-time using the conjugate (CG) gradient algorithm [18].

C. Parallel Simulation of Cloth

Our parallelization approach for the particle-based model for cloth is similar to the one we employed in [2]. In order to provide context, we shall briefly describe the main techniques used in order to efficiently simulate cloth on graphics clusters.

For parallel simulation of cloth we proposed a static data decomposition technique which splits the computation into smaller equally weighted computation processes, with good data locality, between the GPU equipped nodes of a graphics cluster. This means partitioning the model of the cloth into groups of the same size and statically assign each group to a GPU which processes each particle in a different thread. The positions of neighboring particles that belong to different GPUs has to be communicated, therefore, in order to minimize the communication overhead, it is essential to split up the model in a way that ensures minimal length borders. To ensure minimal modifications to the kernels used for the single node version of the application, each node keeps an extended model, consisting of its own patch and neighboring particles that belong to other nodes, so called ghost-points, whose positions have to be synchronized at each step of the simulation.

We proposed a hybrid CPU/GPU parallel approach that allowed us to fully utilize the CPU for synchronization operations while the GPU performed the computational part, by modifying the sequence of steps in the simulation algorithm in a way as to allow these two steps to be out of phase. For centralized distributed visualization, we used a sort-last approach to parallel rendering, with optimized partial frame merging.

IV. Parallel Implicit Time Integration

In this section we will discuss in detail our approach for parallel implicit time integration on a GPU cluster, applied for simulating cloth. First we shall give details about the parallel CG algorithm and libraries used. Then, we will describe how the global left-side matrix of equation 6 (matrix A) is computed, taking into account the particularities of the particle-based model, and we will present our approach for efficiently updating it at every simulation step.

A. Parallel Conjugate Gradient

For a parallel implementation of the CG algorithm, we rely on the ViennaCL [4] library, which aims at providing standard data types for linear algebra operations on GPUs and multi-core CPUs. This open source library can be easily integrated into existing C/C++ applications and thus reduce code changes to a minimum. ViennaCL provides CUDA, OpenCL and OpenMP computing backends, enabling simple, high-level access to the vast computing resources available on parallel architectures such as GPUs. Moreover, ViennaCL operations that use the OpenCL backend can run in user-provided contexts and wrap existing memory with its own data types, making it even easier to integrate with our OpenCL-based application and allowing us to avoid unnecessary memory transfers.

The main bottleneck of the CG algorithm is represented by repetitive matrix-vector multiplications performed in order to minimize a residual and converge to a solution. The parallel CG in ViennaCL is based on a parallel version of sparse matrix-vector multiplication (SPMV). The large sparse A matrix from equation 6 is split into blocks and multiplication is carried out in parallel on all blocks using the GPU. The format used to store matrix A is Compressed Sparse Row (CSR), a storage format optimized for arithmetic operations, row slicing, and matrix-vector products. In order to accelerate convergence, we chose to use the CG algorithm with a simple Jacobi diagonal preconditioner, also supplied by the ViennaCL library. The main challenge encountered was the fact that matrix A changes at each simulation step, thus an efficient method for updating it was required in order to be able to use implicit integration for very large models.

B. Parallel Global Matrix Update

The idea behind efficiently updating the very large sparse A matrix is to employ data parallelism in order to compute and to place local entries into their corresponding location in global memory. For that, we must exploit the particular structure of the Jacobians from equations 5 and 6. They are $(3n \times 3n)$ matrices consisting of (3×3) submatrices. Elements (i, j) of the Jacobian for position and for velocity are defined as:

$$\left(\frac{\partial \mathbf{F}}{\partial \mathbf{x}}\right)_{i,j} = \frac{\partial \mathbf{F}_i}{\partial \mathbf{x}_j}, \left(\frac{\partial \mathbf{F}}{\partial \mathbf{v}}\right)_{i,j} = \frac{\partial \mathbf{F}_i}{\partial \mathbf{v}_j} \tag{7}$$

We can easily observe that these derivatives are non-zero only if the force acting on particle i depends on the position or the velocity of particle j, i.e. particle i and particle j are connected by a spring. Due to particles being connected as exemplified in figure 1, this results in a high degree of sparsity for the two matrices. Taking into account the physical formulation of forces from equations 1 and 2, we define the spring Jacobians for position and velocity as [19]:

$$\mathbf{J}\mathbf{x}_{i,j} = \left(\frac{\mathbf{x}_{i,j} \cdot \mathbf{x}_{i,j}^T}{\mathbf{x}_{i,j}^T \cdot \mathbf{x}_{i,j}} + \left(\mathbf{I} - \frac{\mathbf{x}_{i,j} \cdot \mathbf{x}_{i,j}^T}{\mathbf{x}_{i,j}^T \cdot \mathbf{x}_{i,j}} \cdot \left(1 - \frac{l}{|\mathbf{x}_{i,j}|}\right)\right)\right) \tag{8}$$

and

$$\mathbf{J}\mathbf{v}_{i,j} = \mathbf{I} \cdot \gamma \tag{9}$$

where $\mathbf{x}_{i,j}^T \cdot \mathbf{x}_{i,j}$ denotes the inner product and $\mathbf{x}_{i,j} \cdot \mathbf{x}_{i,j}^T$ denotes the (3×3) matrix which is the outer product. The spring Jacobians are added to the global Jacobians at positions (i, j) and (j, i), and subtracted at positions (i, i) and (j, j).

Using equations 8 and 9 we can now compute individual entries in the global Jacobians. Since $\mathbf{M}^{-1}$ is a diagonal matrix, multiplying the Jacobians with $\mathbf{M}^{-1}$ becomes trivial, thus the computation of each element of A becomes independent, except for elements on the main diagonal, where race conditions may occur. This encourages a scatter approach to updating the contents of matrix A, i.e. compute local elements and put them in their corresponding place in the global structure.

This leaves us with the problem of parallelizing the scatter operation. As we previously mentioned, the elements of the main diagonal must be processed sequentially in order to avoid race conditions, or to use atomic operations on the GPU which are far too slow and should be avoided.

The Compressed Sparse Row (CSR) format stores an initial matrix using three one-dimensional arrays, one containing the non-zero elements, one containing the index of the first nonzero element of each row and another containing column indexes of all non-zero elements. Considering that the structure of the model does not change during simulation, then we can observe that the second and third array of the CSR representation for matrix A do not change. This means that these structures can be precomputed in an initial step before the simulation begins, thus the only problem remaining is to parallelize the computation and insertion of values in the first array at their corresponding positions.

Figure 2 shows the structure of a Jacobian for a (4×4) grid of particles. Each non-zero element of row i (represented in color) is a (3×3) matrix that describes how particle i is affected by neighboring particles. Considering this, if we were to assign each such entry to a different GPU thread, then each thread would have to perform complex conditional logic in order to determine its position in the globally ordered list of neighbors (and thus the corresponding position to update in the CSR array), logic that GPU processors are ill suited for and that would cause a drastic decrease in performance. It is therefore preferred to assign each full row of the Jacobians

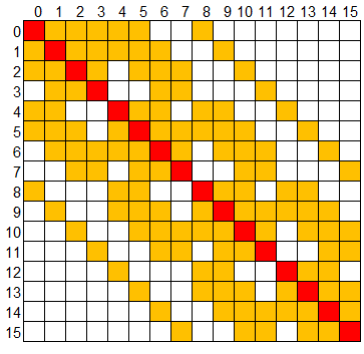


Fig. 2. Jacobian structure for a (4×4) grid of particles

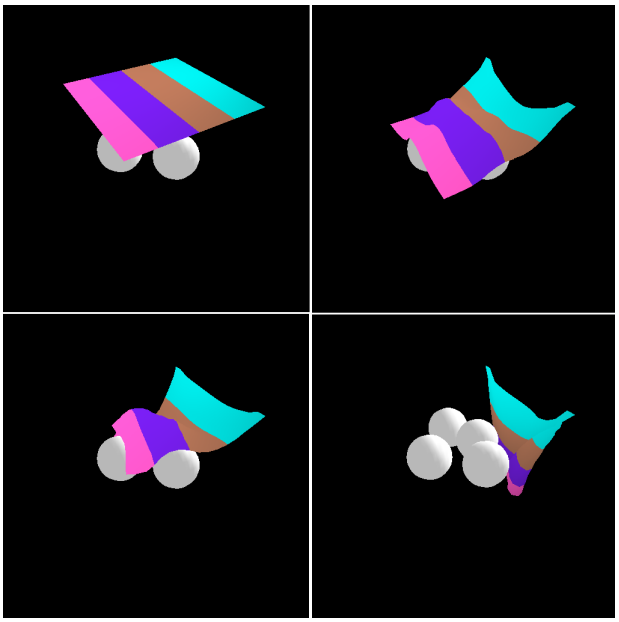


Fig. 3. Experimental scenario - Cloth colliding with spheres

to a different thread. In this way, each thread covers one particle and all its neighbors, and conditional logic is avoided by sequentially iterating through the entire globally ordered neighbor list. Elements on the main diagonal are computed and updated sequentially in a second step.

V. Experiments

The experiments were carried out on a graphics cluster consisting of four nodes with NVIDIA GeForce GTS 250 and NVIDIA GeForce 8800 GTS GPUs and Intel Core 2 Duo E8400 CPUs running on Debian 6.0.5 Squeeze, connected to a Gigabit network. The scenarios devised were aimed at evaluating the performance of the proposed solution taking into account factors such as the resolution of the model (i.e. number of particles) and the number of nodes used for the simulation. As criteria for measuring overall performance we chose the number of Frames per Second (FPS) which was considered relevant for a graphical application and a measure commonly used for evaluating parallel computation, the speedup. Each FPS sampling was performed after double the number of frames of the previously computed frame rate, with an initial threshold of 90 frames. To account for performance variations due to the load on the cluster, the final computed frame rate was the average of 10 successive samples over five consecutive runs. For evaluating the parallel global matrix update approach, we obtained the average computing time for both the sequential approach and the parallel one over a total of 500 frames.

The test scenario considered for all experiments was the simulation of a piece of cloth colliding with other objects in the scene, in our case four spheres as shown in figure 3. Each colored patch was handled by a different cluster node.

A. Parallel Global Matrix Update

The aim of this experiment was to evaluate the increase in performance obtained by using the parallel global matrix update approach when compared to a sequential one. Table I shows the average measured processing time for a complete matrix update using both methods, for model sizes starting at 256 (16×16) up to 65536 (256×256) particles. Due to the fact that the comparative measurements were taken on a single GPU equipped node, model resolutions of 262144 (512×512) or higher could not be simulated due to GPU memory limitations.

Nb. of particles	Sequential time (s)	Parallel time (s)
256	0.193	0.000038
1024	0.914	0.000038
4096	4.57	0.000038
16384	26	0.000038
65536	164	0.000038

TABLE I. Global matrix average update time

As can be observed, the sequential processing time and the parallel processing time are of different orders of magnitude mostly due to the high degree of parallelism of our model. Moreover, for model resolutions that can be simulated on a single GPU, the time for parallel updating the global matrix does not depend on the number of particles. This is due to the fact that all GPUs used in our experiments have the capability to launch jobs of work sizes that are greater than the size of our particle grid. This means that the number of parallel threads that the GPU launches is always equal to the number of particles, thus all particles are processed in parallel. One important drawback when using a sequential approach is the fact that computation for updating the matrix is carried in the CPU and the CG algorithm runs in GPU, thus there is a continuous transfer of data between host and device, which greatly lowers performance. The GPU parallel approach allows us to share the device memory between the updating procedure and the CG algorithm.

B. Parallel Implicit Integration

The aim of this experiment was to evaluate the overall performance improvement of the proposed parallel global matrix update technique coupled with the parallel conjugate gradient algorithm provided by the ViennaCL library when running the simulation on a graphics cluster with up to five

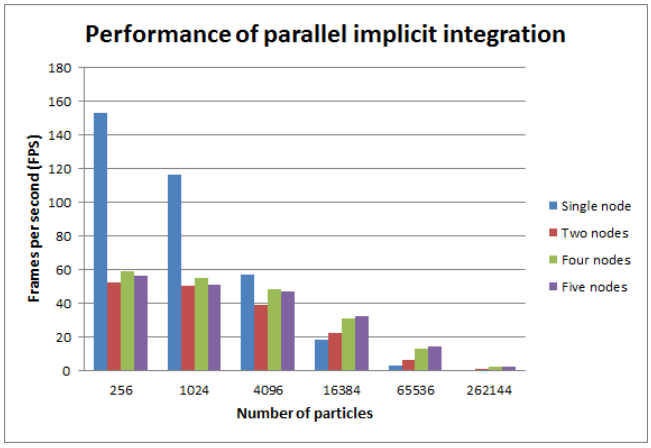


Fig. 4. Performance of parallel implicit integration

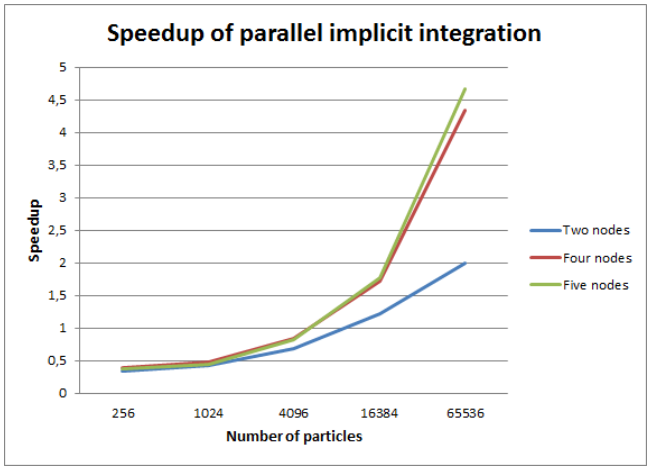


Fig. 5. Speedup

GPU equipped nodes. Figure 4 shows the frame rates obtained using two, four and five nodes, compared to the performance obtained on a single node.

As can be observed, the single node simulation achieves better performance for small model sizes (256 - 4096 particles) due to the fact that synchronizing ghost points and transmitting data for centralized visualization (when using more nodes) takes considerably more time than the numerical integration and the collision detection phases. However, as the resolution for the cloth model increases past 16384 particles, the computation becomes overwhelming for a single node, thus the advantage of model distribution becomes obvious. Model sizes were pushed up to 262144 particles, where all interactivity is lost (i.e. frame rates drop below 10 FPS) and the simulation on a single node can no longer be executed due to memory limitations.

As shown in figure 5, the solution offers good scalability in terms of speedup obtained when running the simulation of five cluster nodes. As more nodes are used, maximum speedup is obtained for a larger number of particles, and a method for determining the optimal number of nodes for a certain model size may also be derived from empirical data, providing that there are a sufficient number of nodes available in the graphics cluster used for testing.

VI. Conclusion

In this paper we presented a graphics cluster-based approach to implicit time integration for particle-based simulations. We focused on an efficient update technique of the large sparse matrix in Compressed Sparse Row (CSR) format that the Conjugate Gradient (CG) algorithm uses, taking into account the particularities of the particle-based model used for simulating cloth. We validated and evaluated the parallel efficiency of our approach through experiments on a GPU cluster with five processing nodes coupling our integration approach with parallelization techniques developed in previous work.

Because scalability is encouraging, we will further explore particle-based modeling on graphics clusters using more processing nodes and employing more complex collision detection techniques.

Acknowledgment

This paper is supported by the Sectoral Operational Programme Human Resources Development (SOP HRD), ID134378 financed from the European Social Fund and by the Romanian Government.

References

[1] X. Provot, "Deformation constraints in a mass-spring model to describe rigid cloth behavior," in *IN GRAPHICS INTERFACE*, 1995, pp. 147–154.

[2] A. Sabou and D. Gorgan, "Physical simulation of 3d dynamical surfaces on graphics clusters," in *Information Communication Technology Electronics Microelectronics (MIPRO), 2013 36th International Convention on*, 2013, pp. 292–297.

[3] L. Verlet, "Computer "experiments" on classical fluids. I. Thermodynamical properties of lennard-jones molecules," *Phys. Rev.*, vol. 159, pp. 98–103, Jul 1967.

[4] K. Rupp, F. Rudolf, and J. Weinbub, "ViennaCL - A High Level Linear Algebra Library for GPUs and Multi-Core CPUs," in *Intl. Workshop on GPUs and Scientific Applications*, 2010, pp. 51–56.

[5] E. Hairer, S. P. Nørsett, and G. Wanner, *Solving Ordinary Differential Equations I: Nonstiff Problems (Springer Series in Computational Mathematics) (v. 1)*, 2nd ed. Springer, Dec. 2009. [Online]. Available: http://www.worldcat.org/isbn/3540566708

[6] E. Hairer, S. P. Norsett, and G. Wanner, *Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems*, 2nd ed., Springer, Ed. Berlin: Springer, 1991.

[7] E. Hairer, C. Lubich, and G. Wanner, *Geometric Numerical Integration.: Structure-Preserving Algorithms for Ordinary Differential Equations.*, ser. Springer series in computational mathematics. Springer-Verlag GmbH, 2002.

[8] M. Hauth, O. Etzmuß, and W. Straßer, "Analysis of numerical methods for the simulation of deformable models," *The Visual Computer*, vol. 19, no. 7-8, pp. 581–600, 2003.

[9] P. Volino and N. Magnenat-Thalmann, "Comparing efficiency of integration methods for cloth simulation," in *Proceedings of the International Conference on Computer Graphics*, ser. CGI '01. Washington, DC, USA: IEEE Computer Society, 2001, pp. 265–. [Online]. Available: http://dl.acm.org/citation.cfm?id=872749.873303

[10] D. Baraff and A. Witkin, "Large steps in cloth simulation," in *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, ser. SIGGRAPH '98. New York, NY, USA: ACM, 1998, pp. 43–54.

[11] J. Bolz, I. Farmer, E. Grinspun, and P. Schröoder, "Sparse matrix solvers on the gpu: Conjugate gradients and multigrid," *ACM Trans. Graph.*, vol. 22, no. 3, pp. 917–924, Jul. 2003. [Online]. Available: http://doi.acm.org/10.1145/882262.882364

[12] J. Krüger and R. Westermann, "Linear algebra operators for gpu implementation of numerical algorithms," *ACM Trans. Graph.*, vol. 22, no. 3, pp. 908–916, Jul. 2003.

[13] D. Bartz, B. Raffin, H. wei Shen (editors, M. Keckeisen, and W. Blochinger, "Parallel implicit integration for cloth animations on distributed memory architectures."

[14] B. Thomaszewski and W. Blochinger, "Physically based simulation of cloth on distributed memory architectures," *Parallel Comput.*, vol. 33, no. 6, pp. 377–390, Jun. 2007.

[15] B. Thomaszewski, S. Pabst, and W. Blochinger, "Parallel techniques for physically based simulation on multi-core processor architectures," *Computers & Graphics*, vol. 32, no. 1, pp. 25 – 40, 2008.

[16] M. Verschoor and A. C. Jalba, "Elastically deformable models based on the finite element method accelerated on graphics hardware using cuda." *Journal of WSCG*, vol. 20, no. 3, pp. 179–188, 2012.

[17] D. H. House and D. E. Breen, Eds., *Cloth modeling and animation*. Natick, MA, USA: A. K. Peters, Ltd., 2000.

[18] J. R. Shewchuk, "An introduction to the conjugate gradient method without the agonizing pain," Pittsburgh, PA, USA, Tech. Rep., 1994.

[19] K.-J. Choi and H.-S. Ko, "Stable but responsive cloth," *ACM Trans. Graph.*, vol. 21, no. 3, pp. 604–611, Jul. 2002.

Frequency Table Computation on Dataflow Architecture

P. Škoda*, V. Sruk**, and B. Medved Rogina*
* Ruđer Bošković Institute, Zagreb, Croatia
** Faculty of Electrical Engineering and Computing, University of Zagreb, Zagreb, Croatia
pskoda@irb.hr

Abstract – Frequency table computation is a key step in decision tree learning algorithms. In this paper we present a novel implementation targeted for dataflow architecture implemented on field programmable gate array (FPGA). Consistent with dataflow model of computation, the kernel views input dataset as synchronous streams of attributes and class values. The kernel was benchmarked using key functions from C4.5 program for decision tree learning. For large datasets with many attributes – over 100,000 items, and over 60 attributes – the realized kernel, clocked at 333 MHz, outperforms the software implementation running on CPU clocked at 3.2 GHz.

Keywords – frequency table, dataflow, field programmable gate arrays, decision tree learning

I. INTRODUCTION

Histogram computation is one of the fundamental procedures used in many data mining algorithms. Its two-dimensional extension – frequency table computation – is a key part of the attribute selection step in decision tree learning algorithms [1]. Decision tree learning algorithms, as well as all other tools used in data mining, are faced with ever-growing size of analyzed datasets. This drives efforts to parallelize existing algorithms, develop new algorithms targeted for parallel and distributed computer systems, and develop new hardware platforms that provide high computational power [2].

One platform that is now gaining more ground in computing applications is field programmable gate array (FPGA). FPGAs are digital integrated circuits designed to be user-configured after manufacturing [3], [4]. An FPGA comprises an array of configurable logic blocks, memory blocks, configurable interconnect, input/output (I/O) blocks, and various specialized arithmetic and communication blocks. These elements together allow implementation of custom computational architectures that can be reconfigured on demand. The computational model that FPGA is most suitable for is the dataflow model.

Majority of computer systems implement a control-flow model of computation. In control-flow, the computation is defined by a sequence of operations that are executed in specified order. In dataflow model, computation is defined by a chain of transformations applied to a stream of data [5]. The dataflow is described spatially in form of dataflow graph. Each operation is represented by a node, and data flow between nodes is represented by edges. Dataflow graph naturally maps to hardware implementation. Operations are translated into hardware blocks that work concurrently, and edges are translated into data links. By dividing the graph into stages separated by registers, it is transformed into a pipelined structure suitable for implementing on FPGAs.

In this paper we present a novel implementation of frequency table computation targeted for dataflow architecture realized on FPGA. The kernel is developed for integration in C4.5 decision tree learning program [6], [7]. Frequency table computation is a key step in decision tree learning algorithms. Its result is used to calculate dataset split quality, usually Gini impurity or an entropy based measure such as information gain, which is then used for attribute test selection [1].

II. RELATED WORK

Frequency table computation on FPGA has been previously published as a part of efforts in implementing decision tree learning algorithms. FPGAs have been proven to be suitable of implementing data mining algorithms, though so far little attention has been directed to decision tree learning algorithms. Only a few attempts of using FPGAs to accelerate execution of decision tree learning have been reported [8], [9].

Chrysos et al. [9] implement frequency table computation on FPGA as part of HC-CART system for learning decision trees. Frequency table computation is implemented in Frequency Counting module. The module receives attribute-class label pairs. The pairs are used to generate addresses for frequency table locations which contents are then incremented by one. The contents of the frequency table are then copied over to other two auxiliary memories, and used to generate item frequencies for computing Gini impurity. In their implementation, all attribute-class labels received by the kernel have to be prepared by program running on CPU, and then preloaded to memory attached to FPGA. Input data is transferred from CPU to FPGA memory many times in the course of program execution. In our work, the data is transferred to FPGA memory (LMem) only once, at the beginning of the program. The kernel reads the data only from FPGA memory, and the only data transfer between kernel and CPU is sending the computed frequency tables to the CPU.

TABLE I. BASIC INFORMATION ON MAXELER VECTIS-LITE FPGA BOARD

FPGA	Xilinx Virtex-6 XCVSX475T
Off-chip RAM (LMem)	6 GiB (6×1 GiB) DDR2-800 SDRAM
On-chip RAM (FMem)	~ 4 MiB Block RAM
CPU ↔ FPGA bandwidth	2 GB/s
LMem ↔ FPGA bandwidth	38.4 GB/s max.

III. DATAFLOW ENGINE ARCHITECTURE

The Data Flow Engine (DFE) is programmed with one or more kernels and a manager. Kernel implements computation. It defines the computational architecture with scalar and stream inputs and outputs. Manager organizes data movement within the DFE. In managers, the kernels are instantiated and links to off-chip RAM (LMem), CPU or other DFEs are defined.

A. *Maxeler Vectis-Lite FPGA platform*

Maxeler platform is designed for implementing compute engines using the dataflow model of computation. The platform comprises an FPGA board, drivers, development tools, and API. Dataflow engine is described in Java, using API which provides objects and methods that can be compiled into hardware structures. The compilers translate the input Java code into VHDL code and create functions that implement interface for CPU code [10].

Maxeler Vectis-Lite board was used for realization of the DFE on FPGA. The board is a PCIe expansion board which houses an FPGA, and SDRAM. Basic information on the board is shown in Table I.

B. *Kernel architecture*

Architecture of the ComputeFreq kernel is shown in

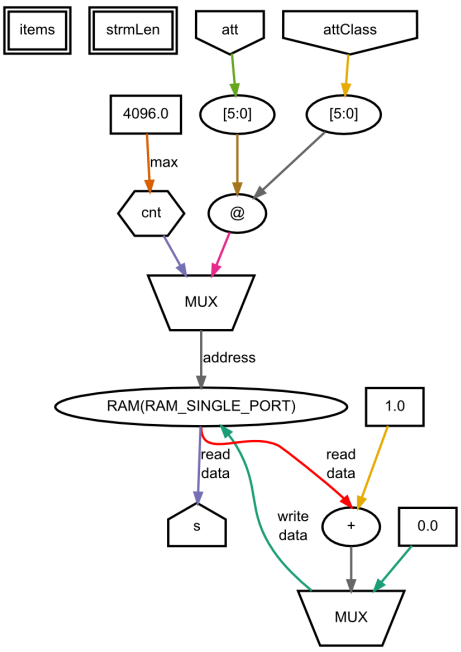


Figure 1. ComputeFreq kernel architecture for up to 64 unique attribute and class values ($N_A = 6$, $N_C = 6$).

TABLE II. FPGA RESOURCE USAGE BY THE DFE

Resource	Used	Total available	Utilization
LUTs	30053	297600	10.10 %
Flip-flops	43416	297600	14.59 %
DSP blocks	0	2016	0.00 %
Block RAM	105	2128	4.93 %

Fig. 1. The kernel has two scalar inputs for parameters: stream length (*strmLen*), and number of items (*items*). Stream length is the number of elements read from LMem, which requires all memory accesses to be made in 96 byte blocks. Number of items is the actual number of items of interest in the stream.

The kernel receives two 32 bit unsigned integer streams as input – one for attribute values (*att*) and the other for class values (*attClass*) of the items. The inputs are sliced so that the low N_A bits are taken from attribute and N_C bits from class value. The sliced inputs are concatenated to form the frequency table index, with.

Frequency table is stored in block RAM (BRAM). The BRAM is addressed by the frequency table index formed from attribute and class values. The value output from the BRAM is incremented by one and written back to the same location. The BRAM is configured as a single port RAM with read-first synchronization. The compiler tool automatically inserts registers in appropriate locations, which leads to latency inside the loop. To accommodate this latency, kernel's input streams are throttled to allow input of one element every m clocks, where m is the length of loop latency.

After all elements of interest are streamed in, contents of the BRAM are streamed out to the CPU through stream output *s* and at the same time reset to zero. This is achieved by using multiplexers to switch BRAM address port from table index to values generated by a counter. In the same manner the BRAM write data port is switched from incremented-by-one value to constant zero. Since there are no loop dependencies in this instance, the output stream is not throttled. One element is output every single clock. At the same time, any remaining elements in the input streams are read un-throttled (one per clock) and ignored. They are all read to ensure proper functioning of the DFE.

C. *Manager*

A single ComputeFreq kernel is instantiated in manager. Stream inputs *att* and *attClass* are linked to LMem, and use linear memory access pattern. Stream output *s* is linked to CPU. Scalar inputs *items* and *srtmLen* are also linked to CPU. The kernel parameters were set for up to 64 unique attribute and class values ($N_A = 6$, $N_C = 6$). The manager clock frequency was set at 100 MHz, while kernel clock frequency was set at 333 MHz.

DFE interface defines inputs for source addresses of data for input streams *att* and *attClass*, values for scalar inputs *items* and *srtmLen*, and destination address for output stream *s*.

With given parameters, frequency table stores 4096 32-bit words. Loop latency of the synthesized design is 5

clocks, which means that the DFE can process up to 66.6×10^6 items per second. FPGA resource usage by the DFE is given in Table II.

IV. Experimental Results

A. Benchmarks and test environment

Kernel was benchmarked by using code from C4.5 Release 8 decision tree learning program [11]. The normal tree building features from the original program were disabled. Only parts required for loading datasets and computing the frequency table were left operational. A function was added for transforming dataset into format appropriate for DFE and writing it to LMem. Time measurement was added to *ComputeFrequencies* function. For execution on DFE, the *ComputeFrequencies* function was modified to use function calls to the DFE, and copy the received result into the original frequency table data structure. From measured execution time, the function's throughput was computed by:

$$T_{a,i} = \frac{a\,i}{t_{a,i}} \tag{1}$$

where a is number of attributes, i is number of items, $t_{a,i}$ is measured execution time, and $T_{a,i}$ is throughput for dataset with a attributes and i items.

A set of 84 datasets was generated for the benchmark. All datasets have only nominal attributes with 63 unique attribute values (values from 1 to 63), and 64 unique class values (from 0 to 63). Number of items vary exponentially from 2×2^{10} to 4×2^{20} (2^{11} – 2^{22}) items, while number of attributes varies exponentially from 1 to 64 (2^0 – 2^6). Attribute and class values are generated randomly using uniform distribution.

Benchmark program was compiled using gcc compiler version 4.4.7. Benchmarks were run on workstation with Intel Xeon E5-1650 CPU running at 3.2 GHz, 16 GiB DDR3-1600 RAM, under Centos 6.5 Linux OS. The workstation is equipped with Maxeler Vectis-Lite FPGA acceleration board.

TABLE III. Measured execution times on CPU

Number of items	Number of attributes / Execution time						
	1	2	4	8	16	32	64
2048	39.62 µs	78.74 µs	156.1 µs	316.6 µs	651.2 µs	1.339 ms	2.859 ms
4096	48.97 µs	97.27 µs	194.7 µs	392.2 µs	841.7 µs	1.766 ms	3.807 ms
8192	68.39 µs	136.1 µs	273.6 µs	551.0 µs	1.202 ms	2.582 ms	6.065 ms
16384	104.5 µs	214.7 µs	428.3 µs	866.7 µs	1.945 ms	4.420 ms	10.94 ms
32768	180.1 µs	364.4 µs	735.0 µs	1.505 ms	3.556 ms	8.116 ms	25.95 ms
65536	341.7 µs	705.5 µs	1.372 ms	2.841 ms	6.663 ms	20.45 ms	108.1 ms
131072	654.0 µs	1.319 ms	2.632 ms	5.561 ms	16.97 ms	55.01 ms	235.7 ms
262144	1.427 ms	2.885 ms	5.858 ms	13.78 ms	36.63 ms	109.4 ms	473.5 ms
524288	3.056 ms	6.161 ms	13.11 ms	28.14 ms	72.70 ms	216.4 ms	947.4 ms
1048576	6.080 ms	12.19 ms	24.57 ms	55.80 ms	144.7 ms	428.8 ms	1.889 s
2097152	12.10 ms	24.49 ms	48.83 ms	110.1 ms	288.7 ms	855.6 ms	3.800 s
4194304	24.30 ms	48.82 ms	97.81 ms	220.5 ms	574.6 ms	1.707 s	7.575 s

B. ComputeFrequencies CPU benchmark results

Execution times on CPU are given in Table III, and are shown in Fig. 2. For input datasets larger than 64×2^{10} items, execution time scales approximately linearly with number of items. There is a sharper increase in execution time visible for datasets with 8 or more attributes, in ranges for 128×2^{10} – 256×2^{10} items for 8 attributes, to 32×2^{10} – 64×2^{10} items for 64 attributes. Function throughput, shown in Fig. 3, more visibly illustrates this decline in performance. Throughput rises with number of items until a peak is reached at 16×2^{10} to 128×2^{10} items, depending on number of attributes. After this point it declines and quickly stabilizes to a constant value. Constant throughput corresponds to linear scaling of execution time with number of items. The largest constant throughput is 173×10^6 items per second for dataset with 1 attribute, and the lowest is 35.4×10^6 items per second for dataset with 64 attributes.

This decline in performance is most likely an effect of

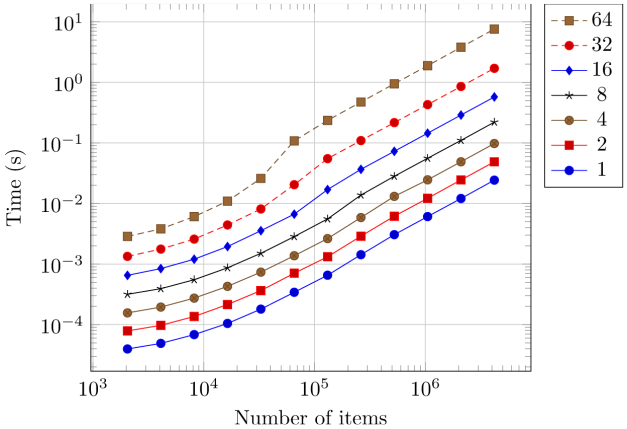


Figure 2. Execution time on CPU as function of number of items, measured for datasets with 1 to 64 attributes

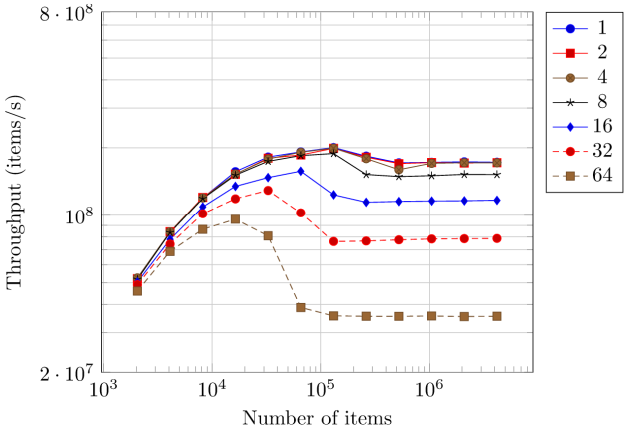


Figure 3. Throughput on CPU as function of number of items, measured for datasets with 1 to 64 attributes

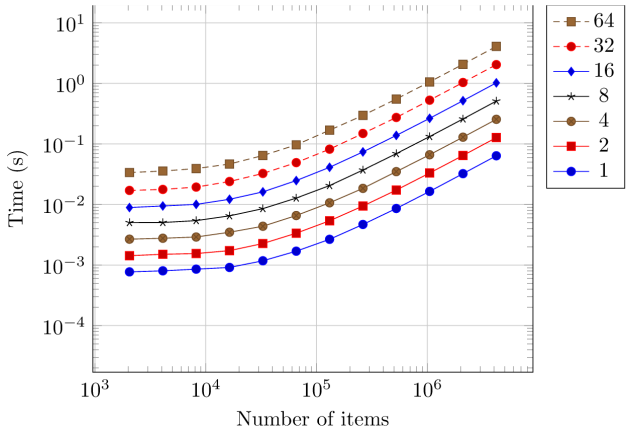


Figure 4. Execution time on DFE as function of number of items, measured for datasets with 1 to 64 attributes

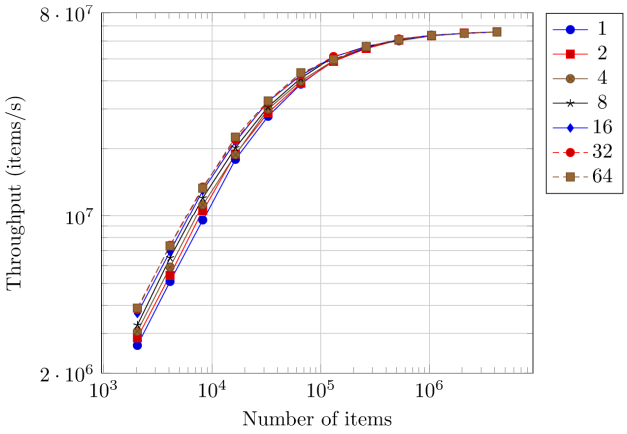


Figure 5. Throughput on DFE as function of number of items, measured for datasets with 1 to 64 attributes

CPU's cache system. Smaller sets are less affected by cache misses than larger sets. In addition, the dataset is stored in attribute major order. Since *ComputeFrequencies* function works by accessing the same attribute over different items, this leads to consecutive accesses to non-consecutive memory locations, which are more likely to result in cache misses.

C. *ComputeFrequencies DFE benchmark results*

Measured execution times on DFE are given in Table IV, and shown in Fig 4. Execution time scales linearly for datasets with more than 128×2^{10} items. While scaling is not linear for datasets with fewer items, there is no performance drop like the one present when executing on CPU. From throughput results, shown in Fig 5, it is visible that there is little difference in throughput between datasets with different number of attributes, and that the main factor is number of items. With 256×2^{10} and more items, throughput is fairly constant at 62.4×10^6 items per second, which is close to the theoretical maximum of 66.6×10^6 items per second. Lower performance with smaller datasets is a consequence of overheads of control and communication with the DFE. This overhead is independent from number of items in the dataset, but scales linearly with number of attributes, since the same set of function calls has to be made to process each attribute.

TABLE IV. MEASURED EXECUTION TIMES ON DFE

Number of items	Number of attributes / Execution time						
	1	2	4	8	16	32	64
2048	771,8 µs	1,428 ms	2,675 ms	5,017 ms	8,904 ms	17,06 ms	33,70 ms
4096	802,7 µs	1,507 ms	2,771 ms	5,041 ms	9,456 ms	17,74 ms	35,69 ms
8192	855,6 µs	1,549 ms	2,904 ms	5,414 ms	10,04 ms	19,37 ms	39,20 ms
16384	912,8 µs	1,730 ms	3,481 ms	6,489 ms	12,18 ms	23,96 ms	46,55 ms
32768	1,176 ms	2,266 ms	4,372 ms	8,542 ms	16,16 ms	32,56 ms	64,36 ms
65536	1,695 ms	3,356 ms	6,573 ms	12,76 ms	24,80 ms	49,13 ms	96,60 ms
131072	2,661 ms	5,376 ms	10,66 ms	20,53 ms	41,12 ms	81,87 ms	168,6 ms
262144	4,680 ms	9,435 ms	18,47 ms	37,05 ms	73,81 ms	148,9 ms	295,7 ms
524288	8,598 ms	17,29 ms	34,88 ms	69,08 ms	138,3 ms	273,5 ms	551,4 ms
1048576	16,45 ms	33,10 ms	66,27 ms	132,3 ms	264,6 ms	527,5 ms	1,056 s
2097152	32,30 ms	64,39 ms	129,2 ms	258,2 ms	517,5 ms	1,034 s	2,067 s
4194304	63,75 ms	127,2 ms	254,9 ms	510,3 ms	1,020 s	2,040 s	4,082 s

D. *Comparison of the results*

To compare the results, ratio of execution time on CPU and time on DFE was calculated. The ratio is shown in Fig 6. Execution on DFE is in most cases significantly slower than on CPU. This is expected since the DFE runs on 333 MHz and works with DDR2 SDRAM clocked at 333 MHz, while CPU runs on 3.2 GHz and works with DDR3 SDRAM clocked at 800 MHz.

For small datasets DFE is at worst 20× slower than CPU. However, DFE's performance comes closer to the CPU's as number of attributes and items increases. For largest datasets (over 1×2^{20} items), DFE is at worst 2.6× slower than CPU. For datasets with 64 attributes, DFE outperforms CPU when number of items is larger than 64×2^{10}. Best performance – 1.80× faster than CPU – is reached when number of items is larger than 256×2^{10}.

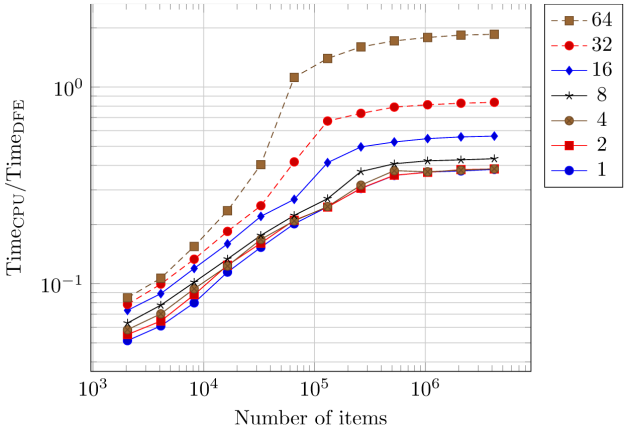


Figure 6. Ratio of execution times on CPU and DFE as function of number of items, measured for datasets with 1 to 64 attributes

V. CONLUSION

In this paper we have presented a novel implementation of frequency table computation on dataflow architecture. The architecture was realized on FPGA platform. From benchmark results it is clear that a single kernel DFE is outperformed by the CPU in most circumstances, but it does provide better efficiency when their clock rates are accounted for. The CPU runs at 3.2 GHz. Its best performance is 173×10^6 items per second which gives 18.5 clocks per item. The DFE runs at 333 MHz and its best performance is at 62.4×10^6 items per second which gives 5.34 clocks per item. Of the 5.34 clocks per item, 4 are consequence of the BRAM-adder-multiplexer-BRAM loop latency.

There are ways to design around the in-loop latency. One approach is by replicating the BRAM-adder-multiplexer structure as many time as needed (in this case it's 5 times), and "demultiplex" the input stream to each structure. Once the input streaming is complete, the final result is computed by summing the partial results from each BRAM. Such design could run with 1.34 clocks per items which would give 249×10^6 items per second. This exceeds maximum performance achieved on CPU.

In future work, a multi-kernel version of DFE will be developed which will allow simultaneous processing of several attributes. The kernel will be modified to lower the in-loop latency. This DFE is planned to be integrated with the C4.5 program as accelerator unit for processing nominal attributes. Given presented results, it is expected that the FPGA design that exploits parallelism will outperform conventional CPU.

ACKNOWLEDGMENT

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7), under grant agreement no. 317882 – E2LP Embedded Computer Engineering Learning Platform.

REFERENCES

[1] L. Rokach and O. Maimon, *Data Mining with Decision Trees: Theroy and Applications*. River Edge, NJ, USA: World Scientific Publishing, 2008.

[2] A. N. Choudhary, D. Honbo, P. Kumar, B. Ozisikyilmaz, S. Misra, and G. Memik, "Accelerating data mining workloads: current approaches and future challenges in system architecture design," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 1, no. 1, pp. 41–54, Jan. 2011.

[3] I. Kuon, R. Tessier, and J. Rose, "FPGA Architecture," *Found. Trends Electron. Des. Autom.*, vol. 2, no. 2, pp. 153–253, 2008.

[4] Mark L. Chang, "Device Architecture," in *Reconfigurable Computing: The Theory and Practice of FPGA-based Computation*, S. Hauck and A. DeHon, Eds. Morgan Kaufmann, 2008, pp. 3–27.

[5] J. B. Dennis, "Data Flow Supercomputers," *Computer*, vol. 13, no. 11, pp. 48–56, Nov. 1980.

[6] J. R. Quinlan, *C4.5: Programs for Machine Learning*. San Mateo: Morgan Kaufmann, 1993, p. 302.

[7] S. Ruggieri, "Efficient C4.5," *IEEE Trans. Knowl. Data Eng.*, vol. 14, no. 2, pp. 438–444, 2002.

[8] P. Škoda, B. Medved Rogina, and V. Sruk, "FPGA implementations of data mining algorithms," in *MIPRO, 2012 Proceedings of the 35th International Convention*, 2012, pp. 362–367.

[9] G. Chrysos, P. Dagritzikos, I. Papaefstathiou, and A. Dollas, "HC-CART: A parallel system implementation of data mining classification and regression tree (CART) algorithm on a multi-FPGA system," *ACM Trans. Archit. Code Optim.*, vol. 9, no. 4, pp. 1–25, Jan. 2013.

[10] O. Pell and V. Averbukh, "Maximum Performance Computing with Dataflow Engines," *Comput. Sci. Eng.*, vol. 14, no. 4, pp. 98–103, Jul. 2012.

[11] J. R. Quinlan, "C4.5 Release 8," 1993. [Online]. Available: http://rulequest.com/Personal/c4.5r8.tar.gz. [Accessed: 23-Jun-2013].

Virtue - a Different Approach To Human/Computer Interaction

Zorislav Šojat, Tomislav Ćosić, Karolj Skala*
* RBI/CIC, Zagreb, Croatia
sojat@irb.hr, tcosic@irb.hr, skala@irb.hr

Throughout the development of computing tools usage and computer science, from the calculating machines up and into the present day computers, the major technical problem to be solved was the reduction of the amount of components and their bulk. This led inevitably to the use of serial processing. Only quite recently we started mass producing and using multi-processing. Though the nature itself is infinitely parallelised, the technical problems of computer development led also to the development of such software tools and programming languages which mirrored the serial nature of computers, so that the serialisation of parallel natural processes is performed by humans, the programmers. However, modern day developments of scientific and every-day needs for computing power have led to the introduction of multiprocessors, GPUs, clusters, grids and clouds of computers, as to ascertain enough processing speed, power and memory for very complex algorithms. However, the human/computer interaction that supports these developments is still heavily based on "classical" computer programming languages, serial programming, and the multi-computer environment is accessed only through programmed interfaces, where all the burden of parallelising the, now well known, serial algorithms is again the job of (human) programmers.

Virtue is a development primarily based on the idea - as there already is so much we know of mathematical, logical and other important basic algorithms used in many fields, and their computer implementation(s) - that we would be able to "raise" the level of computer "understanding" more towards the level of our own, human language communication. This means that, for example, mathematical operations in Virtue are performed not only on integers and reals (floats), but also on complex numbers, quaternions and octonions, and that they are freely intermixed. Or that all logical operations work also on multi-levelled and multi-dimensional logical values (and not just Boolean). These are mathematically well defined operations used very often in visualisations, as well as scientific modelling. Furthermore, another basic underlying idea of the development of Virtue is that the language developed may not destroy a possibly parallel structure of data, which then allows direct and user (human) independent automatic parallel execution. Within such an approach interactive sentences can be "decoded" (executed, performed, saved, used...) on a single processor, a multi-core, multiprocessor, GPU, cluster, grid or cloud system, or any combination of them.

Designed on these basic principles, Virtue is a language which proposes a different approach, by keeping the inherent parallel structure of natural algorithms, and doing the parallel processing by itself, if it is algorithmically possible. Virtue is a syntactically very simple, yet semantically extremely complex language, offering no "reserved words", synonyms, automation of memoisation, multiple word contexts, combined data types of anything Virtue supports (e.g. functions, symbol names, scalars, multidimensional sub-structured arrays etc.), stochastic processing, multivalued and multidimensional logic operations, multidimensional sub-structured file access structures, continuations etc.

Therefore, due to this semantic richness and grammatical simpleness, in Virtue, for example, the text of the algorithm for Conway's "Game of Life" necessitates only 12 language tokens (7 words, 14 numbers in 3 vectors and 2 delimiters) in one sentence.[1]

Further development of the idea allows for development of a more syntactically rich very high level human oriented machine interaction language, which, combined with additional artificial intelligence components, appropriate ergonomic human presentation/sensory interfaces and with the integration of user style association memory, we sincerely hope can help the future development of Computer Science and Usage Practice.

I. INTRODUCTION

The challenges of the modern day world, the development of sciences and the development of social networking change quite radically the way computers are used as compared to the time when most of the basic computer construction principles were innovated. By basic computer construction principles we primarily mean the basic computer architecture based on monolithic processors executing serially specific instructions on individual values. Though many of the challenges of multiprocessing, execution parallelism, big data, cluster, grid and cloud computing are more or less, on the architectural level, solved, and throughout the development of computing a lot of physical and technical problems of multi-processing cooperation were solved and are being solved, there is still a huge lack in the development of appropriate human oriented interfaces with these (primarily hardware/firmware/systemware) complex integrated and/or dispersed computing systems.

Modern age computing needs to solve several important problems. We have to deal with enormous amounts of data which shall be mined, we necessitate High Performance Computing focusing on tightly coupled

[1] `MONADIC (1 1 1 1 0.5 1 1 1 1) [3 3] MONADIC RAVEL SUM (2.5 3 3.5) IDENTICAL ANY; MASK;.` This programme will work for any size of the "Game of Life" board.

parallel jobs for complex scientific and modelling applications, High Throughput Computing focusing on the efficient execution of a large number of loosely-coupled tasks, we need the Internet of Things, where individual appliances interrelate, and we need Personal Assistants...

And finally we have also a stringent need for High Productivity Computing, which stresses the importance of global Human-computer interaction results, i.e. the productivity, and not just the pure performance of the (already programmed) computer. High Performance Computing means speeding up the process of defining a problem to be solved and obtaining the results from the computer, that is shortening the overall problem solving time. That means that High Productivity Computing is primarily oriented towards shortening the time necessary to solve a new problem, that is, the time which is necessary for a Human to coordinate efforts with the Computer in getting an answer to a freshly thought out idea (i.e. problem to be solved).

The Virtue project is actually a product of a wish to integrate the present day computer science knowledge, as much as possible and feasible, into a consistent linguistic framework, with features which enable it to be "levelled up" towards the Human language communication possibilities, as to provide a personalised communication experience.

II. Present State of the Art

Throughout centuries of the development of computing equipment serial processing was more or less the only option, therefore, after the event of modern day computing, from the first half of the 20th century, and the development of programming languages, most of the computer-linguistic effort was put into the development of conventional serial programming languages. And, to cite John Backus from his ACM paper of 1978(!): "Conventional programming languages are growing ever more enormous, but not stronger. Inherent defects at the most basic level cause them to be both fat and weak: their primitive word-at-a-time style of programming inherited from their common ancestor—the von Neumann computer, their close coupling of semantics to state transitions, their division of programming into a world of expressions and a world of statements, their inability to effectively use powerful combining forms for building new programs from existing ones, and their lack of useful mathematical properties for reasoning about programs." "Programming languages appear to be in trouble. Each successive language incorporates, with a little cleaning up, all the features of its predecessors plus a few more." "Each new language claims new and fashionable features, such as strong typing or structured control statements, but the plain fact is that few languages make programming sufficiently cheaper or more reliable to justify the cost of producing and learning to use them." "For twenty years programming languages have been steadily progressing toward their present condition of obesity; as a result, the study and invention of programming languages has lost much of its excitement."

Now it is already 36 years after John Backus made this statement, so his statement would today cite "For almost 60 years...". Because, unfortunately, modern day developments continued down these lines with those very few extremely popular and very programmer unfriendly languages like C, C++, Java, and a whole bunch of their syntactic followers, then the popular "scripting" languages, and the vast overpopulation of all kinds of so called "Application Programming Interfaces" superimposed over the already overcrowded linguistic situation of those languages, including the fact that some of such "language extension" interfaces even claim to be "languages" (e.g. OpenGL claims to be a Graphical Language, but is actually a very complicated, even complex, API). We can actually see that what Backus said so many years ago is getting towards human limits of comprehension possibilities. And this basic problem is actually the main bottleneck of our present day human-computer productivity.

Imagine a mathematically simple and effective visualisation – a four-dimensional hypersphere to be rendered as simple asterisks showing all the dots inside the sphere's radius, and spaces outside, the third and fourth dimension to be shown as two-dimensional slices. By just looking at such a very elementary visualisation the basic structure of the hypersphere can be easily deduced. Now find a C, C++ or Java programmer and ask him to make a short program to show such a sphere to you... And now, after they have shown you this four-dimensional sphere, ask them to show to you how it would look like in 5 or 6 dimensions... Or, better, do not ask them to do any of that for you in such a language, as it would take quite a lengthy time even for the best. (A non-optimised solution which we prototyped in C for 4 dimensions has more then 130 lines of source code![2] An optimized version would take much longer to develop.) And it is a kind of "one-execution is enough" request and programme.

But with the all-penetrating spread of computer equipment, networking, and data collecting and serving there are vast and rapidly growing needs for new computer applications on all levels of the human civilisation.

We started tackling these problems systematically several years ago, and an investigation into basic principles of computer programming, as well as the principles by which we serialize the inherently extremely massively parallel universe around us into serial algorithms was performed - and a new approach taken.

One of the major points which started being obvious during the investigation of the programming principles is that most (or almost all commonly used) computer languages, being primarily serially oriented both in their

[2] The definition of a new Virtue verb "sphere" which solves the above problem for any number of dimensions up to 8 -- taking three scalar numerical objects, expressed as two two-dimensional numbers (complex - 2d, quaternion -3d or 4d, or octonion -5d, -6d, -7d or -8d) for the centre and the space-size, and a real number – the radius -- is this sentence: `TRIADIC SPACE CENTRE MAGNITUDE GREATER '*' MULTIPLY; OPERATOR @sphere SET.`

What it says is that you want to have asterisks wherever the radius is greater than the magnitude of the centered space elements. The verb SPACE (synonymous to INTERVAL) with a scalar numerical object makes a space of indices from 1 (or 1i1, 1i1j1, 1i1j1k1...). The word CENTRE is a simple synonym for the verb SUBTRACT.

semantics and in their syntax – i.e. in their whole grammar, enforce the de-contextualisation of their data structures mostly in their imperative parts. In other words this means that though a complex data structure may be declared in the language, the operations on that data structure are linguistically restricted on individual data elements from that data structure. As a consequence of this approach the context of the individual data elements in a data structure has been erased from the programme, leading to important information loss, and the language executor (compiler, interpreter, assembler…) has theoretically[3] no possibility to 'know' that the user is actually doing parallel processing in a serial way.

Interestingly enough, although obviously a consequence of the above-mentioned centuries of serial computing[4], there are not many computer languages at all which do not do this 'de-contextualisation' of their data structure operations – notably between these are APL and APL-children (APL2, J, K, A+…), and the 'array processing' linguistic branch. Unfortunately, of those few languages which keep the operations context available for the executing processor and therefore allow the processor (usually an interpreter, not the low level processing units) to make intelligent decisions on the automatic parallelisation of each particular operation on (possibly) different data structures[5] there is presently (2014) no publicly (particularly open-source) available parallel implementations we are aware of. Many so-called parallel programming languages actually necessitate explicit programmer description of the parallelistic execution, as well as (for example in Occam) explicit inter-serial-part-of-the-parallel-algorithm communication.

III. Approach

Based on the above, the approach was taken to define a new data processing language, or better to say a new data processing language family, which would allow reasonably simple algorithm description for complex data structures, and which would allow the lowest level "processor" to automatically distribute (parallelise) the operations execution on a set of processing units. Generally the idea is to have a hierarchy of languages, whereas the lowest level one, described on these pages, has to be as close to the computer hardware level as possible, given the imposed complexity, and should actually behave as a **Virtual Executor**. The job of the Virtual Executor is to do all the necessary execution and parallelisation during execution based on the data, i.e. data types and data structures. Therefore it could be said that the resulting Virtual Executor is parallely programme and data structure driven.

The idea of such a language system was born when clusters of computers and grids started being developed on top of single-processor, multiprocessor, multithreading and multicore computers, as well as supercomputers of different architectures.

The major question with these collections of systems is: How do we programme them in a uniform, efficient and reasonably simple way? How do we integrate computers of different speeds, processors, memory sizes, computation widths[6] and byte sexes?

Is it possible to develop a Language system which could cope with the necessities of High Productivity Computing, High Performance Computing, High Throughput Computing, Big Data, Visualisations, Internet of Things... and be seamlessly applicable to an indefinite number of processing stations?

"What might a language look like in which parallelism is the default? How about data-parallel languages, in which you operate, at least conceptually, on all the elements of an array at the same time? These go back to APL in the 1960s, and there was a revival of interest in the 1980s when data-parallel computer architectures were in vogue. But they were not entirely satisfactory. I'm talking about a more general sort of language in which there are control structures, but designed for parallelism, rather than the sequential mindset of conventional structured programming. What if do loops and for loops were normally parallel, and you had to use a special declaration or keyword to indicate sequential execution? That might change your mindset a little bit." (Guy Steele, Dr Dobbs Journal 24 Nov 2005).[7]

With these questions in mind the development of the Virtue system started several years ago.

IV. Virtue

The result of this approach is the first version of **Virtue** – a grammatically extremely simple, yet

[3] Practically it is possible to automatically regain a certain knowledge on the data element context in specific cases, as e.g. when a loop transiting an array has a fixed number of iterations – based on pre-execution knowledge (by the programmer) of e.g. how many elements are in the processed data structure. There are certain "parallelising" compilers on the market for several much used languages, but they can do the "parallelisation" of the algorithm described in a serial language only for very special cases. Theoretically, the lost context information is not regainable.

[4] It is probably important to mention that between humans, e.g. in mathematics, the algorithms (as for example mathematical formulas) are described in a contextfull way, however, due to our own restrictions, we always calculate the results stepwise, usually completing as much as possible work on one element of an array before proceeding to the next, which is usually by applying the given formula completely on individual data points. This inherently human approach may have also lead to the adoption of strictly serial programming languages in the past.

Parallel programming mindframe, on the other hand, starts with the same contextfull algorithm (e.g. formula), but, as opposed to the serial mindframe, executes not the whole formula on the first data structure element, and then again on the next, but executes the first operation from the formula on each data element of the structure (e.g. array) in parallel, before proceeding to the next operation from the formula.

[5] This is, naturally, dependent on a particular language implementation, although it is 'automatic' due to the fact that the actual execution parallelisation has to be developed by the language implementor, and the "programmer" does not have to think at all about how much and which of her (or his) algorithm is executed in parallel, and on how many computing resources.

[6] I.e. 32 or 64 bit, as well as other bit sizes – 8 and 16 for microcontrollers, possibly 128 or other number of bits for special computing equipment, half, normal and double, as well as extended floating point formats etc.

[7] In Virtue there are logical rules which define which combinations of objects can be processed in parallel and which must be processed in serial. These rules are internal and are actually inherently semantic regarding the verbs applied to the objects, i.e. data.

semantically extremely rich Language, which is internally and externally "resizeable". What we mean by "resizeable" is that the amount of words which are understood (i.e. executed) by Virtue can be effortlessly expanded (externally, by defining new words), and that the internal knowledge of computing can be both expanded and shrunken. So for example an embedded version of Virtue would have a reduced amount of words as well as data types, and a Visualisation version would have to know specific words, whereas a Modelling version could have special words for often used algorithms. Furthermore, the "resizeability" of Virtue is automated by it's possibility to encompass a wider range of computers, depending on the scale of the problem to be solved.[8]

Actually *Virtue* is defined to be a *Virtual Executor*, a kind of system-ware/between-ware, a *Language* system taking the role of an *Inter-actor* between a *Computer System* and a *Human*.

Let's look at the particulars of what we want to say with this definition of Virtue's aims:

A. Computer System

A computer system in the above definition is actually any possible hardware or software device which understands the grammar of Virtue. These then could be just simple appliances, or the Computer System could be a grid of supercomputers or whatever else, like server farms, clusters etc.

As said, any computer system can be part of the Virtue linguistic space, as long as it understands the common standardised Virtue Data structure. For new knowledge, i.e. when a computer system with less features (e.g. a Virtual Executor which can process only integer numbers) has to solve problems it can not solve based on these, it will automatically invoke (if possible) another network based Virtual Executor which knows how to process this data and/or words. New words, if they can be processed in a Virtual Executor, would be just transferred as explanations, but data which can not be locally processed will always stay on the computer that was asked to process this part of the overall data structure. From the perspective of the Human user, the data would appear as if it would be locally processed and saved.

And finally, the most important feature of a Computer shall certainly be to know at the lowest, hardware or firmware level of execution, that which in Computer Science, Mathematics, Logic... we already have developed as long proven methods and algorithms, and which are regularly used.

B. Virtual Executor

The Virtual Executor is actually the basic (or very expanded) Virtue linguistic system on a Computer system. It is the Virtual Executor who enables Computer systems to understands the Virtue language – so it is the Virtual Executor who is the collocutor in the Human-Computer interaction.

An important feature of a Virtual Executor is that all instructions and definitions ever written in the Virtue language will suddenly start being 10% faster, without any additional effort (except the installation of a better version of Virtue) if the Virtual Executor is sped up by 10%. Contrary to this, no C written executable programme in the world will suddenly start being faster if a new compiler version is available, which produces a 10% faster code, without the specific effort of recompiling all the executables.

C. Human

The Human is the driving force of Human development. To be able to express all of his Human innovation capacity, and to intelligently behave in the present day information-oriented world, it is necessary for her to be able to constantly process all kinds of data and thoughts.

In this sense a Human presently more and more needs an assistant, which will perform operations like computations, data searching and problem solving, he needs in the intellectual, informational field - something we for a long time use in the physical environment – amplifiers. To move a huge rock, we use an amplifier of our own feeble movements. To solve a huge problem, we need to use an amplifier of our huge problem solving abilities. We actually need this which William Ross Ashby decades ago said, we need a "computer as an intelligence amplifier". And exactly this is what is the final aim of Virtue – interactive communication with the Human, algorithm to Virtue – answer to Human. Or question to Virtue – algorithm to Human.

Virtue is aimed towards allowing the Computer to become an intelligent human assistant.

D. Language

Long years of investigations into Human languages, all the worlds linguistic efforts show that the Language is the primary means of Human external world model forming, and his absolutely prime and most developed communication and understanding system.

Therefore, it is absolutely essential, to be able to actually fully integrate computers into the modern world Human society as intelligence amplifiers, to raise the level of Computer's language understanding towards the level of a full Human language, i.e. allowing the Human-Computer interaction language to be able to provide external world models for both of them.

V. DATA TYPES

Virtue has complex data structures (scalars, arrays of any dimension, array elements which are themselves structures of scalars or arrays…) and diverse data types (booleans; multidimensional fuzzy logicals; characters; numbers – real, complex, quaternion, octonion; addresses; functions; labels; files; or possible specific types like planetary ephemeris, pixels etc.).

[8] The present experimental implementation of Virtue uses only SMP. Work is in progress to encompass TCP/IP communication protocols with other computers running the Virtue system, by which all those processors/computers become a single large compute-assistant. The internal architecture of the experimental implementation is already prepared for multi-computer parallelism.

The inclusion of e.g. functions as data types, inter-mixable with any other data-type, as well as the form in which the programme itself is internally represented - in the exactly same way as data, enables Virtue to use, for example, variable functions whenever necessary, or to define function parallelism (e.g. by applying a vector of functions to a vector of values).

Files are also scalar data types, which means that they can be inside a multidimensional structured array (there is no theoretical limit to the number of open files in Virtue, the same as with the dimensionality of arrays). Each use of an argument which is a File (or contains a File in its structure) will read the data from the file, and provide the data defined by a "prototype record" supplied when the file was opened (i.e. to open a file a name and a prototype record is given) for the pending operation. This linguistically very simple and understandable approach allows for very complex data manipulation across a huge range of input data-sets and processing systems.

Virtue has an almost fully orthogonal approach towards data – any operation that has sense is usable. However, nonsensical operations (like e.g. square root of a character) are strictly disallowed. The execution framework of Virtue is very forgiving, yet strict, in imposing the processing rules for different combinations of data structures and data types.[9]

Further important features of Virtue include the complete lack of reserved words or symbols[10], synonyms, word dependant hierarchical contexts, continuations, and the possibility to use any data whatsoever as a name for something else.[11]

VI. Serial versus Parallel processing

The Virtue Language is designed to keep as many logically parallel algorithmic structures expressed in such a way that the parallelism is obvious. However, an algorithm is, per definitionem, a series of operations, possibly intermixed with decision making.

Virtue tries to keep consistency inside the two sets of operations, the:

- serial operations
- parallel operations

However, it should be noted, that the whole language can be implemented on a purely serial computer, but can also be implemented on parallel data processing equipment (multiprocessor, vector, cluster, grid...). The consistency of semantic distinction between parallel (i.e. parallelisable) and serial (non-parallelisable) instructions in an program on many levels does not a-priory dictate the implementation of Virtue. Although multidimensional logical values are parallel constructs, it would be counter-productive to parallelise a logical operation on 2, 4 or 8 individual values (or wouldn't it?). However operations on conformable multi-dimensional arrays are also parallel constructs, but it would be very advisable to implement them in parallel on several independent processing units.

It is important to understand that not all constructs may be safely executed in parallel, although applied to conformable arrays of enough elements. Although, for example, the application of functions on vectors/arrays may be a parallel operation, it must be serialized if the function has side effects.

Such a strict distinction is necessary, as there is no a-priory order in parallel operations, and there is no communication possible between strictly parallely executing operations.

Already during the analysis and/or during the processing Virtue will introduce into the internal Data Description (in this particular case of a function) a flag indicating if it may be executed in parallel, or the processing must be serialised. So, for example, by using input/output commands or global variables in a function which otherwise could be executed parallely, Virtue will automatically flag that condition, and serialise the execution. In the same time, these conditions in a functional sentence will prevent it from being memoised.

As another example of usage of the Data Description, an argument whose any element was produced by a random number generator (which indicates stochastic processing) applied to a normally memoisable function will prevent the memoisation of this function with that particular argument. This is necessary as otherwise simulations relying on random number generation would not properly work if memoisation is used. It will, naturally, normally work with all data which was not produced by manipulation of a random number.

It is obvious that with such internal structures the Human does not have to take care of the way his problems will be solved on a hardware level – serially, parallely, or in a concurrent combination of the two (for example more processors executing different functions on a vector, and using vector processors to execute parallel operations inside those functions). Virtue will execute SIMD, MIMD, MISD and SISD, even concurrently, as possibly in the above example. This makes the Virtue system widely adaptable to the underlying hardware, as any of the hardware architectures possibly used will excel in at least some aspects of execution.

VII. The Implementation of Virtue

The Virtue Language is fully defined in it's primary form.

The computer implementation is presently in Alpha 0.6 state, and parallelisation is implemented on SMP systems.

[9] For example, a dyadic (two argument) operation on arrays whose dimensions are not equal is prohibited, except if one of the argument array dimensions is a subset of the other argument's array dimensions, or one of the arguments is a scalar.

[10] Except the symbol '::', which reverts any redifined internal Virtue word or symbol to its original meaning, and the symbol '.', which always marks the end of a sentence.

[11] For example the sentence `'A vector of numbers from 1 to 1000' 1000 INTERVAL AT SET.` will save the given string under the name of a vector from 1 to 1000. To recall the string from memory, such a vector shall be given, as for example: `900 INTERVAL (901 902) CATENATE 98 INTERVAL 902 + CATENATE GET.` This sentence will then give the answer "`A vector of numbers from 1 to 1000`".

The experimental implementation is constantly parallely developed and tested on a very wide range of different computers, ranging from the mid-1980-ies Sun3 (16MHz/16MiB and 20MHz/24MiB) workstations up to modern day blades, with various operating systems and their generations (SunOS, Solaris, NetBSD, FreeBSD, Linux, Win), various processors (68k, MIPS, SPARC, PowerPC, AMD, Intel), on 32 and 64 bit processors in single-processor and SMD multi-processor, multi-threading and multi-core computers. Such a wide range of computers, both historically and speed-wise, for the experimental Virtue implementations allows for a development of a very easily adaptable system, and the behaviour of the old Sun3 systems shows that even on them the execution seems very fast for the amount of data which can be represented in the memory of those computers[12].

The internal speed measurements which the Virtue Executor has (and they are important for the future multi-computer implementations, to allow load balancing) have provided us with quite a lot of important data on the behaviour of different computer systems and different processors, so an investigation into the "speed of a computer" is presently being performed, with some results to be presented soon.

VIII. Conclusion

Virtue is a pliable language, shrinkable and extensible, but generically compatible as a Language on all possible levels of implementation. Therefore it could be implemented in hardware, as a specific processor, it could be small and embedded, but it could automatically grow towards huge systems and big data. The idea of a Language even in the inter-computer communication, allows for Virtual Executors which do not recognise a word the user used, to consult (if on network) other known, but probably bigger Virtual Executors about the meaning of the word, and/or help in the execution of a task.

The final aim of this long-term effort is to provide a "human assistant", an intelligent, selflearning and selforganising "assistant" in this more and more complex life environment.

From the level of being a chip, or a programme in a chip, able to monitor, coordinate and regulate processes in complicated systems, up to the level of data mining in distributed environments, Virtue is aimed to provide a seamless integration.

As a mathematical processor, as a Grid Library Application, as a Modelling Tool, Visualisator or a data/calculation preprocessor...

References:

[1] John Backus "Can Programming Be Liberated from the von Neumann Style? A Functional Style and Its Algebra of Programs", Communications of the ACM, Vol 21, No. 8, pp. 613-641, August 1978., http://delivery.acm.org/10.1145/360000/359579/a1977-backus.pdf?ip=93.136.170.6&id=359579&acc=OPEN&key=4D4702B0C3E38B35%2E4D4702B0C3E38B35%2E4D4702B0C3E38B35%2E6D218144511F3437&CFID=415707851&CFTOKEN=76568762&__acm__=1393661168_49edb86547c11a2260a8d1ad8731f4b4

[2] Barendregt, H.P. and Eekelen, M.C.J.D. and Hartel, P.H. and Hertzberger, L.O. and Plasmeijer, M.J. and Vree, W.G. (1987): "The Dutch parallel reduction machine project". Future Generation Computer Systems, 3 (4). pp. 261-270. ISSN 0167-739X, http://doc.utwente.nl/55757/1/dutch.pdf

[3] Borna Bebek, Zorislav Šojat: "Optimal Distribution of Organisational Resources", Diplomatic Academy Year–Book, Edition: Central European Initiative – International Conference "Diplomacy for the twenty-first century: knowledge management", Vol. 2/2000, No. 1, HR: Ministarstvo vanjskih poslova i europskih integracija; 2000

[4] EGI Glossary; European Grid Initiative WIKI, https://wiki.egi.eu/wiki/Glossary_V1, 22/2/2014, 21:56

[5] "Gannett's Law", http://0x07bell.net/WWWMASTER/CrayWWWStuff/Glaw.html

[6] Tomislav Grubesa, Goran Topic, Valentin Vidic, Zorislav Sojat, Karolj Skala: "Towards the Design and Application of Grid Portals", Hypermedia and Grid Systems, MIPRO, Opatija, Croatia; 06/2005

[7] Roland N Ibbett & Nigel P Topham (1996): "HIGH PERFORMANCE COMPUTER ARCHITECTURES - A Historical Perspective", http://homepages.inf.ed.ac.uk/rni/comp-arch/index.html

[8] Sinisa Marin, Robert J. Thorpe, Zorislav Sojat: "A Modular Robot Programming System", 1989; London, GB: RDP Technology.

[9] Sinisa Marin, Mihajlo Ristic, Zorislav Sojat: "An Implementation of a Novel Method for Concurrent Process Control in Robot Programming", Third International Symposium on Robotics and Manufacturing: Research, Education and Application, ISRAM '90, Burnaby, BC/CA; 1990

[10] John Owens "Data Level Parallelism (2)" EEC 171 Parallel Architectures, UC Davis, http://www.nvidia.com/content/cudazone/cudau/courses/ucdavis/lectures/dlp2.pdf 1/3/2014, 6:13

[11] Nikola Pavkovic, Karolj Skala, Valentin Vidic, Zorislav Sojat: "Bioinformatics Application Oriented IT Deployment Model",

[12] The classical double recursive algorithm for calculating the n-th number of a Fibonacci series could be expressed in Virtue for example thus: `MONADIC DUPLICATE 3 < @%a IF JUMP DUPLICATE 1 SUBTRACT Fibonacci! SWAP 2 SUBTRACT Fibonacci! ADD RETURN %a DISCARD 1; @Fibonacci SET.` The sentence, translated into English, says: "Check if the requested Fibonacci series element number is less then 3, if not subtract 1 from this number and calculate that Fibonacci (series element), then subtract 2 from the same number and calculate the Fibonacci, otherwise (if the number was < 3) jump to conclusion, forget that number and just return 1, then remember this phrase, which expects one object, as the word 'Fibonacci').

The calculation of all the first 1500 Fibonacci series numbers (slightly over the limit of IEEE double FP) on a 1987. Sun3/60 (SunOS 4.1.1, original Sun cc compiler, MC68020+MC68881, 20MHz, 24MiB 32-bit RAM - ~3MIPS, 192 kFLOPS) using this algorithm with the memoisation word RESULT - that is inputing the sentence `1500 INTERVAL MONADIC Fibonacci RESULT; EACH.` (in English: "Make an interval vector from 1 to 1500, then for each element of the vector calculate the Fibonacci, but first check if you already remembered the result.") - the vector of those first 1500 (!) Fibonacci numbers will be produced after just 27.827s, whereas the next time the same sentence is entered, the results will come in just 4.461s. Such speed is fully acceptable for normal work! Modern day common computers are several hundred times faster (not several thousand times, as would be expected from the MIPS/GIPS relationship. Just one example: a recently tested 1.67GHz Intel Atom in a reputable notebook with 2GiB 667.0 MHz DDR2 SDRAM memory, on Windows 7/32, Microsoft Visual Studio 2010 Express C compiler, performed the above operations in 0.21s (first time), and in 0.045s (afterwards), a speedup compared to Sun-3 of only around 100 times, despite the high sounding numbers. The reason is the more and more intensively felt "von Neumann bottleneck" - the memory vs. processor speed and channel constraints!)

Parallel Numerics, Theory and Application; Salzburg : University of Salzburg Austria(2005):217-222; 2005

[12] Karolj Skala, Zorislav Sojat: "Towards a Grid Applicable Parallel Architecture Machine.", Computational Science - ICCS 2004, 4th International Conference, Kraków, Poland, June 6-9, 2004, Proceedings, Part III; 2004

[13] K. Skala, Z. Sojat: "Image Programming for Scientific Visualization by Cluster Computing", Autonomic and Autonomous Systems and International Conference on Networking and Services, ICAS-ICNS; 2005;

[14] Karolj Skala, Nikola Pavkovic, Zorislav Sojat: "Scientific Visualization by Cluster Computing", 8th COST 276 Workshop, Trondheim, Norway; 05/2005

[15] Zorislav Šojat: "Jezik kao samoorganizirajući stroj", Interbiro '82, Zagreb, Yugoslavia; 1982

[16] Zorislav Sojat: "Operating System Based on Device Distributed Intelligence", 1st Orwellian Symposium, Baden Baden, Germany; 1984

[17] Zorislav Šojat: "Ustrojenje jezika APL", 1989; Zagreb, YU: Filozofski fakultet Sveučilišta u Zagrebu

[18] Zorislav Sojat, Sinisa Marin: "ISOCOM 20 Filter System User Documentation: Flow Through Filter Language and Graphics Organisation Language", 1992; BTS, Purley, GB.

[19] Zorislav Šojat: "Nanoračunarstvo i prirodno distribuirani paralelizam", Glasilo Instituta Ruđer Bošković. 3(7/8):20-22.; 2002

[20] Zorislav Sojat, Karolj Skala: "Multiple Programme Single Data Stream Approach to Grid Programming", Hypermedia and Grid Systems, Opatija, Croatia; 2004

[21] Guy Steele, Jan-Willem Maessen: "Fortress Programming Language Tutorial", Sun Microsystems Laboratories, 2006; http://stephane.ducasse.free.fr/Teaching/CoursAnnecy/0506-Master/ForPresentations/Fortress-PLDITutorialSlides9Jun2006.pdf

[22] Jack J. Woehr: "A Conversation with Guy Steele Jr.", Dr. Dobbs, April 1, 2005, http://www.drdobbs.com/jvm/a-conversation-with-guy-steele-jr/184406029

Dynamic workflow support in gUSE

E. Kail[1], A. Bánáti[1], K. Karóczkai[2], P. Kacsuk[2,3], M. Kozlovszky[1,2]

[1] Obuda University, John von Neumann Faculty of Informatics, Biotech Lab

Bécsi str. 96/b., H-1034, Budapest, Hungary

[2] MTA SZTAKI, LPDS, Kende str. 13-17, H-1111, Budapest, Hungary

[3] University of Westminster, 115 New Cavendish Street, London W1W 6UW

{kail.eszter,banati.anna,kozlovszky.miklos}@nik.uni-obuda.hu,

{karoczka, kacsuk}@sztaki.mta.hu

Abstract - Scientific workflow systems aim to provide user friendly, end-to-end solutions for automating and simplifying computational or data intensive tasks. A number of workflow environments have been developed in recent years to provide support for the specification and execution of scientific workflows. Normal static workflows can poorly cope with the ever changing status of the existing distributed systems. During workflow enactment unforeseen scenarios may arise, which can cause significant delays, failed executions, or improper results. Manual workflow enactment has its obvious limitations, however automatic failover mechanisms require minimum an accurate information set about the workflow tasks and about the status of the underlying processing infrastructure. Dynamism can be defined at different abstraction levels and in different phases of the workflow lifecycle. In this paper we identify the requirements of dynamic workflows in general and provide a thorough survey about gUSE/WS-PGRADE's dynamic workflow handling capabilities.

Keywords: dynamic workflow, gUSE, workflow management system.

I. INTRODUCTION

Scientific workflow systems are used to develop complex scientific applications by connecting different algorithms to each other. Such organization of huge computational and data intensive algorithms aim to provide user friendly, end-to-end solution for scientists. However normal static workflows cannot cope with the ever changing status of the existing distributed system and the unforeseen scenarios that may arise during workflow enactment. These scenarios or exceptions may impact the workflow at various levels. Different classes of exceptions require different approaches in their handling and recovery in the various phases of the workflow lifecycle. Some cause only resubmitting the workflow, while others may result in changes even in the workflow model. Depending on the requested handling strategy we defined the levels at which the system should support dynamism.

A number of scientific workflow environments have been developed over the past decade to provide support for the specification and execution of scientific workflows (ASKALON [24], Triana [22], Taverna [5], Kepler [23], MOTEUR [25] and gUSE [10]). gUSE has been developed and maintained since 2008 by MTA SZTAKI. The gUSE enables scientific communities to compose and execute a workflow in a scientific application on different DCIs (such as clusters, grids, supercomputers, desktop grids and clouds).

In this paper we give a taxonomy of the dynamism in the context of the workflow lifecycle phases and the defined abstraction levels. According to the detailed taxonomy we give a thorough survey about the gUSE/WS-PGRADE's dynamic workflow capabilities.

The paper is organized as follows. In section II we give a brief overview of the dynamism, found in the literature. Section III identifies the different phases of the workflow lifecycle, introduces the definition and levels of dynamism and presents issues and their handling regarding dynamism. Section IV provides a thorough survey about dynamism in gUSE. Conclusion and future work in section V and the bibliography closes our work.

II. RELATED WORK

Very few papers give overall analysis of dynamism, they rather select one or few segments of it and propose solutions for them in a concrete workflow management system.

In his thesis Bahsi [1] studies the different workflow management systems (Apache Ant, Askalon, DAGMan, Kepler, Pegasus, Taverna, Triana etc.) and compare their capabilities in terms of dynamic and conditional structures they support. He found out that some systems have primitive logic elements that can be used to build advance conditional structures, others already have those high level ones. In addition, although some systems do not have conditional elements, some other mechanisms can be used for implementing the requested conditional behavior. Furthermore some system instead of returning an error code in case of certain failures, they cause the whole workflow to fail. This complicates building dynamic workflows.

Gil at al. [2] summarized the requirements of future scientific applications and the challenges that they present to current workflow technologies. The dynamic workflows and user steering topic was also investigated, and many expectations were determined, but the levels or phases of dynamism were not systematized. For example the workflow have to support dynamism in making decision about which steps to take next based on the available information in different scenarios (external events, workflow interoperability, user steering) or have to support dynamism in the appropriate user interface that enable scientist to make a complex exploratory process. The dynamic management of workflow was also conceived.

Lee et al. [16] in their research use Autonomic Computing methodologies, such as monitoring, analyses, planning and execution (MAPE) model to manage the long run scientific workflow execution processes.

Yu and Buyya [18] in his paper propose a general taxonomy that characterizes and classifies various approaches for building and executing workflows on grids. Their analysis affects the topic of dynamism in workflow design, scheduling and data management but they did not map these solutions to phases or abstraction levels. In their work among others they introduce automatic workflow composition and automatic data movement. In the former case the system itself generates the workflow according to the users' requirements, while in the latter case the system provides the mechanism to transfer intermediate data. They distinguish centralized, mediated and peer-to peer solutions

III. A SCIENTIFIC WORKFLOW LIFE CYCLE

A. Scientific Workflow Life Cycle

The various phases and steps associated with planning, executing, and analyzing scientific workflows comprise the scientific workflow life cycle (WFLC). The following phases are largely supported by existing workflow systems using a wide variety of approaches and techniques. [3]

Hypothesis Generation (Modification): Development of a scientific workflow usually starts with hypothesis generation. Scientists working on a problem, gather information, data and requirements about the related issues to make assumptions about a scientific process. From these data they build a specification which can be modified later during the whole lifecycle, or after the result analysis.

Experiment / Workflow Design: During the experiment an actual workflow is assembled based on this specification. This phase is the workflow development or design phase, which differs from general programming in many ways. It is usually the composition and configuration of a special-purpose workflow from pre-existing, more general-purpose components, sub-workflows, and services. During workflow composition, the workflow developer either creates a new workflow by modifying an existing one or composes a new workflow from scratch using components and sub workflows obtained from a repository. In contrast to the business workflow world, where standards have been developed over the years (e.g., WS-BPEL 2.0 [4]), scientific workflow systems tend to use a language set of internal languages and exchange formats (e.g., SCUFL [5], GPEL [6], and MOML [7]). Reasons for this diversity include the wide range of computation models used in scientific workflows and the initial focus of development efforts on scientist oriented functionality rather than standardization.

Instantiation: Once the workflow description is constructed, scientific workflow systems often provide various functions prior to execution. These functions may include workflow validation, resource allocation, scheduling, optimization, parameter binding and configuration. Workflow mapping is sometimes used to refer to optimization and scheduling decisions made during this phase.

Execution: After the workflow instantiation the workflow can be executed. During execution, a workflow system may record provenance information (data and process history) as well as provide real-time monitoring and failover functions. Depending on the system, provenance information generally involves the recording of the steps that were invoked during workflow execution, the data consumed and produced by each step, a set of data dependencies stating which data was used to

derive other data, the parameter settings used for each step, and so on. If workflow migration or adaptation (i.e.: change the workflow model or the running instance) is enabled or supported during execution (e.g., due to the changing environment), the evolution of such a dynamic workflow may be recorded as well in order to support subsequent event handling.

Result Analysis: After workflow execution, scientists often need to inspect and interpret workflow results. This involves evaluation of the results, examination of workflow execution traces, workflow debugging and performance analysis.

Data and workflow products can be published and shared. As workflows and data products are committed to a shared repository, new iterations of the workflow life cycle can begin.

B. Definition of dynamism

Dynamism on one hand is the ability of a system to react or to handle unforeseen scenarios raised during the workflow enactment phase, in a way to avoid certain failures or to recover from specific situations automatically or with user intervention. The adaptation to new situations may range from resubmitting a workflow to even the modification of a whole workflow model. On the other hand the dynamism is the opportunity to change the abstract or concrete workflow model or to give faster execution and higher level performance according to the actual environmental conditions and intermediary results. We distinguish several levels at the different lifecycle phases of a workflow. The *system level* concerns with those dynamic issues that are supported by the workflow management system. The *composition level* includes the language or the DAG support. With *task level solutions* large scale dynamism can be achieved if the system is able to handle tasks as separate units. *Workflow level* dynamism deals with problems which can only be interpreted in the context of a certain workflow. *User level* gives the opportunity for user intervention.

During design time, the workflow developer creates an abstract workflow model first (for example a DAG) [16][18], which contains the tasks (computational process), the structure of jobs and the data dependency between them. This model is independent from the exact computing infrastructure. To reach the required dynamism at this phase the system should give support at three different levels, namely system level, composition level, and the task level. Table [I.] summarizes the levels connected to the individual phases.

During the instantiation phase the abstract workflow to concrete workflow mapping takes place. [16][18] Depending on the system support and the status of the resources full or partial concrete model can be generated even before or during workflow execution. At this phase the dynamism also at three distinct level can be supported (system, task and workflow level, see table [1]). Scheduling also takes place at this phase.

The three above mentioned levels also appear during execution time. On one hand this is the phase where the various exception handling strategies, the monitoring and logging functions or the provenance based decisions are affected. On the other hand at this time happens resource reallocations, workflow instance or model changing and even the user intervention.

C. Related issues of the different phases of WFLC

The dynamism supported by the workflow management system can be interpreted in 4 phases of the above mentioned time intervals of the workflow lifecycle. (i.e.: design, instantiation, execution, analysis). The result analysis, log mining and the provenance is out of scope of this paper.

Design time

During design time, dynamism can be primarily supported by the modeling language at composition level. Several existing workflow managers have support for conditional structure in different levels. While some of them provide if, switch, and while structures that we are familiar from high level languages, some of the workflow managers provide comparatively simple logic constructs. In the latter case, the responsibility of creating conditional structures left to users by combining those logic constructs with other existing ones [1].

Heinl et al. [8], [9] gave a classification scheme for flexibility of workflow management systems. He defined two groups: flexibility by selection and flexibility by adaptation. Flexibility by selection techniques also should be implemented in the design time but of course they need some system level support. It can be achieved by advance modeling and late modeling.

The advance modeling technique means that the user can define multiple execution alternatives during the design or configuration phase and the completion or incompletion of the predefined condition decides the actual steps processed in run time.

The late modeling technique means, that parts of a process model are not modeled before execution, i.e. they are left as ‘black boxes’ and the actual execution of these parts are selected only at the execution time.

During this phase the system may also support task level dynamism in a sense, that sub workflows, or task from existing workflows should be reusable in other workflows as well. The modular composition of workflows also enables the simple and quick composition of new workflows.

TABLE I. ABSTRACT LEVELS OF DYNAMISM

design time (abstract wf level)	**system level**	**composition level**	**task level**	
	- black boxes - advance modelling technique	- language or graph structure	- modularity, reusability	
instantiation time (concrete wf level)	**system level**	**task level**	**wf level**	
	- breakpoints - incremental compilation - various protocol support - provenance based sched - multi instance activity	- task based sched. - late binding of data	- partitioning to sub workflows - parameter sweep appl. - wf based sched - mapping adaptation	
execution time	**system level**	**task level**	**wf level**	**user level**
	- exception handling (EH) - provenance based decisions - monitoring, logging - dynamic resource allocation	- resource alloca-tion	- alternate task - change the model	- user inter-ven-tion

Instantiation time

Static decision making involves the risk that decisions may be made on the basis of information about resource performance and availability that quickly becomes outdated. [11] As a result with system level support, benefits may appear either from incremental compilation, whereby resource allocation decisions are made for part of a workflow at a time (e.g. [12]), or by dynamically revising compilation decisions of a concrete workflow while it is executing (e.g. [13, 14, 15]). In principle, any decision that was made statically during workflow compilation should be revisited at runtime (e.g. [16])

Another way to support dynamism at system level during instantiation is using breakpoints. To interact with the workflow for tracking and debugging, the developer can interleave breakpoints in the model. At these breakpoints the execution of job instances can be enabled or prohibited, or even it can be steered to another direction. [10]

We also reckon multi instance activities among the above mentioned system level dynamic issues. Multi instantiation of activities gives flexibility to the execution of workflows. It means that during workflow enactment one of the tasks should be executed with multiple instance (i.e.: parallelism), but the number of instances is not known before enactment.

A way to allow flexibility in data management at system level is to support access to object stores using a variety of protocols and security mechanisms. [17]

A task level challenge for workflow management systems is to develop a flexible data management solution that allows for late binding of data. Tasks can discover input data at runtime, and possibly choose to stage the data from one of many locations.

At workflow level using mapping adaptations depending on the environment, the abstract workflow to concrete workflow bindings can change. Lee et al. [16] deal with this issue in details.

If the original workflow can be partitioned into sub-workflows before mapping, then each sub-workflow can be mapped individually. The order and timing of the mapping is dictated by the dependencies between the sub-workflows. In some cases the sub-workflows can be mapped and executed also in parallel. The partitioning details are dictated by how fast the target execution resources are changing. In a dynamic environment, partitions with small numbers of tasks are preferable, so that only a small number of tasks are bound to resources at any one time [3].

In scientific context the most important applications are parameter sweep applications over very large parameter spaces. Practically it means to submit a workflow with various data of the given parameter space. This kind of parallelization gives faster execution and high level flexibility in the execution environment.

Scheduling algorithms can also be task based (task level) or workflow based (workflow level) and with system level support the performance and effectiveness of the algorithms can be improved with provenance based information.

Execution time

In a dynamically changing environment, during workflow enactment unforeseen scenarios may result in various work item failure (due to faulty results, resource unavailability and so on). Many of these failures could be avoided with workflow management systems that provide more dynamism and support certain level of adaptivity to these scenarios.

We categorize the related issues into levels according to table [1].

The first level is made up from the failure of hardware, software or network component, associated with the work item or data resources unavailability. In these cases, either the exception handling may or should include mechanisms to detect and to recover from failures (for example restart the job or workflow or make some other decision based on provenance data), even with provenance based support. In all these cases possible handling strategies should be tracking, monitoring and gathering provenance

information in order to support users in coming to a decision.

The user level dynamism consists of scenarios where the system waits for user steering. Here we can rate the breakpoints, where workflow execution can be suspended and enabled again by the user. Also at this time happens the interpretation of the black boxes (late modelling technique).

Suspending a workflow and then continue with a new task by deviating from the original workflow model also gives more flexibility to the system at workflow level. In Heinl's taxonomy [8, 9] it is defined by flexibility by adaptation. In this case we distinguish adaptive systems and ad hoc systems. While adaptive systems modify process model on instances leaving the process model unchanged, in ad hoc systems the model migrates to a new state, to a new model [9].

IV. DYNAMIC SUPPORT IN WS-PGRADE/GUSE

The WS-PGRADE Portal [10] is a web based front end of the gUSE infrastructure. The structure of WS-PGRADE workflows are represented by directed acyclic graphs (DAG).

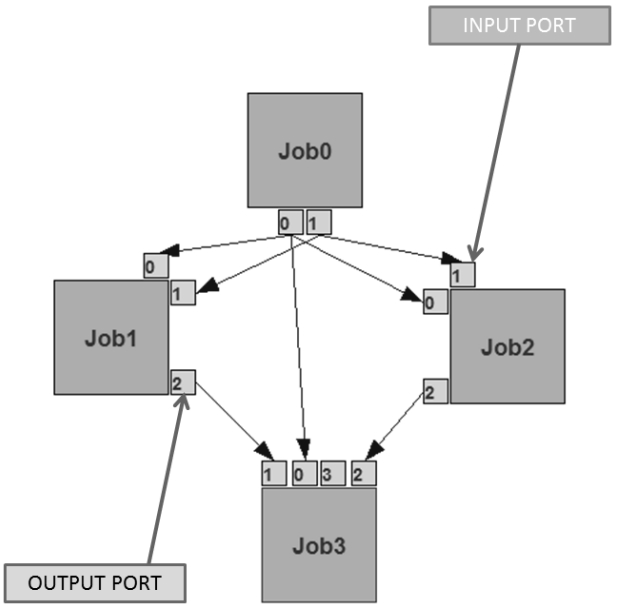
The nodes of the graph, namely jobs are the smallest units of a workflow. They represent a single algorithm, a stand-alone program or a web-service call to be executed. Ports represent input and output connectors of the given job node. Directed edges of the graph represent data dependency (and corresponding file transfer) among the workflow nodes. This abstract workflow can be used in the second step to generate various concrete workflows by configuring detailed properties (first of all the executable, the input/output files where needed and the target DCI) of the nodes representing the atomic execution units of the workflow.

The algorithm configuration determines the functionality of the job, the resource configuration determines where this activity will be executed, and the port configuration determines what the input data of the activity are and how the result(s) will be forwarded to the user or to other jobs as inputs. Fig 1.

After all the properties of the workflow have been set, it can be submitted resulting in an instance of the workflow. A concrete workflow can be submitted several times (for example, in case of performance measurements) and every submission will result in a new instance of the same concrete workflow.

A job may be executed if there is proper data (or dataset in case of a collector port) at each of its input ports and there is no prohibiting programmed condition excluding the execution of the job

Figure I. WORKFLOW COPOSITION IN GUSE



The execution of a workflow instance is data driven forced by the graph structure: A node will be activated (the associated job submitted or the associated service called) when the required input data elements (usually file, or set of files) become available at each input port of the node.

Design time

At composition level on the one hand the gUSE provide the possibility of creating embedding workflows. This enables the usage of the subroutine paradigm where workflow B can be used inside a parent workflow A by specifying that the executable of a certain node in A should be workflow B. [10] On the other hand multiple execution alternatives can be defined with specifying some condition on input parameters. In this way the completion or the incompletion of the condition during execution time determines the actual execution of the concrete workflow.

Instantiation time

At system level gUSE supports the use of breakpoints. The user may suspend or abort the execution of a running instance. A previously suspended workflow instance can be resumed at any time by the user. This way the user can observe the progress of workflow elaboration on the fly and can perform a service call instance level check-pointing when needed.

Parameter study executions are supported in the gUSE at workflow level. From an abstract workflow various concrete workflow can be generated with different input data from the parameter space, and even from one concrete workflow more instances can be produced by saving with unique ID-s in order to support performance analysis.

Multi instance activities are also supported with the configuration of job nodes to be generators and others to act as collectors. A workflow node (job) has generator property if the execution of the associated job or service call instance may produce more than one output data elements associated to given distinguished output port(s) [19]. It gives task level dynamism support.

The incremental compilation, data management flexibility and the late binding techniques are not supported in gUSE, however, all the tasks that a workflow model comprise of can be individually mapped to resources during the configuration phase of the WFLC.

Execution time

At system level exception monitoring is supported in the system (with breakpoints), and in case of software/hardware failures, resource unavailability, running failures, network unreachability the concrete workflows are only resubmitted. So problem solving in some of this cases need an administrator or user to be involved.

At user level the predefined breakpoints provide dynamism to the workflow developer.

In gUSE only the actually running instance may deviate from the original workflow model, leaving the model itself unchanged (Adaptive systems [7, 8]).

V. CONLUSION

In our work we introduced a detailed definition for dynamism that should be supported by dynamic scientific workflow management systems. We systematized the individual issues regarding dynamism into different levels at the various phases of the workflow lifecycle. According to our presented taxonomy we investigated the gUSE dynamic capabilities. In our future work we intend to improve the dynamic support of the SZTAKI gUSE system.

REFERENCES

[1] E. M. Bahsi, *Dynamic Workflow Management For Large Scale Scientific Applications*, PhD Thesis, B.S., Fatih University, 2006.

[2] Y. Gil, E. Deelman, M. Ellisman, T. Fahringer, G. Fox, D. Gannon, C. Goble, M. Livny, L. Moreau, és J. Myers, *Report on the 2006 NSF Workshop on Challenges of Scientific* „Workflow. Examining the challenges of scientific workflows”, Citeseer, 2006.

[3] B. Ludäscher, I. Altintas, S. Bowers, J. Cummings, T. Critchlow, E. Deelman, D. D. Roure, J. Freire, C. Goble, és M. Jones, „Scientific process automation and workflow management”, Scientific Data Management: Challenges, Existing Technology, and Deployment, Computational Science Series, o 476–508, 2009.

[4] D. Jordan and J. Evdemo. Web Services Business Process Execution Language, Version 2.0 (WS-BPEL 2.0), 11 April 2007

[5] The Taverna Project, http://www.taverna.org.uk , 2009.

[6] Y. Wang, C. Hu, and J. Huai. A New Grid Workflow Description Language. In IEEE International Conference on Services Computing , IEEE SCC, pp. 257–260. IEEE Computer Society, 2005.

[7] C. Brooks, E. A. Lee, X. Liu, S. Neuendorffer, Y. Zhao, and H. Zheng. Heterogeneous Concurrent Modeling and Design in Java (Volume 3: Ptolemy II Domains). Technical Report No. UCB/EECS-2008-37, April 2008.

[8] P. Heinl, S. Horn, S. Jablonski, J. Neeb, K. Stein, and M. Teschke. A Comprehensive Approach to Flexibility in Workflow Management Systems. In Proceedings of the International Joint Conference on Work Activities Coordination and Collaboration (WACC ’99), pages 79–88. ACM Press, New York, NY, USA, 1999.

[9] M. Pesic: Constraint-Based Workflow Management Systems: Shifting Control to Users, PhD thesis, Eindhoven: Technische Universiteit Eindhoven, 2008

[10] gUSE user manual: https://sourceforge.net/projects/guse/files/3.4.4/docs/Portal_User_Manual_v3.4.4.pdf/ download (2012). Accessed 28 Oct 2012

[11] K. Lee, N. W. Paton, R. Sakellariou, E. Deelman, A. A. A. Fernandes, és G. Mehta, „Adaptive workflow processing and execution in Pegasus”, Concurrency and Computation: Practice and Experience, vol 21, sz 16, o 1965–1981, 2009.

[12] E. Deelman, G. Singh, M.-H. Su, J. Blythe, Y. Gil, C. Kesselman, G. Mehta, K. Vahi, G. B. Berriman, és J. Good, „Pegasus: A framework for mapping complex scientific workflows onto distributed systems”, *Scientific Programming*, vol 13, sz 3, o 219–237, 2005.

[13] T. Heinis, C. Pautasso, and G. Alonso. Design and evaluation of an autonomic workflow engine. In 2nd International Conference on Autonomic Computing, pages 27–38. IEEE Computer Society, 2005

[14] R. Duan, R. Prodan, and T. Fahringer. Run-time optimisation of grid workflow applications. In Proc. Intl. Conference on Grid Computing, pages 33–40. IEEE Press, 2006

[15] J.-H. Lee, S.-H. Chin, H.-M. Lee, T. Yoon, K.-S. Chung, and H.-C. Yu. Adaptive workflow scheduling strategy in service-based grids. In GPC, pages 298–309. Springer, 2007.

[16] K. Lee, R. Sakellariou, N. W. Paton, and A. A. A. Fernandes. Workflow adaptation as an autonomic computing problem. In 2nd Workshop on Workflows in Support of Large-Scale Science (Works 07) in Proceedings of HPDC, pages 29–34, 2007

[17] K. Vahi, M. Rynge, G. Juve, R. Mayani, és E. Deelman, „Rethinking data management for big data scientific workflows”, in *Big Data, 2013 IEEE International Conference on*, 2013, o 27–35.

[18] J. Yu és R. Buyya, „A taxonomy of scientific workflow systems for grid computing”, *Sigmod Record*, vol 34, sz 3, o 44–49, 2005.

[19] T. Gottdank: WS-PGRADE Portal Cookbook; Solutions and Examples for WS-PGRADE Users 2012-2014 MTA SZTAKI LPDS

[20] J. Wainer és F. de L. Bezerra, „Constraint-Based Flexible Workflows”, in Groupware: Design, Implementation, and Use, J. Favela és D. Decouchant, Szerk Springer Berlin Heidelberg, 2003, o 151–158..

[21] The Kepler Project, http://www.kepler-project.org.

[22] The Triana Project, http://www.trianacode.org

[23] The Askalon Project, http://www.askalon.org

[24] The moteur Project, http://www.shiwa-workflow.eu

Formal Checking of WS-BPEL Orchestrations

J. Matković*, K. Fertalj **
* Elektroprivreda HZ-HB d.d. Mostar/Sektor za informatiku i telekomunikacije, Mostar, BiH
** Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva, Zagreb, Hrvatska
jelena.matkovic@ephzhb.ba

Abstract - This paper is focused on understanding techniques used to formally verify behaviour of business processes developed using WS-BPEL standard.

The first part of the paper gives an overview of WS-BPEL control-flow constructs. We analyze offered constructs and try to derive control-flow scenarios that can be built out of them. Derived complex control-flow scenarios are further discussed. Possible execution misbehaviour, which cannot be discovered by pure static analysis of WS-BPEL code, is identified. Misbehaviour can be any of the following: unreachable code, deadlock states, non-progress cycles (code is repeated without passing through progress state), or any other unwanted scenario, like unwanted execution order of activities, unwanted order of messages being exchanged and similar execution misbehaviour.

The second part of the paper describes techniques used for formal verification of any code in general with an emphasis on research how these techniques are exploited to verify WS-BPEL code. The final part of the paper concludes with the discussion how successful described work was in verification of possible identified misbehaviour from the first part of the paper.

I. INTRODUCTION

WS-BPEL [1] (in further text BPEL) is a main XML based paradigm that can be used to combine multiple, stateless web services into complex workflow processes. Even when there is only sequenced behaviour present, workflow structures can contain many complex paths that can be hard to compute by heart. When parallelism is added to the behaviour of the workflow system, situation becomes more complex and additional techniques are required to check the behaviour in early phases of the development. Techniques for formal verification come in handy in these situations.

This paper is concentrated on techniques used for formalization of BPEL processes (*called orchestrations*) and tries to give an overview of the existing work related to it. The paper is organized as follows. Section II presents offered BPEL activities and concludes with potential problems that can occur inside of BPEL executions. Section III introduces mechanisms used for formal verification in general. Section IV introduces two approaches related to formal verification of BPEL processes using mechanisms described in section III and it gives a comparative analysis of those approaches. Section V concludes the paper.

II. WS-BPEL ACTIVITIES

A. WS-BPEL basic activities

WS-BPEL (*BPEL in further text*) is widely accepted paradigm for integrating software applications (*exposed as services*) into complex business processes that span within and across organizational boundaries. Interactions in BPEL processes are expressed as compositions of communication activities (*invoke, receive, receive/reply, invoke/receive*) used for modelling different communication patterns between BPEL processes and integrated services (*in further text partner services*). Communication between BPEL processes and partner services can be *one-way*, *synchronous request-response* and *asynchronous request-response*. *One-way* pattern is utilized when a BPEL process invokes *one-way* operation of a partner service or when a partner service invokes *one-way* operation of a BPEL process. Synchronous *request-response* communication is utilized when a BPEL process invokes *request-response* operation of a partner service by sending message to a service, waiting for a response and receiving the response. Asynchronous *request-response* communication is utilized when a BPEL process invokes *request-response* operation of a partner service but it continues with its execution and does not wait for a response. In order for asynchronous communication to be performed, a BPEL process must have registered *one-way* operation that is used to asynchronously receive a response from a partner service when the response is ready. Communication activities come under basic BPEL activities together with *empty* (*do nothing*), *wait*, *throw* (*generate error*), *compensate* (*invoke compensation handler*) and *exit* activities.

B. WS-BPEL structured activities

Besides pure message exchange scenarios between BPEL processes and their partner services, the idea of BPEL is to capture business logic as well. This is carried out using structured activities that allow parallel, sequential and conditional execution, event and exception handling, fault and compensation handling and data manipulation.

Sequence activity allows all its child activities to be performed sequentially in the order in which they are

written. *Flow* activity allows all its child activities to be performed in parallel. Besides parallelism, flow activity supports conditional execution of its child activities using *links* that synchronize execution of a *source* activity and a *destination* activity. A destination activity cannot execute unless at least a source activity executes successfully. *Links* mechanism allows construction of extremely complex net of conditional executions. Cyclic dependencies are forbidden but everything other is allowed.

Pick activity is used for selective event processing. It contains several branches each of which corresponding to the occurrence of a specific event. Which branch is going to be selected and executed depends on the type of the event that has been raised. After the event raises and its corresponding branch starts its execution, other events can no longer be accepted by that pick activity. Types of the events accepted by the pick activity are: message events (*receipt of a special message type*) and time events (*expiration of a specific time-out*). *ForEach* activity executes its child scope activity N times sequentially or in parallel.

Other structured activities are well-known from most programming languages and thus they do not require detailed introduction: *switch* (*conditional routing*), *if* (*conditional behaviour*), *repeatUntil* (*repeatable execution*) and *while* (*repeatable execution*).

Scope is a special structured activity that enables segmentation of business logic into special units used for attaching event, fault and compensation handlers related to that business logic segment. Attached handlers contain logic which is executed when an event or a fault inside of a scope occurs and they are executed in parallel or after the termination of the scope. *Event handlers* are related to processing of an event and they execute in parallel with the scope. Events can be exceeded timeout or receipt of a specific message type. Time event handlers can have only one active raised instance (when the timeout is reached), while message event handlers are raised upon every receipt of a related message type and can have many raised instances during scope execution. Special type of time event handlers (*repeatEvery*) can have multiple active instances raised in parallel with the scope and they occur repeatedly after every duration period exceeds.

Fault handlers are related to processing an error that occurred in the execution of a scope they are attached to. When a fault occurs, it is caught by the corresponding fault handler and when that happens, the scope switches from its normal processing mode to its fault handling mode. Fault handlers can be defined for a custom fault type explicitly or they can be defined implicitly as default fault handlers. If faults are not handled successfully by fault handlers, they are re-thrown to the parent scope. The scope, in which a fault occurred and whose fault handler has been activated, are considered to have ended abnormally and they cannot be compensated.

Compensation is an odd term introduced with the BPEL specification. Compensation refers to application-specific logic that attempts to undo the work of a previously successfully completed scope. As with fault handlers, *compensation handlers* can be defined explicitly or implicitly. Compensation handler of a successfully completed scope is invoked by *compensate* activity from a fault handler or a compensation handler attached to the parent scope.

Regular termination of a scope begins when fault handler of a scope starts its execution. Termination of a scope means that all activities contained within it must be terminated. Forced termination of a scope begins when there is a need for terminating scope that is in a normal running mode. *Termination handlers* can be attached to a scope in order to override default termination handlers.

Control links (*mentioned previously*) can cross boundaries of all of these structured activities excluding repeatable constructs *repeatUntil*, *forEach*, *while* and *compensation handlers*.

C. WS-BPEL data manipulation activities

BPEL processes exchange data with their partner services. Exchanged data are manipulated inside of executions and can be used to identify the right instance of a BPEL process. *Variables* are constructs used for storing data and they can be defined as *XML types*, *XML elements* and *WSDL message types*. *Query languages* are used for retrieving parts of variables to get or set their values. Data manipulation is performed using *expressions*. *Assign* activity is used to assign values to variables using *data assignments*. It can be used to copy values from variables to variables or to insert new values generated using data expressions. Default query and expression language in BPEL is *xPath*, but it is allowed to define own query and expression languages. BPEL specification defines rules for binding WS-BPEL variables in *XPath* and it also defines some restrictions for usage of *XPath*.

D. Misbehaviour of WS-BPEL processes

BPEL processes are workflow processes that can have a complex net of parallel and conditional behaviour which is the main reason for trying to formalize BPEL processes. Conditional behaviour is introduced with the notion of links whose behaviour is going to be described in more detail now. Link connects a source and a target activity (Figure [1]). After the source activity completes, transition condition of the link (*Boolean expression*) is evaluated and its return value is set as a link status. After the link status is set, it is 'forwarded' to the target activity. When the target activity is ready for execution, it must wait to 'receive' statuses of all of incoming links. After all those statuses are received, join condition of the target activity (*Boolean expression*) is evaluated to determine if the target activity will be executed or not. Links can be used to develop extremely complex net of conditional behaviour and it is very hard and annoying to draw a map of all possible executions in order to verify them. It is much easier to use help of some formal verification techniques to do so.

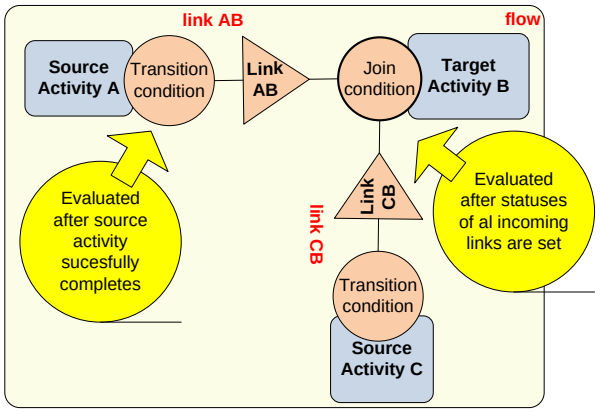


Figure 1. Links mechanism in WS-BPEL

Parallel behaviour is introduced with *flow*, parallel *forEach* and *event handler* activities. When parallel behaviour is present, it is important to verify all possible executions to see if there are executions with unwanted race conditions. Race conditions occur when shared resources are manipulated concurrently which can lead to possible data corruption. If race conditions are allowed, then semaphores are introduced to control access to shared resources. But still it is necessary to check if there are induced deadlocks or process starvations. An example of an unwanted race condition in BPEL processes would be a situation when some of variables referenced by the transition condition are modified in the concurrent path. As opposed, if this race condition is allowed, then the result of the transition condition evaluation may depend non-deterministically on the timing of behaviour among concurrent activities [1]. An example of a strongly forbidden (*by BPEL specification*) race condition is an existence of competing message consuming activities (*two or more receive activities that consume the same message must not be concurrently enabled*). An unwanted race condition would be also a situation when one parallel component initializes shared resource while the other reads it (*and reading of unutilized resource is erroneous*). This could lead to situations that one parallel component is trying to read a resource while the other parallel component did not finish an initialization of the resource. Example of this scenario in a BPEL process is a try to read a *partnerlink* before its *partnerRole EPR* is initialized which results with an error.

It is of big importance to formalize BPEL executions, in order to check absence of deadlocks, unwanted race conditions, process starvations and of any other system misbehaviour and to see if computational correctness of the process is preserved in its every execution.

III. FORMAL VERIFICATION

Formal verification is performed to check properties of or to simulate executions of distributed systems. The term '*distributed system*' refers to logically or to physically distributed systems. Logically distributed systems comprise parallel and asynchronous functional behaviour that shares common resources. Formal verification is applied not to distributed systems only but to sequential systems that contain complex synchronized functional behaviour.

Testing is a traditional way of verifying behaviour of a system but without 100% correctness. Classical testing techniques of modular systems are performed in such a way that functionality of each module is tested individually. After testing functional correctness of each module, integration testing is performed. If modules are executed sequentially and system's behaviour is not complex, then classical testing techniques of verifying functional correctness can be satisfactory, but even in these cases, formal verification methods can be complementary techniques to classical testing techniques especially when behaviour of the system is complex.

But what if functional modules of the system do not execute exclusively sequentially but there is parallel execution as well?! Parallel execution introduces new situations that can lead to computational incorrectness. Some of them have already been mentioned in the paper and those are: race conditions, data corruptions, delays, process starvations, deadlocks etc. With parallel and asynchronous behaviour state space can be extremely large so it is almost impossible to compute all possible combinations of speeds of parallel executions. Classical testing techniques will pass through some of possible executions but not through all and that is the main reason for appliance of formal verification techniques. Drawbacks of formal verification techniques are their computational complexity, time and memory consumption and not being user-friendly but they can give an answer with 100% correctness.

A. *Formal verification using finite state automata*

Finite state automata [2] and algorithms for traversing their states [3] form a strong mathematical basis upon which systems can be formally described and analyzed (*Figure 2*, *Figure 3*). In order to simulate or verify behaviour of the system, it needs to be formally described first. The first step of formal description is identification of parallel communicating components inside of the system's behaviour. After they are identified, each parallel component is expressed as a standalone extended finite state automation using *Promela* syntax. *Promela* is a language syntax used by Spin [4] which is the most

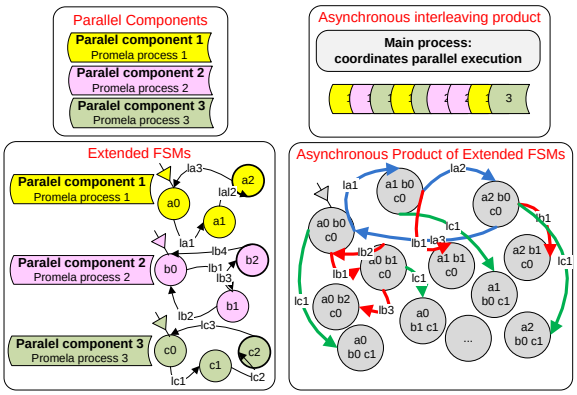


Figure 2. Mathematical basis of representing system's behaviour used in Spin

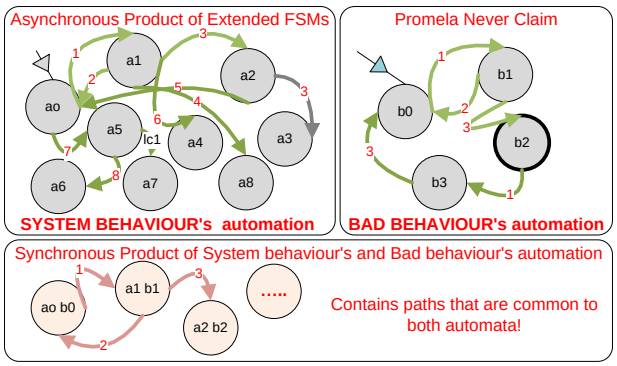


Figure 3. Mathematical basis of finding bad behaviour used in Spin

popular tool for formal verification based on finite automata. In order to express peace of code in *Promela*, it is important to define mapping rules from source syntax into *Promela* syntax which is the main challenge. When *Promela* processes are ready, they are executed in parallel and asynchronously, exchanging data using global variables or message channels. This is called *asynchronous interleaving product* and it corresponds to asynchronous product of extended finite state automata. A pure finite state automation of the system (*its whole state space*) is generated from the *asynchronous interleaving product* by its execution. Algorithms for simulation and verification [3] operate on pure finite state automation of the system (*state space of the system*).

1) Checking properties on finite state automata

Algorithms that traverse *state space of the system* are used for checking two types of properties: *liveness* and *safety* properties. *Liveness* properties must eventually be satisfied (*system eventually needs to be in an accepting state*) while safety properties must always be satisfied (*every state of the system needs to be an accepting state*).

Nested depth-first search algorithm (NDFS) is used for checking if *liveness* properties are satisfied and the procedure for its appliance is described hereafter. First, finite state automation of bad behaviour (*which contains cycles only with states where liveness property is not satisfied*) is specified using linear temporal logic (LTL) and afterwards it is transformed into *Promela* syntax (*called Promela never claim*). *Promela never claim* is executed synchronously with the finite state automation of the system. Synchronous execution of two automata corresponds to synchronous product of automata. A synchronous product of automata is a new automation that contains only those executions that are common to both automata. So after a synchronous product of the system's automation and *Promela never claim* is produced, *nested depth-first search algorithm* [3] is performed over the resulting synchronous product automation. Accepting states in the synchronous product automation are taken from *Promela* never claim, and those are exactly the states where a liveness property is not satisfied. Nested depth-first search algorithm checks if there are accepting states reachable from the start state and then it checks if those are reachable from itself. If there are such states, it proven that there is an example of the system's execution path where system never traverses an accepting state, i.e. it returns an example of the finite (*with stuttering extension*) or the infinite execution where a *liveness* property is not satisfied.

Some states in the system's automation can be marked as progress states. Progress labels are used to mark states that must be traversed eventually in every infinite execution of the system. The system is correct if it always eventually passes through the progress states and the system is not correct if there are non-progress cycles. Traversing progress states eventually is an example of a *liveness* property and it is checked in above described manner.

Mechanism for checking safety properties is the same as one used to check *liveness* properties, except a different algorithm over synchronous product is used. Safety properties are checked using regular *depth-first search algorithm* (DFS) [3] and can be checked on finite (*no stuttering extension needed*) and infinite executions. *Promela never claim* automation (*which contains violation of a safety property*) describes finite behaviour containing states where safety property is not satisfied. *Depth-first search algorithm* over synchronous product finds those states where safety property is not satisfied (*states where Promela never claim automation ends with its execution*).

2) Default checkings

When checking *liveness* or safety properties of the system, existence of deadlock states is implicitly checked as well. Deadlock states are the states in which the system cannot make a transition to any other state whilst at the same time that space is not marked as an end state. Behaviour of the system (*its extended finite state automation*) expressed as an interleaving asynchronous product of the parallel components (*asynchronous product of component extended finite state automata*) is considered to be in an end state if every parallel component s in its end state and all exchanged messages between them are consumed. Default end state of every parallel component is its curly brace, but it is allowed to define other states as end states.

Existence of unreachable states is implicitly checked as well when verifying liveness or safety properties of the system. Unreachable states represent the states of the parallel components that have never been traversed in the interleaving asynchronous product.

B. *Formal verification using Petri nets*

Petri nets [5] are graphical mathematical model for formalizing behaviour of parallel and asynchronous systems. Petri nets are applied to many systems because of their generality and tolerance. Any workflow system that contains parallel activities can be described by Petri nets. Complexity of resulting models is one of the main disadvantages of Petri nets. There are many offered tools for drawing, analysis and simulation of Petri models. Petri nets are nothing but directed, weighted graph with two types of nodes: *places* and *transitions*. *Arcs* are used to connect nodes (*places to transitions and transition to places*) and they are marked with their *weight factor* which is a number of parallel arcs. Places are marked with non-negative integer numbers called *tokens*. A marking of

a net is M-vector, where M is a number of states and M(p) is a token of a state p. States represent conditions for events to happen while transitions represent actual triggering of events. Every transition has its input and output places representing its pre-conditions and post-conditions. The main rule of Petri nets is the *rule of transition enabling and firing* which is used to simulate dynamical behaviour of the system. The rule of transition enabling and firing functions in the following way:

- Transition t is enabled if every input state p of transition t is marked with at least w (p,t) tokens where w(p,t) is a weight factor of an arc leading from p to t.
- Enabled transitions t can fire but it does not have to which depends on the actual occurrence of an event.
- Firing of enabled transition t removes w (p,t) tokens from every input state p of transition t and adds w (t,p) tokens to every output state p of transition t, where w (t,p) is a weight factor of an arc leading from t to p.

Petri nets can be used for different kinds of analysis since there are different interpretations of places and transitions (Table 1). One of the most common problems that can be answered with Petri nets is a *reachability* problem. The firing of an enabled transition will change the token distribution in a net according to the transition rule. An initial marking of a net is denoted with M_0 .A marking M_n is said to be reachable from M_0 if there exists a sequence of firings that transforms M_0 to M_n. The set of all possible markings reachable from M_0 forms its state place. The reachability problem for Petri nets is the problem of finding if M_n is in the set of all possible markings reachable from M_0.

The concept of *liveness* in the terms of Petri nets is closely related to the absence of deadlocks. A Petri nets is said to be live if no matter what marking has been reached from M_0, it is possible to ultimately fire any transition of the net by progressing through some further firing sequence which means that a live Petri guarantees deadlock-free.

Boundness is also checked with Petri nets. A Petri net is said to be bounded if for every place in every marking, the number of tokens does not exceed some finite number. A Petri net is safe if it is 1-bounded. *Reversibility* of a Petri net is its possibility of reversing to some home state from any marking (*any state*). *Persistence* of a Petri net assumes that for any two enabled transitions, the firing of one transition will not disable the other. A transition in a persistent net, once it is enabled will always stay enabled until it fires and this kind of nets is used in the context of analysing parallel systems. For more detailed explanation of Petri nets mechanism, the reader is referred to [5].

IV. FORMAL VERIFICATION OF WS-BPEL PROCESSES

Chapter III described techniques used for formal verification of workflow systems in general, while this chapter analyzes work related to formal verification of BPEL processes using techniques described in chapter III. The first part of the chapter describes work related to formalization of BPEL using finite state automata, while the second part of the chapter describes formalization of BPEL using Petri nets.

A. Formal verification of WS-BPEL using Petri-net

In [6] authors verified execution of standalone BPEL processes using Petri-nets. Petri-net model is obtained using developed mappings from BPEL control flow constructs solely into Petri-net constructs. Authors abstracted data manipulation aspects of BPEL execution, so their work has a goal to verify control flow solely. Mapped control flow construct are basic activities and structured activities (*sequence, flow, switch, pick, while*).

Scope, as a special structured activity used for event and exception handling, was also mapped to Petri-net. It is worth mentioning that mappings for scopes were defined even with event handlers, fault handlers and compensation handlers attached to them. Mapping for event handlers were made with the limitation that bounded number of message event handlers can be raised in parallel with the scope. They introduced this limitation in order not to get an unbounded net which is difficult to manage and which requires computationally expensive techniques.

When considering fault handlers, there are defined mappings for execution of all fault handling scenarios: scopes with general fault handlers, scopes with fault handlers related to process defined faults (*generated using throw activity*), scopes containing activities with *suppressJoinFailure* set to 'no', handling a fault occurred within a fault handler itself, handling a fault occurred within a scope with event handlers attached and dead path elimination for control links leaving a fault handler.

When considering compensation handlers, there are defined mappings for handling compensation invoked by the compensate activity and mappings for handling faults which happened during execution of a compensation handler itself.

When considering termination of activities, there are defined general mappings for termination of basic and structured activities (*with both values of suppressJoinFailure attribute*). Special mappings were defined for termination of *pick* and *while* structured activities because of their specific behaviour. There are also mappings for terminating hierarchy of copes and terminating a process due to execution of an exit activity.

TABLE I. SOME TYPICAL REPRESENTATIONS OF PLACES AND TRANSITIONS [5]

Input places	***Transitions***	***Output places***
Preconditions	Event	Post conditions
Input data	Computation step	Output data
Input signals	Signal processor	Output signals
Resources needed	Task or job	Resource released
Conditions	Clause in logic	Conclusion(s)
Buffers	Processor	Buffers

All mappings were defined for normal and skipping behaviour of activities, with and without control links attached to the activities. Since authors in [6] analyzed control-flow constructs in detail and have developed complete mapping for them, their formalization can be used to understand and analyze possible control-flow behaviours of BPEL processes but data layer of BPEL processes is totally abstracted. Authors explained that their formalization can be used for three scenarios:

- *Detection of unreachable code* - They use two techniques to detect activities that will never be executed, namely *relaxed soundness* which explores the whole state space and *transition invariants* which reduce state space but whose results are not necessary complete.
- *Detection of competing message consuming activities* – Competing message consuming activities are those activities that are active at the same time and that are waiting for the same type of message to come to the process. BPEL does not allow this scenario since it produces a race between those activities and this is strictly not allowed. In order to detect the existence of such activities, the whole state space must be searched.
- *Garbage collection of queued messages* – For every basic activity they annotate which type of messages can be consumed after activity's completion. So after this kind of analysis is done, the result is annotated BPEL process with markings for every basic activity. These markings can be used by BPEL engine afterwards to delete messages in an input queue which cannot be consumed anymore.

B. Formal verification of WS-BPEL using finite state machines

In reference [7] authors formalized asynchronous communication among multiple BPEL processes (*named peers*). Every peer has an attached input FIFO queue which it uses to consume messages sent to it by other peers. Authors named this kind of communication among peers as a *composite web service*. Their formalization (*of a composite web service*) consists of a conversation schema (*set of peers and types of messages exchanged between them*) and of each peer's behaviour. They introduced a *virtual watcher* that records *conversation*, i.e. the sequence of messages exchanged between peers. Exchanged messages are XML files and their types correspond to XML schemas. For each peer p_i, its behaviour is formalized using guarded automata (*extended automata*) with the following components: set of all incoming and outgoing message types for the peer, set of local variables, set of all states, an initial state, final states, and set of transitions. Local variables are represented as XML files as well. For every peer there are three types of transitions: receive transitions (*the receiving message is consumed from the input queue*), local transitions (*an internal action inside of the peer is performed*) and send transitions (*the peer appends the sent message to the input queue of the receiving peer*). Guards of the transitions in guarded automata are expressed as *XPath* expressions.

Formal behaviour of each peer (*expressed as a guarded automation*) is generated using mapping rules from BPEL into guarded automata. Guarded automata of the peers are later translated into *Promela* processes using mapping rules from guarded automata into *Promela*. Input queues of peers are translated to global message channels which are used by *Promela* processes to communicate asynchronously. A bound is put on the sizes of the input queues in order to get limited state space.

Authors used LTL formulas to check whether certain properties are present in the behaviour of the system. Atomic propositions (*used in LTL formulas*) are exchanged messages and predicates over exchanged messages. LTL formulas express if order of exchanged messages between BPEL processes is as wanted or if the right message response is returned according to a message content being sent as an input.

C. Comparative analysis of the two approaches

It can be concluded that authors in [7] did not abstract

TABLE II. COMPARISON OF THE APPROACHES IN [6] AND [7]

Property	Approaches to formal verification	
	Petri net approach [6]	***FSM approach [7]***
Appliance to an individual BPEL process or to communicating BPEL processes	Individual BPEL process	Communicating BPEL processes
Sequence of exchanged messages	Not reported	Yes
Complete coverage of all offered controls flow constructs with correct syntax	Yes	Not described
Concurrency inside of one individual BPEL process (*flow child activities, multiple raised event handlers, parallel forEach..*)	Yes	Not described
Data component (*data manipulation, data usage inside of BPEL execution*)	No	Yes
Race conditions (*multiple activities, that share the same resources, are active*)	Not reported	Not reported
Deadlocks, competing message consuming activities, garbage collection of input queue of a BPEL process	Contributions of the work	None reported
Usage of formal models for complete or partial automatic correction of BPEL code	Not reported	Not reported
"Not reported" means that the property is not reported in the paper and it is assumed it is not implemented.		

data component of BPEL processes as authors in [6] who were focused on detailed control-flow analysis solely. But while focusing on inter-communication between processes and preservation of data component, authors in [7] did not pay much attention to detailed formalization of control-flow semantics of each BPEL process (peer). They did not describe how they solved concurrency inside of execution of each peer (BPEL process), nor they were interested in detailed formal verification of a single BPEL process which can have complex behaviour with multiple parallelism and conditionals executions. Thus their work can be applied to verify inter-communication (*without abstraction of data component*) of multiple BPEL processes which execute in parallel and communicate asynchronously, but with no parallel execution inside of each BPEL process.

On the other side, authors in [6] analyzed control-flow component of BPEL specification in detail and they made a complete formalization of control-flow semantics of BPEL specification but with complete abstraction of data layer. Work in [6] refers to an individual BPEL process. Table 1 shows comparison analysis of the two approaches.

V. CONCLUSION

This paper gave an insight into the nature of BPEL specification. It was reported that BPEL processes can comprise many concurrent and conditional behaviour which, if not written properly, can lead to unwanted race conditions, deadlocks, process starvations and other system failures. Business logic can also become corrupted.

In order to override these problems, formalization techniques are use to check if there are unwanted scenarios in the behaviour of BPEL processes. Two approaches are introduced in this paper and it is concluded that they solve some problems but they are incomplete. According to the results of the comparative analysis (*Table 2*), they are somehow complement to each other.

Future work related to BPEL formalization is going to analyze BPEL specification in more detail in order to collect and organize as many as possible types of BPEL misbehaviours. Suitable approach for formalization of standalone BPEL processes and detection of unwanted misbehaviours will be suggested.

LITERATURE

[1] Web Services Business Process Execution Language Version 2, OASIS Standard, April, 2007., Available at: http://docs.oasis-open.org/wsbpel/2.0/wsbpel-v2.0.html

[2] Finite-State Machines, Available at: http://www.cse.chalmers.se/~coquand/AUTOMATA/book.pdf

[3] On Nested Depth First Search, Available at: http://spinroot.com/gerard/pdf/inprint/spin96.pdf

[4] Software Model Checking with SPIN, Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.227.8263&rep=rep1&type=pdf

[5] Petri Nets: Properties, Analysis and Applications, Available at: http://embedded.eecs.berkeley.edu/Research/hsc/class.F03/ee249/discussionpapers/PetriNets.pdf

[6] C. Ouyang et. al., "Formal semantics and analysis of control flow in WS-BPEL," Journal of Science of Computer Programming, vol. 67, issue 2-3, 2007, pp. 162-198.

[7] X. Fu, T. Bultan and J. Su, "Analysis of interacting BPEL web services," in Proceedings of the 13th international conference on World Wide Web, New York, 2004, pp. 621 - 630.

Hybrid MPI-OpenMP parallel optimization for thermo-fluid problems

M. Depolli
Jožef Stefan Institute/Department of Communication Systems, Ljubljana, Slovenia
matjaz.depolli@ijs.si

Abstract

This article deals with the minimization of the energy transport through the isolated enclosure by means of obstructing the fluid flow. The differentially heated cavity is partially covered with non-permeable obstacles that affect the natural convection flow. We seek the optimal position and shape of the obstacles to minimize the heat flux through the domain with the minimal cumulative size of the obstacles, i.e. we maximize insulation with minimal material needed. The task is performed by evolutionary multi-objective optimization of a local thermo-fluid solver, which comprises energy transport and incompressible fluid dynamics. An effective parallel algorithm is presented, in which the physical model is solved by a local meshless numerical method on shared memory architecture, coupled by optimization algorithm that exploits the distributed computer parallelism.

I. INTRODUCTION

In this article, a parallel simulation-based multi-objective optimization, i.e. the optimization is done on a simulated results, is presented. The principal incitement in this work is a minimization of energy losses in a differentially-heated air-filled cavity by placing obstacles in the cavity. The cavity is cooled on the top and heated on the bottom edge (Dirichlet boundary conditions), with left and right walls perfectly isolated (Neumann boundary conditions). The simulations begin with a stationary fluid, but the instable configuration of the case (heavier fluid on the top of a lighter one) soon leads to a pronounced natural convection flow patterns. The energy transport over the domain, i.e. energy losses, is therefore not governed only by diffusion but also convection. The most straightforward solution of the problem at hand would be to simply fill the whole domain with a good non-permeable insulator. However, our goal here is to minimize both the energy losses and insulating material consumption, therefore we only partially fill the domain with non-permeable obstacles to alter the air flow.

There are numerous examples of similar systems where one would be highly interested in such optimization, e.g. designing windows or other insulating elements for buildings, optimizing heat storage systems, optimizing heat distribution within rooms, etc.

If we want to solve the problem within reasonable time constraints, we need an effective implementation of numerical simulation and optimization tools. Solving the problem faster can be can be achieved by harnessing more processing power. The available processing power can be increased either by increasing the processing unit's clock frequency or by increasing the number of processing units. The clock frequencies are approaching their physical limits, therefore increasing the number of processing units is a more viable solution. Parallel processing, available today in most desktop computers or computer servers, can compensate for the lack of performance of a single processing unit only if an efficient parallelization of the computational method is known. Various application programming interfaces (APIs) for parallel programming are used to maximize the performance of parallel processing. Nowadays, the most widely used APIs for parallel programming are MPI for distributed-memory systems, and Ptreads and OpenMP for shared-memory systems [1]. There are several publications regarding the parallelization of different numerical schemes for various applied problems [2-6].

In this paper we present the efficiency of a shared-memory parallel simulator coupled with an evolutionary optimization on a modern cluster of multi-core computers. The numerical simulation is built on the popular variant of local point interpolation meshless method [7] where the collocation with the radial basis functions is used [8] combined with a local pressure-velocity [9] coupling to compute solenoidal flow.

Optimizer implements the AMS-DEMO algorithm [10], which is an evolutionary algorithm [11] for multi-objective optimization [12] of real-valued functions. Evolutionary algorithms are a subgroup of stochastic optimization algorithms [13] and can solve problems, for which only the cost (or evaluation) function is known.

II. PROBLEM DEFINITION

This paper deals with a straightforward natural convection simulation that includes energy transport, Navier-Stokes equation, mass continuity and Boussinesq approximation. There have been several numerical solutions of that type of problems published in the past [14-17]. The domain is a closed air-filled cavity with differentially heated horizontal walls with temperature difference ΔT and insulated vertical walls. Walls are Non-permeable and no-slip, i.e. velocity on all boundaries is set to zero. The problem comprises three coupled Partial Differential Equations (PDEs): mass (1), momentum (2) and energy conservation (3) equations and the Boussinesq approximation (4) used for the body force in the

momentum equation. The problem is thus stated with the following system

$$\nabla \cdot \mathbf{v} = 0 \,, \tag{1}$$

$$\rho \frac{\partial \mathbf{v}}{\partial t} + \rho \nabla \cdot (\mathbf{v}\mathbf{v}) = -\nabla P + \nabla \cdot (\mu \nabla \mathbf{v}) + \mathbf{b} \tag{2}$$

$$\rho \frac{\partial (c_p T)}{\partial t} + \rho \nabla \cdot (c_p T \mathbf{v}) = \nabla \cdot (\lambda \nabla T) \tag{3}$$

$$\mathbf{b} = \rho [1 - \beta_T (T - T_{\text{ref}})] \mathbf{g} \tag{4}$$

where $\mathbf{v}(u, v)$, P, T, λ, c_p, $\mathbf{g}$, ρ, β_T, T_{ref}, μ and $\mathbf{b}$ stand for velocity, pressure, temperature, thermal conductivity, specific heat, gravitational acceleration, density, coefficient of thermal expansion, reference temperature for Boussinesq approximation, viscosity and body force, respectively.

The problem at hand is solved with a strong-form local meshless numerical method [18], by a methodology that has been found a reliable and effective tool for solving complex multiphase problems [19-21]. The basic idea of the method is to use local approximations of the considered field to evaluate differential operators. The size and shape of local support, i.e. the subset of computational nodes used in the local approximation, is not restricted in any way. Although there is no restriction regarding the dimension of the system, we use the collocation approach, i.e. the number of support points is the same as the number of basis functions [22]. The temporal discretization is done through a two-level explicit time stepping.

To maintain the locality of the solver we employ the local pressure-velocity coupling [9], where instead of solving pressure correction Poisson equation, the pressure is simply corrected with the divergence of intermediate velocity. The local pressure-velocity coupling is similar to the artificial compressibility method that has been recently under intense research [23, 24] and also to the SOLA approach [25].

III. Simulation-Based Optimization

The obstacle setup is optimized by the means of evolutionary algorithm AMS-DEMO (Asynchronous Master-Slave Differential Evolution for Multi-objective Optimization) [10]. AMS-DEMO is a stochastic algorithm that works on a fixed-size set of solutions. Solutions are encoded as positions and sizes of the five obstacles (four floating point numbers per obstacle). Each solution is assessed on two criteria. First is the ratio of the domain space that is covered by obstacles, and second is the average vertical heat flux in the domain, normalized by the heat flux through an empty domain. Criteria are to be minimized, that is, minimal amount of material is to be used, and the energy losses generated by heat flux are also to be minimized.

Starting set of solutions (encoded as vectors) is randomly initialized, all further solutions are recombined from existing solutions. Optimization procedure consists of iterations over the whole population, where each solution is selected to act as a parent in recombination and crossover procedures once. Recombination creates a candidate solution, by calculating the vector difference between two randomly selected solutions, scaling it, and adding it to a so called parent solution. Crossover is then applied between the parent and the candidate, which has a predefined chance of selecting individual vector elements from either the candidate (crossover probability) or the parent (*1* – crossover probability), forming a new candidate solution. After formation, each solution is evaluated via simulation under both criteria and added to the working set of solutions (the population). Optimization procedure keeps the best solutions in population while it discards the rest. Best solutions are selected under the Pareto dominance criteria, which states that one solution is better than the other (dominates the other), when it is better under one criterion and equal or better under all other criteria. If given a pair of solutions, neither one solution dominates the other, the pair is said to be incomparable. A set of incomparable solutions, for which no known solution dominates any of them is said to be the Pareto optimal set or Pareto optimal front. The multi-objective optimization algorithm returns the Pareto optimal set as the result of optimization. If more solutions are incomparable than the population size, then the crowding distance [26] is calculated for each solution and the solutions with the lowest crowding distance are removed.

IV. Parallelization

The simulation-based optimization principle allows for two complementary parallelization principles to work together. The meshless simulator can be very efficiently parallelized for multi-core shared-memory machines, and can achieve even super-linear speedups [27]. It is less suitable for parallelization on distributed-memory systems, such as computer clusters, where the parallelization is more complex and in much less efficient. Therefore for the presented setup, the meshless simulator is limited to run on all cores of one computer but not on multiple computers.

AMS-DEMO, on the other hand, excels at minimizing the data dependencies of an evolutionary algorithm, and uses asynchronous message-passing to achieve near-linear speedups even on heterogeneous computer systems. It is therefore set up to use all the available computers but run only one instance of evaluation function – the meshless simulator – per computer.

Combining both algorithms on a cluster, AMS-DEMO efficiently feeds distributed machines with jobs, while individual machines efficiently employ all their cores for individual simulations. The result is an excellent usage of all available processing resources, subsequently enabling very high speedups. Otherwise impractical job of optimizing a complex fluid-flow problem via full simulation is thus feasible on a computer cluster.

V. Results

The experiments were run on a computer cluster, composed of 32 worker computers with a single quad-core Intel Xeon 5520 processor each. Although the computers are interconnected with 8 Gigabit Ethernet connections,

only one connection was used, with all the used computers connected to the same switch. The number of used computers varied between 1 and 32 for the experiments. Since the optimization algorithm is stochastic, each experiment was repeated 3 times to ensure the consistency of the results.

The simulations were left to converge to a steady state, characterized by the stopping criteria – temperature change between two consecutive time steps is lower than 0.01% in all nodes – or to run for a limited amount of time and then return an infeasible result, assuming the natural convection resulted in either chaotic or periodical solution [28]. The non-stationary solutions are excluded from consideration.

A. Simulation results

The simulation starts with fluid at rest and at constant temperature. Then the lower wall starts heating while the upper wall starts cooling down the surrounding fluid. The changes in temperature affect the density of the fluid resulting in a natural convection flow. In Figure 1 the steady state in the obstacle free domain is presented. The principal characteristics of flow can be seen, i.e. a global vortex, fueled by the differences between the top and bottom wall temperatures, fills the majority of the domain. Adding the obstacles obstruct the flow is obstructed (Figures 2 and 3) and consequently the convective part of the energy transport diminishes. By positioning the obstacles in an clever way, the heat flux over the domain can be greatly reduced at the relatively low cost in terms of material consumption (coverage).

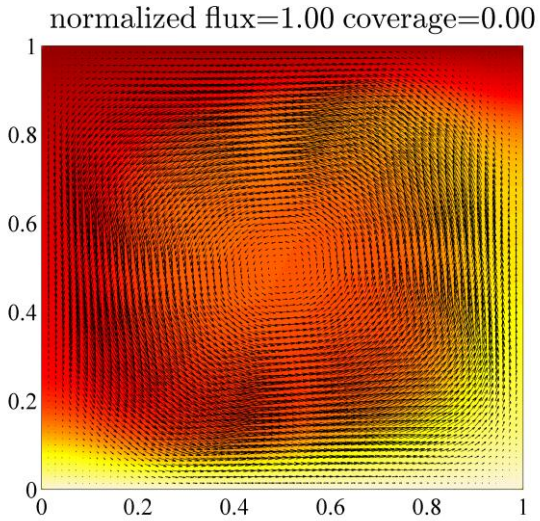


Figure 1. The temperature distribution marked by color and fluid flow marked by arrows inside the domain without obstacles.

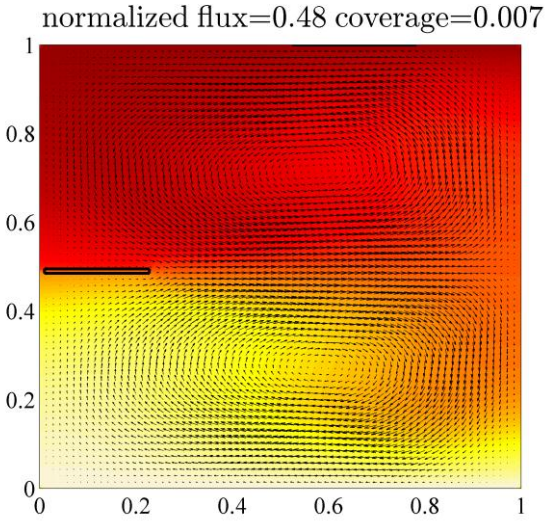


Figure 2. The temperature distribution marked by color and fluid flow marked by arrows inside the domain with obstacles positioned for the flux lowered by about half.

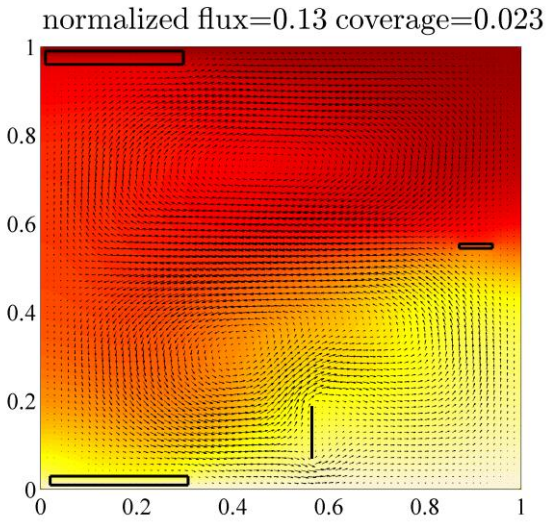


Figure 3. The temperature distribution marked by color and fluid flow marked by arrows inside the domain with obstacles positioned for the minimal flux.

B. Optimization Results

Optimization settings were selected by trying out several parameter sets and selecting the one that worked best on a computationally simplified test. The following optimization settings were selected:

- Population size of 50;
- Stopping criterion – 100.000 performed evaluations;
- Rand/1/bin scheme for differential evolution [12];
- Crossover probability 0.2;
- Scaling factor 0.5;

First, the convergence of the optimal solution set obtained by the AMS-DEMO optimization, is explored. To visualize the convergence of solutions, hypervolume indicator [12] is used to measure the quality of a set of solutions. In the given maximum number of evaluations, the non-dominated set of solutions mostly converged, as the hypervolume indicator shows in Figure 4. Although

the limits to which the solutions converged seem very different from the hypervolume plots alone, the plot of final non-dominated fronts in Figure 5 shows a different picture. A large portion of solutions – with the normalized flux of 0.2 or more – is covered quite similarly in all experiment repetitions. Only those solutions with the flux lower than 0.2 differ greatly among the runs. Half of the runs (three) do not find good solutions for low fluxes while the other half seem to have no such problems.

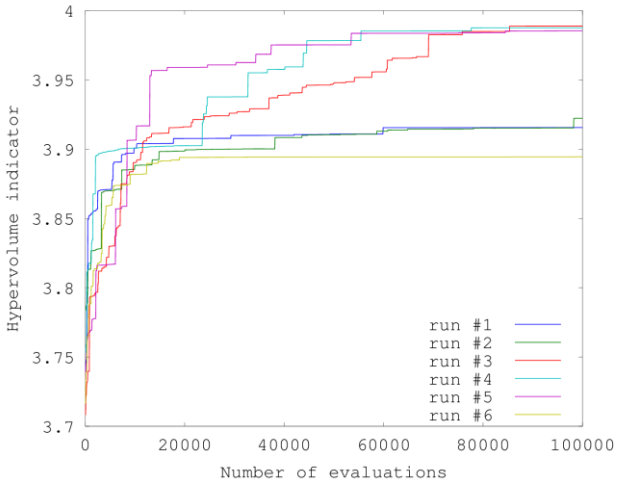


Figure 4. The hypervolume indicator for 6 test runs of the AMS-DEMO algorithm. Clearly visible are 2 sets of 3 runs, hypervolume of one set converging to clearly higher values than the other.

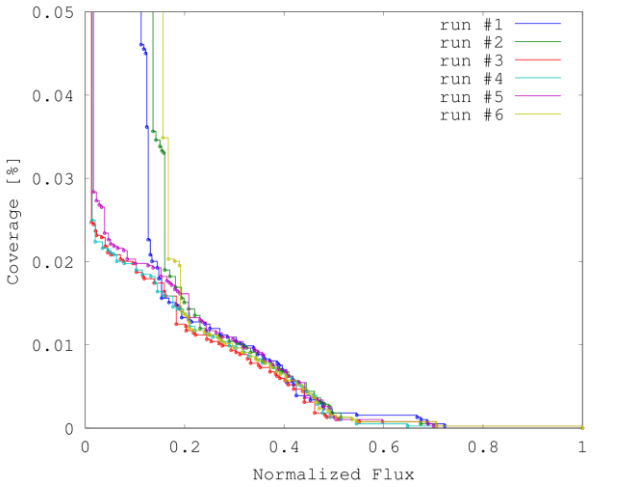


Figure 5. The final nondominated fronts of solutions (after 100.000 evaluations) for the same test runs as in Figure 4. The fronts make it clear that the runs, where hypervolume converged to lower values, had problems exploring the solutions with flux < 1.5 W/m.

Figure 6 shows a single algorithm run from another perspective – as the coverage of the dominated objective space with progressing number of performed evaluations. Dominated objective space contains all possible solutions that are dominated by at least one known solution. As can be seen from the figure, the optimizer quickly finds solutions that reduce the heat flow by about a half, but it takes a much larger number of evaluations after that for the solutions to evolve to allow only very low flow.

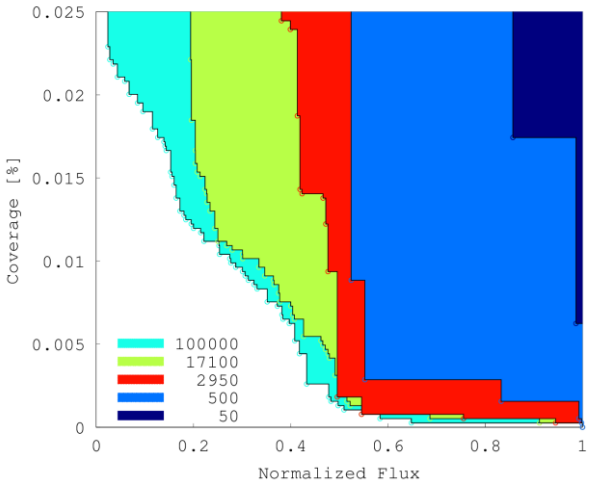


Figure 6. The evolution of the dominated objective space, for five checkpoints within a single optimization run. Checkpoints are made at after each shown number of completed evaluations. Spaces are plotted as overlapping polygons, with only their bottom-left border visible, dotted with solutions marked as colored circles. Note that theoretically, objective spaces extend to infinity at the top and right sides of graph.

We also compare the solutions according to their parameters, that is, we look at the obstacles within the domain for a single final non-dominated front of solutions in Figure 7. Patterns of the obstacle positioning are immediately visible.

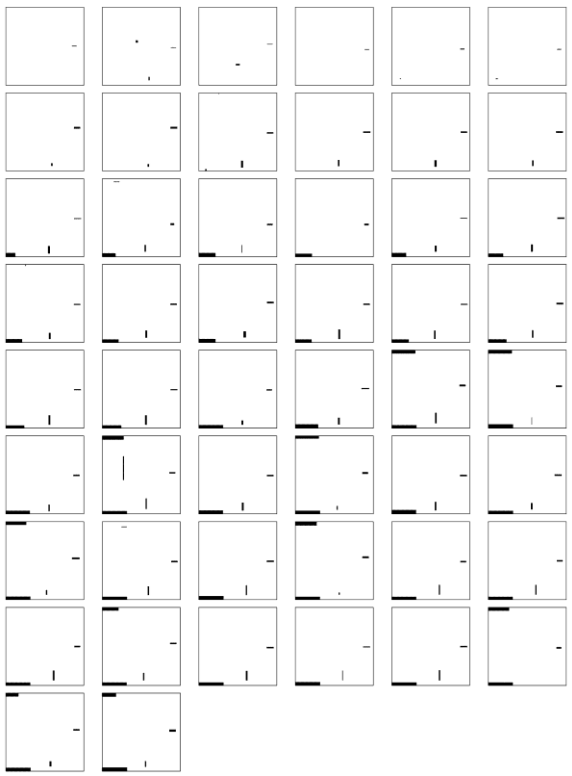


Figure 7. The nondominated solutions that were found during a single optiization run. Several patterns of obstacle positioning are clearly visible.

C. *Execution performance*

In this work, we are mostly interested in the efficiency of the parallelization, which we measure through parallel speedup. We divide parallel speedup into two categories,

one is the speedup of the simulator using all the cores of a single computer, and the other is the speedup of the optimizer, which distributes multiple simulator runs across multiple computers.

We measure the speedup of the simulator on 4 cores compared to running on a single core. For the selected domain settings of the simulator, we measure its parallel speedup on 4 cores to be 3.34. This is the speedup of the simulation, including also the serial procedures, such as reading from input, writing to output, and several other pre- and post-processing steps. For the calculation of speedup, we measure the whole execution time of the simulator executable, since the optimization procedure comprises of multiple calls to it. The simulator is otherwise capable of higher speedups, especially when longer simulations take place, reducing the relative ratio of the serial procedures. Speedups on multiple cores can, supplied an architecture with non-shared caches, even become super-linear [27], but do not in the presented case, since the non-shared caches of the used architecture are very small, compared to the number of nodes in the simulated domain.

Speedup of the optimization is much harder to calculate, since the parallel optimization does not traverse the same search space as the serial optimization does. As shown in [10], the convergence ratio of the optimization procedure is inversely correlated to the number of parallel worker processes that are executing evaluations. In the presented case, the number of parallel worker processes equals the number of computers. It is therefore expected that the convergence rate is lower for each added computer, yet measuring it exactly is much harder. In [10], it is shown that since the optimization is stochastic, the change of rate is only statistically significant when a large number of runs are made (more than 20), and when the number of parallel evaluations approaches the population size. For the presented case, since the population size is always larger than the number of computers and only a handful of repetitions are made for each test, the potential drop in the convergence rate is undetectable. Note that testing convergence rate more precisely is very time demanding, since it requires repeating the referential test on a single computer, which takes extremely long time to complete. Figure 8 shows the average parallel speedup calculated from 3 repetitions of an experiment. Since the convergence rate is probably reduced but we are not sure by how much, we only show the weak-speedup [10], which only takes the number of performed evaluations into the account, but not the convergence ratio. The jagged appearance of the graph is due to the stochastic nature of the optimization, mixed with varying-length simulation, and the limited number of repetitions.

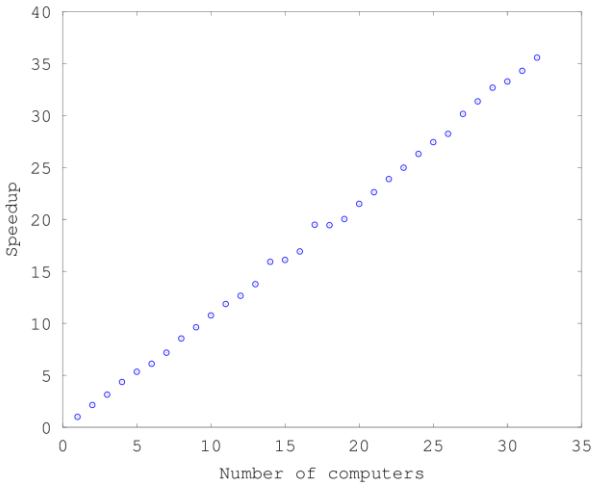


Figure 8. Weak speedup of the parallel optimization.

Altogether, the weak speedup of simulation based optimization is 123 at 32 computers with 4 cores each. The efficiency of such setup is almost ideal.

VI. CONCLUSIONS

In this paper we demonstrate parallel execution performance of coupled simulation and evolutionary optimization algorithm. From the execution point of view, the coupling bonds very well when executed on a computer cluster, with each application employing parallelization on a different hardware layer. On a 128-core cluster, a two orders of magnitude lower execution times were achieved than would be on a single core.

The execution time plays a crucial role in realistic optimization processes. For successful optimization process bigger optimization populations, as well as more dense computational nodal distributions in simulations are required. The execution time increases dramatically with refining those two parameters. In this paper we demonstrated the scalable Hybrid OpenMP/MPI parallel implementation of simulation based optimization that could be employed also to more complex optimization cases.

REFERENCES

[1] Pacheco PS. An Introduction to Parallel Programming. Burlington: Morgan Kaufmann Publishers; 2011.
[2] Singh V. Parallel implementation of the EFG Method for heat transfer and fluid flow problems. Comput Mech. 2004;34:453–63.
[3] Zhang L, Wagner GJ, Liu WK. A parallelized meshfree method with boundary enrichment for large-scale CFD. J Comput Phys. 2002;176:483–506.
[4] Trobec R, Šterk M, Robič B. Computational complexity and parallelization of the meshless local Petrov-Galerkin method. Comput Struct. 2009;87:81-90.
[5] Rabczuk T, Bordas S, Askes H. Meshfree Methods for Dynamic Fracture. Computational Technology Reviews. 2010;1:157-85.

[6] Kosec G, Zinterhof P. Local strong form meshless method on multiple Graphics Processing Units. CMES-Comp Model Eng. 2013;91:377-96.
[7] Wang JG, Liu GR. A point interpolation meshless method based on radial basis functions. Int J Numer Meth Eng. 2002;54:1623-48.
[8] Šarler B. From global to local radial basis function collocation method for transport phenomena. Advances in Meshfree Techniques. Berlin: Springer; 2007. p. 257-82.
[9] Kosec G, Šarler B. Solution of thermo-fluid problems by collocation with local pressure correction. Int J Numer Method H. 2008;18:868-82.
[10] Depolli M, Trobec R, Filipič B. Asynchronous Master-Slave Parallelization of Differential Evolution for Multiobjective Optimization. 1998 Ieee International Conference On Robotics And Automation, Vols 1-4. 2012.
[11] Eiben AE, Smith JE. Introduction to Evolutionary Computing. Berlin, Heidelberg: Springer-Verlag; 2003.
[12] Zitzler E, Thiele L, Laumanns M, Fonseca CM, Fonseca VGd. Performance Assessment of Multiobjective Optimizers: An Analysis and Review. IEEE Transactions on Evolutionary Computation. 2002;7:117–32.
[13] Burke EK, Kendall G. Introduction to Stochastic Search and Optimization. Hoboken: John Wiley & Sons; 2003.
[14] de Vahl Davis G. Natural convection of air in a square cavity: a bench mark numerical solution. Int J Numer Meth Fl. 1983;3:249-64.
[15] Prax C, Sadat H, Salagnac P. Diffuse approximation method for solving natural convection in porous Media. Theor App T. 1996;22:215-23.
[16] Divo E, Kassab AJ. Localized meshless modeling of natural-convective viscous flows. Numer Heat Transfer. 2007;B129:486-509.
[17] Wan DC, Patnaik BSV, Wei GW. A new benchmark quality solution for the buoyancy-driven cavity by discrete singular convolution. Numer Heat Transfer. 2001;B40:199-228.
[18] Trobec R, Kosec G, Šterk M, Šarler B. Comparison of local weak and strong form meshless methods for 2-D diffusion equation. Eng Anal Bound Elem. 2012;36:310-21.
[19] Kosec G. Simulation of multiphase thermo-fluid phenomena by a local meshless numerical approach. Informatica. 2012;36:227-8.
[20] Kosec G, Šarler B. Simulation of macrosegregation with mesosegregates in binary metallic casts by a meshless method. Engineering Analysis with Boundary Elements. 2014:10.1016/j.enganabound.2014.01.016.
[21] Šarler B, Kosec G, Lorbiecka AZ, Vertnik R. A meshless approach in solution of multiscale solidification modeling. Mater Sci Forum. 2009;649:211-6.
[22] Šarler B. Towards a mesh-free computation of transport phenomena. Eng Anal Bound Elem. 2002;26:731-8.
[23] Malan AG, Lewis RW. An artificial compressibility CBS method for modelling heat transfer and fluid flow in heterogeneous porous materials. Int J Numer Meth Eng. 2011;87:412-23.
[24] Arpino F, Massarotti N, Mauro A, Nithiarasu P. Artificial compressibility based CBS solutions for double diffusive natural convection in cavities. Int J Numer Method H. 2013;23:205-25.
[25] Hong CP. Computer Modelling of Heat and Fluid Flow Materials Processing. Bristol: Institute of Physics Publishing; 2004.
[26] Deb K, Pratap A, Agarwal S, Meyarivan T. A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Transactions on Evolutionary Computation. 2002;6:182–97.
[27] Kosec G, Depolli M, Rashkovska A, Trobec R. Super linear speedup in a local parallel meshless solution of thermo-fluid problems. Comput Struct. 2014;133:30-8.
[28] Kosec G, Šarler B. Solution of a low Prandtl number natural convection benchmark by a local meshless method. Int J Numer Method H. 2013;23:22.

Delivering Bioinformatics MapReduce Applications in the Cloud

Lukas Forer*, Tomislav Lipić**, Sebastian Schönherr*, Hansi Weißensteiner*,
Davor Davidović**, Florian Kronenberg*, Enis Afgan**
* Division of Genetic Epidemiology, Medical University of Innsbruck, Innsbruck, Austria
** Center for Informatics and Computing, Ruđer Bošković Institute, Zagreb, Croatia
Enis.Afgan@irb.hr

Abstract - The ever-increasing data production and availability in the field of bioinformatics demands a paradigm shift towards the utilization of novel solutions for efficient data storage and processing, such as the MapReduce data parallel programming model and the corresponding Apache Hadoop framework. Despite the evident potential of this model and existence of already available algorithms and applications, especially for batch processing of large data sets as in the Next Generation Sequencing analysis, bioinformatics MapReduce applications are yet to become widely adopted in the bioinformatics data analysis. We identify two prerequisites for their adaptation and utilization: (1) the ability to compose complex workflows from multiple bioinformatics MapReduce tools that will abstract technical details of how those tools are combined and executed allowing bioinformatics domain experts to focus on the analysis, and (2) the availability of accessible and flexible computing infrastructure for this type of data processing. This paper presents integration of two existing systems: Cloudgene, a bioinformatics MapReduce workflow framework, and CloudMan, a cloud manager for delivering application execution environments. Together, they enable delivery of bioinformatics MapReduce applications in the Cloud.

I. INTRODUCTION

Due to the advent of Next Generation Sequencing (NGS) [1], which can be described as the ability to decode the human genome in a massively parallel way, the amount of data in genomics has increased rapidly over the recent years. This remarkable increase in the volume of data is forcing affected institutions to consider novel approaches for storing the data and improving data analysis algorithm performances. However, for the bioinformatics domain experts without a background in computer science, utilizing these algorithms is often a challenging task because they require access to advanced compute infrastructure setup.

Well-established workflow systems in bioinformatics data analysis, such as Galaxy [2] or Taverna [3], enable scientists to access a large set of computational tools through an accessible web interface. Domain experts are able to compose complex pipelines consisting of multiple tools, while the systems keep track of all the parameter options and allow the analysis to be reproduced or shared. These systems abstract the details of how the available tools are run, and allow the domain expert to focus on the data analysis rather than the technical details required to run a tool or load the data. While these systems represent an indispensable value to today's research community, they are limited to the traditional tools (i.e., sequential, MPI, embarrassingly parallel) and traditional computational infrastructures (i.e., dedicated or cloud-based compute clusters and standalone workstations).

Given the large volume of data generated by NGS, novel algorithm parallelization methods are becoming increasingly important. One example of such a method is MapReduce [4][5], a straightforward programming model that allows efficient data parallelization. The MapReduce programing model (simply called MapReduce) allows users to process large amounts of data without understanding the underlying job execution environment complexity. In general, MapReduce provides a scalable way to parallelize large amounts of data using many inexpensive computational nodes and can simply be described as a general data-processing tool. However, not every algorithm can be efficiently parallelized with MapReduce. For example, parallelizing iterative algorithms with a lot of inter-process communication are better parallelized using the message-passing (MPI) model instead of MapReduce.

So far, the support for the MapReduce parallelization model in the field of bioinformatics has been available only on a per-tool basis (see Section 2). Such individual tools typically cover one aspect of an otherwise larger data analysis pipeline. While beneficial, these solutions require domain experts to manually compose different tools in order to ensure a complete pipeline. What is currently lacking is the ability to interactively chain multiple such tools and allow the domain experts to focus on the data analysis rather than on the mechanics of the pipeline. Although, the previously mentioned and popular workflow systems enable this level of abstraction for the traditional tools, currently they do not offer the support for the tools based on MapReduce parallelization model.

Realizing such a workflow platform requires a number of features, including: a graphical platform to integrate and execute available MapReduce tools, the option to easily reproduce experiments, the ability to extend the platform with additional or alternative tools, and a simplified access to the required computational infrastructure. In this paper, we describe each of these components and present a system that fulfils mentioned requirements. Moreover, we give an overview of bioinformatics applications that implement the

MapReduce model, some of which have already been integrated into the presented system.

II. BIOINFORMATICS MAPREDUCE APPLICATIONS

The most widely used open-source implementation of MapReduce programming model for large data batch processing is Apache Hadoop [6]. Hadoop also includes several sub-projects such as a distributed file system (HDFS) used for data storage in combination with MapReduce and Hadoop Pig, a high-level dataflow language to simplify the generation of MapReduce jobs. Within MapReduce, the user is responsible to write the *map* and *reduce* functions according to the algorithm's requirements. The MapReduce itself then achieves parallelization, data distribution, load balancing, and fault-tolerance on cluster architecture by appropriately invoking the *map* or *reduce* functions. Despite its potential, particularly for batch processing applications in bioinformatics, only a limited number of algorithms have adapted this model.

Table I gives an overview of currently available libraries and applications based on the MapReduce paradigm in the bioinformatics context. Utilizing these solutions often requires significant technical expertise, which may present a challenge for domain experts without a background in computer science or without access to the necessary MapReduce cluster architectures. To facilitate their adoption, several features that have already been delivered by graphical workflow systems such as Galaxy or Taverna, must be also fulfilled for the MapReduce applications. This includes a graphical platform to integrate and execute available MapReduce workflows, the possibility to reproduce experiments easily and a simplified access to a MapReduce cluster in private and public clouds. Libraries such as SeqPig [8] or BioPig [9] are helping biologists to use the aforementioned paradigms by abstracting the underlying Hadoop framework and providing high-level Apache Pig functions (or User Defined Functions (UDFs)). Nevertheless, a combination of algorithms to workflows, a standardized way to import/export data and an execution platform for algorithms in public or private cloud infrastructure is still lacking.

III. BIOINFORMATICS MAPREDUCE WORKFLOWS

Cloudgene [18] presents a web-based platform to create and execute workflows consisting of Hadoop MapReduce, Hadoop Pig and command line-based programs. It can be seen as an additional layer between Hadoop MapReduce and the end user that hides the complexity of the framework. Therefore, Cloudgene allows integrating different programs or algorithms within one platform that can be easily combined to workflows. The general architecture of Cloudgene is presented in Figure 1. As depicted, the client communicates with the server through a REST API. The server itself consists of three components: a workflow engine, a workflow manager, and a data manager. The workflow engine processes a previously generated workflow, specified on client side by utilizing Cloudgene's workflow definition language (WDL) and it is responsible for executeing MapReduce steps on a Hadoop cluster. Job details, including all metadata (e.g., status of a job), are analyzed by the workflow manager and then provided to the client. Finally, the data manager component has the responsibility of interacting with the Hadoop distributed file system (HDFS). In order to submit jobs or interact with HDFS, Cloudgene must be installed on a Hadoop

TABLE I. BIOINFORMATICS MAPREDUCE APPLICATIONS OVERVIEW

Area	Program	Description	Cite
Hadoop MapReduce libraries for Bioinformatics	Hadoop BAM	Manipulation of aligned next-generation sequencing data (supports BAM, SAM, FASTQ, FASTA, QSEQ, BCF, and VCF)	[7]
	SeqPig	Processing NGS data with Apache Pig; Presenting UDFs for frequent tasks; using Hadoop-BAM	[8]
	BioPig	Processing NGS data with Apache Pig; Presenting UDFs	[9]
	Biodoop	MapReduce suite for sequence alignments / manipulation of aligned records; written in Python	[10]
DNA - Alignment algorithms based on Hadoop	CloudBurst	Based on RMAP (seed-and-extend algorithm) Map: Extracting k-mers of reference, non- overlapping k-mers of reads (as keys) Reduce: End-to-end alignments of seeds	[11]
	Seal	Based on BWA (version 0.5.9) Map: Alignment using BWA (on a previously created internal file format) Reduce: Remove duplicates (optional)	[12]
	Crossbow	Based on Bowtie / SOAPsnp Map: Executing Bowtie on chunks Reduce: SNP calling using SOAPsnp	[13]
RNA - Analysis based on Hadoop	MyRNA	Pipeline for calculating differential gene expression in RNA; including Bowtie	[14]
	FX	RNA-Seq analysis tool	[15]
	Eoulsan	RNA-Seq analysis tool	[16]
Non-Hadoop based Approaches	GATK	MapReduce-like framework including a rich set of tools for quality assurance, alignment and variant calling; not based on Hadoop MapReduce	[17]

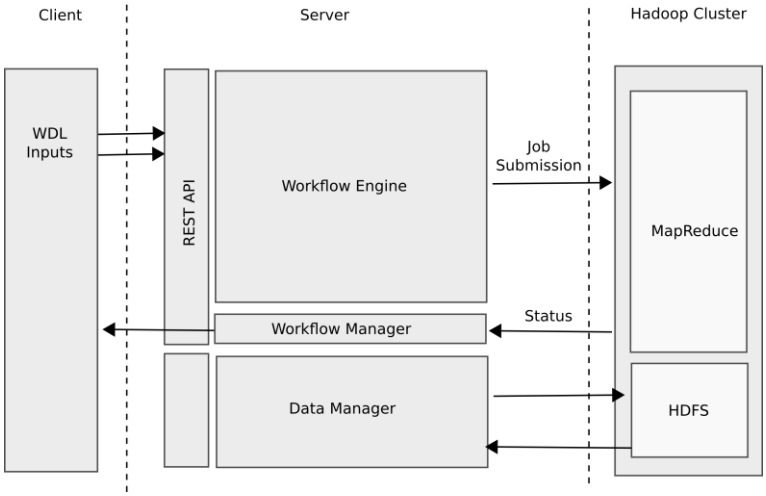


Figure 1. Cloudgene Architecture

cluster namenode.

Existing applications can be integrated into the workflow platform by utilizing Cloudgene's plugin interface. No adaptation to the source code is needed, while only a simple WDL manifest file including a header, input parameters, output parameters and the definition of the workflow itself need to be created. Figure 2. shows the integration process of CloudBurst [11] into Cloudgene. The manifest file includes all the aforementioned parameters. When launching Cloudgene, the manifest file is loaded and the client interface is automatically rendered using information from the file.

A project like Cloudgene lives from its integrated use cases. Therefore, several tools (Crossbow, MyRNA, CloudBurst and Seal) have been already integrated in Cloudgene and will be extended with additional applications in the future. Especially with Hadoop 2 (based on YARN), new algorithms utilizing other programming models than MapReduce will be most likely developed.

IV. BIOINFORMATICS MAPREDUCE WORKFLOWS IN THE CLOUD

Underlying a workflow system such as Cloudgene, functional compatible cluster architecture is a prerequisite for executing MapReduce jobs. Unfortunately, small to medium sized research institutes can hardly afford the acquirement and maintenance of own computer systems with adequate performance. A possible solution comes in the form of cloud computing, which opens the opportunity to use compute and storage resources or services on demand. Cloud computing provides the possibility to rent computer hardware from different providers (e.g., AWS, HP) and those can be used to analyze necessary datasets.

However, laborious challenges need to be addressed to make these resources available to the researchers. Specifically, although cloud computing provides a way to acquire computational resources on demand, the resources provided are either virtual machines on the Internet or specific programming libraries, which are unusable for domain experts because they require considerable configuration and ongoing management. A viable analysis solution thus needs to be accessible and deployable

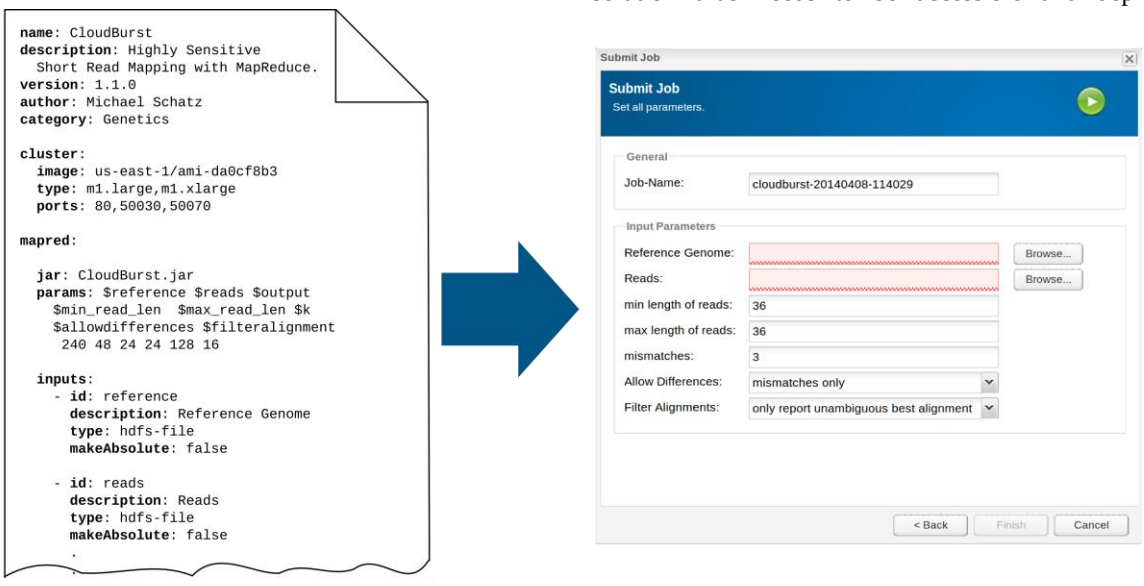


Figure 2. The integration process of CloudBurst into CloudGene

without informatics expertise while efficiently and automatically using dynamically scalable resources.

Commercial companies also offer higher-level services that provide functional MapReduce frameworks. Examples include AWS Elastic Map Reduce (EMR) service or the Rackspace Hortonworks. While these options represent a workable solution, they are also forcing the system to use a particular cloud infrastructure provider, thus locking the user down. Open source Hadoop cluster distributions, such as the Cloudera CDH or MapR M series offer alternatives to the Apache Hadoop distribution. These distributions need to be installed, configured, and managed on computational infrastructure comparable to the low-level cloud resources (i.e., virtual machines).

In response, as part of Cloudgene, Cloudgene-Cluster has been developed as a web-based system to acquire and configure a MapReduce cluster in the cloud. Cloudgene-Cluster allows the orchestration of a full working Hadoop MapReduce cluster, the installations of additional services such as Apache Pig and user specific libraries or software. On top of the cluster, Cloudgene's workflow engine is provided. The end result is that a user is able to access Cloudgene in the cloud without any limitations to an in-house cluster architecture or having to setup their own cluster. Although Cloudgene-Cluster provides an acceptable solution for providing a Hadoop cluster for the Cloudgene workflow application to utilize, its functionality is limited and, more importantly, borderlines the scope of an otherwise workflow system. Instead of developing a full-featured Hadoop manager within the workflow application, it is more beneficial to delegate resource orchestration step to another application.

To help in this regard, we have previously constructed a software system called CloudMan [18], which makes it possible to easily procure and configure a functional data analysis platform on a cloud infrastructure. The procured platform delivers a scalable cluster-in-the-cloud and a data analysis environment preconfigured with a number of applications. With its ability to be launched and managed via a web browser on a number of clouds, customized as necessary, and easily shared with collaborators, CloudMan makes it possible to readily utilize cloud resources in a research environment [20]. Notably, CloudMan provides access to the Galaxy application, preconfigured with dozens of bioinformatics tools and hundreds of gigabytes of reference genome data [21]. Beyond these user-level features exposed in a web browser, CloudMan offers a set of application execution environments: a batch scheduler environment via Sun Grid Engine (SGE), a MapReduce environment using Hadoop and SGE integration, and a support for federated job execution environment via HTCondor [22]. These enable a range of workloads to be readily executed atop procured cloud resources.

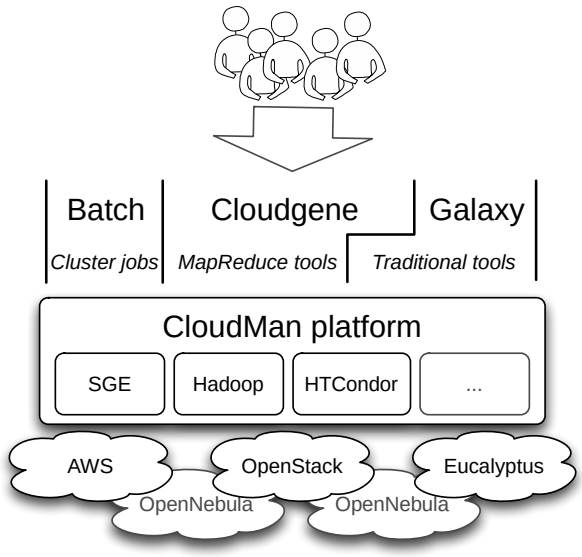


Figure 3. A layered view of the applications composing a flexible, functional tools ecosystem

Internally, CloudMan implements a service-oriented architecture that allows arbitrary tools to be described as services. Once implemented, these services are easily deployed within the CloudMan platform making the tools available (to the user or other services/tools). CloudMan further implements a service dependency management framework that allows services to specify other services as their prerequisites. This makes it suitable for wrapping the Cloudgene workflow application as a CloudMan service. Because CloudMan provides a number of cloud orchestration and scaling features as well as includes a Hadoop-based application execution environment [22], by specifying Hadoop as its prerequisite, CloudMan will ensure a functional Hadoop cluster to be available for the Cloudgene workflow application to utilize when started. This service-to-tools ecosystem is visualized in Figure 3.

Delegating the setup of the parallel compute environment to CloudMan opens the door for Cloudgene application to incorporate and make use of additional, big data, computational models (e.g., real-time data streaming). Comparable to the Hadoop environment, rather than needing to provide an implementation for provisioning the desired environment, such environment may be provided as a CloudMan service that can be simply requested by the Cloudgene application. The given environment will still need to be provided with the CloudMan platform. However, the expectation is that it will be possible to reuse much of the functionality already provided by the platform (e.g., cloud resource management), thus simplifying the service development process. Lastly, CloudMan platform already integrates a number of bioinformatics tools and a workflow system (Galaxy). As mentioned, Galaxy implements an execution model for the running jobs via the traditional batch scheduler. Adding Cloudgene into the platform, as a workflow system for MapReduce type applications, presents an opportunity to integrate the two job execution models.

V. CONCLUSION

The MapReduce model has proven to be a simple but effective programming model for implementing algorithms for the analysis of large datasets on large clusters and is thus very well suited for the Cloud. However, the usage of this paradigm is still limited to a small number of highly qualified domain experts.

Improving accessibility of solutions that utilize this concept is one of the first steps in democratizing access to the technology and reaping its benefits. In this paper, we summarize a number of software solutions that exist in the domain of bioinformatics that utilize the MapReduce programming model. In order to facilitate their utilization and integration, we then describe Cloudgene as a graphical workflow engine that allows these existing solutions to be easily chained together. This facilitates development of open-ended analyses and promotes acceptance of the technology. Finally, we identify a number of technical issues that still exist when trying to utilize this technology in the Cloud context and present an overview of a solution to overcome a number of those.

ACKNOWLEDGMENT

This work was, in part, supported by the "Scalable Big Data Bioinformatics Analysis in the Cloud" grant from the Croatian Ministry of Science, Education, and Sport and the Austrian Federal Ministry of Science and Research (BMWF) and by the FP7-PEOPLE programme grant 277144 (AIS-DC).

REFERENCES

[1] Next Generation Sequencing (NGS), http://res.illumina.com/documents/products/illumina_sequencing_introduction.pdf, visited on 10th February, 2014.

[2] Goecks, Jeremy, Anton Nekrutenko, James Taylor, and T. Galaxy Team. "Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences." Genome Biol 11, no. 8 (2010).

[3] Wolstencroft, Katherine, Robert Haines, Donal Fellows, Alan Williams, David Withers, Stuart Owen, Stian Soiland-Reyes et al. "The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud." Nucleic acids research (2013).

[4] Pireddu, Luca, Simone Leo, and Gianluigi Zanetti. "MapReducing a genomic sequencing workflow." Proceedings of the second international workshop on MapReduce and its applications. ACM, (2011).

[5] Zou, Quan, et al. "Survey of MapReduce frame operation in bioinformatics." Briefings in bioinformatics (2013).

[6] Taylor, Ronald C. "An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics." BMC bioinformatics 11.Suppl 12 (2010)

[7] Niemenmaa, Matti, Aleksi Kallio, André Schumacher, Petri Klemelä, Eija Korpelainen, and Keijo Heljanko. "Hadoop-BAM: directly manipulating next generation sequencing data in the cloud." Bioinformatics 28, no. 6 (2012): 876-877.

[8] Schumacher, André, Luca Pireddu, Matti Niemenmaa, Aleksi Kallio, Eija Korpelainen, Gianluigi Zanetti, and Keijo Heljanko. "SeqPig: simple and scalable scripting for large sequencing data sets in Hadoop." Bioinformatics 30, no. 1 (2014): 119-120.

[9] Nordberg, Henrik, Karan Bhatia, Kai Wang, and Zhong Wang. "BioPig: a Hadoop-based analytic toolkit for large-scale sequence data." Bioinformatics 29, no. 23 (2013): 3014-3019.

[10] Leo, Simone, Federico Santoni, and Gianluigi Zanetti. "Biodoop: Bioinformatics on hadoop." In Parallel Processing Workshops, 2009. ICPPW'09. International Conference on, pp. 415-422. IEEE, 2009.

[11] Schatz, Michael C. "CloudBurst: highly sensitive read mapping with MapReduce." Bioinformatics 25, no. 11 (2009): 1363-1369.

[12] Pireddu, Luca, Simone Leo, and Gianluigi Zanetti. "SEAL: a distributed short read mapping and duplicate removal tool." Bioinformatics 27, no. 15 (2011): 2159-2160.

[13] Langmead, Ben, Michael C. Schatz, Jimmy Lin, Mihai Pop, and Steven L. Salzberg. "Searching for SNPs with cloud computing." Genome Biol 10, no. 11 (2009).

[14] Langmead, Ben, Kasper D. Hansen, and Jeffrey T. Leek. "Cloud-scale RNA-sequencing differential expression analysis with Myrna." Genome Biol 11, no. 8 (2010).

[15] Hong, Dongwan, Arang Rhie, Sung-Soo Park, Jongkeun Lee, Young Seok Ju, Sujung Kim, Saet-Byeol Yu et al. "FX: an RNA-Seq analysis tool on the cloud." Bioinformatics 28, no. 5 (2012): 721-723.

[16] Jourdren, Laurent, Maria Bernard, Marie-Agnès Dillies, and Stéphane Le Crom. "Eoulsan: a cloud computing-based framework facilitating high throughput sequencing analyses." Bioinformatics 28, no. 11 (2012): 1542-1543.

[17] McKenna, Aaron, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella et al. "The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data." Genome research 20, no. 9 (2010): 1297-1303.

[18] Schönherr, Sebastian, Forer Lukas, Weißensteiner Hansi, Kronenberg Florian, Specht Günther and Anita Kloss-Branstätter. "Cloudgene: A graphical execution platform for MapReduce programs on private and public clouds." BMC Bioinformatics 13, no. 1 (2012):200

[19] Afgan, Enis, Dannon Baker, Nate Coraor, Brad Chapman, Anton Nekrutenko, and James Taylor. "Galaxy CloudMan: delivering cloud compute clusters." BMC bioinformatics 11, no. Suppl 12 (2010).

[20] Afgan, Enis, Brad Chapman, and James Taylor. "CloudMan as a platform for tool, data, and analysis distribution." BMC bioinformatics 13, no. 1 (2012): 315.

[21] Afgan, Enis, Brad Chapman, Margita Jadan, Vedran Franke, and James Taylor. "Using cloud computing infrastructure with CloudBioLinux, CloudMan, and Galaxy." Current Protocols in Bioinformatics (2012): 11-9.

[22] Kowsar, Yousef, and Enis Afgan. "Support for Data-intensive Computing with CloudMan." In MIPRO. 2013.

Microspheres as sensors to measure instant refractive index variation during phase transition of low melting substances

Hong-Bo Fu, Yao-Xiong Huang*
Department of Biomedical Engineering, Ji Nan University, Guang Zhou, China
tyxhuang@jnu.edu.cn

Abstract- Microspheres were used as sensors to perform real time in situ accurate measurement on the instantaneous refractive index / density variation during the liquid-solid phase transition of low melting substances. The measurements were carried by immersing microspheres in the substances in liquid state and analyzing their phase-contrast microscopic images during the phase transition from liquid to solid states. In this paper, the principle of the method and the experiments of using the method for different applications will be described.

I. Introduction

Understanding the phase behavior of low melting substances such as paraffin wax, agar gel, and polystyrene are of industrial importance. However the present methods have limitations to monitor the instantaneous process of the phase transition of the materials, especially unable to detect if the substances are spatially homogeneous or not during the phase transition. Considering microspheres can be used as a powerful sensory tool for different purposes [1,2] [3]. Here we describe a simple, innovative and accurate method for the measurement of the refractive index (RI) variation during the phase transition of some low melting point substances. RI was determined simply by immersing microspheres with known RI in the substances in liquid state and analyzing their phase-contrast microscopic images. Since the refractive index of a substance is a function of its density and other thermodynamic parameters, this method can be also used for density variation and other parameters monitoring during the phase transition of the substances.

II. Principle

When a microsphere with a RI of n_2 is immersed in a medium of n_1 ($n_2 > n_1$) and illuminated by a plane light wave propagating along its optical axis, due to the refraction effect, no parallel paraxial ray incident on the sphere's first surface will emerge to the periphery of its second surface, thus a dark ring appears in the image of the microsphere. The ratio between the radius of the central bright spot in the image of the sphere *r*, and the radius of the sphere *R*, is found to have a relation with n_2 and n_1 as follows[3]:

$$n_1 = n_2\left(\frac{3}{8}\sqrt[3]{X^2+\sqrt{X^4-X^6}}+\frac{3X^2}{8\sqrt[3]{X^2+\sqrt{X^4-X^6}}}+\frac{1}{4}\right)^{\frac{1}{2}} \quad (1)$$

Where $X = r/R$. Therefore, by using a microsphere with known refractive index n_2 and measuring both *r* and *R* , one can easily obtain the refractive index of its surrounding medium n_1.

III. Experimental

Several low melting point substances including paraffin wax, agar gel, and polystyrene were tested using the technique. The substances were set at the temperatures higher than their melting points at first, so all of them were in liquid states. Then some silica microspheres with diameters of 50-150 μm were immersed into the substances. The phase contrast images of the spheres were taken at a

frame rate of 25 f/s using a digital camera set in a phase contrast microscope as the temperature was lower down during their liquid-solid phase transition. Thereafter the values of *r* and *R* were measured and the variation of their RI during the phase transition was obtained by the calculation using equation 1.

IV. Results and discussion

As reported previously, the technique has a satisfied accuracy in determining RI of media. The standard deviation of the measurement depends on the precision of the image size measurement and is estimated to be 0.002–0.004 when using a 1M pixels camera for imaging. It can be improved by increasing the resolution of the imaging camera and using super-resolution method for measuring both *r* and *R* to as small as 10^{-6} [3]. By recording images with a video rate of 25frame/s, our technique was proved to be able to monitor the dynamic process of the liquid-solid phase transitions for the three substances. Fig.1 shows the typical images of the microspheres in the three substances when they had transited into solid states. Fig.2 gives the variation of the RI of paraffin wax vs. temperature during the phase transition. We can see that the technique can monitor the instantaneous variation of the material's properties. Since optical refraction takes place at the speed of light, the instant variation of the RI in the surrounding medium of a microsphere can immediately induce a change in the radius *r* of the image of the sphere. Therefore, by using a high speed camera for imaging, our technique can catch up very rapid RI/concentration variation, thus it is able to determine the phase transition point very accurately (see fig.2). On the other hand, as we can see in figure 1 that, by immersing several microspheres at different locations, we can simultaneously determine the RI/density distribution in various sites of the substances. So by this technique, one can also determine if the substances are spatially homogeneous or not during the phase transition. Compared with the other techniques, our method is not only of the advantages of simplicity, accuracy and fast response, but also there is almost no further interference on the sample after inserting the microsphere into it. The only needed operation is to capture the image at any desired moment, so easy and rapid real time in situ monitoring analysis on phase transition is possible. Furthermore, it can determine the instantaneous variation of the sample properties in micro-regions as small as several ten microns.

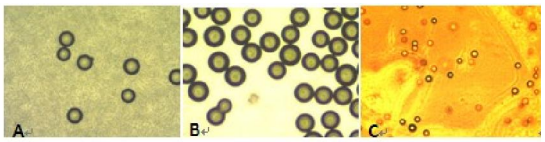


Figure1. The phase contrast images of microspheres in: (A) paraffin wax, (B) agar gel, and (C) polystyrene.

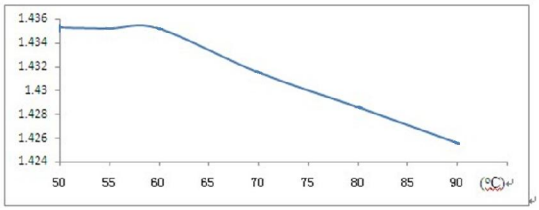


Figure2. The variation of RI vs. temperature observed using the microsphere method on paraffin wax.

V. Conclusion

The technique described here is an ideal one for many applications in which local RI/density at different sites in substances should be monitored from time to time, and even phase transition or other critical phenomena may take place. It can be applied to many fields of studies including material analysis, environmental monitoring, pharmaceutical and specialty chemical industries.

References:

1. J.R Kettman, T Davies,et al.," Classification and properties of 64 multiplexed microsphere sets", Cytometry, 33, pp.334–343, 1998

2. K. J. Vahala, "Optical microcavities," Nature, **424**, pp.839-846, 2003

3. Y L Yue, P Wang, YX. Huang, "Easy method to determine refractive indices of microspheres and in micro-regions of inhomogeneous media", Biosensors and Bioelectronics, 30, pp.216-222, 2011

Multilingual Cloudlet-based Dictionary

Aleksandar Bahtovski
VIP Operator

Skopje, Macedonia
Email: abahtovski@gmail.com

Marjan Gusev
Ss. Cyril and Methodius University
Faculty of Information Sciences and Computer Engineering
Skopje, Macedonia
Email: marjan.gushev@finki.ukim.mk

***Abstract*—Cloud Computing is a technology that moves the World forward. It helps the mobile users to run compute demanding applications on a remote cloud and access the results on their mobile devices. Usual processing of the cloud based applications is realized via a WAN connection. However, this approach is inefficient for real-time applications due to relatively huge WAN latency. Cloudlets solve this problem by exploiting the benefits of Cloud Computing without being WAN-limited. The cloudlet is a small cloud located near the mobile user. Cloudlets can perform the same processing tasks as a remote cloud by using Virtual Machines (VMs).**

In this paper we introduce the cloudlet approach to design a Multilingual Dictionary, which is a Cloudlet based application specifically developed to be processed on a near cloudlet as well as on a distant cloud. It will be used for translating words entered through a mobile device by using a VM deployment on the cloud/cloudlet and displaying the results on the mobile device. The initial version will support translation between two languages. Afterwards, it will be expanded to support multilingual translation.

***Index Terms*—Cloud Computing, Virtualization, Cloudlet, Dictionary, Virtual Machines**

I. Introduction

Mobile devices are becoming an essential part of everyday human life by being the most important and most proper communication tool. They have brought various functionalities to our fingertips. Mobile users can use mobile applications (iPhone apps, Google apps, Windows Phone apps, etc.) executing them on their mobile devices. However, mobile devices suffer from resource scarcity. They have limited resources and because of that, most of today's applications cannot be executed on the mobile device.

Alternative methods have been analyzed to obtain more resources and allow mobile users to use a wide variety of mobile applications [1],[2]. One of the alternative methods is using a cloud [3]. Cloud enables mobile users to use its resources and run applications, which provides mobile users the opportunity of running resource-demanding applications. It acts as an agent between the original providers and the mobile device [4].

Cloud resources allows service providers and users to adjust their computing capacity depending on how much is needed at a given time or given task. However, sometimes, the usage of cloud resources to execute mobile application can be a bad solution. Clouds are located far away from the mobile user and WAN delays can make this approach inefficient for real time applications. To solve this issue Satyanarayanan introduced the concept of cloudlets as a small cloud located near the mobile user [5].

A cloudlet is a trusted, resource rich computer, which can do the same tasks as a distant cloud [5]. This technology concept enables the mobile users to utilize nearby computers to obtain the benefits of cloud computing without WAN delays, congestions, failures. With this approach it is much easier to achieve better application processing due to the cloudlet's proximity.

The focus of this article is to measure the processing time and execution of a multilingual dictionary application running on a mobile device. The application will use cloud or cloudlet resources in order to translate the words and display the results back to the mobile device. Execution times of the word translation using cloud or cloudlet will be compared and conclusion will be made at the end of the article. In Section II we explain the reasons for research in this field, and the motivation for this article. Section III gives some related work done in this field. The use of mobile cloud computing as augmented execution is explained in Section IV. The architecture of the new concept is introduced in Section V including description of the augmentation process and some implementation details. Finally, the last sections are devoted to discussion about the new concept and conclusions of the implementation.

II. Motivation

With today's growth of the mobile market, there are lot applications which cannot be processed using only the mobile device. Some of the applications need additional resources in order to run smoothly [7], [8], [9]. For example, a lot of mobile users play games, taking pictures, upload videos, manage their personal data, such as Dropbox [10], Docusign [11], etc.

Mobile devices are not capable to run such application on their own. Additional resource should have been used in order for execution of these mobile applications to be faster. Using this starting point, the introduction of clouds or cloudlets is of huge help. There are numerous applications developed to use external resources for processing and present the result back on the mobile device. LinkedIn [12], Boomerang [13], SlideRocket [14], TripIt [15], UEFA.com [16], Zomato [17],

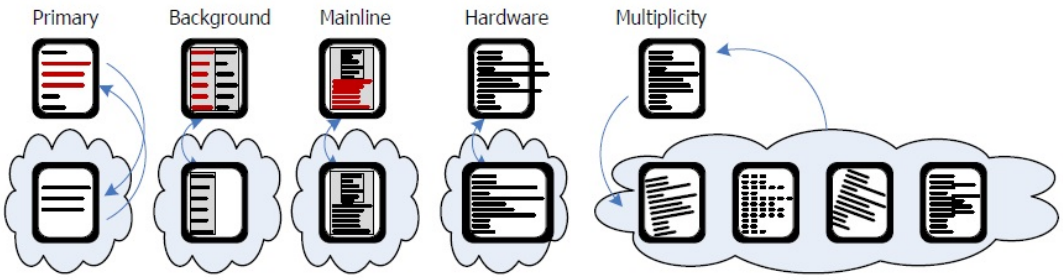


Fig. 1. Categories of Augmented Execution [6]

or Dropbox [10] are just some of the applications which are using cloud or cloudlet for execution.

The main idea of this article is to develop a multilingual dictionary which can use a cloud and cloudlet for processing. The user can enter a word for translation and view the corresponding meaning of the word in another language. Mobile device will be used to enter the words and see the results. Computing (translation) will be done on an external resource, cloud or cloudlet.

III. Related work

There are lots of real time applications which are using cloud or cloudlets to process big data. Most of them are applications for Weather forecast [18], Betting [19], Poker games [20], multilingual dictionaries, etc.

Remote execution of resource intensive applications for resource poor hardware is a well known approach in mobile computing. All remote execution applications run simple tasks on the mobile device and computation intensive tasks at the remote server. There are several research papers on saving power via remote execution, such as Rudenko et al. [21], Flinn et al. [22] and Satyanarayanan et al. [22].

Cyber foraging uses surrogates (untrusted and unmanaged public machines) opportunistically to improve the performance of mobile devices [23]. For example, both data staging and Slingshot use surrogates. In particular, Slingshot [24] creates a secondary replica of a home server at nearby surrogates. ISR provides an ability to suspend on one machine and resume on another machine by storing virtual machine images in a distributed storage system [6].

Coign automatically partitions a distributed application composed of Microsoft COM components. Few students from International Institute of Information Technology Bangalore have made a multilingual dictionary which can use cloudlet for processing and translation of the words [25]. Also Achanta et al. present a model of cloudlet-based multilingual dictionaries [26]. They are using a VM installed on the cloudlet to initiate the dictionary and then use it over a wireless network. Our approach is based on their ideas to realize the solution.

IV. The Using of Mobile Cloud Computing in Augmented Execution

Mobile cloud computing means that processing of the mobile based application will be done on a portable device. Mobile devices can execute resource intensive applications on a distant high performance computer server and support the client user interactions with application over the Internet. This may lead to increased WAN latencies due to distance between the mobile user and the remote cloud. WAN latency is the most important parameter for real time applications, since they can affect usability in a critical path of user interaction by degrading the crispness of system response even for trivial applications.

Thousands of users of the same applications access the cloud for processing and therefore the cloud-based approach can become inefficient. Solution of this problem is using cloudlets and doing the processing of the application on it. In the cloudlet-based approach, the application will first try to access the nearby cloudlet and realize the required processing. If there is no cloudlet nearby, the processing will be done on the cloud. In the worst case scenario, the mobile device should use its own resources to process the application and present the results to the user.

By using this approach, we will develop a mobile application solution that will be processed in a cloudlet. In case there is no cloudlet nearby, this task will be redirected on a remote cloud.

A. Augmented Execution

First of all we need to explain what augmented execution is in order to realize why external resources are required to overcome smartphones capabilities. Augmented execution refers to a technique used to overcome the limitations of smartphones in terms of computation, memory and battery [6]. Chun and Maniatis [6] propose an architecture which is seamlessly offloading execution from smartphones to external computational infrastructure where cloned replica of the smartphone's software is running. This provides the illusion that the mobile user has a more powerful device than they actually have.

The mobile device hosts computation and memory demanding applications. Some of the computation tasks can be offloaded to the cloud where cloned system image of the mobile device is running. Augmentation for asynchronous operations can operate in the background, while for synchronous operations it can be performed by blocking progress on the smartphone until the results arrives from the clone in the cloud. For concurrent operations, augmentation can also be

concurrent in the clone cloud or even speculative with ability to undo the operation on the smartphone according to the result from the clone.

Figure 1 shows five categories of possible augmented execution for mobile phones [6]. The *primary functionality* refers to functions where computer hungry applications are automatically split, user interface and low processing is done at the phone side, while high computation is offloaded to the infrastructure synchronously. The *Background functionality* does not need to interact with the users in a short time scale. This is functionality which typically happens in the background. Entire processes can be marked by the user or automatically interfered as background processes and migrated to the infrastructure. In this case, offloaded functionality can take the role of a virtual client. Even when the smartphone is turned off, background tasks continue to run. Later when the smartphone turns on, it can synchronize its state with the infrastructure [6].

Mainline functionality can be placed between primary and background augmentation. Here, the user may opt to run a particular application in a wrapped fashion, altering the method of its execution but not its semantics. Unlike background augmentation, mainline augmentation is interspersed in the execution of the application. *Hardware augmentation* compensates for fundamental weaknesses of the smartphone platform, such as memory caps or other constraints, and hardware peculiarities. Using faster hardware significant savings can be observed; faster processor and much more RAM memory will reduce the costs and make execution faster. The last category is augmentation through *multiplicity*. It uses multiple copies of system images executed in different ways. This approach can help running data parallel applications. Also this can allow applications to evaluate all possible choices for a decision before making that decision. With this functionality much energy is wasted at the infrastructure side to save a little bit on the mobile device [6].

Similar approach for using VM technologies to execute computation extensive software from mobile device to external infrastructure is explained by Satyanarayanan [5].

B. Virtual Machine Synthesis

Two different approaches can deliver Virtual Machine state to infrastructure. The first one is VM migration where an already executing VM is suspended, its processor, disk and memory state are transferred and finally VM execution is resumed at the destination from the exact point of suspension [5].

Similar approach of using VM technologies executing computation intensive software from mobile device is by using Dynamic VM Synthesis [5]. In the dynamic VM synthesis architecture, a mobile user exploits VMs to instantiate a customized service software on a nearby cloudlet and uses the service over Wi-Fi. In this case, a VM is dynamically delivered to the infrastructure. Mobile devices request small VMs overlay from the nearby cloudlet infrastructure which already owns the base VM from which this overlay was derived as shown on Figure 2 [27].

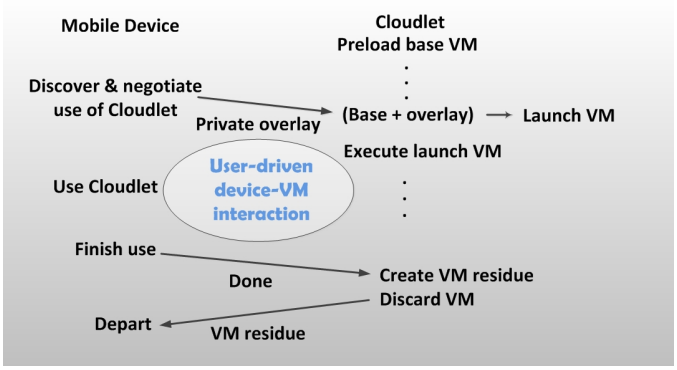


Fig. 2. Dynamic VM synthesis

The infrastructure applies the overlay to the base to derive the VM which starts executing in the precise space in which it was suspended. Our solution of a multilingual dictionary application is based on a cloudlet, which is a server that receives entered words from a mobile device, performs translation and returns the output for word synthesis. If the cloudlet is a cluster, VM's could be rapidly cloned to exploit parallelism.

The Dynamic VM synthesis process [26] is presented on Figure 3. The solution consists of the following:

- A small overlay VM is presented in the mobile device. A base VM is present in the cloudlet infrastructure.
- The dictionary overlay VM from the mobile device is delivered to the cloudlet where base VM is already presented?
- The cloudlet infrastructure applies the dictionary overlay VM to the base VM in order to launch VM.
- Launch VM command using cloudlet execute the application in the exact state from which the overlay was derived.
- The software in the Launch VM is a server that receives the input text from the mobile device, performs the translation and returns results to the mobile device.
- After the application is completed, the result is sent back to the mobile device.

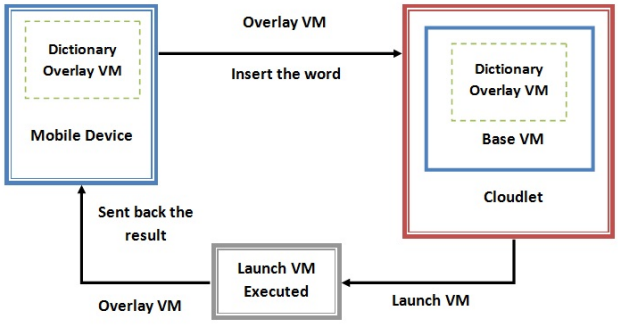


Fig. 3. Dictionary Dynamic VM synthesis diagram

C. Virtual Machine Overlay Creation

To develop this application we intend to use a Virtual Box, that is hosted Virtual Machine Manager for Linux. The base VM hosts a minimally configured guest OS.

The process of VM Creation [26] is presented in Figure 4 and is explained by the following actions:

- The tool creates VM overlays using base VM and executes install-script in the guest OS and the result is VM that has been configured for use by the mobile device.
- Next, the tool executes resume-script in order guest OS to launch the desired application and bring it to a state ready for user interaction. This VM is called "Launch VM" and is suspended. It can be resumed at runtime without delays of application launch.
- After creating the "Launch VM", the tool differentiate its memory and disk images with those of base VM to obtain the VM overlay.
- The last step is to compress and encrypt this overlay.

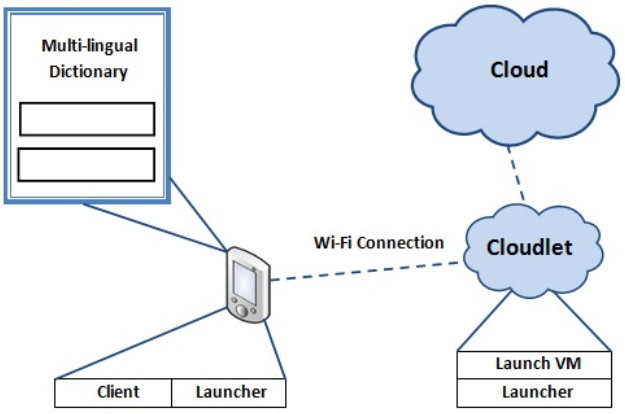


Fig. 4. Dictionary Virtual Machine Overlay Creation Algorithm

A TCP connection between the device and cloudlet is established after the creation of a VM. It is used for sending the input requests and receiving the results. After the mobile device sends the input request and the VM overlay, the cloudlet executes a script and applies the overlay to the base VM. The suspended VM is then launched and is ready to provide services to the mobile device. Once the suspended VM is in running state, the application is executed and the output is sent back to the mobile device.

V. Architecture of a Cloudlet Solution

In this section we explain the architecture of the newly proposed cloudlet-based multilingual dictionary solution.

A. Modeling the augmentation process

The augmentation process of the multilingual dictionary application is modeled by using an external cloudlet infrastructure. The new proposed model builds on an idea to augment smartphone application by utilizing heterogeneous computing

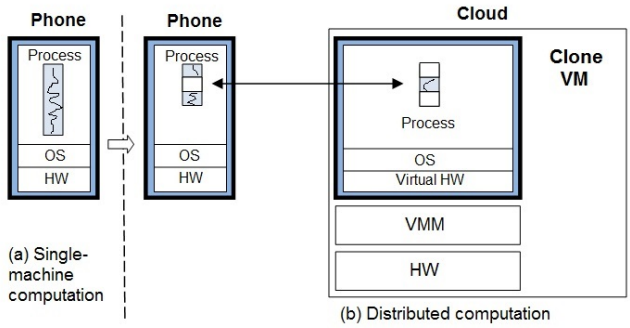


Fig. 5. Transformation from single to distributed execution model

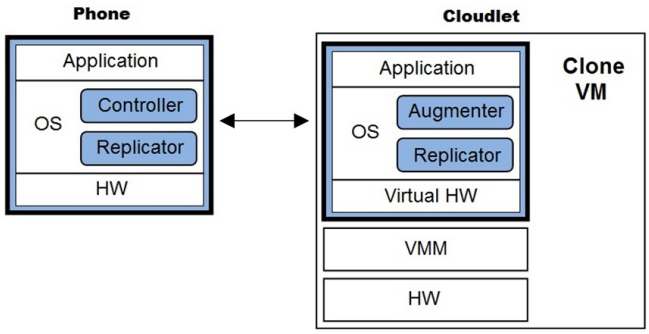


Fig. 6. High-level architecture of a cloudlet solution

platforms through cloning and computing transformation [26]. To do that, the system semi automatically transforms a single machine execution into a distributed execution where resource intensive execution is run in powerful clones. Figure 5 illustrates the high-level system model transformation from a single execution model into a distributed execution model.

Figure 6 presents our solution for augmented execution of the multilingual dictionary. It is performed in the following four steps:

- First clone of the smartphone is created within the cloud/cloudlet.
- The state of the primary phone and the cloud/cloudlet is synchronized periodically or on-demand.
- The whole application or parts of it are executed in the clone, on request or automatically.
- Results of the execution are sent back from the clone to the smartphone.

B. High level architecture model

The high-level view of architecture for the cloudlet-based multilingual dictionary is presented on Figure 6. All system components are running inside the operation system.

The Replicator is synchronizing the changes in phone software and state of the clone. Controller running in the smartphone invokes an augmented execution and returning its results back to the mobile device. While coordinating

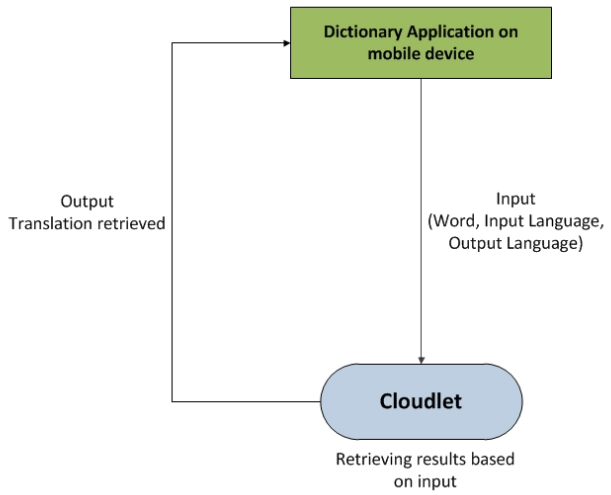


Fig. 7. Workflow Concept of the Cloudlet-based Multilingual Dictionary

the augmentation it interacts with the Replicator in order to synchronize the states. Augmenter running in the clone manages the local execution and returns the results.

C. Implementation Details

This Section contains implementation details on how the dictionary application will run on a mobile device. The user should enter the following input parameters: Input language in which the word will be entered, Output language to which the word has to be translated and the word which has to be retrieved, as presented in Fig. 7.

After the word is inserted and all other input parameters are specified, the cloudlet parses the data of the input language dictionary, looking in the database for specific input word and retrieves the correct meaning in the output language specified. Once the correct meaning is retrieved, it is sent back to the mobile device.

D. Client Cloudlet Interaction

Interaction between the client and the cloudlet is the main task. First step in this interaction is the request from the client to the cloudlet. After receiving the cloudlet's response, the client establishes the connection with it. Each client has a copy of the cloudlet configuration in the form of a patch file.

After the connection is established the cloudlet checks if it has the correct file. If it does, than the client does not have to send the file. Otherwise the client sends the patch file. After this, client sends the input file, containing input language, output language and the word for translation. When this file is copied to the guest OS, the execution starts and runs the dictionary application at the guest OS.

The translated word generated by the application is written as an output file and sent back to the client (mobile device).

VI. Discussion

The benefit of the cloudlet-based solution for the multilingual dictionary is performance based. We have argued that the current approach of using the mobile phone resources for a multilingual dictionary is relatively slow due to limited mobile phone resources. The cloud based approach using WAN connection is expected to be better and the cloudlet based solution will eventually offer the best performance.

Cloudlet approach is cost effective compared to cloud services. Resources required to establish a connection to the cloudlet are less compared to cloud. Application execution on the mobile device consumes a lot of the mobile's device resources (battery, memory, processing power). Also, a cloudlet uses Intranet or wi-fi which is free and available locally, compared to the distant cloud which requires an Internet connection.

Most of the people today that need translation of words or phrases are using web based dictionaries. We have realized user requirements research and performance analysis concluding about the average translation time to translate a word or phrase. Use of a mobile device with a wi-fi connection increases the average translation time to a double value of the average translation time.

According to the average response time delays for clouds and cloudlets we strongly believe that our approach will speedup the process and achieve better results. The expectations of average translation times for a word or phrase to be at most half the average translation time. This will be of great benefit for people who needs fast translation of the word and minor use of mobile's device resources.

Our solution is based on ideas introduced by Achanta et al. [26]. However, our solution does not plan to use datasets as in the Achanta's solution. To speed up the process of translation we introduce a table-based approach with predefined values for words and phrases with their corresponding translation. This approach saves time due to the time demanding operations that install and use a certain number of datasets on the cloudlet.

VII. Conclusion and Future work

Cloudlets are becoming the most important part in mobile application execution. With their use the user can execute resource rich applications without using mobile's device resources. In this paper we give a solution design of cloudlet based application, its architecture, execution and creation.

We explained details of the new approach to build a cloudlet based mobile application for multilingual dictionary by using the dynamic VM synthesis approach. The new model is presented with description of the VM overlay creation algorithm and the workflow concept. The described overall system architecture builds on the realization of augmenter and replicator modules for the cloudlet solution and the real multilingual dictionary application.

In the future we plan to deploy these solution for Android based mobile devices and compute execution times in both cases, with using cloudlet and cloud for execution of the application. From the results we hope that we can verify our theory that cloudlet offers better execution environment than the cloud and that running application on the cloudlet is much faster than running it on the cloud.

REFERENCES

[1] N. DI, "Gearing resource-poor mobile devices with powerful clouds: Architectures, challenges, and applications," *IEEE Wireless Communications*, p. 3, 2013.

[2] R. Kemp, N. Palmer, T. Kielmann, and H. Bal, "The smartphone and the cloud: Power to the user," in *Mobile Computing, Applications, and Services*. Springer, 2012, pp. 342–348.

[3] G. Huerta-Canepa and D. Lee, "A virtual cloud computing provider for mobile devices," in *Proceedings of the 1st ACM Workshop on Mobile Cloud Computing & Services: Social Networks and Beyond*. ACM, 2010, p. 6.

[4] K. Kumar and Y.-H. Lu, "Cloud computing for mobile users: Can offloading computation save energy?" *Computer*, vol. 43, no. 4, pp. 51–56, 2010.

[5] M. Satyanarayanan, P. Bahl, R. Caceres, and N. Davies, "The case for vm-based cloudlets in mobile computing," *Pervasive Computing, IEEE*, vol. 8, no. 4, pp. 14–23, 2009.

[6] B.-G. Chun and P. Maniatis, "Augmented smartphone applications through clone cloud execution." in *HotOS*, vol. 9, 2009, pp. 8–11.

[7] Officescope. (2014, February) Officescope. [Online]. Available: http://www.officescope.com/

[8] Producteev. (2014, February) Producteev. [Online]. Available: https://producteev.com/

[9] Mailboxapp. (2014, February) Mailboxapp. [Online]. Available: http://www.mailboxapp.com/

[10] Dropbox. (2014, February) Dropbox. [Online]. Available: https://www.dropbox.com/

[11] Docusign. (2014, February) Docusign. [Online]. Available: http://www.docusign.com/

[12] Linkedin. (2014, February) Linkedin. [Online]. Available: https://www.linkedin.com/

[13] Boomerang. (2014, February) Boomerang. [Online]. Available: https://play.google.com/store/apps/details?id=com.baydin.boomerang&hl=en

[14] Sliderocket. (2014, February) Sliderocket. [Online]. Available: http://www.sliderocket.com/

[15] Tripit. (2014, February) Tripit. [Online]. Available: https://www.tripit.com/

[16] UEFA. (2014, February) Uefa. [Online]. Available: http://www.uefa.com/

[17] Zomato. (2014, February) Zomato. [Online]. Available: http://www.zomato.com/tr

[18] Weather. (2014, February) Weather. [Online]. Available: http://www.weather.com/mobile/

[19] Skybet. (2014, February) Skybet. [Online]. Available: http://www.skybet.com/mob

[20] Pokerstars. (2014, February) Pokerstars. [Online]. Available: http://www.pokerstars.com/mobile/

[21] A. Rudenko, P. Reiher, G. J. Popek, and G. H. Kuenning, "Saving portable computer battery power through remote process execution," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 2, no. 1, pp. 19–26, 1998.

[22] J. Flinn, D. Narayanan, and M. Satyanarayanan, "Self-tuned remote execution for pervasive computing," in *Hot Topics in Operating Systems, 2001. Proceedings of the Eighth Workshop on*. IEEE, 2001, pp. 61–66.

[23] R. Balan, J. Flinn, M. Satyanarayanan, S. Sinnamohideen, and H.-I. Yang, "The case for cyber foraging," in *Proceedings of the 10th workshop on ACM SIGOPS European workshop*. ACM, 2002, pp. 87–92.

[24] Y.-Y. Su and J. Flinn, "Portable storage support for cyber foraging," in *2005 International Workshop on Software Support for Portable Storage*. San Fracisco, CA, USA, March 2005.

[25] A. A. Avinash, M. Sahitya, and V. Thomas, "Cloudlet-based multi-lingual dictionaries," International Institute of Information Technology Bangalore, Tech. Rep. IIITB-TR-2012-002, 2012.

[26] V. S. Achanta, N. T. Sureshbabu, V. Thomas, M. L. Sahitya, and S. Rao, "Cloudlet-based multi-lingual dictionaries," in *Services in Emerging Markets (ICSEM), 2012 Third International Conference on*. IEEE, 2012, pp. 30–36.

[27] D. Kovachev, Y. Cao, and R. Klamma, "Mobile cloud computing: a comparison of application models," *arXiv preprint arXiv:1107.4940*, 2011.

Mjolnirr: private PaaS as distributed computing evolution

D.I. Savchenko* and G.I. Radchenko*
*South Ural State University, Chelyabinsk, Russia
dmitry@mjolnirr.com, radchenkogi@susu.ac.ru

Abstract - Cloud computing became very popular during the last few decades. It provides convenient access to remote computing resources for individual users and organizations. However, there are still security issues arise if the private data is transmitted to the public cloud for processing. This issue can be resolved with private cloud systems. In this paper, we propose the design and implementation of a Mjolnirr private cloud platform for development of the private PaaS cloud systems. It provides infrastructure for cloud applications development, including software developer kit and message brokering system. Proposed cloud platform can be deployed on resources of a distributed computing system, including idling resources of personal computers. For a developer, a Mjolnirr application is represented as a collection of independent components communicating by a message exchange communication protocol.

Keywords - PaaS, Mjolnirr, Cloud computing, distributed system, message passing

I. INTRODUCTION

Cloud computing became very popular computing paradigm during the last decade [15], [25]. This approach to resource provisioning offers convenient access to computing resources, which makes it easy to use them for custom services development.

Before the introduction of the cloud-computing concept, there was an another distributed computing concept called "grid computing" [11]. It is still used by the scientific community to solve extra-large and resource-intensive scientific tasks. The name "grid" originated from the metaphor of the power grid. However, unlike electricity devices, integration of new resources to the grid-computing environment is not trivial.

Meanwhile, PaaS (Platform as a Service) [9] solutions provide the ability to create custom applications [6]. However, the problem is that the most PaaS solutions are deployed on remote hosting platforms. The owner of the application may not know exactly where his information is stored, as well as he cannot be sure about the safety and security of his information [20]. These issues might be solved by the private PaaS platform [10].

In this paper, we propose the design and implementation of the private cloud platform called "Mjolnirr". There are several solutions available now that provide private IaaS (VMWare Private Cloud, OpenStack) and PaaS (Yandex Cocaine, AppFog, Stackato) levels of service. Comparing to these solutions, main features of our approach are integrated messaging subsystem, flexible workload management and native UNICORE [24] grid environment integration module.

This paper is organized as follows. In section II we present the concept of the Mjolnirr cloud platform. In section III we describe the results of the analysis of existing cloud platforms and compare them with the Mjolnirr platform. In section IV we describe the architecture of the Mjolnirr platform and the component communication protocol. In section V we describe the implementation of the Mjolnirr platform and demonstrate the process of custom applications development. In section VI we describe the platform performance evaluation. In section VII we summarize the results of our research and further research directions.

II. MJOLNIRR PLATFORM CONCEPT

The Mjolnirr private cloud platform provides development of a Java-based private PaaS cloud infrastructure. A Java application or library can be implemented as a Mjolnirr-based service. The Mjolnirr platform provides infrastructure for cloud applications development, including software developer kit, message-brokering system. For a developer, an Mjolnirr application is represented as a collection of independent components communicating by message exchange. This approach allows development of flexible and scalable cloud applications.

Mjolnirr also provide integration with the UNICORE grid environment through the DiVTB [18] platform. The DiVTB (Distributed Virtual Test Bed) platform provides a task-oriented approach for solving specific classes of problems in computer-aided engineering through resources supplied by grid computing environments. Thus, Mjolnirr can be used to provide infrastructure both for scientific projects with the grid systems and in a business infrastructure.

During our research, we should:

- analyze technologies of cloud systems and private cloud platforms development;
- develop Mjolnirr platform architecture;

The reported study was partially supported by Grant Fund for Assistance to Small Innovative Enterprises in Science and Technology: "UMNIK" Programme, research project No. №0000829 and by RFBR, research project No. 14-07-00420-a.

- implement Mjolnirr platform;
- evaluate the performance and scalability of the Mjolnirr platform.

III. ANALYSIS OF EXISTING SOLUTIONS

An investigation shows that C-level executives and IT managers in enterprise companies have issues with integration of cloud computing in their data processing [5]. One of the most serious concerns is the possibility of data breaches. A cloud provider can give access to the company's private data (accidentally or intentionally) or may bring harm to the data owner [14], [21].

For the prevention of such privacy threats, it is possible to use the encryption of data that is stored in the cloud [19], [22]. However, this approach is effective only when cloud is used just for storage. If the data is processed in the cloud, it become available decrypted in the memory of the host, where the processing occurs [4]. In addition to this drawback, the owner of the data does not control the location of his virtual machine, so it can be moved to the physical computer with the virtual machine that contains malware. It may cause the ban of the virtual machine or forfeiture of a computer containing these virtual machines.

Nowadays, there are two ways to ensure the data security in the cloud. The first way is called "trusted computing". It ensures security of virtual machines in the cloud [7]. Nevertheless, your data cannot be completely safe. For example, in present IaaS clouds, the virtual machine can be dynamically moved to the other host, but the concept of trusted computing provides security only for virtual machines running on the same host. Otherwise, the concept of TCCP (Trusted Cloud Computing Platform) [12] solves this problem by creating the safe environment for running virtual machines. In fact, neither of these approaches solves the problem of VMs placement on the same host with a malicious VM.

Another way to deal with the security issues is to deploy the cloud infrastructure on the private hardware. In this case, the owners themselves will protect all the data. However, buying and maintaining of the hardware is more expensive than rent of computing resources [9], [10].

The simplest way is to create a private cloud system, and there are several different products existing in this area. They can be divided into two categories according to the level of abstraction – infrastructure (IaaS) solutions (e.g. OpenStack and VMWare Private Cloud), and platform (PaaS) solutions (e.g. Yandex Cocaine, AppFog and Stackato).

OpenStack [17] is a popular open source package for private cloud systems provision. It provides infrastructure layer based on virtual machines abstraction. Software developers should install web servers, load balancers and other specific services manually. It offers great opportunities, but IaaS maintenance is expensive. VMWare Private Cloud [16] also provides IaaS-level solution, but on a commercial basis. VMWare Private Cloud offers advanced virtualization capabilities, as well as, the commercial support.

Yandex Cocaine and AppFog platforms provide the ability to create private PaaS solutions based on application containers [2], [3]. These platforms allow development of a "Heroku-like" [13] application hosting. They provide a number of built-in modules and a server infrastructure. Stackato [1] provides all advantages of the mentioned solutions and provide local application store. Some of them use Docker [8] containers to provide isolated environment to the applications, but Docker can be used only on Linux hosts. However, all of the above solutions consider custom applications as a monolith and ignore its internal structure that is why it is not possible to automatically and effectively balance the load on the individual subsystems applications. In addition to this drawback, none of these solutions considers end-user workstations as computing resources providers.

We decided to create a Mjolnirr cloud platform, which solves these problems as follows:

- Using popular open source libraries and programming languages, as well as the opportunity to work not only on the server platform, but also on the

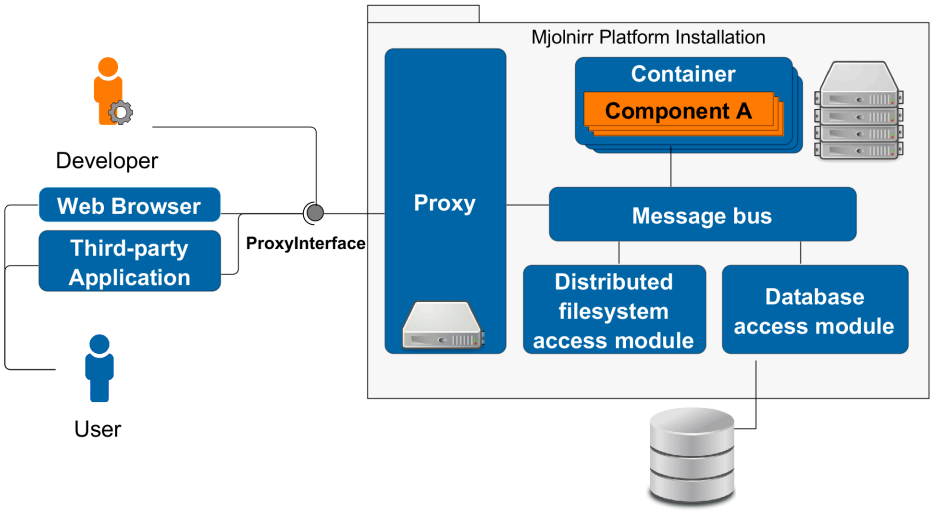


Figure 1: Mjolnirr platform architecture

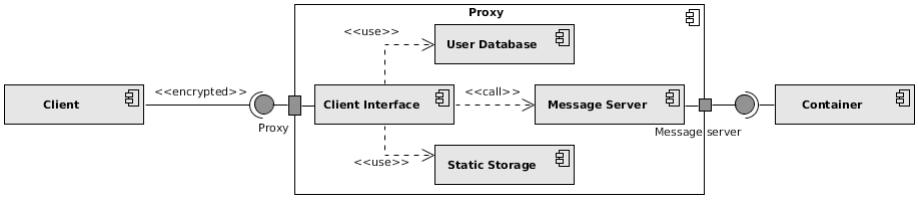


Figure 2: Mjolnirr proxy architecture

personal computers, using idling resources of the computer system, will provide cost reduction.

- Using popular programming languages, developer tools and SDK will provide ease of application development.

- Integration of new resources ease will be provided by the application architecture modularity and custom components reusability.

IV. ARCHITECTURE

Mjolnirr platform architecture is presented at the figure 1. The Mjolnirr platform consists of the following components.

- *Proxy* provides access to the cloud system for the external clients and manages the communication between cloud application components. Proxy is the only component that accessible from the external network;

- *Containers* are deployed on computing nodes and responsible for hosting of cloud applications components and message transmission. It can be deployed both on the personal computer and on the nodes of dedicated server;

- *Components* are custom applications, created to run in cloud environment. Each component has a unique name. UI components (applications) are multi-page applications, it means that applications has different representations for different actions, like different web-pages of one web-application;

- *Client* (browser or third-party application) is an application, which render user interfaces and provide user interaction with cloud applications, hosted in Mjolnirr platform. All clients should use encrypted channel and client certificates to communicate with the proxy.

A. Proxy

The Proxy (see fig. 2) provides access to Mjolnirr system from the external network. The Proxy performs the following actions:

- it stores and provides static resources (page layout descriptions, images etc.) of the deployed cloud applications in the Static Storage;

- it acts as a messaging server for the components of cloud applications;

- it handles client's requests to the Client Interface;

- it performs the authorization and authentication of users.

The external Proxy interface provides the following methods as RESTful API:

- *getUI (applicationName, pageName)* – returns the page layout description;

- *getResourceFile (applicationName, resourceName)* – gets the static file;

- *processRequest (componentName, methodName, args)* processes a client request, redirecting it to the first free suitable component. This method can be called directly from the application page (e. g. with JavaScript).

B. Container

The container (Fig. 3) provides cloud application component hosting. The container provides an API for remote components instances method invocation. The Mjolnirr installation can have any number of containers.

Any Mjolnirr-based application consists of independent components, which use a built-in messaging system, implemented in the basis of the Publisher-Subscriber pattern. The Proxy is responsible for message queue maintenance. The Message Server of the Proxy provides publisher-subscriber messaging service for cloud application components. Mjolnirr Containers subscribe to Message Channels that operates as a broadcast delivery – any message sent to the Message Channel will be transmitted to the subscribers of this channel. Each cloud application instance is subscribed on two types of Message Channels:

- Component Public Channel: every instance of the

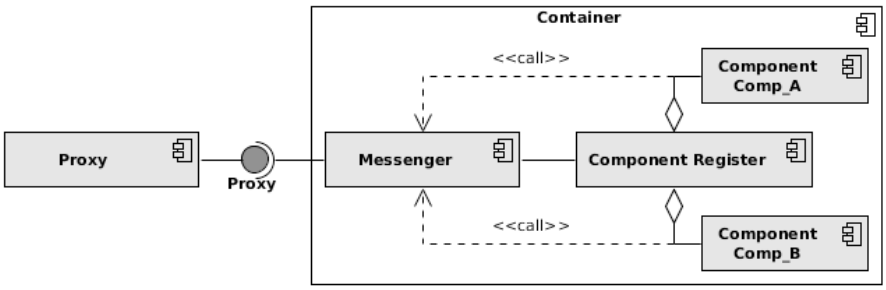


Figure 3: Mjolnirr container architecture

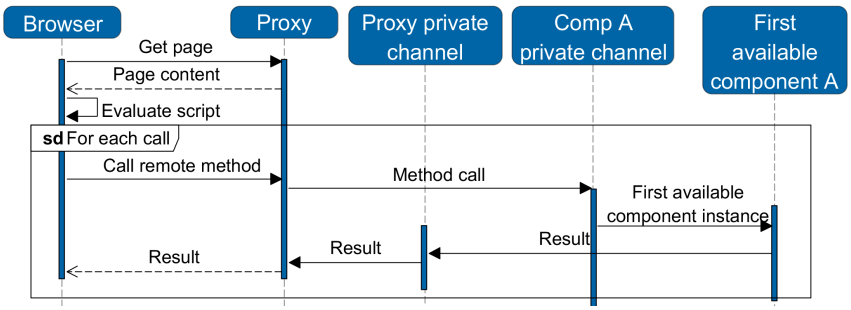


Figure 4: Mjolnirr container architecture

cloud application component is subscribed on this public channel. This a listener channel – when any message come to this channel, the appropriate component instance will be invoked.

- Instance Private Channel: provides direct communication between instances.

When container starts, it performs several initialization steps. The order of the container initialization:

- Container registers in the proxy database and receives the list of components to load in response;
- For each component from the list:
  - The container checks it's local cache, for each missing package container downloads it from the proxy;
  - Container runs the component;
  - Container subscribes on the Component Public Chanel and Instance Private Channel for the loaded component.

Container provides messaging API to the hosted component. Typical message transmission sequence looks like shown below (Fig. 4):

- when one component calls another component remotely, the container sends the call to a Component Public Channel (for example, when the *Comp_A* component calls the *Comp_B* component, the application *Comp_A*, will send the request to the channel *CPC_B*);
- The first available instance of the component in the cloud system processes the request;
- The response is returned to a private IPC channel of the instance that sent a message.

In addition, container has an opportunity to work in stand-alone mode. In this mode, the container does not support communication with other containers and acts as a stand-alone computing system (container and proxy at the same time).

C. Components

From the developer's point of view, Mjolnirr cloud application is a collection of independent components communicating by message exchange. Components represented as a package that contains the following information:

- manifest, that provides the interface of the component, including description of provided methods and their parameters;
- executables to handle incoming requests;
- static files, used in pages rendering (images, page layout descriptions and scripts) for UI provision.

Each component can be:

- Application Component: provides the user interface definition, scripts, styles and UI-specific actions. Optionally contains complex logic.
- Module Component: represents a single entity in the domain logic of the application. The Module Component provides data processing and storage, but doesn't provide interface and static files.

D. Clients

Mjonirr platform uses HTML and JavaScript to represent end-user interface in common Internet browser. But it's possible to consume Mjolnirr resources in third-party application because of using commonly used and standardized protocols.

V. Implementation

A. Components

Each custom component must have a manifest: an XML file that contains all the necessary information about the component's name, the name of the main class, list of methods and a list of parameters for each method. Sample manifest for the Calculator service is shown below.

```
<hive>
<application
  name="Calculator"
  classname="org.exmpl.Calculator">
  <methods>
   <method
     name="calculate"
     type="java.lang.String">
    <parameters>
     <parameter
       type="java.lang.String"/>
     </parameters>
    </method>
  </methods>
</application>
</hive>
```

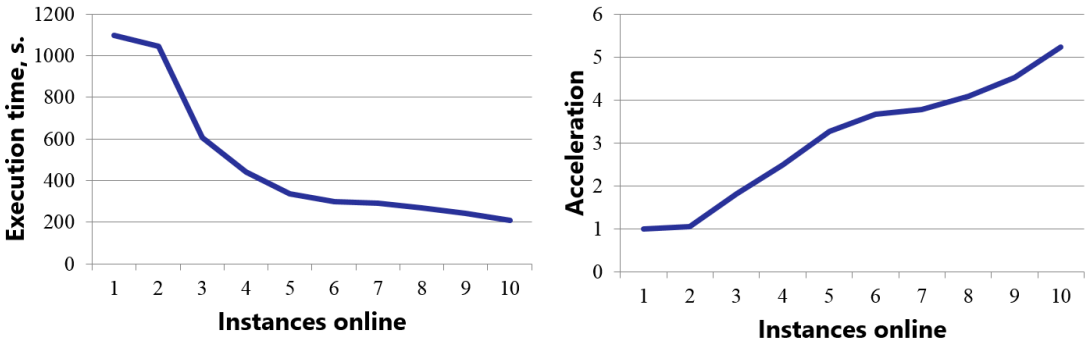


Figure 5: Performance evaluation

The container parses the manifest for each loaded component.

As stated above, the manifest must have the name of the main class as a fully qualified name of the facade class of the described components. Each facade class must have an implementation of the method "*initialize*". The container calls this method immediately after instantiating the component and passes the component *context there*. This context contains information about the working directory of the current application instance and the directory containing the configuration files.

B. *Messaging subsystem implementation*

Mjolnirr platform messaging subsystem is implemented on the basis of the open source HornetQ queue, which is built into the Proxy component. HornetQ was chosen for the following reasons:

- it can work in embedded mode;
- it provides high queue management performance [23].

The container provides messaging API to the components. Example of using messaging interface is shown below.

```
Communicator communicator =
        new HornetCommunicator();

YourResultClass result =
Communicator
     .sendSync(
        componentContext,
        "componentName",
        "methodName",
        argsList,
        YourResultClass.class);

communicator.send(
   componentContext,
   "componentName",
   "methodName",
   argsList,
   new Callback<YourResultClass>() {
       public void
       run(YourResultClass res) {}
   });
```

C. *UNICORE integration*

The Mjolnirr platform provides a UNICORE 6 integration module. This module uses DiVTB Server for communications with grid environment. It enables Mjolnirr applications to use UNICORE grid resources easily.

The concept of DiVTB allows users to split the supercomputing simulation in two phases - building experiment and launching the test stand. The main advantage of the DiVTB concept is the ability to use several times once set up the experiment.

UNICORE integration module provides the following methods:

- *uploadTestBed* – used for test bed archive uploading;
- *indexTestBed* – indexes specific text bed;
- *indexAllTestBeds* – indexes all available test beds;
- *createExperiment* – create experiment from one of the test beds;
- *startExperiment* – start experiment with specific parameters;
- *getStatus* – get experiment status (started, finished or failed);
- *getResults* – get experiment results.

VI. PERFORMANCE EVALUATION

Mjolnirr platform was evaluated using text-processing experiment. 1 gigabyte of text data was divided on 100 parts and distributed between available worker components for processing. Each worker divided text on words and counted frequency for each unique word. Pieces of work were distributed automatically – each worker polled Message Bus to receive new task. The results of the experiments are shown on the figure 5.

We used 1 virtual node with 2 CPU cores and 2048 MB RAM for proxy node and 10 virtual nodes with 1 CPU core, 512 MB RAM as worker nodes. Virtual machines were installed on the computing server node

with the following characteristics: 2 Intel Xeon X5680 CPU (6 Cores, 3.33 GHz) 12 GB DDR3 RAM.

Experiments have shown that the platform is stable. Average execution time on **10** containers was **219** seconds, comparing to **1107** seconds using 1 worker. Thus, acceleration of parallel word frequency counter task reached **5.3** on **10** worker nodes.

VII. CONLUSION

In this article, we described the design and implementation of a private cloud platform called Mjolnirr, which allows creating distributed cloud applications on private computing resources. Main features of the described platform are advanced messaging system and distributed computing support.

As a further development of the Mjolnirr platform we will investigate and implement application-level migration support, integration with the advanced resource monitoring systems, flexible adaptation to load changes, advanced system security and application store. The application store will reduce the number of duplicate software products and simplify the creation of individual business infrastructure to meet the needs of a particular company.

REFERENCES

[1] Activestate.com. *Stackato: The Platform for the Agile Enterprise ActiveState*. [Online] Available from: http://www.activestate.com/stackato [Accessed 22 Jan 2014].

[2] Api.yandex.com. *Cocaine. Cocaine technology description*. [Online] Available from: http://api.yandex.com/cocaine [Accessed 22 Jan 2014].

[3] Appfog.com. *AppFog - Platform as a Service for public clouds*. [Online] Available from: https://www.appfog.com/ [Accessed 22 Jan 2014].

[4] Armbrust M, Fox A, Griffith R, Joseph A, Katz R, Konwinski A, Lee G, Patterson D, Rabkin A, Stoica I, Others. A view of cloud computing. *Communications of the ACM*. 2010; 53 (4): 50--58.

[5] Avanade.com. *Global Study: Cloud computing provides real business benefit | Avanade*. [Online] Available from: http://www.avanade.com/us/about/avanade-news/press-releases/Pages/Global-Study-Cloud-Computing-Provides-Real-Business-Benefits-But-Fear-of-Security-and-Control-Slowing-Adoption-page.aspx [Accessed 22 Jan 2014].

[6] Chang W, Abu-Amara H, Sanford J. *Transforming enterprise cloud services*. Dordrecht: Springer; 2010.

[7] Christodorescu M, Sailer R, Schales D, Sg, Urra D, Zamboni D. Cloud security is not (just) virtualization security: a short paper. 2009;: 97--102.

[8] Docker.io. *Docker: the Linux container engine*. [Online] Available from http://www.docker.io/ [Accessed 03 Feb 2014]

[9] Fanning K, Centers D. Platform as a service: Is it time to switch?. *Journal of Corporate Accounting & Finance*. 2012; 23 (5): 21--25.

[10] Fortis T, Munteanu V, Negru V. Towards a service friendly cloud ecosystem. 2012;: 172--179.

[11] Foster I, Kesselman C. *The Grid 2*. Burlington: Elsevier; 2003.

[12] Garfinkel T, Pfaff B, Chow J, Rosenblum M, Boneh D. Terra: A virtual machine-based platform for trusted computing. 2003; 37 (5): 193--206.

[13] Heroku.com. *Heroku | Cloud Application Platform*. [Online] Available from: https://www.heroku.com/ [Accessed 22 Jan 2014].

[14] Kamara S, Lauter K. Cryptographic cloud storage. *Proceedings of Financial Cryptography: Workshop on Real-Life Cryptographic Protocols and Standardization*. 2010;: 136-149. [Accessed 22 Jan 2014].

[15] Mell P, Grance T. The NIST definition of cloud computing (draft). *NIST special publication*. 2011; 800 (145): 7.

[16] Peng J, Zhang X, Lei Z, Zhang B, Zhang W, Li Q. Comparison of several cloud computing platforms. 2009;: 23--27.

[17] Pepple K. *Deploying OpenStack*. Sebastopol, CA: O'Reilly; 2011.

[18] Radchenko G, Hudyakova E. Distributed Virtual Test Bed: an Approach to Integration of CAE Systems in UNICORE Grid Environment. *MIPRO 2013 Proceedings of the 36th International Convention*. 2013;: 183-188. [Accessed 22 Jan 2014].

[19] Ren L, Yu X, Fletcher C, Van Dijk M, Devadas S. Design Space Exploration and Optimization of Path Oblivious RAM in Secure Processors.. *IACR Cryptology ePrint Archive*. 2013; 2013: 76.

[20] Rhoton J, Haukioja R. *Cloud computing architected*. [Tunbridge Wells, Kent]: Recursive Press; 2011.

[21] Ristenpart T, Tromer E, Shacham H, Savage S. Hey, you, get off of my cloud: exploring information leakage in third-party compute clouds. 2009;: 199--212.

[22] Santos N, Gummadi K, Rodrigues R. Towards trusted cloud computing. 2009;: 3--3.

[23] Spec.org. *All Published SPEC SPECjms2007 Results*. [Online] Available from: http://www.spec.org/jms2007/results/jms2007.html [Accessed 22 Jan 2014].

[24] Streit A. UNICORE: Getting to the heart of Grid technologies. *eStrategies*. 2009; 3: 8-9. [Accessed 22 Jan 2014].

[25] Vaquero L, Rodero-Merino L, Caceres J, Lindner M. A break in the clouds: towards a cloud definition. *ACM SIGCOMM Computer Communication Review*. 2008; 39 (1): 50--55.

VISUALIZATION

Visualization Support for Code Development in EUROfusion Integrated Modelling

Leon Kos and ITM-TF contributors*
University of Ljubljana, Mech. Eng., LECAD, Ljubljana, Slovenia
leon.kos@lecad.fs.uni-lj.si

***Abstract*—The need for diverse visualization tools within integrated modelling is based on the requirement of physics codes to be coupled under the tailored database model from which these visualization tools read data. We discuss these visualization requirements for fusion integrated modeling, and the existing visualization tools developed within the scope of the European Integrated Modelling (EU IM) framework. The datastructure model provides fundamental description for data exchange between codes that are developed in a variety of programming languages. From this description in XML schema definition one can generate required language interfaces and database access layer routines. The persistent storage database has to efficiently support codes running on HPC with memory caching mechanisms. Visualization tools can be treated as one of the codes running in-situ or as post-process. The complexity of the fusion datastructure and several visualization tools requires that scientists describe standard visualizations in XML schema directly in order to instruct visualization tools what are meaningful visualizations available in the very database. Besides standard visualizations there are specific visualizations that cannot be easily described (as mapping or axes linking) without algorithm. For the custom visualizations we provide Python snippets collected in user-shared library. Interfaces to scientific workflow engine Kepler for visualization with VisIt visualization tool and Matplotlib are presented.**

I. Introduction

The Code Development for Integrated Modelling EUROfusion project aims to provide to the European fusion community a suite of validated codes for the interpretation of present experiments and predictive modelling of future reactors. This effort builds on the European integrated modelling platform and framework developed within the ITM-TF[1]. Verification and validation of predictive and interpretative simulation runs require that modelling platform allows combining experimental and simulation-data at various stages throughout scientific workflow development. Traceability (provenance) in Kepler[2] workflow engine used within the EU IM framework should provide easy verification and validation at the modelling level. Sharing workflows between scientists is "easy part" of the *workflow-based provenance* and must be accompanied with often *big data* that orchestrated codes needs to trace. Data model seems to be one of the challenges that cannot be universally addressed for all domains. Fusion-oriented framework DataSpaces [3] separates data model into server–client model with distributed hash tables with the goal of providing transparent memory-to-memory transfer (staging) to applications. For "expensive to get" data, *lookup* and metadata is provided in the DataSpaces three-layered architecture. The EU IM data model currently relies on a single-layered Universal Access Layer[4] (UAL) that concentrated on accessing a tailored (tokamak) datastructure description[5]. Data gathered within CPOs (Consistent Physical Objects) is used for "partitioned" interchange between codes in "strongly" coupled workflows. Similarly to CPOs, ADIOS system [6] provides *groups* that describe collections of data written at once. ADIOS as a HPC library for input/output (IO) provides simplified and configurable write strategy to parallel codes. The EU IM approach uses UAL library routines in upgraded fusion codes for data exchange through a structured CPOs database. Staging and write strategy is a part of the UAL configuration and can range from direct file access, memory-cached or parallel [7] access. The UAL back-end responsible for storage uses MDSplus and HDF5 database formats. Both formats are commonly used in scientific and experimental (fusion) environments. As mentioned before, single-layered UAL is providing simple IO with minimal number of functions to codes for fetching and storing CPOs. UAL adaptation to hashed client–server model as in DataSpaces, DART [8] asynchronous data communication layer or using ADIOS should not change the UAL front-end routines used in codes. Code specific data transport configuration and provenance can be governed by the workflow engine (e.g. as XML file used in ADIOS). For distributed workflows where database access is not easy to arrange on must rely on serialization of CPOs for exchange between compute sites. One such approach in fusion modelling is MAPPER project [9] that can couple codes with enveloping them into communication kernels and configure into workflow.

The EU IM effort turned its focus into creation of workflows for specific physics applications that can easily include similar physics codes for verification and comparison in integrated modelling. Key principle for these workflows is not UAL but rather *datastructure* description (ontology) consisting of CPOs. Consistency enforced by CPO data model requires that existing codes need to be "adapted" to be able to retrieve/store results in database. This task requires developer's knowledge of the code and CPOs description that needs to be mapped to each other in enforced units (MKSA and eV).

* See the Appendix of G. Falchetto et al. 2014, Nucl. Fusion **54** 043018 doi:10.1088/0029-5515/54/4/043018

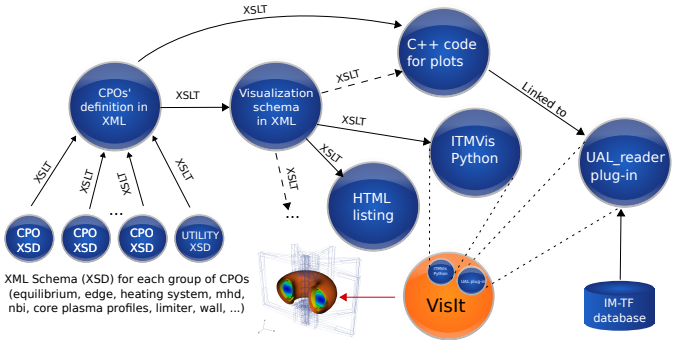


Fig. 1. XSLT flowchart for standard representation transformations used for visualization.

After (necessary) adaptation, code can be executed as a black-box within *scientific workflows* that couples different codes in a compatible way and sequence. Similarly to CPOs, ITER Integrated Modelling & Analysis Suite (IMAS) [10] foresees the development of a datastructure composed of Interface Data Structures (IDSs) as standardized entities for use between physics components. The collection of CPO descriptions form together a complete data model for diverse simulations that can also contain imported experimental data. This approach enables a direct comparison of results with the experiment and/or use experimental data as an input. The implementation phases and continuous development of the EU IM framework are simplified by the fact that the CPO data structures are described by an XML schema definition (XSD), which can be easily modified. This semi human-readable description allows rigorous validation, creation of data bindings and translations for different purposes. The XSD data-structure description is mainly used to generate CPO definitions in an XML format obtained by applying the XSL translation [11] (XSLT) language. Consistency of CPO definitions is assured by the XSD description, so that the derived CPO definitions in XML are consistent too. From the CPOs schema one can automate (through XSLT processing) generation of "include" definitions and UAL linking for all supported languages. From each CPO defined with XSD we then translate into single *CPODef.xml* file that includes complex and common structure types defined in *utility.xsd*. Figure 1 on the left shows this initial transformation from many CPO definitions into single CPOs' definition in XML. The fundamental purpose of *CPODef.xml* is a description of the EU IM persistent storage database. As most of the CPOs are time dependent, the EU IM database stores *slices* of each CPO during simulation iterations or experiment sampling. Slicing occurs at different time scales for each CPO depending on the physics involved [12, see Fig. 12 therein].

Described data model and its access is a basis for visualization tools that are aimed to support users in all stages of modelling framework usage. To support the diagnostic of the time-dependent data in CPOs, several visualization tools are used depending on the scales and *representation* of the data that is regularly used in simulations. For many simulations, custom visualization are created using general-purpose graphics libraries. To avoid such custom approaches and to provide *standardized* sets of visualizations EU IM aims to provide tools that can be used for visualizations without the need for scientists to manually program each *plot* and allow creation of reusable custom visualizations.

TABLE I
THE EU IM VISUALIZATION TOOLS CAPABLE OF UAL DIRECT ACCESS

EU IM Tool	Graphical plot config (user level)	Non-interactive processing	Input data pre-processing	Publication ready figures
ITMVis	no (basic)	yes	yes	yes - matplotlib
VisIt	yes (basic)	yes	limited	yes - raster gr.
Matlab	yes (expert)	yes	yes	yes - builtin
ISE	yes (basic)	no	no	no

II. VISUALIZATION SUPPORT IN THE EU IM FRAMEWORK

Graphical inspection of data in variety of ways becomes more difficult with increasing data complexity. As time is a key physical quantity of tokamak modeling, this quantity is included in nearly every CPO and can be assumed as an additional dimension when time-varying properties are inspected. Usually, time-dependent visualization is done by specifying a time point or selecting a cycle (frame or time-slice) from the UAL database. From that point data can be retrieved and visualized with different tools in a variety of ways. Visualization tools capable of accessing UAL data directly (without exporting or converting) were developed to provide required visualizations. Three approaches are used by the EU IM for producing visualization: (i) programming languages, (ii) general tools, and (iii) specialized tools with capabilities as shown in Table I. Scripting languages like Python and Matlab are suited for all kinds of data processing. However, their use requires development effort. Non-scripting languages (C++, Java, Fortran) for producing visualizations are even less attractive. They are used for application programming and within visualization tools. In fact, there are no general purpose tools available that can cover all aspects of usage. The Integrated Simulation Editor (ISE) was designed with Kepler [13] workflow control in mind in a "study"-like manner and currently only provides simple 2D signal editing and visualization.

A. ITMVis

One goal of the EU IM effort is to consolidate and extend the visualization capabilities of the platform, providing developers with tools for (automated) creation of specialized plots that can be easily distributed to end users and tools for creating plots for code/workflow development and debugging. Eventually, users will benefit from easy-to use, interactive tools for data exploration, easily available custom visualizations for specific physics problems/workflows and tools that are capable of producing publication-ready plots. Towards this goal the EU IM developed a structured 2D Python visualization library — ITMVis which unifies the existing approaches, provides extensibility and enables reuse of custom visualizations for various target visualization tools (back-ends). The ITMVis library encompasses the following:

1) A standard front-end consisting of command line Python scripts for accessing UAL database by prescribing run, shot, time, user, tokamak and CPOs needed for visualization.
2) A structured repository of code fragments for computing data required for plot, possibly combining data from various CPOs. For each plot *metadata* and *plotdata* function is provided.
3) A collection of "standard" processing functions which can then be shared among multiple visualization scripts.
4) Collects and makes code owned/developed by specific projects available to everybody in a central place and provides a plot sharing mechanism for users to reference exported visualization scripts.

Developers/users need to concentrate only in data processing and plot description by providing *metadata* and *plotdata functions*. Metadata describes the plot and verifies data availability so the user is informed whether a specialized plot can be produced or not. Only if a database contains the required data, the *plot data* function can be executed. To detect modules containing metadata and plot data functions in the Python search path, some function naming conventions are prescribed. Backends implemented as outputs are text, VisIt and matplotlib [14] library with several graphical formats. ITMVis implemented a Python library can be included in all Python scripts that can regularly include matplotlib along with UAL wrapper library that requires NumPy/SciPy stack.

B. VisIt

For complex 3D visualizations the EU IM effort adopted *VisIt* [15] visualization software (open-source, maintained by LLNL) upgraded with custom *UAL_reader* plugin developed under EUFORIA/JRA4 [16] project. VisIt is popular and extensible visualization software that provides various representation modes and data post-processing in interactive and non-interactive (batch or scripting) mode. With custom plugin is able to read CPOs from the the EU IM database directly. VisIt as a general visualization tool is capable of complex visualizations that are primarily 3D. Experimental data that is imported into UAL by the EU IM tool *exp2itm* mostly comes in 1D time-dependent (TD) form and can be inspected by surplus of graphing tools. VisIt role in workflows is primarily dedicated to inspection of 3D or even 4D data resulting from "integrated modeling". Normally, insight into 3D data is extracted by variety of *operators* at selected cycle (moment in time). Animation and *time-slider* can be useful TD tools for general overview of inspected phenomena. VisIt uses *FlipBook animation* for showing time-varying databases. Controls for moving though cycles/time are similar to many software players (or VCR). UAL plug-in for VisIt (UAL_reader) enables VisIt fusion data reads through the UAL in order to build the resulting visualization. In order to connect VisIt with UAL two components of the plugin was developed as shown in Fig. 2. Standard visualizations are fetched directly from the database. For custom plots ITMVis library is included through plugin-embedded Python interpreter and accesses database by Python

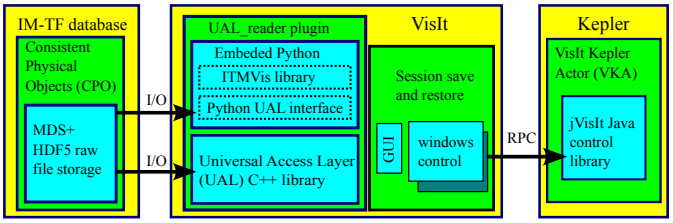


Fig. 2. Schematics and communication of basic components for visualization with VisIt under Kepler workflows.

UAL interface.

VisIt is built on top of *Analysis and Visualization Toolkit* (AVT) that has strong correlation to VTK Toolkit made by Kitware. Data structures from UAL needs to be represented in AVT compatible way. When CPO is read into memory all data is available and information on how this data can be "naturally" represented in coordinate system is being amended by physicists. "Natural" *representation* of data of interest is related to conventions and physicist's way of thinking. For example, one needs to decide which field represent x and y-axis for *1D Curve*. A representation is, in fact, a description of the visualization. There can be multiple representations of the same data. The *UAL_reader* plugin for VisIt (see Fig. 1) utilizes *representation tags* that are included in the XSD files describing the CPOs. This means that datastructure "knows" how it can be visualized. This is similar to the concept of *views* in SQL, which are stored in a database directly (except that these would be stored directly in the database). Representation tags are just a description that does not enlarge storage but facilitates views on data. Each representation consists of name, mesh type, links for the axes, and data attributes. The representation name is arbitrary and is shown as the last name in the VisIt path when adding plots. The data attribute can be either a scalar or a vector and describes the kind of data mapped on the mesh represented by one or more links. Meshtype describes the type of variable mapping in space. Several meshtypes exist: i) *curve* as a polylines plot, ii) *axis1D0* X axis given as a link1 vs. a bunch of curves from a matrix; iii) *mesh1D0* Contours (closed curves) as a (2D) surface mesh, iv) *rectlinear0* or *curvilinear0* meshes in 2D and 3D. Each meshtype can have a variable number of links (link1, link2 and/or link3) and thus different kind of visualizations that are constructed as VTK objects in VisIt. Links provide axis mapping to corresponding vectors or matrices that constitute a mesh. The topology name at the end of the meshtype means (0D - points, 1D - lines, 2D - surface, 3D - volume). The number at the end distinguishes variations of the same meshtype (0, 1, ...). If no topology is given, a general description is assumed (2D and 3D).

UAL_reader plugin is a basis for all kinds of visualizations with VisIt. As shown in Fig. 2 customized plots can be accessed through ITMVis library and *standard* plots by direct translation from datastructure. Both parts of the plugin are complementary and can be used together in a dashboard-style setup for *European Transport Solver* [12] shown in Fig. 3 where some difficult to combine data were rather prepared by custom plots.

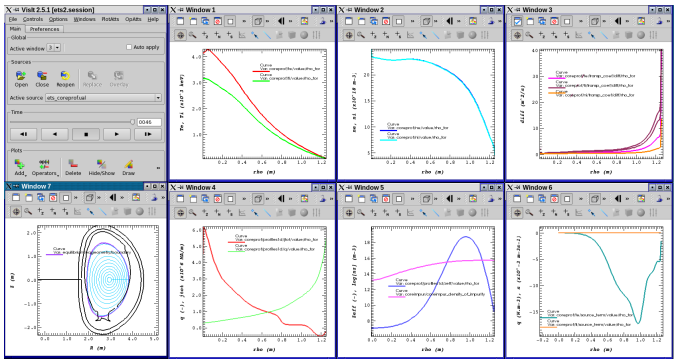


Fig. 3. European Transport Solver (ETS) convergence monitoring with VisIt.

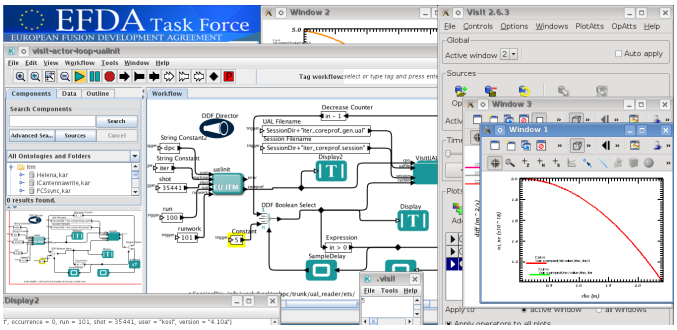


Fig. 4. The VisIt Kepler Actor – VisItUAL in a workflow within a loop reading data through the Universal Access Layer. Multiple VisIt windows configured with session save/restore mechanism are shown and updated during iterations.

C. Visit Kepler Actor

The EU IM selected the Kepler [13] scientific workflow engine which enables graphical programming of workflows. Data flow between processes (called *actors*) is modelled by wires. Through wires only small amount of data (signals) are passed, usually just single numbers, strings and structures with CPO names, which identify database entries. The Actual data exchange between IM actors encapsulating fusion codes is then implemented by direct access to the "real" data through the UAL library.

For inclusion of VisIt in the Kepler workflows the VisIt Kepler Actor (VKA) was developed. The VKA in Fig. 2 can restore session created previously within VisIt but with new source database. Session saved within VisIt graphical user interface (GUI) consists of all windows, applied plots, operators and attributes. This means that user can configure session under VisIt GUI inside or outside Kepler. Then user builds a particular visualization for the quantities of interest. Finally, user stores a VisIt session file which contains all the information needed to restore the state of VisIt when the session file is generated. In the second stage, the user launches Kepler, inserts the VKA to his workflow and add as parameter of the VKA the session file generated at stage 1. When the workflow is executed, once the VKA is activated by the workflow, one or several visualization windows will pop up with the data contained in the incoming database. The resulting visualization is rendered with the parameters contained in the session file. This means that session files are special tailored descriptions of visualizations with expected inputs. Normally most scripting languages (IDL, Matlab, gnuplot, ROOT, ...) that require visualization also generate scripts for visualization by coding. VisItUAL is a similar approach but without coding as user can with GUI save XML session file. Such session can then be shared and collected for visualizations of interest and used as standalone or within Kepler workflow. Such approach is useful for fusion codes and experiments that output complete results in UAL database but researchers are interested in specific phenomena and selected variables and ranges that needs to be monitored within larger scientific workflow supported by Kepler. For UAL database and CPO support composite actor Coupling VisIt to UAL is supported by UAL_reader database plugin. VisIt by design needs to open database provided as filename. UAL database plugin thus needs .ual file provided before it can open UAL within VisIt. Creation of this file can be manual with one line specifying CPO, shot and run. After UAL database is opened with VisIt than one can create a session file for restore within Kepler workflow. Fig. 4 shows such a workflow with user supplied visualization settings. Location of the *ualfile* is given for automatic creation at the specified location. This location can be the same as previously manually generated session file. VKA allows visualizations of multiple CPOs and includes its own GUI written in Java for controlling interactive work under Kepler and restoring VisIt GUI. It can also be used in a synchronized loop as in Fig. 4 that cycles through prescribed time steps found in CPOs.

D. The visualization schema

The *ITMVis* library tries to follow the VisIt data description by splitting visualizations of CPOs into meta-data and plot data. This separation is a natural choice for all tools that want to prepare a list of possible visualizations depending on data availability. It should be noted that not all CPOs are filled with data when running a particular simulations. The ITMVis library concentrates on *custom plots*, while the UAL_reader plugin works with *standard representations*. Standard representations are currently used for data visualization within some CPOs, where the structure of the data fields allows this. For visualizations using data from multiple CPOs, the *ITMvis* approach provides scripting/post-processing capabilities that can combine results and output them through different backends. VisIt combined with the *UAL_reader* plugin allows both representations (standard and custom) by embedding a Python interpreter in the plugin itself. Standard representations were not included in *ITMVis* initially, although this was possible by direct interpretation of *CPOdef.xml* or by Python code generation as in *UAL_reader* (where C++ code representing plots is generated from *CPOdef.xml* directly). The process of XSL translation shown in Fig. 1 with solid lines could be repeated for ITMVis too. From the experience gained in developing templates for C++ code used in *UAL_reader*, where 220000+ source lines are generated, we came to a conclusion that introducing an *intermediate* XML description that extracts representation data from the *CPOdef.xml* will

reduce the complexity of the XSLT process when applying it to several programming languages. This *visualization schema* is still in XML, except that it contains the standard representations described more naturally in a custom XML schema that is easily converted further with XSLT. The complexity of the XSLT code is thus reduced and distributed among XSL translations to and from *visualization schema*. Such an approach unifies standard CPO representations and can be extended with additional fields in the intermediate XML format for easier translation as required.

III. Discussion

Presented visualization tools discuss mostly coupling for the purpose of simulations at the workflow level which is in our case Kepler. Naturally, visualization tools described can also be used as a standalone. There are other uses for such tools which are handled differently. For example, control panels and simulation instrument-frameworks tends to create specialized dashboard-style GUI. Dashboards usually include additional control buttons with custom actions. Workflow parameter-control is included in Kepler by the *Runtime Window*, where built-in output-plotters are included as well. This is mostly insufficient for complex visualizations. For custom "publication-quality" plots one can use general Python actor developed by the EU IM and then use ITMVis with matplotlib. Note that ITMVis inclusion can be difficult in non-native Python implementations such as Jython (Python for the Java Platform) used in Kepler-provided Python actor. VisIt with corresponding plugin can mimics multi-instrument (cf. Fig. 3) as a multi-window setup that can be saved as a session and reused often. Still, this is not single dash-window although VisIt allows coding of custom controlling GUIs with jVisit.

VisIt uses remote procedure calls (RPC) mechanism for inter-process communication (IPC). In some sense the VKA is a complete IPC application as it includes its own-coded GUI using Java widgets and jVisIt for controlling communication. Communication with RPC in Fig. 2 is in the direction from VisIt to VKA and not opposite as one may expect [17]. The VKA therefore acts as a server and VisIt attaches as a observer to it with a random *hostkey* that is generated when VisIt client is launched by the VKA. Controlling VisIt is thus possible only by launching from the VKA. It is not possible to attach VisIt to the running workflow on demand. That is the reason for existence of *VisIt controller* GUI included with the VKA. To allow controlling of the visualizations at the workflow run time.

VisIt allows *in-situ* connection [18] for simulation code steering. It requires modification of the instrumented code to include VisIt-library *libsim* and provide meta-data and current state as mesh objects in a similar way as the (UAL) reader. EU IM framework provides a tool *ual_connector* that connects UAL database and VisIt in such way. Caveat of this tightly in-situ coupling for our analyses is that only current code state can be requested. This rules-out most time-dependent visualizations (e.g. for animations). Usage of libsim is therefore limited to the the initial usage intent. Getting time slider to work still requires UAL_reader post-processing.

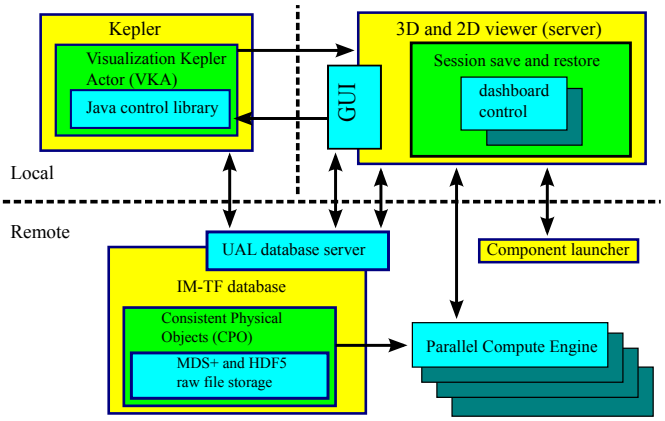


Fig. 5. Visualization architecture with remote database and computing. 2D and 3D viewer should allow remote connects and disconnects.

Obviously, different circumstances often merit different solutions. Concentrating on the past experiences in the field we can draft the visualization architecture in Fig.5, where the following properties for a general purpose 2D and 3D visualization tool (viewer) are sought:

1) It works as standalone application with usual (server-side) remote visualization engines. Local client may choose server-side or client-side rendering, depending on the client display graphics capabilities. Server side rendering with protocol like VirtualGL may be helpful for low-end clients. Current visualization tools (VisIt, Paraview) mostly use *legacy OpenGL code* for display and this impedes effective remote client-side OpenGL rendering that can be provided by *modern* OpenGL 2.1+ with GLSL. Most notable client-side display improvements of modern OpenGL are expected for 3D graphics manipulation.
2) Preferable, connections to remote visualization server may be session-aware allowing disconnects and reconnects from the viewer. Similar functionality is now exists in various remote desktop protocols where server holds a session that can be restored from different clients. Component launcher on remote side may took over of the visualization state handling and pass it to clients at session restore requests.
3) Session as a viewer configuration including visualization pipeline is saved into session file that can be restored at later time on request by Kepler.
4) Viewer as a client should act as a server too. Allowing connections and reconnects from Kepler. Kepler may launch viewer if there is no previously opened one. Viewer with its GUI can be launched as standalone and still allowing for Kepler to attach to it if having proper port and hostid assigned in the workflow.
5) GUI that comes with should be able to associate itself to running Kepler workflow for the purpose of workflow

steering. Actor for workflow instrumentation may not be the same as the Visualization Kepler Actor.

6) Viewer with mixed 2D and 3D windows can be tiled in a single canvas (dashboard) or built of multiple windows. *GUI Designer* for views may allow placing custom controls that can be user scriptable.

Clearly, present tools are not compliant to the above "desires". Reasons for most are linked to past computing architectures. Moreover, 3D visualization is closely linked to computer graphics capabilities. Our implementation with VisIt and developed tools resembles some of the above *requirements*.

Engineering still struggles to present results in plots with a single variable dependence. Rich 2D-graphing may seem not a good idea to be included into 3D visualization tool. However, limitations of VisIt and Kepler graphing brought custom 2D ITMVis library and virtually duplicated the effort needed. Not everything is well aligned in database and for that scripting can be found handy. Providing publication quality graphing is quite favorable for all tools. Offline rendering in such tools can further ease scripting. Scientists often prefer ease over performance. This fact should be considered in visualization tools by allowing graphical programming whenever possible and saving this configuration as visualization flowchart (session). Additionally, there are 3D cases (MHD, diagnostics, ..) where for comparison layout of the wall structure CAD model and some intersection (poloidal) from code needs to be presented. CAD inclusion in visualization tools can be done with premeshing of Bézier and NURB surfaces. Integrating a CAD "reader" would certainly ease visualization tool usage.

IV. Conclusion and outlook

Visualization in a mission to assist researchers in achieving their research goals needs to provide proper analytics with tools that tackle specific domain. Our introduction of the *visualization schema* into the *datastructure* unifies and reduces efforts required to develop different visualization tools. A set of *standard* visualizations provided by the visualization schema enables users to browse through CPOs available. In cases of custom requirements, scripting for plots can be used by *ITMVis* library that was embedded in VisIt UAL database reader. How to provide more sophisticated (algorithmic) mappings in the schema is ongoing question. Discussion on the present state and requirements for visualization tools may be considered as a general view for integrated modelling problematics and a starting point for a future work.

Acknowledgment

This work, supported by the European Communities under the contract of Association between EURATOM-MESCS was (partly) carried out within the framework of the Task Force on Integrated Tokamak Modelling of the European Fusion Development Agreement. The views and opinions expressed herein do not necessarily reflect those of the European Commission.

References

[1] "EFDA Integrated Tokamak Modelling Task Force website," http://www.efda-itm.eu/, 2014.

[2] I. Altintas, O. Barney, and E. Jaeger-Frank, "Provenance collection support in the Kepler scientific workflow system," *Lecture Notes in Computer Science*, vol. 4145 LNCS, pp. 118–132, 2006.

[3] C. Docan, M. Parashar, and S. Klasky, "DataSpaces: An interaction and coordination framework for coupled simulation workflows," in *HPDC 2010 - Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing*, 2010, pp. 25–36.

[4] G. Manduchi, F. Iannone, F. Imbeaux, G. Huysmans, J. Lister, B. Guillerminet, P. Strand, L.-G. Eriksson, and M. Romanelli, "A universal access layer for the integrated Tokamak Modelling Task Force," *Fusion Engineering and Design*, vol. 83, no. 2-3, pp. 462–466, 2008.

[5] F. Imbeaux, J. B. Lister, G. T. A. Huysmans, W. Zwingmann, M. Airaj, L. Appel, V. Basiuk, D. Coster, L.-G. Eriksson, B. Guillerminet, D. Kalupin, C. Konz, G. Manduchi, M. Ottaviani, G. Pereverzev, Y. Peysson, O. Sauter, J. Signoret, and P. Strand, "A generic data structure for integrated modelling of tokamak physics and subsystems," *Computer Physics Communications*, vol. 181, no. 6, pp. 987–998, 2010.

[6] J. Cummings, J. Lofstead, K. Schwan, A. Sim, A. Shoshani, C. Docan, M. Parashar, S. Klasky, N. Podhorszki, and R. Barreto, "EFFIS: An end-to-end framework for fusion integrated simulation," in *Parallel, Distributed and Network-Based Processing (PDP), 2010 18th Euromicro International Conference on*, Feb 2010, pp. 428–434. [Online]. Available: http://info.ornl.gov/sites/publications/files/Pub24705.pdf

[7] A. Galonska, P. Gibbon, F. Imbeaux, Y. Frauel, B. Guillerminet, G. Manduchi, F. Wolf, and ITM-TF contributors, "Parallel universal access layer: A scalable I/O library for integrated tokamak modeling," *Computer Physics Communications*, vol. 184, pp. 638–646, 3 2013.

[8] C. Docan, M. Parashar, and S. Klasky, "Enabling high-speed asynchronous data extraction and transfer using DART," *Concurrency and Computation: Practice and Experience*, vol. 22, no. 9, pp. 1181–1204, 2010. [Online]. Available: http://coewww.rutgers.edu/www4/cacweb/TASSL/Papers/dart_hpdc.pdf

[9] O. Hoenen, L. Fazendeiro, B. Scott, J. Borgdorff, A. Hoekstra, P. Strand, and D. Coster, "Designing and running turbulence transport simulations using a distributed multiscale computing approach," in *EPS 2013, Europhysics Conference Abstracts, 37D*, 40th EPS Conference on Plasma Physics. European Physical Society, 2013. [Online]. Available: http://ocs.ciemat.es/EPS2013PAP/pdf/P4.155.pdf

[10] S. Pinches, "The ITER integrated modelling programme," in *55th Annual Meeting of the APS Division of Plasma Physics*. Denver, Colorado: APS, November 11-15 2013, vol. 58, no. 16. [Online]. Available: http://meetings.aps.org/link/BAPS.2013.DPP.PO4.1

[11] I. Herman, "Overview of XSLT and XPath," http://www.w3.org/Consortium/Offices/Presentations/XSLT_XPATH, W3C, 2006.

[12] D. Coster, V. Basiuk, G. Pereverzev, D. Kalupin, R. Zagoórksi, R. Stankiewicz, P. Huynh, F. Imbeaux, and ITM-TF contributors, "The european transport solver," *Plasma Science, IEEE Transactions on*, vol. 38, no. 9, pp. 2085 –2092, sept. 2010.

[13] I. Altintas, C. Berkley, E. Jaeger, M. Jones, B. Ludascher, and S. Mock, "Kepler: An extensible system for design and execution of scientific workflows," in *SSDBM '04: Proceedings of the 16th International Conference on Scientific and Statistical Database Management*. Washington, DC, USA: IEEE Computer Society, 2004, p. 423.

[14] "matplotlib – a Python 2D plotting library," http://matplotlib.sourceforge.net/, 2014.

[15] "VisIt official homepage," http://visit.llnl.gov/, Lawrence Livermore National Laboratory, 2014.

[16] M. Haefele, L. Kos, P. Navaro, and E. Sonnendrücker, "Euforia integrated visualization," in *PDP 2010 - The 18th Euromicro International Conference on Parallel, Distributed and Network-Based Processing*, Pisa, Italy, 2010, pp. 498–502.

[17] LLNL, "VisIt design overview," an overview of the design of a scientific data visualization and analysis tool. [Online]. Available: http://portal.nersc.gov/svn/visit/trunk/docs/Design.doc

[18] B. Whitlock, J. Favre, and J. Meredith, "Parallel in situ coupling of simulation with a fully featured visualization system," *Proceedings of the Eurographics Symposium on Parallel Graphics and Visualization*, pp. 101–109, 2011. [Online]. Available: http://portal.nersc.gov/svn/visit/trunk/docs/Presentations/EGPGV2011_InSituPaper.pdf

Fast Motion Estimation Algorithm Using Dual Bit-plane Matching Criteria

Changryoul Choi and Jechang Jeong
Dept. of Electronics and Communication Engineering, Hanyang University,
Seoul, Republic of Korea
denebchoi@gmail.com & jjeong@hanyang.ac.kr

Abstract - The constrained two-bit transform (C2BT) is recently proposed for low complexity motion estimation (ME) to reduce the high computational complexity of the typical two-bit transform (2BT) while maintaining the ME accuracy. And the bit-inverted Gray-coded bit-plane matching (BGCBPM) is also proposed to enhance the ME accuracy. In this paper, we propose a fast ME algorithm exploiting these two bit-plane matching (BPM) criteria. Using the low complexity image transformations of typical images into bit-planes and their corresponding matching criteria and another hybrid matching criterion, we can efficiently determine multiple candidate motion vectors and increase the ME accuracy substantially. Experimental results show that the peak-to-peak signal-to-noise ratio (PSNR) difference between the proposed algorithm and the typical sum of absolute differences (SAD) based full search algorithm is only 0.03 dB on average with negligible computational complexity increase.

I. INTRODUCTION

Since the multimedia data are abundant everywhere, the video compression technology is indispensable for reducing the multimedia data for transmission and storage in many applications. In video compression, the motion estimation (ME) plays one of the most important parts since it can reduce the temporal redundancy of multimedia data substantially using the temporal correlation between neighboring frames and deployed in many video coding standards [1-3]. In typical ME process, a current frame is partitioned into small rectangular blocks and a motion vector is calculated within its search range in the reference frames by estimating the most similar block according to some mathematical matching criterion such as the sum of absolute differences (SAD) or the sum of squared differences (SSD). This algorithm is called as a block matching algorithm (BMA). Although the full search BMA (FSBMA) can find an optimal motion vector according to some matching criterion, its computational complexity is so huge that it is not adequate for real time applications. To reduce the computational burden, many techniques have been proposed in the literature [4-15]. Among them, there proposed some techniques that use different matching criterion instead of the classical SAD or SSD to make the faster computation of the matching criterion itself exploiting the bit-wise operations [8-15]. These algorithms are called as bit-plane matching (BPM) based ME. Some advantages of these techniques over the typical SAD or SSD based ME algorithms can be summarized as follows: fast computation of the matching criterion and reduced memory bandwidth in the interim of ME process. These techniques include one-bit transform (1BT) [8], two-bit transform (2BT) [10], constrained 1BT (C1BT) [9], truncated Gray-coded BPM (TGCBPM) [11], weightless TGCBPM (WTGCBPM) [12], constrained two-bit transform (C2BT) [14], bit-inverted Gray coded BPM (BGCBPM) [15], etc. Although the BPM based ME succeeded in reducing the computational complexity, its ME accuracy is relatively poor compared with the SAD based ME, resulting in degraded reconstructed images. To remedy this situation, multiple candidate motion searches were proposed in [13].

In this paper, we propose a low complexity multiple candidate ME algorithm based on C2BT and BGCBPM. Since the image transformations of C2BT and BGCBPM are very low, we can efficiently determine the respective candidate motion vectors. And by using the hybrid matching criterion of C2BT and BGCBPM, we can determine another candidate motion vector. Adopting the multiple candidate motion search strategy into those candidate motion vectors dramatically enhance the ME accuracy.

The rest of this paper is organized as follows: In Section II, we review some previous works pertinent to our proposed algorithm. In Section III, the proposed algorithm is presented. Experimental results and analyses are given in Section IV. Finally, Section V provides conclusions.

II. PREVIOUS WORKS

A. C2BT based ME

In transforming original image frames into two-bit representations, the typical 2BT needs to calculate variances of the local blocks. This calculation of variances inevitably involves multiplication operations and renders the computational complexity of typical 2BT somewhat high [10]. By exploiting advantages of the typical C1BT and the typical 2BT, C2BT significantly reduces the computational complexity of transformation of image frames into two-bit representations while maintaining the ME accuracy [14]. By replacing the multiplication operations in variance calculation of 2BT with the simple thresholding technique, the two-bit representations of C2BT are attained as follows:

$$C_1(i,j)=\begin{cases}1, & I(i,j)\ge\mu\\ 0, & otherwise\end{cases}$$
$$C_2(i,j)=\begin{cases}1, & |I(i,j)-\mu|\ge D\\ 0, & otherwise\end{cases} \quad (1)$$

where μ is the mean of the local threshold window around the transforming block, D is a threshold, $I(i, j)$ are the pixel values of the transforming block and $C_1(i, j)$ and $C_2(i, j)$ are the two-bit representations of C2BT. The typical transformation block size is 8×8 as is the case of the typical 2BT. However, unlike the typical 2BT, the local threshold window is set to 32×32 to replace the multiplication (division) operations with the shift operations. And the corresponding matching criterion, the number of non-matching points (NNMP) is as follows:

$$\begin{aligned}
&NNMP_{C2BT,1}(m,n)=\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}\{C_1^t(i,j)\oplus C_1^{t-1}(i+m,j+n)\}\\
&NNMP_{C2BT,2}(m,n)=\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}\{C_2^t(i,j)\oplus C_2^{t-1}(i+m,j+n)\}\\
&NNMP_{C2BT,3}(m,n)\\
&=4\times\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}[C_2^t(i,j)\bullet\{C_1^t(i,j)\oplus C_1^{t-1}(i+m,j+n)\}]\\
&NNMP_{C2BT,4}(m,n)\\
&=4\times\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}[C_2^{t-1}(i+m,j+n)\bullet\{C_1^t(i,j)\oplus C_1^{t-1}(i+m,j+n)\}]\\
&NNMP_{C2BT}(m,n)=\sum_{i=1}^{4}NNMP_{C2BT,i}(m,n)
\end{aligned} \quad (2)$$

where $\oplus$ and $\bullet$ denote the Boolean XOR and AND operation, respectively.

B. BGCBPM based ME

TGCBPM and WTGCBPM use Gray-code mapping as transforming image frames into bit-planes, which is very simple compared to other bit-plane transformation algorithms using complex filtering operations. Let the gray-level of the pixel at location (m, n) be represented as follows:

$$f(m,n)=a_{K-1}2^{K-1}+a_{K-2}2^{K-2}+\cdots+a_1 2^1+a_0 2^0 \quad (3)$$

where K represents the pixel-depth and a_k ($0\le k\le K$-1) takes on either 0 or 1. Then the corresponding Gray-code representation is given by

$$\begin{aligned}
g_{K-1}&=a_{K-1}\\
g_k&=a_k\oplus a_{k+1},\quad 0\le k\le K-2
\end{aligned} \quad (4)$$

where $\oplus$ denotes the Boolean XOR operation and a_k is the k-th bit representation. After this transformation, TGCBPM and WTGCBPM use respective NNMPs as matching criteria. Compared with the previous BPM based ME algorithms, TGCBPM and WTGCBPM based ME show significant gains in terms of the ME accuracy [11-12].

Since all the matching criteria of BPMs try to replace the SAD or SSD with the bitwise operations for reduction of computational complexity, their output must be similar to the SAD or SSD in order to find a more accurate motion vector. To this end, BGCBPM was proposed [15]. In BGCBPM, the typical Gray code mapping is bit-inverted as follows:

$$h_k=\sim g_k,\quad NTB\le k\le K-1 \quad (5)$$

where K represents the pixel-depth, ~ represents the Boolean NOT operation, NTB is the number of truncated bits and g_k is the k-th Gray bit representation. And the corresponding matching criterion is as follows:

$$\begin{aligned}
&NNMP_{gram,1}(m,n)=\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}\sum_{k=NTB}^{K-1}h_k^t(i,j)\oplus h_k^{t-1}(i+m,j+n)\\
&NNMP_{gram,2}(m,n)\\
&=2^{K-NTB}\times\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}[h_{K-2}^t(i,j)\bullet\{h_{K-1}^t(i,j)\oplus h_{K-1}^{t-1}(i+m,j+n)\}]\\
&NNMP_{gram,3}(m,n)\\
&=2^{K-NTB}\times\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}[h_{K-2}^{t-1}(i,j)\bullet\{h_{K-1}^t(i,j)\oplus h_{K-1}^{t-1}(i+m,j+n)\}]\\
&NNMP_{BGCBPM}(m,n)=\sum_{i=1}^{3}NNMP_{gram,i}(m,n)
\end{aligned} \quad (6)$$

where $\bullet$ denotes the Boolean AND operation, K represents the pixel-depth, NTB is the number of truncated bits, the motion block size is $N\times N$, and $-s\le m, n\le s$ is the search range. The kth most significant bit of the bit-inverted Gray-coded image frame of time t is represented as $h_k^t(i, j)$. In this case, the actual distortion between two pixels in terms of the SAD is proportional to the absolute difference between these transformed symbols and the corresponding matching criterion of BGCBPM actually shows this relationship, resulting in increased ME accuracy [15].

C. Multiple Candidate Motion Searches

In [13], the authors proposed multiple candidate motion searches based on two different matching criteria. They exploited the almost identical operations between two different matching criteria and did the dual motion searches with negligible computational complexity increase. The multiple candidate motion searches based on the weighted 2BTs (MCW2BT) can be summarized as follows [13]:

1) Find two best motion vectors using two different matching criteria, respectively.
2) If two best motion vectors with the respective matching criteria are the same, declare it as the best motion vector for the current block and go to 4).
3) Calculate SADs of the two best motion vectors

and declare the motion vector with less SAD as the best motion vector for the current block.

4) Go to the next current block.

III. PROPOSED ALGORITHM

To improve the overall ME performance, we adopt the strategy in [13] of multiple candidate ME exploiting the low complexity image transformations of C2BT and BGCBPM. Table I shows the average PSNR results of C2BT (D = 30) and BGCBPM when the motion block size is 16×16 and the search range is ±16. As can be seen from the table, the average PSNR performance of BGCBPM with NTB=5 is slightly better than that of C2BT. However, if we look into the average performance of each sequence, sometimes that of C2BT outperforms that of BGCBPM with NTB=5 (for example, the sequence of container). Therefore, there is enough room for improving the ME accuracy using the multiple candidate search.

TABLE I. AVERAGE PSNR RESULTS OF C2BT AND BGCBPM

	C2BT	**BGCBPM (NTB = 6)**	**BGCBPM (NTB = 5)**
football	23.71	23.64	23.95
akiyo	42.53	42.08	42.61
foreman	32.28	32.00	32.83
hall	33.62	33.22	34.03
container	38.20	37.67	37.70
dancer	30.94	30.48	31.16
table	28.51	28.31	28.62
news	36.86	36.81	37.04
Average	**33.33**	**33.03**	**33.49**

Besides these two candidates, we also propose to use another matching criterion which is a hybrid matching criterion as follows:

$$NNMP_{HYBRID}(m,n) = NNMP_{C2BT}(m,n) + NNMP_{BGCBPM}(m,n) \qquad (7)$$

Then, the proposed multiple candidate motion searches based on C2BT and BGCBPM can be described as follows:

1) Find three best motion vectors using three different matching criteria, respectively.

2) If three best motion vectors with the respective matching criteria are the same, declare it as the best motion vector for the current block and go to 4).

3) Calculate SADs of the three best motion vectors and declare the motion vector with the least SAD as the best motion vector for the current block.

4) Go to the next current block.

Note that the calculations of SADs are needed only when three best motion vectors are different, which is very rare as will be seen in the experimental results.

IV. EXPERIMENTAL RESULTS

The performance of the proposed algorithm was compared with 1BT, C2BT, BGCBPM (NTB=5 and NTB=6), AM2BT, and FSBMA using the measure of SAD. The first 100 frames of 17 CIF (352×288) sequences are used as test sequences. All the searching processes were in spiral order.

Table II shows the average PSNR comparison results when the motion block size is 16×16 and the search range is ±16. The average numbers of SAD calculations per motion block of AM2BT and the proposed algorithm are also shown in the table. Note that the maximum number of SAD calculations in one motion block is two for AM2BT and three for the proposed algorithm, respectively.

From the table, we can see that the performance of the proposed algorithm outperforms other BPM based ME algorithms. To be specific, the average PSNR of the proposed algorithm is better than that of 1BT by 1.19dB, that of C2BT by 0.35dB, and that of AM2BT by 0.28dB. The SAD calculations of the proposed algorithm are needed about 1 out of 5 motion blocks on average. The PSNR gap between the proposed algorithm and the FSBMA is within 0.03dB, which is very small, therefore we can regard the performance of the proposed algorithm is almost the same as that of the FSBMA.

V. CONCLUSIONS

In this paper, we propose a fast ME algorithm using two different BPM based ME algorithms. Using the low complexity image transformations of typical images into bit-planes and their corresponding matching criteria and another hybrid matching criterion, we can efficiently determine multiple candidate motion vectors and increase the ME accuracy substantially. Experimental results show that the PSNR difference between the proposed algorithm and the typical SAD based FSBMA is within 0.03dB, which is very small.

TABLE II. PSNR PERFORMANCE COMPARISON OF THE ALGORITHMS WHEN THE MOTION BLOCK SIZE IS 16×16 (SEARCH RANGE = ±16)

sequences	1BT	C2BT	BGCBPM (NTB = 6)	BGCBPM (NTB = 5)	AM2BT	Proposed (NTB = 6)	Proposed (NTB = 5)	FSBMA
stefan	25.12	25.56	25.63	25.71	25.68 (23.95)	25.73 (0.62)	25.74 (0.62)	25.75
football	22.64	23.75	23.64	23.95	23.96 (56.76)	24.01 (0.84)	24.03 (0.68)	24.00
akiyo	41.66	42.58	42.08	42.61	42.57 (0.57)	42.79 (0.08)	42.81 (0.08)	42.84
foreman	31.69	32.32	32.00	32.83	32.60 (3.26)	33.15 (1.05)	33.31 (1.03)	33.43
mobile	23.50	23.79	23.82	23.87	23.84 (37.24)	23.91 (0.36)	23.92 (0.33)	23.92
hall	32.13	33.68	33.22	34.03	33.81 (2.82)	34.27 (0.85)	34.28 (0.86)	34.34
coastguard	29.09	29.49	28.09	29.49	29.55 (7.85)	29.59 (0.68)	29.61 (0.41)	29.62
container	37.57	38.20	37.67	37.70	38.16 (0.59)	38.32 (0.15)	38.30 (0.26)	38.33
bus	23.86	24.59	24.66	24.84	24.79 (35.79)	24.88 (0.72)	24.91 (0.62)	24.90
dancer	29.67	30.99	30.48	31.16	31.45 (11.37)	31.78 (1.54)	31.87 (1.57)	32.14
mother and daughter	37.58	39.27	37.48	39.49	39.58 (1.18)	39.93 (0.78)	39.99 (0.75)	40.12
tempete	27.01	27.53	27.46	27.61	27.63 (19.10)	27.68 (0.46)	27.69 (0.41)	27.70
table tennis	27.44	28.54	28.31	28.62	28.65 (10.61)	28.79 (0.84)	28.83 (0.81)	28.87
flower	25.73	25.95	25.86	25.98	25.97 (22.36)	26.02 (0.62)	26.02 (0.55)	26.03
children	28.05	29.01	29.11	29.17	29.16 (17.64)	29.27 (0.25)	29.27 (0.23)	29.24
paris	30.16	30.57	30.53	30.67	30.58 (7.33)	30.72 (0.29)	30.73 (0.26)	30.71
news	35.58	36.89	36.81	37.04	37.05 (3.05)	37.30 (0.23)	37.32 (0.26)	37.33
Average	**29.91**	**30.75**	**30.40**	**30.87**	**30.88 (15.38)**	**31.07 (0.61)**	**31.10 (0.57)**	**31.13**

REFERENCES

[1] Z. He, C. Tsui, K. Chan, and M. Liou, "Low-power VLSI design for motion estimation using adaptive pixel truncation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 10, no. 5, pp. 669-678, Aug. 2000.

[2] *Information Technology - Coding of Audio Visual Objects - Part 2: Visual*, JTC1/SC29/WG11, ISO/IEC 14496-2 (MPEG-4 Visual), 2002.

[3] Advanced Video Coding for Generic Audiovisual Services, *ITU-T Recommendation H.264*, May 2005.

[4] W. Li and E. Salari, "Successive elimination algorithm for motion estimation," *IEEE Trans. Image Process.*, vol. 4, no. 1, pp. 105-107, Jan. 1995.

[5] X. Q. Gao, C. J. Duanmu, and C. R. Zou, "A multilevel successive elimination algorithm for block matching error estimation," *IEEE Trans. Image Process.*, vol. 9, no. 3, pp. 501-504, Mar. 2000.

[6] B. Montrucchio and D. Quaglia, "New sorting-based lossless motion estimation algorithms and a partial distortion elimination performance analysis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no.2, pp. 210-220, Feb. 2005.

[7] C. Choi and J. Jeong, "New sorting-based partial distortion elimination algorithm for fast optimal motion estimation," *IEEE Trans. Consumer Electron.*, vol. 55, no. 4, pp. 2335-2340, Nov. 2009.

[8] B. Natarajan, V. Bhaskaran, and K. Konstantinides, "Low-complexity block-based motion estimation via one-bit transforms," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 7, no. 5, pp. 702-706, Aug. 1997.

[9] O. Urhan and S. Erturk, "Constrained one-bit transform for low complexity block motion estimation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 4, pp. 478-482, Apr. 2007.

[10] A. Erturk and S.Erturk, "Two-bit transform for binary block motion estimation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 7, pp. 938-946, Jul. 2005.

[11] A. Celebi, O. Akbulut, O. Urhan, and S. Erturk, "Truncated Gray-coded bit-plane matching based motion estimation method and its hardware architecture," *IEEE Trans. Consumer Electron.*, vol. 55, no. 3, pp. 1530-1536, Aug. 2009.

[12] A. Celebi, H.-J. Lee, and S. Erturk, "Bit plane matching based variable block size motion estimation method and its hardware architecture," *IEEE Trans. Consumer Electron*., vol. 56, no. 3, pp. 1625-1633, Aug. 2010.

[13] C. Choi and J. Jeong, "Low complexity weighted two-bit transforms based multiple candidate motion estimation," *IEEE Trans. Consumer Electron*., vol. 57, no. 4, pp. 1837-1842, May 2011.

[14] C. Choi and J. Jeong, "Constrained two-bit transform for low complexity motion estimation," *International Conference on Consumer Electronics, ICCE 2013*, Las Vegas, USA, Jan. 2013.

[15] C. Choi and J. Jeong, "Bit-inverted Gray coded bit-plane matching for low complexity motion estimation," *unpublished.*

Edge Detection in Segmented Non-convex Space Polygon

I. Kožar*, D. Lozzi-Kožar ** and Ž. Jeričević***
* Faculty of Civil Engineering /Department of Computer Modeling, Rijeka, Croatia
** Croatian Water Authority, Rijeka, Croatia
*** Engineering Faculty/Department of Computer Engineering, Rijeka, Croatia
ivicak@gradri.hr

Abstract - Survey data in use in Civil Engineering consists of points described with XYZ coordinates. Closed shapes, like lakes described with a number of such points are immediately recognizable to humans although points are grouped in unconnected and unrelated lines. Automatic processing of those closed shapes is required if computer analysis is to be performed, i.e. we would like to produce a finite element mesh of the lake. In order to perform meshing one has to be able to determine whether a point is inside or outside of the space polygon, i.e. polygon edges have to be detected. Special computer procedures have been developed for edge detection. The most challenging part is detection of overlaps of edge lines which has been solved with specially modified Fourier transform procedure. Above procedures are applied on a real-life example for creating a finite element mesh of lake "Botonega" in Istria.

I. INTRODUCTION

Formation of finite element meshes is among the most important tasks in computer modeling (with finite elements used for discretization of the continuum problem). Great effort is devoted to generation of finite element meshes that have to obey some rules to be useful: mesh must correspond to the element type that will be used in the analysis regarding number of nodes per element side, number of nodes and elements has to be kept within given limits, ration of element sides should be between one and about two (distorted elements should be avoided).

Generation of finite element meshes is especially difficult for natural (not man-made) objects since they are highly irregular. Large natural objects like lakes present an additional difficulty by being huge and thus difficult to measure. Traditionally, data is collected through geodesic survey and the main concern is representation of large areas of terrain [1]. In civil engineering it is common to gain information about objects through geodesic survey, too. But here one is more interested in object limits then in areas. Results are groups of uncorrelated points suitable for human interpretation (humans immediately recognize interesting shapes and make conclusions about objects). Mesh of points for computer simulation is quite often produced manually [2], [3].

This work presents a procedure that starts with results of a geodesic survey and finishes with data suitable for automatic finite element mesh generation. Real-life data of Lake "Botonega" in Istria serve as an example.

II. EDGE DETECTION

A. *Data preparation*

The starting point in analysis is data from geodesic survey. In our example they consist of points in XYZ coordinate system describing edges and bottom of the lake. Details about data format could be found in our previous paper [4].

The main problem with geodesic data is that they are grouped and each group is not ordered in a way that forms a closed polygon. Namely, survey starts and ends at convenient points and the survey procedure is rarely performed in one "run". The result is a group of unordered points, see Fig.1.

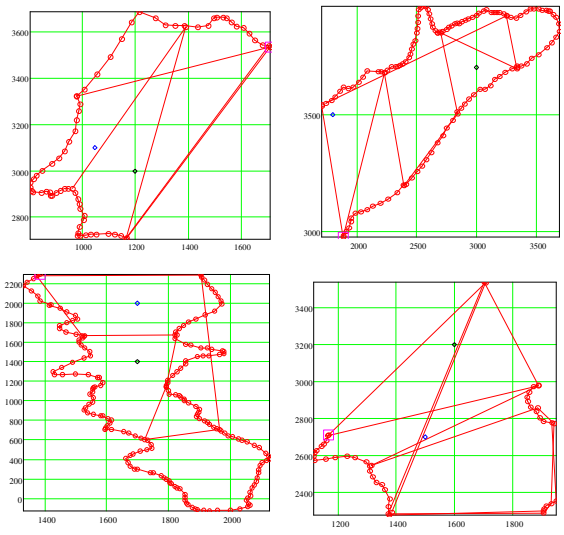


Figure 1. Ordering of points in the geodesic surway (points projected onto the surface of the land). Note the imaginary internal points.

Of course, human can immediately recognize a closed polygon that is represented when points are plotted but it is not so with a computer. Points have to be ordered so that in a vector that contains them the first point is the starting point and the last point is the ending one and the

points are connected neighbor to neighbor. That vector represents a closed polygon when the first and last point are connected.

Proper ordering of points in one group is achieved by applying an original algorithm that starts from a point inside the polygon. As a starting point we have taken center of gravity of unordered points. Area of the imaginary triangles is calculated, increase in the area sorts points in the counter-clockwise direction along the edge, see Fig.2.

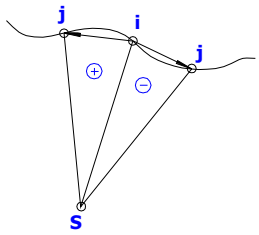


Figure 2. Edge points are oredered regarding the sign of the surface of an imaginary triangle.

Marching through the points is achieved with a proper selection of the next point applying a minimum distance algorithm.

For non-convex polygons this simple rule is not sufficient because not all points could be seen from one pole. In that case new internal point is generated as the gravity center of not visible points and the rule is applied again. The procedure is repeated until there are no "unreachable" points. Since there is known kinematical relation between internal points, proper ordering can be established. Process is repeated until periodicity is detected (using DFT). Resulting polygon is presented in Fig.3.

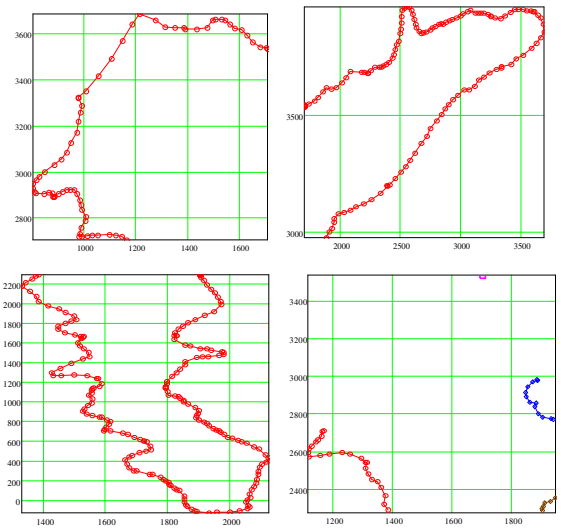


Figure 3. Ordering of points in the geodesic surway. Polygon is not necessarily convex.

Application of the mentioned procedure transforms data from Fig.1 into Fig.3. The basic problem with combination of group of points into a closed polygon can be illustrated with Fig.4.

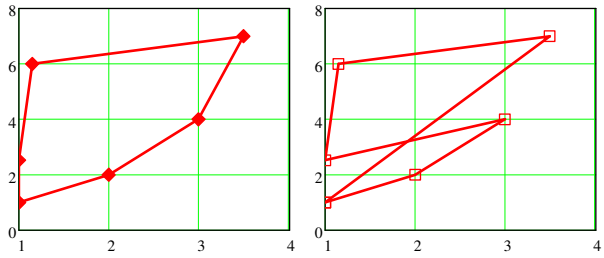


Figure 4. Connecting two groups of points to form a closed polygon.

In Fig.4 one ordering of points is "correct" and the other one is not. It can be easily proved that closed polygon has minimal area when calculated from partial sums along ordered points. However, application of this simple rule is not practical. For rotated points one would have to interchange x and y coordinates. That problem is not present if we take DFT of vector of points and compare their power spectra. It can be shown that the power spectra sum is proportional to the polygon area as described above. Consecutively applying the rule one can combine points into a closed polygon as shown in Fig.5.

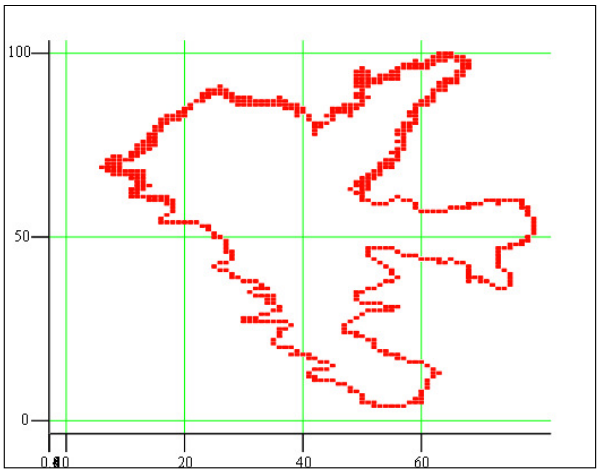


Figure 5. Closed non-convex polygon in normalized coordinates. Note that some points are multiple.

The whole lake "Botonega" is modeled using five polygons that are rotated and scaled in normalized coordinates. The resulting ordered points are presented in Fig.5. Take note that all four polygons from Fig.3 are present in Fig.5 (although scaled and rotated).

Note that due to rotation and scaling in normalized coordinates the edge is not everywhere represented with one line of points.

B. Detection of internal points

Edge of the lake is represented with non-convex polygon from Fig.4. The fact that the edge is not a single line of points (observe multiple points in Fig.4) is the source of confusion in edge determination procedure. Usual polygon fill algorithms [5] or [6] could not correctly

detect edges and fill the polygon (in order to determine inside points).

Edge detection is performed in normalized coordinates by dividing the data matrix into rows and columns. In each row edges are represented with 1 and the rest of terrain with 0 as in Fig.5.

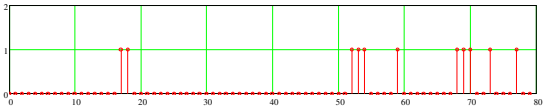


Figure 6. Representation of edges in data matrix. Note that some points are multiple.

Proper edge detection algorithm takes into account multiple points on the edge as well as existence of interior domains. Multiple points are taken into account by forming pairs of changes of data value. Internal domains are resolved by taking into account data from the previous and the next row. In Fig.6 we see red and blue line corresponding to in-line and neighbor rows parameters.

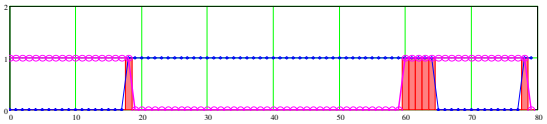


Figure 7. Row of the data matrix as interpreted by edge detection algorithm.

The result of the edge detection algorithm is completely filled non-convex polygon from Fig.4 i.e. internal points have been successfully detected.

C. Mesh generation

With internal points determined we could proceed with generation of the finite element mesh. A computer program has been generated that loops through the data matrix. Only if a point is inside the polygon, loop through the depth is executed and a series of finite elements is generated.

In the example of Lake "Botonega" we have used mesh sized 40 m by 30 m by 1.5 m in depth. Elements of this size would be too distorted so the mesh has to be scaled. Scaling is performed on the differential equation level using the similarity principle. The resulting mesh has 10402 elements and 83216 nodes and is shown in Fig.7. In order to perform calculations, this mesh has to be further processed, i.e. coincident nodes have to me merged. In each node a virtual sphere of given radius is inserted and all inside nodes are coalesced resulting with 15158 remaining nodes. Afterwards nodes and elements are consecutively renumbered. Details could be found in [7].

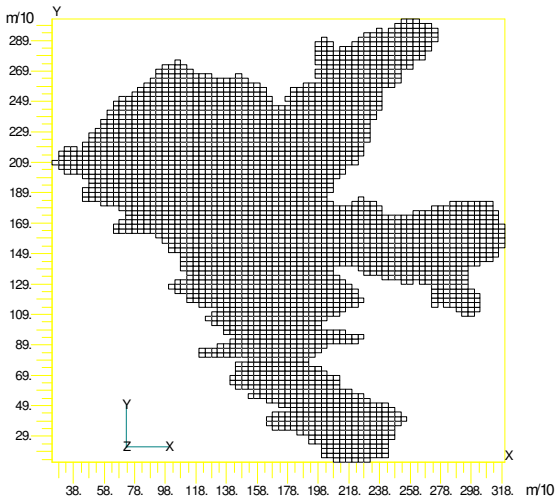


Figure 8. The resulting top layer of finite element mesh of Lake "Botonega".

Successful use of the obtained finite element mesh can be seen in Fig.8. Here we have presented thermal stratification in the lake as a result of calculations on the finite element model from Fig.7.

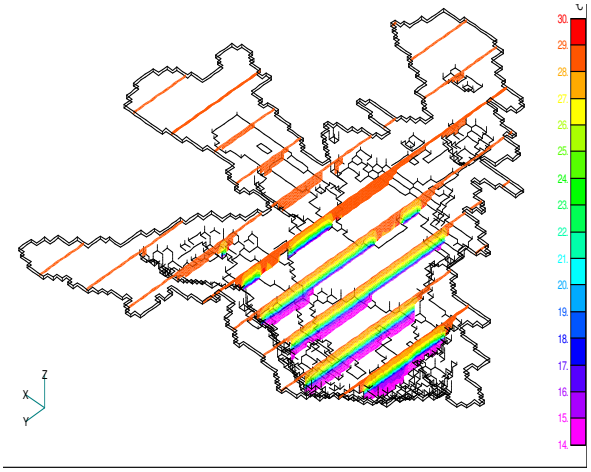


Figure 9. Calculation example using the resulting finite element mesh of Lake "Botonega".

III. CONLUSION

We have briefly presented the procedure for generation of a finite element mesh from uncorrelated geodesic data whose first part that describes interpolation of geodesic data is given in [4]. In this part we have presented formation of edges of the domain and determination of inside points.

ACKNOWLEDGMENT

This paper is supported in part by the University of Rijeka through grant no. 13.05.1.1.02.

REFERENCES

[1] Hjelle Ø., Dæhlen M., Triangulations and Applications, Springer, 2006.

[2] Rathbun S.L., Spatial Modelling in Irregularly Shaped Regions: Kriging Estuaries, Environmetrics, 9, pp. 109-129, 1998.

[3] Rivest M., Marcotte D., Pasquier P., Sparse data integration for the interpolation of concentration measurements using kriging in natural coordinates, Journal of Hydrology, 416–417, pp. 72–82, 2012.

[4] I. Kozar, D. Lozzi/Kozar and Z. Jericevic, "Limit kriging in finite element environmental modeling," in MIPRO 11, P. Biljanovic and K. Skala, Eds. Opatija: MIPRO, 2011, pp. 342–345.

[5] Wayne O. C., Filling Polygons, http:// ezekiel.vancouver. wsu. edu/ ~cs442/ lectures/ raster /polyfill / poly.pdf.

[6] Finley D. R., Efficient Polygon Fill Algorithm, 2007., http://alienryderflex.com/ polygon_fill/

[7] D. Lozzi-Kožar, Three-dimensional Modeling of Thermal Distribution in Lake "Botonega" (in Croatian), Ph.D. thesis, The University of Split, Split, Croatia, 2012.

Visualization of Meteorological Data

I. Hakštok, Ž. Mihajlović
Univesity of Zagreb, Faculty of Electrical Engineering and Computing, Zagreb, Croatia
ivan.hakstok@fer.hr, zeljkam@zemris.fer.hr

Abstract - This paper is focused on the methods used to visualize meteorological factors based on the given data. Properties of the factors for which the data can be gathered were described and the methods of visualizing those properties have been examined. The program solution is centered around interaction with the weather data. The data itself is based on real measurements. It is possible to select an exact point in time for which the data exists, and that data is then visualized on a 3D model of a building, which allows the user to get an idea of the weather state. The model changes lighting, shadows, clouds, opacity and the rotation of a wind vane, based on the data. The data is also represented by graphs, which allow the user to track data changes over a period of time. Finally, since the solution uses large amount of data, a performance analysis was done to point out advantages and disadvantages of the visualization methods used.

I. Introduction

Meteorology is the scientific study of the atmosphere, and as such it is one of the most needed areas of science for everyday lives. It encompasses a vast array of factors which influence the behavior and appearance of the atmosphere.

The need for visualizing meteorological data comes from the fact that it's not easy for people to picture certain aspects of that data using numbers alone. However, those numbers are linked to the state of the atmosphere and visualizing them can give people a better and more clear understanding on the way they work and influence each other. This can be used in order to help plan construction of buildings or placing solar panels, where the effects of meteorological factors on surfaces need to be observed for longer time spans.

An important thing to notice in the visualization of the given data is that not all meteorological factors can be visualized directly, but it's possible to visualize the way they affect the surroundings, such as temperature causing heat blurs and mirages [1]. This paper describes some of the already existing ways to visualize given meteorological data using 2D and 3D graphics, along with graphs and charts.

II. Meteorological Data

Some of the most important factors in meteorology are the temperature, humidity, atmospheric pressure, speed and the direction of the wind, and also the amount of rainfall. In order to visualize those factors, it's important to know how they interact and what effect they have on their surroundings.

It's possible to divide the meteorological factors into those that can be visualized directly and those that can't. Something like temperature of the air, which is the measure of the amount of heat energy present in the air, is not a visible phenomenon, but the effects of the temperature on the surroundings are noticeable: heat haze, ice and mirages are just some of the manifestations of the current air temperature.

Wind speed and direction are also not visible phenomena because humans can't see the movement of molecules in the air, but their interaction with the objects such as trees, leaves, flags and windmills can be observed and measured. Winds are a direct consequence of differing amounts of air pressure in two different places.

Some of the visible factors are cloud coverage, amount of rainfall and insolation. Those can be visualized directly, without the need for visualizing intermediate objects or phenomena. However, they are influenced by other factors which must be taken into consideration to some degree.

A. Air Temperature

Air temperature is the most fundamental meteorological factor. It describes the amount of heat energy of air in a certain location. In terms of visualization, temperature can be visualized using graphs and heat maps, but in a 3-dimensional environment model, direct visualization isn't possible. However, air temperature does bring forth some interesting phenomena whose direct visualization in such a model is possible.

One of those phenomena is the appearance of so called mirages [1]. Mirages are a result of two layers of air with different density, which is a direct consequence of one layer being at a higher temperature than the other one. When air temperatures are high, heating of surfaces also occurs, and the air mass right next to the surface gets heated even further whereas the air further from the surface does not. The higher the air temperature is, the less dense the said air is, according to the ideal gas law [2]. The law states that

$$pV = nRT. \quad (1)$$

Assuming the volume, indicated by V in (1), is constant, the pressure p is inversely proportional to the temperature T.

Since the aforementioned two layers are of different density, light reflection occurs at the borderline of the two

layers when observed from a certain angle [3]. This results in a mirror effect for the viewer.

Another phenomenon affected by the temperature is the rainfall. Rainfall can vary between rain, snow and hail, and the type is largely dependant on the temperature. More of this will be discussed in the section focusing on visualizing rainfall.

B. Wind

Air molecules by themselves aren't visible to the human eye, but the effect of their movement on the surroundings is easily noticeable and possible to visualize. Wind speed is measured using anemometers. Therefore, the anemometer can be used as a way of visualizing the strength and speed of the wind. Because there are various different types of anemometers, it's necessary for visualization to select a type that's easy to model and visualize.

Since the wind direction should also be taken into account when visualizing wind data, the most suitable type would be the windmill-shaped anemometer [4]. Such anemometer is a combination of a wind vane and an anemometer. Wind speed affects the rotation of the windmill part, and wind direction affects the direction of the wind vane. Such rotations (around an arbitrary axis parallel to the xy plane for the windmill, and an arbitrary axis parallel to the z-axis for the wind vane) are easy to model. It's safe to assume that the speed of the windmill wing tip is equal to the speed of the wind multiplied by a certain coefficient. The angular velocity of the windmill would then be

$$\omega = \alpha v/r, \tag{2}$$

where v is the speed of the wind, and r the length of a wing. The coefficient can be any number, so one can be chosen depending on the other parameters and what is best suited for the model itself.

If the vector of the wind velocity is assumed to be parallel to the surface, i.e. the xy plane, the axis of the windmill rotation is then a line whose vector of direction is equal to the wind velocity vector. Following the assumption about the vector being parallel to the xy plane, wind direction can be represented by the angle between the velocity vector and an arbitrary starting vector. The starting vector is usually the direction of North, placed at 0 degrees. East is placed at 90 degrees, South at 180 degrees and West at 270 degrees.



Figure 1. Example anemometer model

C. Cloud Coverage

Speaking in the terms of meteorology on Earth, clouds are visible masses of water droplets suspended in the atmosphere [5]. Appearance of clouds happens almost daily, and the amount of cloud coverage is measurable. Scales for measuring cloud coverage typically range from 0 to 8 or from 0 to 10.

There exist various ways of modeling and visualizing clouds. Clouds can be represented either by solid models, by particle systems [9] or textures. In the terms of meteorology, since measurements of cloud coverage are taken from stations usually located on the ground, one can assume that the need for visualizing clouds will also happen in a model where the viewer is close to the ground, and therefore the most efficient solution is to use different textures for different amounts of cloud coverage.

Textures can by dynamically generated based on actual data or generated as an interpolation between minimum and maximum cloud coverage [7]. An example of texture interpolation using a modeling program called Blender3D can be seen in Fig. 2. Another option is to have pre-rendered textures for each degree of cloud coverage. This reduces the amount of calculation that needs to be done during the rendering of a model, but it's lacking in flexibility. Considering measurement scales for cloud coverage are relatively short and quantized, limitations of pre-rendered textures don't influence the look of the model as much as they would for a more dynamic model.

D. Insolation

Insolation is the amount of solar radiation received by the surface. It is influenced by cloud coverage to a lesser extent. Insolation amounts vary greatly throughout a year because they are not solely dependent on the position of the sun.

Position of the sun is important in a model involving meteorological data, but it can't be calculated by measuring insolation. It's therefore necessary to calculate the position based on the current date. A possible formula for calculating the position is described in [8]. The formula yields two coordinates in a horizontal coordinate system which has the viewer in the center, the solar azimuth γ, which represents the angle between north and the projection of the solar vector on the surface, and the zenith distance θ_z, which represents the angle between the zenith and the solar vector. Using these two coordinates, the exact position can be visualized. However, the distance of the sun from the viewer must be scaled along

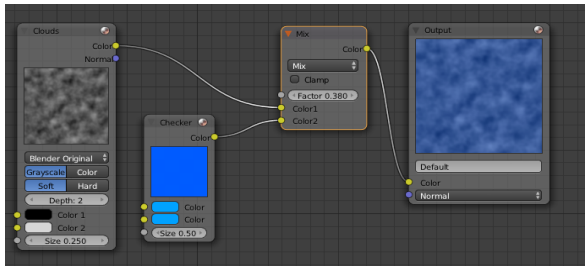


Figure 2. Texture interpolation between two textures in Blender3D

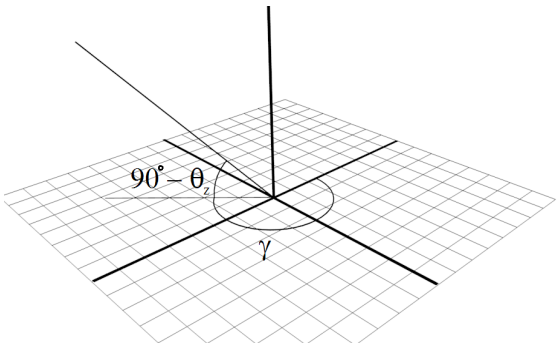


Figure 3. Solar coordinates in a horizontal coordinate system

with the illumination intensity in order to avoid using high distance values inside the model.

E. Rainfall

Rainfall can be visualized using a particle system. Measured amounts of rainfall can be related to the density of the particle system, while the vertical velocity of the particles can be assumed to be constant relative to the viewer and equal to the terminal velocity [10].

Type of rainfall varies between rain, snow and hail. Those types have different characteristics and causes, but it is safe to say that their terminal velocity is similar. The most prominent cause for the different types is the air temperature. Since the visualization is done based on real data, it is hard to deduce the type of rainfall from the data. It can be assumed that snow occurs at temperatures below zero whereas rain occurs at temperatures above zero. Hail is unpredictable and mostly short-term, making it redundant in terms of visualizing data collected on an hourly basis.

Another important factor in the visualization of rainfall is the influence of the wind. Strong winds make the rain and snow fall diagonally or sideways, increasing their horizontal velocities.

Particle systems make it easy to change the parameters for visualizing rainfall, such as particle velocity, vector of movement and the type and color of particles. If the clouds are represented with solid models, such models can serve as particle emitters. However, since clouds are often represented using textures, particle systems with invisible emitters are necessary.

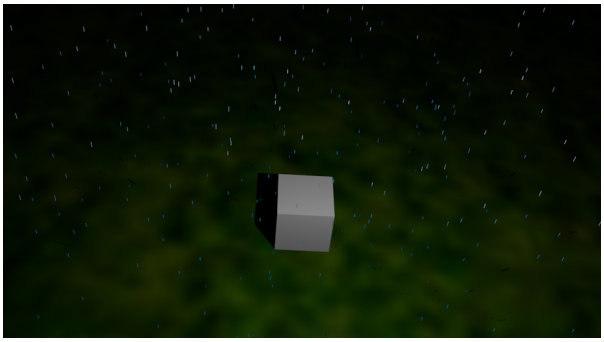


Figure 4. Rain as a particle system

F. Visibility

Visibility is related to the measurement of the distance in which an observer can make a distinction between objects or light sources. It is usually measured in kilometers and defined as the largest distance in which a black object can be seen in a light background [11].

Low visibility is usually caused by fog. Since an observer can move freely through the fog, it needs to be visualized in a way which distributes the fog relative to the observer. It can be visualized using particle systems or textures.

One downside of a particle system for fog visualization is the high amount of particles needed at low visibility. However, in modern computers, particle system fog visualization doesn't present a big problem and it results in more realistic scenes when combined with volumetric lighting [12].

Another way of visualizing fog is by using layers of semi-transparent textures. This method is fast and simple, but it cannot visualize advanced phenomena such as light scattering inside the fog.

III. 3D Model With Simulated Meteorological Effects

In order to demonstrate some of the more simple methods for visualizing meteorological data, a program solution has been made as an example. The program solution supports both rendering a 3D environmental model based on meteorological data at a certain point in time and displaying data for large periods using graphs.

A. Input Data

Available input data consists of real measurements taken at hourly intervals during a larger period of time, and it's contained inside a text file. Text file header contains information about the meteorological station that measured the data: name, geographic coordinates (latitude, longitude) and height above sea level. Each line contains measured data, measured at the interval of one hour. Measured data consists of year, month, day, hour, average wind speed, wind direction, maximum wind speed, air temperature, relative humidity, air pressure, insolation, rainfall quantity, rainfall length, cloud coverage and visibility. Data can easily be accessed for a certain date and time.

B. Graphs

Graphs provide an insight into data, and allow for tracking of changes over a certain period of time. Range of the data selected for the graphs can be arbitrary, which allows for greater flexibility when observing data. An example graph is shown in Fig. 5. A viable option would be selecting the sampling frequency for the horizontal data range, but since performance is not an issue, such option is not considered.

All graphs consist of the date being displayed on the x axis and the value for the measured data being displayed on the y axis. One exception is the wind direction graph,

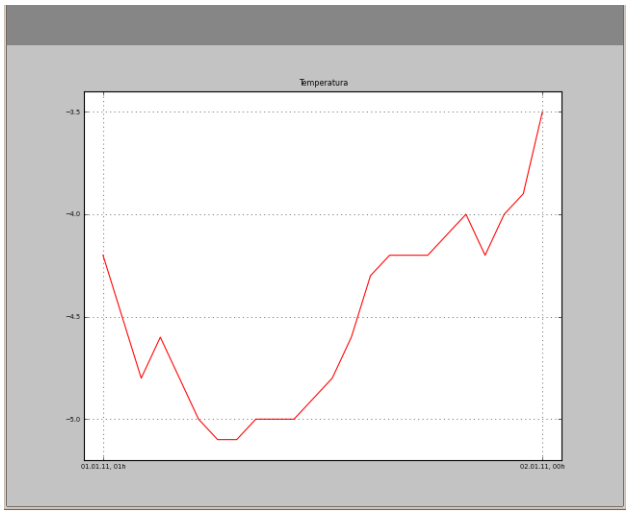


Figure 5. Air temperature graph

which instead displays the frequency of appearance of a wind direction in a polar graph.

C. 3D Graphic Model

Modeling a scene in order to visualize data consists of modeling the scenery and modeling the visual representations of data.

The scenery can be modeled after the real world, e.g. a building is defined. Because the scenery is mostly static, size of the scenery does not play a large part in the performance of the model, but again that depends on the complexity of the methods used for visualization.

For this example, the tools used were Blender3D [6], for the modeling of scenery and additional objects, and Ogre3D [13], for displaying a rendered scene inside a window. Ogre3D uses scene and resource Managers to organize assets necessary for rendering a scene. Resource manager's concerns are importing textures, materials and scene configuration, while the scene manager organizes models, cameras, particle systems and lighting.

Inside the scene manager, it is necessary to initialize a frame listener. Frame listener tracks changes in data and is usually called once per frame, but the calling frequency can be changed. For this example, only change that happens each frame is the rotation of the anemometer. Other changes happen whenever the read input has changed.

In order to visualize the sky and clouds, it is necessary to add a sky dome, as pictured in Fig. 6. A sky dome is a hemisphere placed over the scene. Textures can be placed on the dome which results in the visualized scene to be surrounded with the sky texture.

Since there's a need for detailed tracking of the wind direction and speed, anemometer changes can be followed. That makes necessity for one camera to be focused on displaying only the anemometer. Simultaneous display of multiple cameras can be achieved using multiple viewports, and in this case one viewport is nested inside the other, as can be seen in Fig. 7.

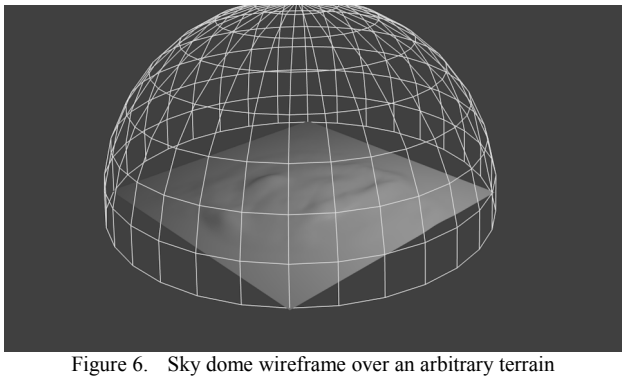


Figure 6. Sky dome wireframe over an arbitrary terrain

Another factor that plays a big role in the rendering of the scene is lighting. Position of the sun can be calculated, and therefore a directional light source can be placed inside the sky dome at the calculated coordinates. However, in small models, light source may continue providing ambient lighting for the model even after sunset. To counter this, ambient lighting must be set to darker values after sunset. Atmospheric refraction can also be simulated, which results in the sky having an orange hue near the time of sunset [14].

Ambient lighting is also dependant on the cloud coverage. The dependancy can be assumed as linear, even though it depends on the distribution of the clouds rather than the exact percentage of the sky covered.

IV. Conclusion

Simpler visualization methods are far from realistic looking in terms of visualizing meteorological data, but provide an insight into the state of the atmosphere at which the data were collected.

There are lots of ways in which various meteorological factors are related to each other, and therefore it is hard to make an accurate model based on small amounts of data. Still, based on the given data, some relations can be noticed and visualized.

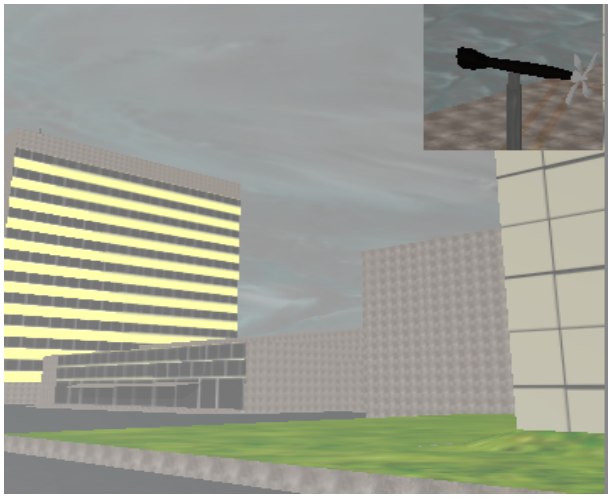


Figure 7. Dual-viewport display

Acknowledgments

This work has been financially supported by the Croatian Science Foundation under grant No. I-4463-2011 (MICROGRID) and by the European Community Seventh Framework Programme under grant No. 285939 (ACROSS). This support is gratefully acknowledged.

References

[1] A. T. Young, An Introduction to Mirages, http://mintaka.sdsu.edu/GF/mirages/mirintro.html, 2013.

[2] R. Hochstim, Gases, http://www.premed411.com/pdfs/CHM456.pdf, 1996.

[3] Reflection and Refraction of Light, http://www.physics.uc.edu/~sitko/CollegePhysicsIII/22-ReflectRefract/ReflectRefract.pdf, 2014.

[4] Anemometer, http://www.saylor.org/site/wp-content/uploads/2011/04/Anemometer.pdf, 2013.Weather Terms Glossary, http://www.erh.noaa.gov/box/glossary.htm, 2014.

[5] Weather Terms Glossary, http://www.erh.noaa.gov/box/glossary.htm, 2014.

[6] Blender 3D, http://www.blender.org/, 2014.

[7] R. Ruiters, R. Schnabel, R. Klein, "Patch-based Texture Interpolation", Eurographics Symposium on Rendering (Proc. of EGSR), vol. 29, 2010, pp. 1421-1429

[8] M. Blanco-Muriel, D. C. Alarcón-Padilla, T. López-Moratalla, M. Lara-Coira, "Computing the Solar Vector", Solar Energy Vol. 70, No. 5, pp. 431-441, 2001.

[9] W. T. Reeves, "Particle Systems— a Technique for Modeling a Class of Fuzzy Objects", ACM Transactions on Graphics (TOG), Vol. 2, No. 2, pp. 91-108, April 1983.

[10] G. B. Foote, P. S. du Toit, "Terminal Velocity of Raindrops Aloft", April 1969.

[11] W. C. Malm, Introduction to Visibility, Cooperative Institute for Research in the Atmosphere (CIRA, NPS Visibility Program), Colorado State University, Fort Collins, CO, 2000. http://www.epa.gov/air/visibility/pdfs/introvis.pdf, 2013.

[12] B. Tóth, T. Umenhoffer, "Real-time Volumetric Lighting in Participating Media", EUROGRAPHICS 2009.

[13] Ogre API Reference, http://www.ogre3d.org/docs/api/html/, 2014.

[14] M. E. Thomas, R. I. Joseph, Astronomical Refraction, http://techdigest.jhuapl.edu/TD/td1703/thomas.pdf, 1996.

3D volume localization using miniatures

Luka Šajn[1], Miroslav Radojevic[2] and Tomaž Dobravec[1]
[1]Faculty of Computer and Information Science
University of Ljubljana, Slovenija
[2]Biomedical Imaging Group of Erasmus MC
University Medical Center Rotterdam, Netherlands

***Abstract*—The prediction of the position of a given volume sample in a full body atlas, also known as a volume localization, is a part of an initial stage of image retrieval in most of the dedicated CAD systems. In this paper we present two methods for volume localization, namely histogram matching and classifier regression. Since the histogram matching method ignores the spatial orientation, it is used when the orientation of the volume cubes are not the same. On the other hand the classifier regression is much faster and can be used as a quick estimation and as a tool to reduce the scope of the initial problem. Both presented methods were tested on a dataset with 3962 volumes of a human body atlas. The accuracy and the speed of execution was compared and is presented in this paper.**

I. Introduction

3D volume localization problem is a problem of estimating the atlas coordinates of a given sample 3D volume (i.e. object, miniature). This is usually done by comparing the characteristics of a sample volume with the characteristics of other volumes with known coordinates. The coordinates of the most similar volume are returned as a result, as shown in Fig. 1.

To achieve this, one can use the method which compares the sample volume with the other volumes from the dataset and the coordinates are estimated as the interpolation of the coordinates of the most similar volumes, weighted by the degree of similarity. Another approach attempts to estimate the coordinates using classifier regression. Part of the problem is to find the matching volumes - this task can be viewed as the contents based image retrieval (CBIR) problem. One of the models to solve the CBIR is to treat the images as collections or histograms of the features [2] and compare the distances (dissimilarities) between the feature vectors. The resulting feature is the one with the lower distances, hence higher similarity. Potential features used for grey-level images

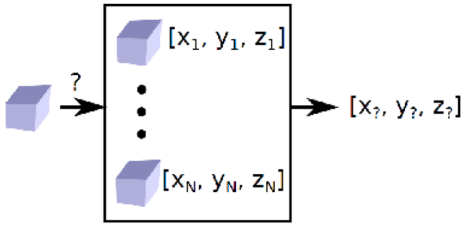


Fig. 1. Estimating the coordinates of the most similar volume

can be the very pixels, features that correspond to human visual perception, texture features such as global descriptors or Gabor features, or local descriptors such as SIFT or corner detectors [2]. Shape features such as Fourier descriptors, moment invariants and finite element models were surveyed in [12], graph based shape features were presented in [1]. A method for retrieving 3D datasets based on Local Binary Patterns (LBP) features [4] was introduced in [11] and compared with features such as 3D Wavelet Transforms and 3D co-occurence matrices. In this paper we use and compare two methods for the coordinate prediction, namely the histogram matching, which is presented in Section II-A, and the classifier regression, presented in Section II-B. The results of the comparison are given in Section III. We conclude the paper in Section IV.

II. Methods

Two methods for the coordinate prediction are used: histogram matching and the classifier regression.

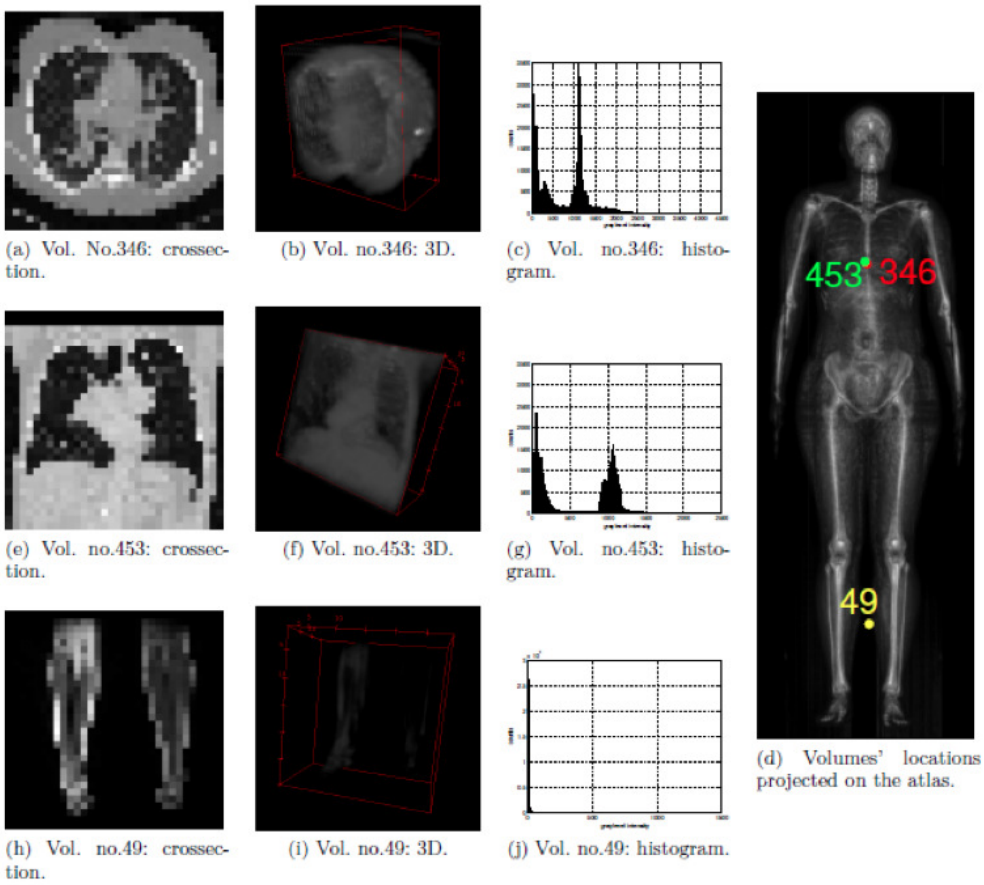
A. Histogram matching

The orientation of the sample volume is usually not the same as the orientation of the volumes in the dataset. Hence, it is necessary either to encounter for their proper alignment, or to use a measure independent to the orientation of the volumes. Such measures include histogram, rotation invariant image moments and local features. The first method presented in this paper compares the histogram of the samples with the histograms of all the other miniatures. The drawback of using histograms is the loss of spatial intensity distribution information. Some other rotation invariant feature extraction method such as SIFT or LBP in 3D [4] would be potentially suitable to overcome the problem of describing the spatial relations between voxels.

Figure 2 depicts three different volume samples in both 3D and cross-section version, their histograms and the detected locations within the human body atlas using the histogram matching method.

1) Histogram dissimilarity measures: There are several cross-bin and bin-to-bin dissimilarity measures to distinguish among histograms being compared [6]. Generally, cross-bin distances such as Earth Movers Distance (EMD) tend to be more robust and more discriminative than bin-to-bin measures when comparing two histograms [9].

Let L and M represent histograms being compared and each value $M(i)$ the count of the number of observations that fall



(a) Vol. No.346: crossection. (b) Vol. no.346: 3D. (c) Vol. no.346: histogram. (d) Volumes' locations projected on the atlas. (e) Vol. no.453: crossection. (f) Vol. no.453: 3D. (g) Vol. no.453: histogram. (h) Vol. no.49: crossection. (i) Vol. no.49: 3D. (j) Vol. no.49: histogram.

Fig. 2. Three volumes with their histograms and detected location within human body atlas.

into one of the disjoint intensity intervals. If n be the total number of observations and k the total number of bins, then the following holds:

$$n = \sum_{i=1}^{k} M(i)$$

Among the most popular bin-to-bin measures let us mention: histogram intersection, L_1 and L_2 norm, χ^2, and Jeffreys Divergence with calculation formulas (in respective order)

- $h(L,M) = \frac{\sum_j min(L(j),M(j)))}{\sum_j M(j)}$
- $L_1(L,M) = \frac{\sum_j abs(L(j)-M(j))}{N}$
- $L_2(L,M) = \frac{\sum_j (L(j)-M(j))^2}{N}$
- $\chi^2(L,M) = \frac{\sum_j (L(j)-M(j))^2}{\sum_j L(j)+M(j)}$
- $\mathrm{Jf}(L,M) = \sum_j \left(L(j) log \frac{2L(j)}{L(j)+M(j)} + M(j) log \frac{2M(j)}{L(j)+M(j)} \right)$

An important representative of the cross-bin histogram comparison method is called the Earth Movers Distance (EMD) [8], [9]. Fast implementation of the cross-bin histogram comparison EMD was used in our tests [8].

2) Weighting the coordinates: We have performed the tests by calculating all of the inter-histogram dissimilarities. After this calculation, ten most similar volumes were queried out of the base and used to estimate the coordinate. Estimation was based on weighted averaging where weights sum up to unit and each weight corresponds to the similarity level d_i and follows negative exponential function with slope determined by normalized variance of 10 selected dissimilarities.
The coordinates of the ten best matches (ranked with i) were weighted by exponentially decaying weights, derived from the variance of their dissimilarity measures. The resulting coordinates $[x_{est}, y_{est}, z_{est}]$ were calculated by

$$[x_{est}, y_{est}, z_{est}] = \sum_{i=1}^{10} w_i [x_i y_i z_i]$$

where $w_i = e^{-\lambda i}$ and $\lambda = max\left(\frac{d_i - mean(d_i)}{var(d_i)}\right)$.

Such weighting was used to overcome the influence of the outliers on the estimation since outliers can cause higher variance in general as shown in Fig. 3.

B. Classifier regression

Another approach to predict the coordinate values uses classifier regression. Classifiers such as 10-NN and Random Ferns

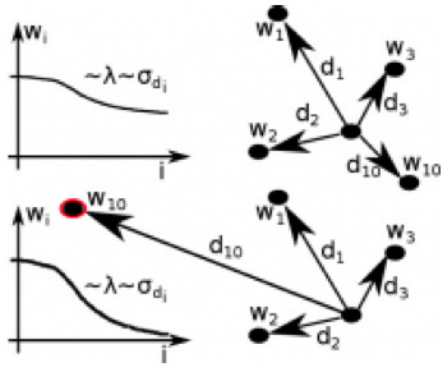


Fig. 3. The weights of the outliers are smaller than the weights of the neighboring candidates.

[7] use the image features and carry out the regression instead of usual crisp classification. PRTools Matlab library [3] is used for 10-NN implementation while Piotrs Matlab Toolbox [10] is used for Random Ferns implementation. Classifier scores are used to interpolate coordinates of those that were classified as being close to the test sample in a way typical for each classifier. Namely, averaging of the 10 nearest neighbors coordinates will be 10-NN regression. Similarly, in the case of SVM classifier usage, coordinates of the most similar ones separated by SVM will be interpolated. Features used for classifier regression are: mean, standard deviation, median, volume (actual number of pixels in the volume), centroid (1st order image moments), central moments (2nd order image moments), and the voxels themselves can be used as features, too. In cases when voxel intensities are used (Random Ferns classification), binary features need to be extracted - hence number of combinations can grow high. Therefore, volumes were resized using Gaussian pyramid before calculating the binary features.

III. Results

We tested the methods on a dataset which consist of 3926 volumes (miniatures representing body regions). Each of the volumes was resampled to $32 \times 32 \times 32$ image cube. Voxel intensities of the image cube are real values ranging from 0 to 4095. Each volume corresponds to certain spatial regions of the body as it was shown with dots (center points of the region) in Fig. 5. Volumes are actually resembling different regions of the body and they are taken from different orientations (transverse, coronal, sagital).

Results are presented for both approaches introduced in the Section II. "Histogram comparison" turns out to be quite precise at finding similar patches. However, the choice of the dissimilarity metric is very important. As mentioned, cross-bin dissimilarity is preferred since it is more robust. The performance of different metrics when matching histograms for query volume (object nr. 1540) shown together with its top 10 most similar patches from the database. Process finds 10 most similar volumes and interpolates their coordinates.

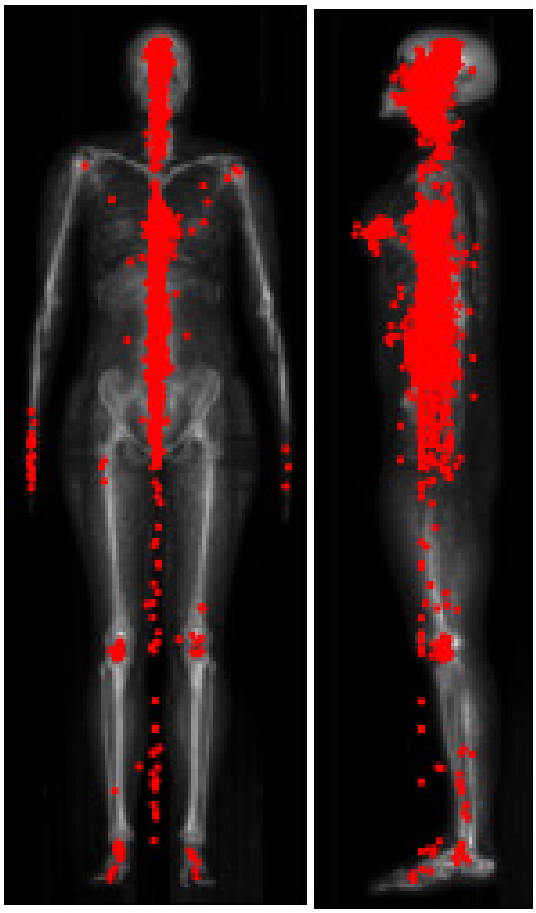


Fig. 4. Human atlas with 3926 volumes.

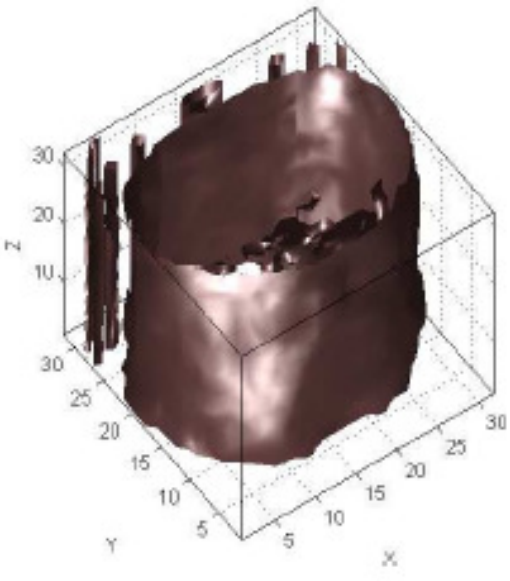


Fig. 5. Voxel representing a region of a human body.

Dissimilarity measure	h	L_1	L_2	χ^2	Jf	EMD
Estimated position error	16.5792	3.7436	129.8316	1.3632	1.2895	**1.2466**

TABLE I

ESTIMATED POSITION ERRORS FOR DIFFERENT HISTOGRAM COMPARISON MEASURES (INTERSECTION, L1 AND L2 NORM, χ^2, JEFFREY DIVERGENCE AND EMD) FOR THE OBJECT NR. 1540

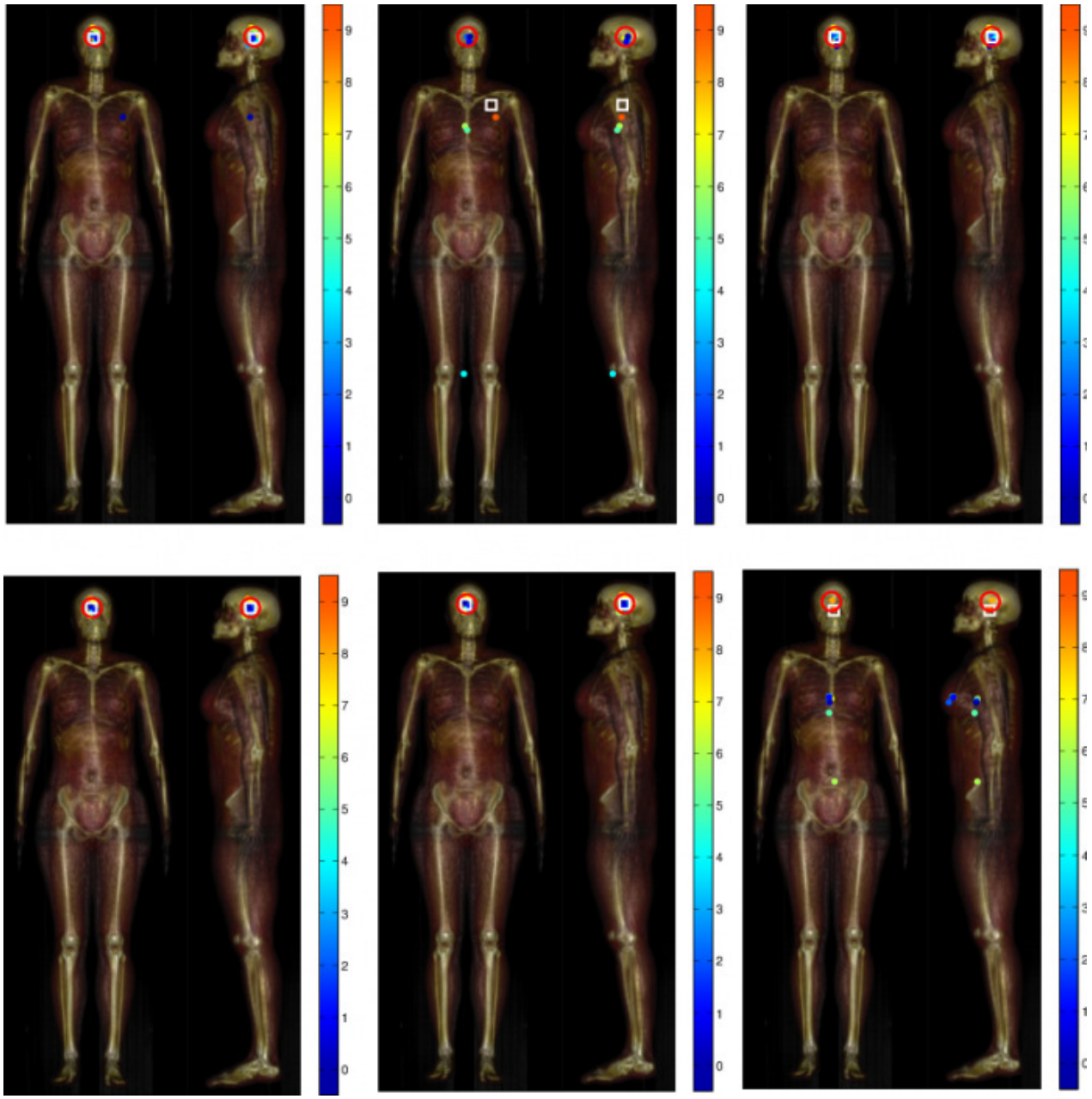


Fig. 6. Results of volume matching using histogram comparison for dissimilarity measures L1 norm, L2 norm, EMD, χ^2, Jeffrey divergence, and Intersect (from left to right, top to bottom).

EMD dissimilarity measure turns out to be the one that performs the best both in terms of accuracy and robustness, however it is slower and features used for matching are dependent on the database. Downside of its usage is computational cost, nevertheless fast implementation of EMD [8] used in this experiment provides satisfactory performance. What is still a potential problem with this approach is the fact that it is dependent on the data from the training set - since it is based on comparing the input with the training dataset. In that sense - classifier regression based approach with invariant

statistical/shape/moment based features would be interesting. Figure 6 show the results of volume matching using histogram comparison for dissimilarity measures: L1 norm, L2 norm, EMD, and χ^2, Jeffrey divergence, and intersection (from left to right, top to bottom). Predicted value is represented with white square, while the true value is described with red circle. Top 10 matching positions were shown as dots in jet color space - ranging from red ones that represent high ranking till the blue ones with low ranking. Estimation fails in case metric was not properly chosen. EMD turns out to be the one that performs the best both in terms of accuracy and robustness.

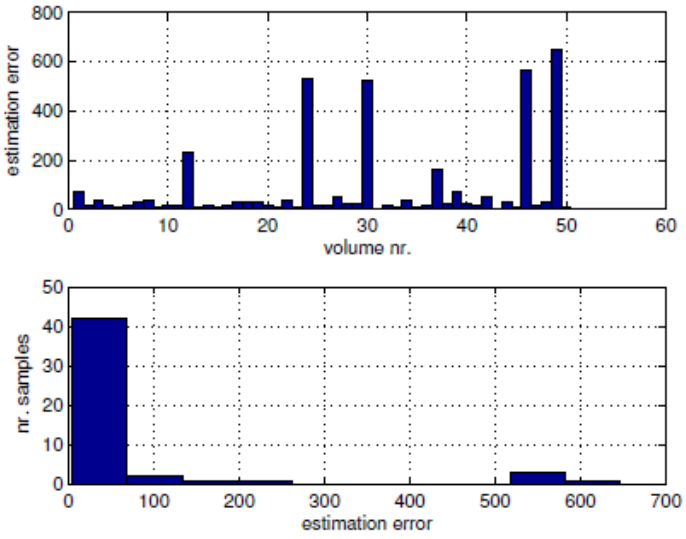


Fig. 7. Statistics for the absolute distance error between the position estimate and the actual position.

Finally, histogram matching was performed individually on 50 randomly chosen objects while the rest were used for matching with the test sample each time. Figure 7 shows the statistics for the absolute distance error between the position estimate and the actual position. Majority of discrepancies fall into 0-50 voxel distance range which suggests that the object was targeted. Usually, distances up to 50 absolute voxel values do properly guess the object since the main task is rough localization of the body region.

"Classifier regression" was based on independent features (such as volume mean, standard deviation, image moments, volume). As stated before - classifier score was the basis for the data regression. 10-NN classifier was used to find the neighbors in feature space and average the coordinates associated with them in order to carry out the regression. This process works much faster since it takes short time to extract the features and there is no need for comparison. It accomplishes the retrieval but the performance is lower as shown in the figure 8.

Regression accuracy is worse than the one obtained using histogram comparison and one of the causes is probably plain averaging of the feature-space-nearest-neighbor coordinates

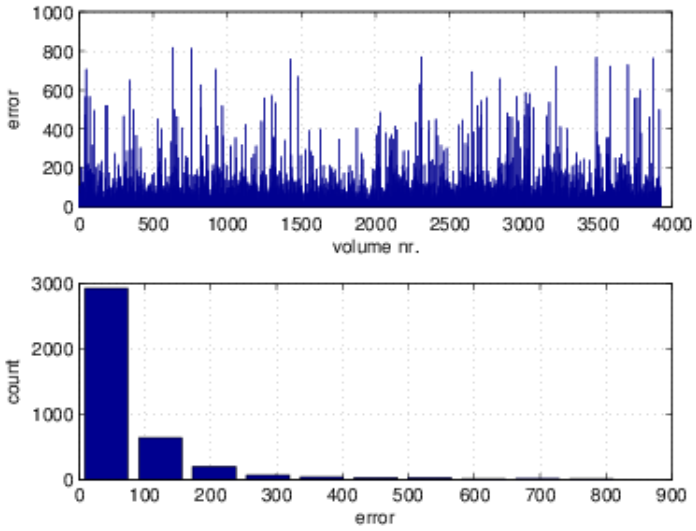


Fig. 8. Results of classifier regression (10-NN) based on independent features

(no weighting). The features used can influence the choice of the nearest neighbors. Finally, Random Ferns [7] were tested on position regression. They performed slightly slower than 10-NN regression, but had an overall lower average error calculated over 50 randomly chosen miniatures. Voxel intensities were used as the source for making binary features that this classification algorithm uses. Usage of statistic/shape features that were supplied to 10-NN did not perform as well as voxel intensities did. Binary features were chosen to be -1 or +1 for each possible pair of voxels. In case the first one from the pair was higher - feature was set to +1 and -1 was for the opposite case. Each miniature needed to be resized to $8\times8\times8$ using Gaussian pyramides and even after the reduction - quarter of the total number (every 4th) of the voxels was used for matching in the experiment due to processing time constrains. Results proved to be promising in spite of all the constrains.

IV. Conclusion

This paper summarizes the results obtained while developing 3D content based image localization framework. Two approaches have been examined - one based on comparing the histograms and another one based on the classifier regression. 10-NN and Random Ferns classifiers were used for the regression. The best performance in terms of accuracy was achieved by comparing histograms using the EMD (Earth Movers Distance) method. Classification regression using 10-NN proved to be slightly less precise but significantly faster, and the same conclusion can be made for Random Ferns that found their best performance by taking voxel values as features and compared the 3D patches for similarity. Average error on position prediction calculated on 50 random samples was lower than the one obtained using 10-NN. Nevertheless, Random Ferns classification was significantly constrained with reducing the feature number for the computational time

purposes, therefore its performance can certainly improve. Future work can include clustering of the coordinates into crisp classes so that further evaluation of position prediction performance can be estimated in a more appropriate way. 10-NN regression can be improved by using weighted average of the nearest neighbors or some additional features. Random Ferns regression can perform better with more elegant way of reducing features it uses. Finally, local feature extraction methods such as 3D SIFT or 3D LBP can be used.

REFERENCES

[1] M.-C. Chang and B.B. Kimia. *Measuring 3d shape similarity by graph-based matching of the medial scaffolds.* Computer Vision and Image Understanding, 115(5):707 720, 2011. Special issue on 3D Imaging and Modelling.

[2] T. Deselaers, D. Keysers and H. Ney. *Features for image retrieval: an experimental comparison.* Information Retrieval 11:77107, 2008. 10.1007/s10791-007- 9039-3.

[3] R.P.W.Duin, P.Juszczak, P.Paclik, E.Pekalska, D.deRidder, D.M.J.Tax, and S.Verzakov. *Pr-tools4.1, a matlab toolbox for pattern recognition.* http://prtools.org, 2007.

[4] J.Fehr and H.Burkhardt. *3d rotation invariant local binary patterns. In Pattern Recognition, 2008.* ICPR 2008. 19th International Conference on, pages 1–4, dec. 2008.

[5] J. Flusser, T. Suk, and B. Zitov. *Moments and Moment Invariants in Pattern Recognition.* John Wiley & Sons, Ltd, 2009.

[6] F.D. Jou, K.C. Fan, and Y.L. Chang. *Efficient matching of large-size histograms.* Pattern Recognition Letters, 25(3):277 286, 2004.

[7] M. Ozuysal, M. Calonder, V. Lepetit, and P. Fua. *Fast keypoint recognition using random ferns. Pattern Analysis and Machine Intelligence*, IEEE Transactions, 32(3):448 461, march 2010.

[8] O. Pele and M. Werman. *Fast and robust earth movers distances.* In Computer Vision, 2009 IEEE 12th International Conference on, pages 460 467, 29 2009-oct. 2 2009.

[9] O. Pele and M. Werman. *The quadratic-chi histogram distance family.* Computer Vision ECCV 2010, volume 6312 of Lecture Notes in Computer Science, pages 749762. Springer Berlin / Heidelberg, 2010.

[10] D. Piotr. *Piotrs image and video matlab toolbox (pmt).* http://vision.ucsd.edu/ pdol- lar/toolbox/doc/index.html.

[11] Y. Qian, X. Gao, M. Loomes, R. Comley, B. Barn, R. Hui, and Z. Tian. *Content-based retrieval of 3d medical images.* In eTELEMED 2011, The Third International Conference on eHealth, Telemedicine, and Social Medicine, pages 712, 2011.

[12] Y. Rui, T.S. Huang, and S.F. Chang. *Image retrieval: Current techniques, promising directions, and open issues.* Journal of Visual Communication and Image Representation, 10(1):39 62, 1999.

Evaluation of Image Processing Algorithms on ARM Powered Mobile Devices

Damir Demirović*, Amira Šerifović-Trbalić*, Naser Prljača*, Philippe C. Cattin†
*Faculty of Electrical Engineering,
University of Tuzla, Tuzla, Bosnia and Herzegovina
{damir.demirovic,amira.serifovic,naser.prljaca}@untz.ba
†Medical Image Analysis Center,
University of Basel, Basel, Switzerland
{philippe.cattin}@unibas.ch

***Abstract*—Image processing plays an important role in medical image analysis. The most popular methods for image processing and analysis are very resource hungry, which leads to some disadvantages in their applications even on a powerful desktop computers. On the other side, modern mobile devices are equipped with powerful processors with an efficient instruction architecture. This lead to better performance per watt than a desktop CPUs. This work investigates the performance of a widely used medical analysis algorithm implemented on a modern mobile devices and desktop CPU. The results obtained with ARM NEON instructions show speed improvements up to 2 times. As this research shows mobile devices cannot yet compete with powerful desktop CPUs, even with using highly optimizations or multiple threads. In the last part of the paper conclusions are drawn for acceptable image input and parameter sizes.**

I. Introduction

In Chandrakasan et al. [1] was one of the first arguments in favor of multicore microprocessors. Reason for using multicore processors is important for power savings, so dual-core microprocessor running at half of frequency of a single core microprocessor use about 0.6 times less power.
The trend of using more, but less powerful, processor cores dominates microprocessor industry in the last decade. Performance of mobile devices depends of power efficiency, and can be expressed as performance per watt.
In this work, we investigate the possibilities and limitations of well known Demons algorithm implemented for mobile Android device. Previous studies have investigate various image processing algorithms implemented on a mobile devices.
One of the first known work about image processing on Android mobile device were given by Wells [2]. This paper analyze the computation time of an image processing algorithm.
Mitra et al. [3] analyze the performance of various image processing algorithms implemented on ARM devices, and compares the results with an Intel CPUs, using SIMD (Single Instruction, Multiple Data) instructions.
In Luxemberger et al. [4] optical flow algorithms on ARM powered Android mobile phones were developed.
In Pulli et al. [5] authors compare performances of various algorithms on mobile Tegra 3 platform using OpenCV, and optimizations like NEON instructions, GLSL (OpenGL Shading Language), and TBB (Threading Building Blocks).
Welch [6] investigates SIMD instructions on a Cortex-A8 processor with NEON co-processor, using two image processing algorithms: bilinear interpolation and distortion. SIMD instructions exploits several pixels at once for bilinear interpolation algorithm with speed up factor of 2, and distortion algorithm optimize memory access by prefetching data, with a speed up factor of 3.
In Holgersson [7] author implements four different implementations of biquads IIR filters for audio processing, and evaluates results with/without fixed-point processing and NEON optimizations. Their optimizations is based on parallel filter run, memory operation efficiency and fixed-point processing.
In Wang et al. [8] authors implemented Edge-based Image Decomposition TV-L1 optical flow algorithm on a mobile phone running Android platform. The algorithm calculates optical flow between the two images.
OpenCV library [9] is a widely used open source computer vision and machine learning software library highly optimized for desktop processors. In recent time OpenCV was ported to mobile platform including Android, and includes optimized code for ARM processors. Several books about using OpenCV in Android exists today (see for example [10]).
All previously mentioned implementations use powerful instructions of the modern ARM processors for mobile devices, which increase their usage in a highly computational applications for various signal processing and computer vision tasks. To our knowledge there is no implementation of Demons registration method for mobile devices.
The main contribution of this work is implementation of the Demons registration algorithm (Thirion [11]) on ARM CPUs running Android.

II. Methods

A. Demons registration

The goal of image registration (see for example Zitova [12]) is to obtain a mapping from one coordinate system of one image to the coordinate system of the other image. The one of the most widely used deformable registration algorithm is Demons registration algorithm (Thirion [11]). This algorithm is widely known and yet simple to use (Figure 1). It has linear computational complexity (i.e. depends linearly with the input image pixel number). Also, it can be easily parallelized, where each pixel can be computed separately, there is no data dependence between the neighbors (see for example Muyan-Ozcelik [13]). Demons algorithm forces can be expressed with the Optical flow equation (Horn [14]).

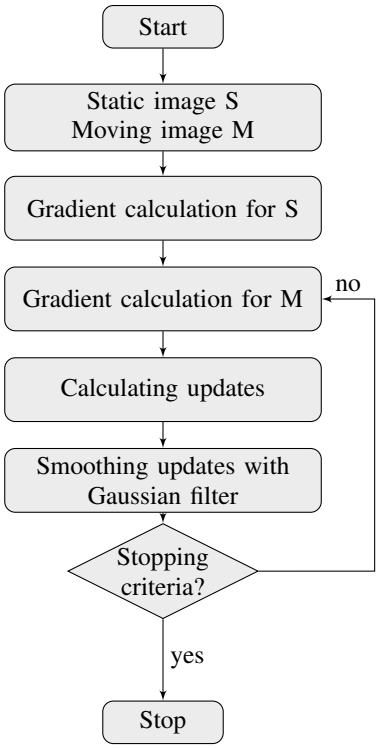


Fig. 1. Flowchart of Demons algorithm

B. Implementation for an Android device

Android platform is now a world dominant in mobile platform market. Android as a development platform use the Java Standard Edition with additional Android classes. Applications are developed in the Java language using Android Software Development Kit (SDK). For a low level interaction with hardware, a modified Linux kernel is used. For performance critical applications Native Development Kit (NDK) is unavoidable [15]. Therefore, part of the applications can be built with C/C++ language. NDK uses Java Native Interface (JNI) which is a standard way for Java platform to interact with the native code.

1) Optimizations for ARM platform: Several optimizations can be used with NDK. For Tegra 3 and Samsung Exynos platforms optimizations for ARM V7 instructions can exploit a specific NEON instructions.

2) SIMD Instructions: SIMD (Single Instruction, Multiple Data) operations use multiple data elements at once. These instructions dates back to Intel MMX (from 1996) set which gains popularity in multimedia applications. Intel MMX instructions yields significant speed-up comparing to the sequential instructions. This usually means the more processing power for the roughly the same energy consumption. Processors used in mobile applications today are usually 32-bit, with recent trend to switch to 64-bit. For example, ARM Cortex A9 processor (used in Tegra 3 chip) is a 32-bit processor. Therefore, operations using 8-bit and 16-bit data are performed in 32-bit registers. These operations can be executed at once. This is significant in multimedia applications and image processing tasks where the same operations can be executed on many image elements at once. Combination of SIMD instructions and parallel code execution is usually optimal for these algorithms.

TABLE I. OPTIMIZATION PARAMETERS FOR GCC COMPILERS

Platform	Optimization string
Desktop	-O2
ARMv7	-march=armv7-a -mtune=cortex-a9 -mfloat-abi=softfp -mfpu=neon
ARMv6	-march=armv6 -marm -mfloat-abi=softfp -mfpu=vfp

3) ARM Neon Architecture: ARM NEON [16] introduced instructions for ARMV7 for using with several data types. NEON implementation is 128-bits SIMD architecture extension which dates back for Cortex-A series processors. It has 32 registers, 64-bits wide, and dual view as 16 registers 128-bits wide. Datatypes can be signed/unsigned 8-bit, 16-bit, 32-bit, 64-bit, and single precision floating point.

C. Porting image registration library for Android platform

We use our developed C++ library for image processing and registration, which is a rather simple library for image loading, image processing and registration tasks. For image visualization Fast Light Toolkit (FLTK) [17] can be used. For the purposes of this study, library was ported from desktop to Android platform using Android NDK. Porting code with NDK is a relatively straightforward task.

III. RESULTS

For the evaluation of registration speed we chose two MRI liver images (Figure 2) obtained from [18]. Two slices (resampled to 256×256) were extracted from the dataset and can serve as the input in all our experiments. We use Demons registration algorithm (Figure 1) to align these images. As a result, deformation field is obtained, which is a vector field showing a mapping between each corresponding pixels in the two images. Single threaded version of Demons algorithm

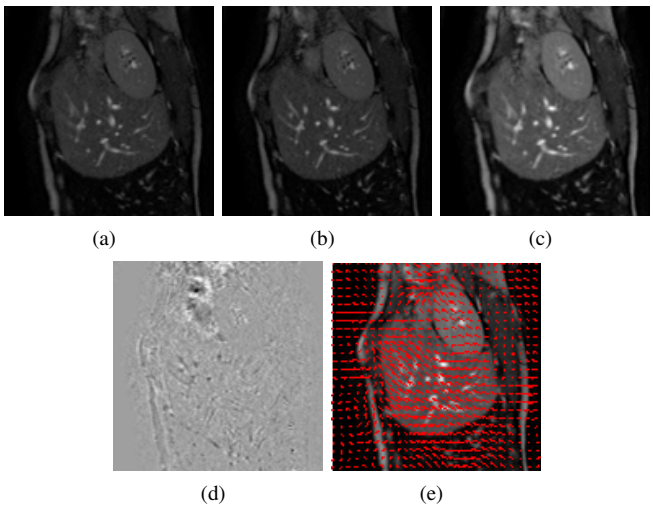


Fig. 2. An example for registration of liver MRI a) fixed image b) moving image c) registered image d) difference between registered and fixed image e) deformation field overlayed on static image

TABLE II. DEVICES SPECIFICATION

Device #	Processor/Model	Threads/Cores/GHz	Cache L1/L2/L3	Memory (GB)	SIMD	Power (W)
1	Intel® Core™ i5-2500	4/4/3.3	4x32/4x256kB/6MB	4	SSE4.2	95
2	Intel(R) Xeon(R) E5450	4/4/3	4x32kB/2x6MB/	8	SSE4.1	80
3	Intel Celeron M ULV	1/1/900MHz	1/512kB7/-	512MB	No	5
4	Samsung GALAXY S3	4/4/1.2	31kB+31kB/1MB-	1	Neon	3 (est)
5	Nexus7 2012	4/4/1.2	31kB+31kB/1MB-	1	Neon	1.1
6	Wildfire S	4/4/1.2	31kB+31kB/1MB-	512MB	No	1

TABLE III. RESULTS FOR DEMONS ALGORITHM (IN SECONDS).

	Parameters		Device #					
Image size	width	σ	1	2	3	4	5	6
256×256	7	1.0	1.5	1.2	34.3	8.6	9.7	30.1
	31	5.0	12.8	18.2	422.7	91.2	103.3	1033.8
	61	10.0	42.2	63.2	1549.9	338.8	379.6	4175.5
	121	20.0	144.7	245.9	5558.1	1351.2	1514.3	17414.5

were used for all experiments. It can be shown that 50 iterations is enough for Demons to converge. For each device several Demons Gaussian parameters are applied (Table III). All calculations have been carried using a single-precision numbers. For a chosen σ we calculate the filter, in pixels, with the following formula $width = 2 * (3 * ceil(\sigma)) + 1$ where the *ceil* is defined as $ceil(x) = \lceil x \rceil$. Four registrations has been runned on the each device (Table II). Obtained results are shown in the Table III, and in Figure 3.

The results can be grouped by device type to: desktop and mobile. Desktop devices for this study were, a small laptop, home PC and server machine, the last two typically used for image registration. Mobile devices varied in speed from older and newer cell phones, and a newer tablet. For the selected platforms we use the following software.

1) Desktop devices: For experiments on the desktop devices we use GCC 4.6.3 compiler on the corresponding machine with GNU/Linux with compiler optimizations (Table I). These results can be taken for ground truth.

2) Mobile devices: For developing for the mobile devices we used the Android NDK version r9 on GNU/Linux, with the GCC 4.6.3 compiler with the corresponding optimizations as shown in Table I.

All simulation results are given in Table III and Figure 3. As we can see from the obtained results Demons registration is iterative and CPU-intensive algorithm. Registrations can run in real-time for a smaller image dimensions and kernel widths, and their usage is restricted in medical practice. From the all runned experiments one can see that performance for the smaller Gaussian kernel sizes are almost identical. This means that for a relatively smaller deformations, smaller kernel width is needed. Due to a bigger cache size and better architecture wider kernel results are in favor of desktop CPUs. Poor results for Wildfire S is due to an older ARM architecture as well as a lower frequency. The results depends on the ARM CPU and it is slower from 1.1 to 9 times.

Exploiting multiple cores can bridge the gap between mobile and desktop CPUs for smaller Gaussian kernels.

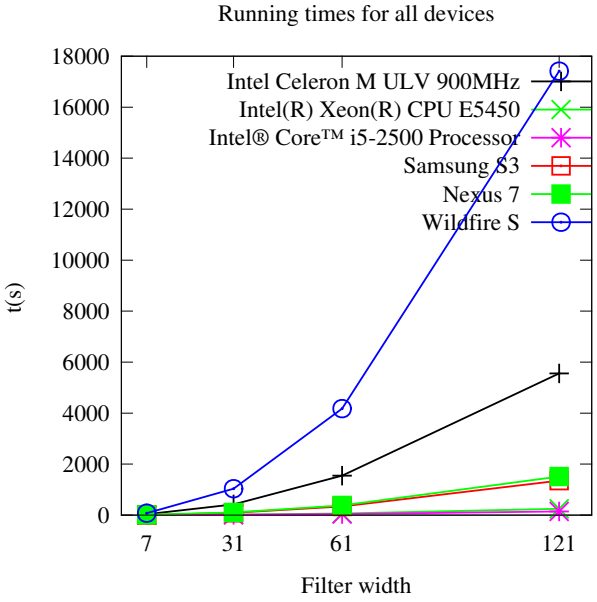


Fig. 3. Registration results for all devices

IV. DISCUSSION

This paper evaluates the performance of Demons image registration implemented on mobile ARM processors. For the experiments we choose a pair of liver MR images, and register them using a several pair of Gaussian parameters. Each experiment were carried on several desktop and mobile devices. Results shows, in general, that mobile devices cannot compete with desktop CPUs in highly demanding tasks like image registrations. Mobile CPUs can be used as an alternative with smaller Gaussian kernel widths, which corresponds to smaller deformations in the input images. The best results for the mobile devices were about 9 times slower than best running time on the desktop CPU.

Further research in this direction can evaluate tradeoff between registration precision and execution times regarding floating point optimizations in a way similar to the method in Holgersson [7]. Regarding speeding up the algorithm one possibility is multicore implementation of Demons algorithm on the ARM CPUs, i.e. implementation using TBB.

V. Acknowledgments

This work has been supported by the SCOPES project of the Swiss National Science Foundation (http://www.snf.ch)

References

[1] A. Chandrakasan, M. Potkonjak, R. Mehra, J. Rabaey, and R. Brodersen, "Optimizing power using transformations," *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, vol. 14, no. 1, pp. 12–31, 1995.

[2] M. Wells, "Mobile image processing on the google phone with the android operating system, online http://www.3programmers.com/mwells/documents/pdf/finalretreived: 10-01-2014," Tech. Rep., 2009.

[3] G. Mitra, B. Johnston, A. P. Rendell, E. McCreath, and J. Zhou, "Use of simd vector operations to accelerate application code performance on low-powered arm and intel platforms." in *IPDPS Workshops*. IEEE, 2013, pp. 1107–1116. [Online]. Available: http://dblp.uni-trier.de/db/conf/ipps/ipdps2013w.html\#MitraJRMZ13

[4] A. Luxenburger, H. Zimmer, P. Gwosdek, and J. Weickert, "Fast pde-based image analysis in your pocket," in *SSVM*, A. M. Bruckstein, B. M. ter Haar Romeny, A. M. Bronstein, and M. M. Bronstein, Eds., pp. 544–555.

[5] K. Pulli, A. Baksheev, K. Kornyakov, and V. Eruhimov, "Real-time computer vision with opencv," *Commun. ACM*, vol. 55, no. 6, pp. 61–69, Jun. 2012. [Online]. Available: http://doi.acm.org/10.1145/2184319.2184337

[6] E. Welch, D. Patru, E. Saber, and K. Bengtson, "A study of the use of simd instructions for two image processing algorithms," in *Image Processing Workshop (WNYIPW), 2012 Western New York*, 2012, pp. 21–24.

[7] S. B. Holgersson, "Optimising iir filters using arm neon," Master's thesis, 2012.

[8] B. Wang, Q. Zhu, H. Xiong, and C. Luo, "Accelerate tv-l1 optical flow with edge-based image decomposition and its implementation on mobile phone," in *Proceedings of the 10th International Conference on Mobile and Ubiquitous Multimedia*, ser. MUM '11. New York, NY, USA: ACM, 2011, pp. 144–151. [Online]. Available: http://doi.acm.org/10.1145/2107596.2107614

[9] G. Bradski and A. Kaehler, *Learning OpenCV: Computer Vision with the OpenCV Library*. Cambridge, MA: O'Reilly, 2008.

[10] J. Howse, *Android application programming with OpenCV*. Birmingham: Packt Publ., 2013.

[11] J.-P. Thirion, "Image matching as a diffusion process: an analogy with maxwell's demons," *Medical Image Analysis*, vol. 2, no. 3, pp. 243–260, 1998.

[12] B. Zitová and J. Flusser, "Image registration methods: a survey," *Image and Vision Computing*, vol. 21, pp. 977–1000, 2003.

[13] P. Muyan-Özçelik, J. D. Owens, J. Xia, and S. S. Samant, "Fast deformable registration on the gpu: A cuda implementation of demons." in *ICCSA Workshops*, M. L. Gavrilova, O. Gervasi, A. Laganà, Y. Mun, and A. Iglesias, Eds. IEEE Computer Society, 2008, pp. 223–233. [Online]. Available: http://dblp.uni-trier.de/db/conf/iccsa/iccsa2008-w.html\#Muyan-OzcelikOXS08

[14] B. K. P. Horn and B. G. Schunck, "Determining optical flow," *ARTIFICAL INTELLIGENCE*, vol. 17, pp. 185–203, 1981.

[15] A. NDK, "http://developer.android.com/tools/sdk/ndk/index.html," *accesed 10.01.2014*.

[16] *ARM Limited, Introduction NEON, Development Article*, 2009.

[17] "Fltk library." [Online]. Available: http://www.fltk.org/index.php

[18] *Organ Motion from 4DMRI*. http://www.vision.ee.ethz.ch/~organmot/chapter_4dmri.shtml.

Natural Eye Gaze Computer Interaction for Medical Oculography Diagnosis: Current Status and Future Prospects

I. Grubišić*, I. Grbeša*, T. Lipić*, K. Skala*, O. Zrinšćak**, R. Iveković** and Z. Vatavuk**
* Ruđer Bošković Institute, Zagreb, Croatia
** "Sisters of Mercy" University Hospital Centre
University Department of Ophthalmology, Zagreb, Croatia

Abstract - With recent technological developments, modern computing systems are becoming more and more powerful. Common computers are capable of calculating more than 50,000 million instructions per second. Currently the biggest obstacle in computer usage is slow communication between the human and the computer. With innovation and development of novel (natural based) sensors capable of capturing positions and natural motion of human body a new, natural, way of interacting with computer is enabled.

This paper gives an overview of eye tracking technology, and presents its application in nature based human-computer interaction (HCI) systems by extending standard controllers, such as the keyboard and the mouse. Novel approaches enable HCI closer to the communication patterns of human beings. In addition, the usage of those concepts in applications for eye diseases and disorders diagnostics and rehabilitation is discussed.

Keywords: eye tracking, computer interaction, hci, oculography, medical diagnosis.

I. INTRODUCTION

The use of standard interaction devices (such as mouse and keyboard) for communication between human and computer is already addressed as one of the biggest obstacle in modern computer usage [1], [2]. In this research field even HORIZON 2020, the EU Framework Programme for Research and Innovation, recognized problem and necessity for improvement. In the year of 2014, one of the first opened calls in HORIZON 2020 Programme, is the special topic Multimodal and Natural computer interaction, in the section ICT 2014 - Information and Communications Technologies.

Recent development of intelligent sensors for capturing and recognizing natural human behaviours enabled novel approaches to answer to this problem by making human-computer interaction (HCI) a natural process. During last few years several different controllers emerged as products affordable for general population.

Microsoft Kinect controller is a multi-modal sensor which contains a RGB camera, an infrared (IR) emitter and an IR depth sensor, a multi-array microphone and a 3-axis accelerometer. With all of these inputs Kinect is able to determine its current orientation, to find the location of the sound source and the direction of the audio wave, to identify human in the scene, to find the location of human body parts, to recognize human body poses and gestures [3]. All of the calculations were performed by computer embedded into the device.

The Leap Motion controller is a small USB peripheral device produced by Leap Motion, Inc. This sensor contains two monochromatic IR cameras and three infrared LEDs. With its smaller observation area, higher resolution and acquisition rate the Leap Motion sensor differentiates from the Kinect. Unlike the Kinect, which is more suitable for whole-body tracking, the Leap Motion is used for acquiring positions, orientations and gestures of human hands and fingers in very high resolution and precision [4].

Eye tracking sensor also emerged as one of the most promising device for HCI and similar applications. During last few years several companies developed eye tracking sensors using different technology, however the devices were still too expensive for general public. In year 2013 Tobii announced a new device Tobii EyeX, a high accurate eye tracking device with affordable price for general population. The true benefit of eye tracking is opportunity to gather data for HCI even for the population with disabilities to use their limbs. The other important feature is a capability to use standard interaction methods, using keyboard and mouse, and gain an additional information to enhance the interaction. Based on those advances, the aim of this paper is to address new capabilities for HCI and new approaches for diagnostics and rehabilitation in the field of medical oculography.

II. EYE TRACKING TECHINQUES

The most popular and also commercially available method to measure the rotation of the eye is "corneal-reflection/pupil-centre" [5]. It is an optical tracking method without direct contact to the eye. This method uses benefits of corneal, which reflects a light directed into the eye. In order to decrease dazzling, infrared light is emitted to the eye. Light enters the retina and a large proportion of it is reflected back, making the pupil appear as a bright, well defined disc known as the "bright pupil" effect. This effect is achieved if the infrared light emitter is placed close to the optical axis of the infrared camera

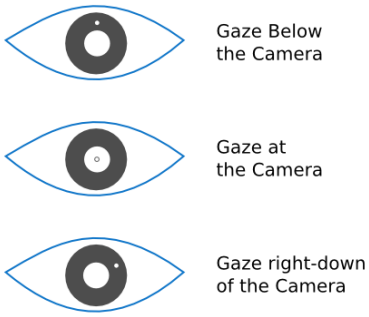


Figure 1. Corneal reflection

device. Setting emitter away from the optical axis will cause the pupil to appear dark. Additional data that can be collected with infrared camera is corneal reflection which is appearing as a small, but sharp glint Figure 1. Based on these two facts and data collected during calibration procedure, a point of regard (gaze point on the surface) or a gaze direction can be calculated using trigonometric calculations.

The complexity of estimation point of regard is depended on number of cameras and light sources. The point of regard can be estimated using one camera and one light source only if the head is completely stationary. Estimation of the point of regard with free head movements is enabled by using one camera and multiple light sources followed with multiple point calibration procedure. Theoretically, it is possible to have one-point calibration procedure to estimate point of regard if multiple cameras and multiple light sources are used. It has been shown that main sources of gaze estimation errors are the discrepancy between the shape of real corneas and the spherical corneal shape assumed in the general theory, and the noise in the estimation of the centres of the pupil and corneal reflections [6].

Limiting factors of the bright pupil method for eye detection are size of the pupil, environmental light and retinal reflection properties. The last one brings most concern and depending on these properties different approach is used, bright pupil or dark pupil technique [7]. This study has found correlation between pupil size and pupil intensity leading to conclusion that subjects with larger pupil sizes are producing higher intensity values. In general, bright pupil method gives higher precision values than dark pupil method. However, it has been shown that dark pupil method provides better results than bright method among ethnic groups, especially for African and Indian population. This bright/dark pupil effect raised the question of how applicable device would be in ethnic diverse country such as South Africa [8].

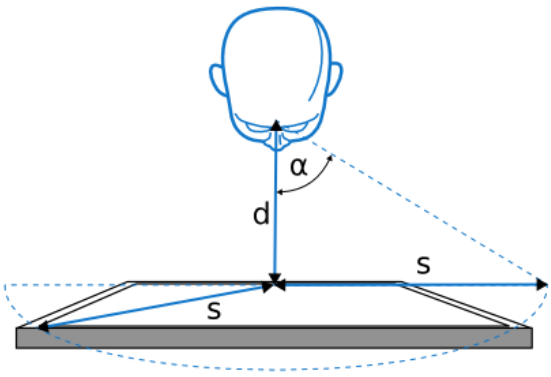


Figure 2. Optimal distance from eye tracker

Before the usage of the eye tracker the calibration procedure is needed to be carried out. During the calibration procedure, the participants are asked to look at the corresponding points also known as the calibration dots displayed on the projected area which is called Active Display Area (monitor, TV screens, projected displays etc.). Number of calibration points and calibration setup affects the accuracy of the results. In general, the more calibration dots, the better the accuracy. Calibration is usually performed on five or nine points. It is very important to have the correct distance to the monitor or the tracked object, otherwise the eye tracker cannot track the entire area. If the monitor or object is placed too close to the participant there is a risk that the gaze data for the corners of the display area of the monitor will not be collected [9]. To calculate optimal distance *d* from eye tracker following formula is used:

$$d = s / \tan(\alpha), \quad (1)$$

where *s* represent the distance from the bottom centre corner of display area to the upper right/left corner, α - optimal gaze angle which value depends on type of eye tracking device (known parameter). During calibration procedure, for each calibration dot an image is taken by the eye tracker, and then analysed to extract parameters for a psychological 3D eye model.

The gaze angle error is expressed as deviation in degrees between two lines where origin is determined by the position of the eye, calculated from images captured by eye tracker. Second point of the first line is determined by the subject's actual gaze point in the active display area, and the second point of the other line is the gaze point measured by the eye tracker.

Another parameter that effects on accuracy and precision of the eye tracker device is environmental illumination [9]. According to Tobii X2-30 eye tracker specification manual it is clearly visible that best accuracy and precision is achieved at illumination of about 300lx what represents the general ambient light levels in offices, classrooms, laboratories and public areas. Illumination above and below this value reduce precision and accuracy of the eye tracker. The main reason for the error is pupil dilation [10].

Constant exposure to the infrared light emitted by the eye tracker has a tiring effect on the eyes which is another concern to observe. The daily use of this technology could be limited if medical research shows harmful effect on the human eyes. Participants wearing glasses or even sunglasses had a better and enhanced bright pupil detection, only problems occurs during reflection from glasses or strong edges of the glasses [8].

III. CHARACTERISTICS FOR EYE TRACKER USAGE

The aim of eye tracking devices is to collect subject's eye activity while they are performing a specific task. The most common types of metrics used in analysis of human behaviour of the eye are fixation and saccades. Fixation are defined as inaction state of the eye above certain period of time. The threshold for this period is time varies typically between 100-200ms. Important attributes that describe fixation are: fixation duration, number of fixations and time to first target fixation. Graphically fixation is presented as circle determined by 2D coordinates with radius defined by duration of fixation. Lines that connect fixation points represent the movement of the eye from one focus to the other and are called saccades. The sequence of fixation and saccades are called a scan path. Collection of scan paths collected for certain period of time represent gaze plot. Generally gaze plots aren't appropriate for analysis for longer duration of time, because graph becomes overwhelmed with plotted data. As common type of visualization for this kind of problem is used heat map. Heat map gives the distribution of density fixation points, which mean heat maps belong to the class of time aggregated diagrams.

Currently eye tracking devices are mostly used in following research fields:

- Psychology studies
- Linguistic studies
- Human-computer interaction research
- Social interaction studies
- Research and studies using simulators
- Copy testing, digital campaigns and TV commercials
- Web and software usability studies of mobile interfaces and devices
- Studies of interactive television, gaming, virtual reality environments and control panels
- Usability studies of physical objects and industrial design
- Real-world interface studies that require a small eye tracker, such as studies of an ATM a ticket machine, or a control pane

IV. NATURAL EYE GAZE COMPUTER INTERACTION

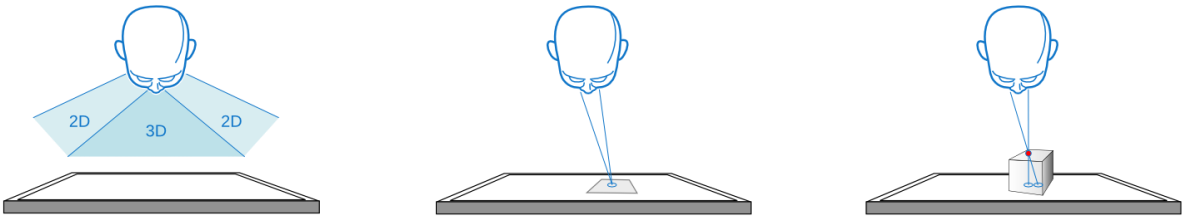
Communication between human and machine goes in two opposite ways - first direction represents visualization of data to be presented by machine to human, the opposite direction goes from human to the machine as commands. To understand how human interact with virtual environment on the computer we need to explain how human perceive a real content in natural environment. Each eye perceive scene from its specific field of view and creates a single 2D image. Left and right eye fields of view intersect in frontal area where two different images are gather for the same scene Figure 3. (a). Using information of left-right eye distance and location of the objects in left/right eye images human brain can calculate depth map (distance of objects in scene).

Interaction with 2D virtual content is enabled by tracking the eye gaze location on the monitor. This location is represent with pixels (x and y coordination of the point) or metrics (mm). Using this information machine can navigate through the content or recognize specific gestures as commands Figure 3. (b).

Unlike the 2D content for interaction with 3D content the depth value is necessary. Depth value can be obtained by extracting two different gaze points of each eye separately Figure 3. With information of 3D position of pupil-centre and 2D gaze point we can calculate gaze vector for each eye. Two vectors should intersect in a single point. This point is a real gaze point and have 3D position Figure 3. (c). This point also represents the focus of the user observation.

The base idea in creation advance HCI is to systematically collect and analyse this eye tracking data to learn behaviours of the user. With this information physical and emotional state of the user can be determined, and intentions detected. This information enables creation of optimal user-specific representation (visualization) of data structures to make interaction with context more natural.

It is important to emphasize that eye tracker cannot be used as primary input device for navigation. It is shown that eye movement input is faster than other current input media [11], [12] but the people are not accustomed to operate only by moving their eyes. Naturally human eye is an input sensor and is not adjusted to trigger actions what leads to conflicts [13]. Usage of the eye tracker for eye-gazed based interaction, problem is in determination of delta time to register input, if this time is too short it leads to Midas-Touch problem and if delta time it too long leads



a) Human perception b) Interaction with 2D virtual content c) Interaction with 3D virtual content

Figure 3. Perception and interaction with 2D and 3D virtual content

to tiredness [14]. Adding the fact how we interpret a perceptually uncertain stimulus [15], additionally raises the question of limited usage of the eye tracker.

V. MEDICAL OCULOGRAPHY DIAGNOSTICS

Taking into account the speed of sampling eye tracker device and its accuracy under optimal condition of working, it can be used in the field of ophthalmology as the instrument in examining cognitive visual function, vision deficiencies and to study oculomotor behaviour of the eyes. It can be used as instrument that replaces and automate current methodologies that relay on observation and as instrument that provide identification of diseases at a much earlier stage of progression. A lot of research has been done in investigating a relationship of disease on the stability of the eye. It has been shown that there is no difference in fixation stability between 12 young observers at mean age of 22 and 12 older observers at mean age of 70 years old [16]. Fixation stability in patients with age-related macular degeneration has been studied using eye tracker system thankfully its higher digital sampling rate rather than the scanning laser ophthalmoscope. Another benefit was that eye tracker allows non-invasive measurements free of artificial support and diagnosis a patient condition in more natural way than that of the scanning laser ophthalmoscope [17]. Saccadic eye movements during tracking, fixation and resting were investigated in schizophrenic and normal participants. It has been shown that the frequency of saccades was higher in schizophrenic's participants than in normal subjects in all of the experimental conditions [18]. Endeavour on replacement the currently used methodologies that relay on observation with automated ones can be seen more and more frequently in the field of ophthalmology. Such example is current research which goal is to investigate the usage of eye tracker as device for detection of strabismus and amblyopia at early stages of development [18], [19]. A feature of high sampling rate in eye tracking can be used to measure the frequency and amplitude of the nystagmus. The investigation of eye pupil dilation under various stimuli attracted a lot of attention among researchers. This phenomena has been connected with many different human emotional and physical states. The most interesting conclusions are:

- pupil dilation can be used as a measure of attention [20],
- visual detection and recognition responses with pupil dilation [21] and
- pupil dilation during visual emotional stimulations can be a measure of affective states of non-communicative individuals [22], [23].

Eye tracker can be used in research as instrument for studying others eyes diseases and also as instrument for diagnostic irregularities in eyes behaviour which would lead to detailed examination of certain eye disorders. Examination the sensitivity of the eye to the light and glare, colour vision changes, peripheral and central vision are some of the irregularities witch can be detected. After surgery of the eye it is usually necessary to conduct a set of eye exercises which is another potential usage of the eye tracker. This would lead to reduction of the load on the medical personnel to focus their importance on emergency cases.

VI. CONLUSION

With new advances in sensor technology a new approaches for HCI can be developed. New controllers should be used in addition with standard HCI input devices (keyboard and mouse). Introduced eye tracking device captures movement of eyes, which is considerably faster than other current input media. However eye tracker cannot be used as primary input device for navigation because people are not accustomed to operate only by moving their eyes and other problems mentioned previously, such as Midas-Touch problem.

Medical usage of eye tracking device is still novel and popular thematic in research. Such technology can approve current methods for oculomotor diagnostics to be carried out automatically. With this advances some eye disorders can be diagnosis in much earlier stages.

REFERENCES

[1] M. Fetaji, S. Loskoska, B. Fetaji, M. Ebibi, "Investigating Human Computer Interaction Issues In Designing Efficient Virtual Learning Environments", Bci 2007, Sofia, Bulgaria, http://halshs.archives-ouvertes.fr/docs/00/19/00/68/PDF/313-324.pdf

[2] A. I. Rudnicky, K. F. Lee, and A G. Hauptmann, "Survey of current speech technology", Communications of the ACM, 37(3): 52-57, 1992

[3] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-Time Human Pose Recognition in Parts from a Single Depth Image", Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2011

[4] F. Weichert, D. Bachmann, B. Rudak, D. Fisseler, "Analysis of the Accuracy and Robustness of the Leap Motion Controller", Sensors, 13, 6380-6393, 2013

[5] H. J. Goldberg, A. M. Wichansky, "Eye tracking in usability evaluation: A practitioner's guide", in J. Hyönä, R. Radach, & H. Deubel (Eds.), "The mind's eye: Cognitive and applied aspects of eye movement research" (pp. 493-516). Amsterdam: Elsevier. 2003

[6] E. D. Guestrin, E. Eizenman, "General theory of remote gaze estimation using the pupil center and corneal reflections", Biomedical Engineering, IEEE Transactions on (Volume:53 , Issue: 6), 2006

[7] K. Nguyen, C. Wagner, D. Koons, M. Flickner, "Differences in the infrared bright pupil responses of human eyes", Eye Tracking Research and Applications Symposium (ETRA 2002), pages 133-138, 2002

[8] R. C. Coetzer, "Development of a robust active infrared-based eye tracking system", University of Pretoria, 2011

[9] Zhu, Z. Fujimura, K., Ji, Q. "Real-time eye detection and tracking under various light conditions", Proceedings of the 2002 symposium on Eye tracking research and applications, pages 139-144, 2002

[10] Tobii Technology AB, "Tobii X2-30 Eye Tracker User's manual", 2014

[11] C. Ware and H. T. Mikaelian, "An evaluation of an eye tracker as a device for computer input", Proceedings of the ACM CHI+GI'87 Human Factors in Computing Systems Conference, (pp. 183-188), 1987

[12] L. E. Sibert and R.J.K. Jacob, "Evaluation of Eye Gaze Interaction", Proceedings of the ACM CHI 2000 Human Factors in Computing Systems Conference, pp. 281-288, 2000, Addison-

Wesley/ACM Press. (Available at http://www.cs.tufts.edu/~jacob/papers/chi00.sibert.pdf [PDF])

[13] S. Zhai, C. Morimoto and S. Ihde, "Manual and gaze input cascaded (MAGIC) pointing", CHI '99. ACM Press, 246-253, 1999

[14] R. J. Jacob, "What you look at is what you get: eyemovement-based interaction techniques", CHI '90, ACM Press, 11-18, 1990

[15] M. O. Hartendorp, S.V. der Stigchel, I. Hooge, J. Mostert, T. de Boer, A. Postma, "The relation between gaze behavior and categorization: Does where we look determine what we see?" J Vis 13, 2013

[16] W. Kosnik, J. Fikre, R. Sekuler, "Visual Fixation Stability in Older Adults", Invest Ophthalmol Vis Sci 27:1720-1725, 1986

[17] C. Bellmann, M. Feely, M. D. Crossland, S. A. Kabanarou, G. S. Rubin, "Fixation Stability Using Central and Pericentral Fixation Targets in Patients with Age-Related Macular Degeneration", Ophthalmology 2004;111:2265–2270, 2004

[18] Y. Matsue, T. Okuma, H. Saito, S. Aneha, T. Ueno, H. Chiba, H. Matsuoka, "Saccadic eye movements in tracking, fixation, and rest in schizophrenic and normal subjects", Biol Psychiatry, 1986 Apr;21(4):382-9.

[19] R. A. Pulido, Ophthalmic Diagnostics Using Eye Tracking Technology, School of Information and Communication Technology (ICT) KTH Royal Institute of Technology Stockholm, Sweden. (Master of Science Thesis) February, 2012

[20] B. Hieks, W. J. M. Levelt, "Pupillary dilation as a measure of attention: A quantitative system analysis", Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands, 1993, 25 (1): 16-2

[21] C. M. Privitera, L. W. Renninger, T. Carney, S. Klein, M. Aguilar, "The pupil dilation response to visual detection", SPIE-IS&T Electronic Imaging, SPIE Vol. 6806, 68060T, 2008

[22] D. Al-Omar, A. Al-Wabil, M. Fawzi, "Using Pupil Size Variation during Visual Emotional Stimulation in Measuring Affective States of Non Communicative Individuals", Universal Access in Human-Computer Interaction. User and Context Diversity Volume 8010:253-258, 2013

[23] B. Lance, S. C. Marsella, "The Relation between Gaze Behavior and the Attribution of Emotion: An Empirical Study", Springer-Verlag Berlin Heidelberg LNAI 5208:1–14, 2008

CTI

International Conference on

TELECOMMUNICATIONS & INFORMATION

Steering Committee

Chairs: Stjepan Golubić, Ericsson Nikola Tesla, Zagreb, Croatia
Branko Mikac, University of Zagreb, Croatia
Vlasta Hudek, Rezultanta, Zagreb, Croatia
Dragan Jevtić, University of Zagreb, Croatia
Davor Šarić, T-Croatian Telecom, Zagreb, Croatia

Members: Željka Car, University of Zagreb, Croatia
Tihana Galinac Grbac, Faculty of Engineering, Rijeka, Croatia
Zdravko Jukić, Croatian Post and Electronic Communications Agency, Croatia
Miroslav Karasek, IREE, Czech Academy of Sciences Prague, Czech Republic
Dragan Knežević, VIPnet, Zagreb, Croatia
Marko Lacković, University of Zagreb, Croatia
Dina Šimunić, University of Zagreb, Croatia
Antonio Teixeira, Institute of Telecommunications, Aveiro, Portugal

INVITED PAPER

Green ICT for Sustainability: A Holistic Approach

Slavisa Aleksic
Vienna University of Technology/Institute of Telecommunications, Vienna, Austria
slavisa.aleksic@tuwien.ac.at

Abstract – In the last three decades, the information and communication technology (ICT) sector has been growing very fast. There is no such example in human history that development of a technology has changed our way of life in such a rapid and fundamental manner. ICT has become an integral part of our everyday life including social interactions, business processes, technology and ecology. Due to the fact that ICT not only promises enormous potentials but also carries risks, it is extremely important to carefully evaluate and assess ICT systems and applications regarding their potential for improving the global energy productivity in a sustainable manner.

This paper concentrates on technologies and methods for achieving high-performance and energy-efficient communication networks. Main achievements and trends towards highly energy-efficient network infrastructure are briefly reviewed. A novel holistic framework able to efficiently treat energy, entropy, and exergy flows in heterogeneous systems through combining thermodynamic and communication approaches is presented.

I. INTRODUCTION

Although information and communication technology (ICT) has been making an enormous progress for more than thirty years and has already influenced many areas of our everyday life, its potential benefits and risks for the environment have gained the interest of the scientific research and broad community only since recently. Even though many extensive studies on energy efficiency of ICT systems have been carried out during the last several years, the influence of ICT on the environment has still not been sufficiently assessed and understood.

On the one hand, broad and intensive use of advanced ICT applications and services promises substantial improvements in many branches such as in industry, logistics, trade, healthcare, and education as well as in society. Furthermore, ICT applications can be used to optimize various processes and, consequently, to support new strategies and mechanisms for a sustainable exploitation of natural resources.

On the other hand, the ever increasing number of ICT equipment and intensive usage of ICT services lead to a continuous increase of ICT-related energy consumption. Additionally, the short lifetime of devices and services cause an increased usage of resources, production intensification, and more hazardous e-waste, which can harm the environment.

The aim of this paper is to briefly summarize recent progress in analyzing communication networks from the energy efficiency perspective and to discuss possibilities of developing and applying a holistic approach in assessing and studying environmental sustainability of different network concepts and ICT applications.

The rest of the paper is structured as follows. The next section summarizes the main trends and projections regarding communication networks, data centers and high-performance computers with particular emphasis on energy consumption. In Section III, we address the methods proposed so far to reduce energy consumption and increase energy efficiency of network infrastructures. Section IV discusses different metrics for energy efficiency. A holistic approach for analyzing sustainability of networks is briefly presented and discussed in section V. Finally, Section VI summarizes and concludes the paper.

II. RECENT TRENDS IN COMMUNICATION NETWORKS WITH PARTICULAR EMPHASIS ON ENERGY EFFICIENCY

The significance of considering energy consumption when developing new technologies and concepts for future networks becomes evident when observing the current situation and projections for the next five to ten years. Already in 2007, the ICT carbon footprint has been estimated to count for about 2 % of the total carbon emissions worldwide, which is comparable to the carbon footprint of the aviation sector [1]. At that time, the estimated ICT-related electricity consumption was above 17 TWh, which corresponds to the production volume of 15 medium-scale nuclear power plants. In 2020, the contribution of the ICT sector to the global carbon emissions is expected to increase to above 4 % [2]. It is worth noting that about 51 % of the ICT-related energy consumption is used by telecommunication infrastructure and data centers. However, it has been estimated that ICT-enabled solutions offer the potential to reduce overall annual emissions by 14.2 % of the projected total in 2020 [2]. According to a more recent estimation, this potential has been revised to 16.5 % [3]. Hence, both the ICT-related energy consumption and the abatement potential grow with time and it is not yet clear if ICT will rather become an important environment pollution source or an enabler for a more efficient and ecological resource management.

A. Increasing Bandwidth Demand

The continuous growth of Internet traffic has led to rapidly increasing capacity provided by network

infrastructure. Observations have shown that the main contributor to the traffic increase in the Internet is the traffic from residential customers [4]. This is mostly because of introduction of new bandwidth-hungry applications for residential users but also due to the fast growth of the number of broadband subscribers [5] as it is evident from Fig. 1.

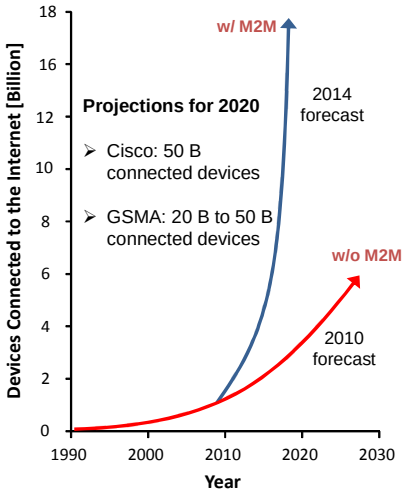


Figure 1. Recent trends in number of devices connected to the Internet [6]. M2M: Machine-to-Machine Communication.

Due to the concurrent growth of Internet traffic and the number of subscribers, both the number of network elements and their capacity are also expected to increase in the future. Recent introduction and wide penetration of smartphones and tablets confirm this trend. Furthermore, according to the vision of the "Internet of Things", it is expected that in the foreseeable future, a huge number of smart autonomous devices will communicate via the Internet, which is referred to as Machine-to-Machine (M2M) communication. As indicated in Fig. 1, it is projected that 50 Billion of smart autonomous devices will be connected to the Internet in 2020 [4]. For this reason, energy efficiency of both network and end-user equipment has become a very important issue.

Historical development and some projections for future trends concerning the growth of Internet traffic and the increase in router capacity and line data rate are presented in Fig. 2. In the past, user traffic was increasing by about 100 % per year. Within the last few years, this growth has slowed somewhat, so that the traffic volume in global communication networks is currently increasing by approximately 50 % per year [7]. In order to keep track with this increasing demand for bandwidth, the capacity of underlying network components has to increase too. However, when considering IP routers, the capacity per rack of equipment has been increasing with a lower rate than Internet traffic. As can be seen from Fig. 2, the capacity per rack was increasing by about 75 % per year for more than ten years. The increase in capacity has slowed in the last few years, which is mostly due to increased energy density and limited cooling rate. The increase of port (line) data rate has been also slowed down to some extent in the last several years. While in the past line data rate increased by a factor of four each three years or by a factor of ten per approximately six years, the step from 40 Gbit/s to 100 Gbit/s, which corresponds to an increase by a factor of 2.5, took more than three years. In order to keep the same increasing rate as during the last decade, the next generation optical transceivers operating at the line data rate of 1 Tbit/s should be already available.

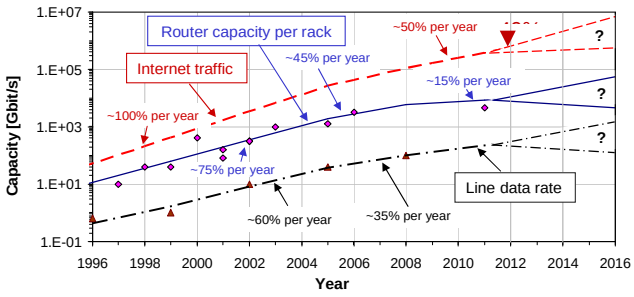


Figure 2. Recent trends in Internet traffic and router capacity [8].

Fig. 3a shows the estimated total power consumption of electronic routers for capacities up to 100 Tbit/s with some examples of electronic high-performance routers represented by the blue dots in Fig. 3a [9]. According to this figure, the router's power consumption increases linearly with bandwidth and each new generation of high-capacity electronic routers consumes more power than the previous one.

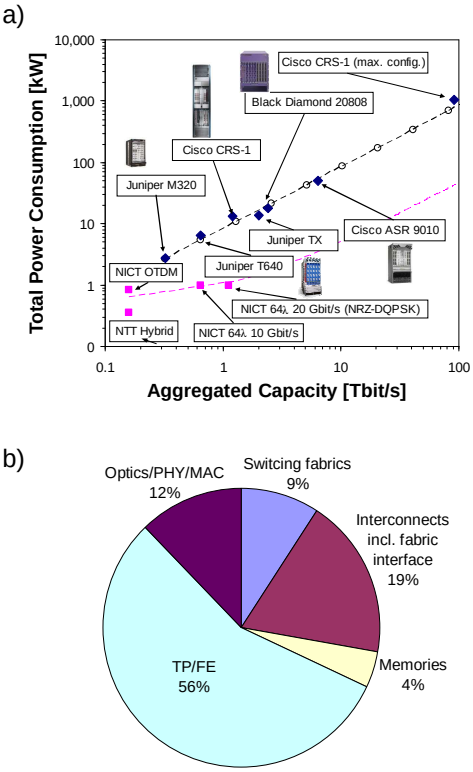


Figure 3. Power consumption of switches and routers [9]. MAC: Medium Access Control, PHY: Physical Layer Chip, TP: Traffic Processing, FE: Forwarding Engine.

About 60 % of the total power consumed by a router is related to IP traffic processing (see Fig. 3b – TP/FE and

Memories). Thus, the most potential for reducing the high power consumption of routers lies in simplifying the network concept in order to reduce the need for the complex and power hungry switching and processing at the packet level. A few examples of optically switched routers are also presented in Fig. 3a by the magenta dots and the corresponding dashed line. It should be noted here that data processing and storage elements realized using current optical technology usually have a quite limited complexity. Consequently, achievable performance of the optically switched routers is not as high as that of their electronic counterparts. Accordingly, the power consumption of an optical router is considerably lower than that of an electronic router assuming the same capacity.

B. Data Centers and Cloud Computing

Data center traffic has been showing an exponential increase mainly due to the introduction of emerging cloud computing applications. It has been estimated that for every byte of user data transmitted over the Internet, 1 GByte are transmitted within or between data centers [10]. While the amount of data-center traffic crossing the Internet is projected to reach 1.3 zettabytes per year in 2016, the amount of traffic within data centers has already reached 1.8 zettabytes per year [11], and by 2016 will grow by a factor of 3.6 to about 6.6 zettabytes per year. This corresponds to a compound annual growth rate (CAGR) of 31 % from 2011 to 2016. The key driver to this growth is cloud computing traffic that is expected to increase six-fold by 2016, becoming nearly two-thirds of the total data center traffic.

Another issue rising with the increase in the data center traffic is energy consumption. The direct electricity used by data centers has shown a rapid increase in the last years. The worldwide electricity use in data centers doubled from 2000 to 2005 and, from 2005 to 2010, it increased by about 56 % [12,13]. It has been estimated that data centers accounted for 1.3 % of the worldwide electricity use in 2010, being an important contributor to the worldwide energy consumption.

C. High-Performance Computing

Similar trends can be observed for large data centers and high-performance computers. Some examples of recently developed systems and projections for power consumption of high-performance computers (HPCs) are shown in Fig. 4. The red dots represent a selection of currently existing HPC systems, of which two illustrative examples are named. It is evident form Fig. 4b that already today large HPC systems consume several MW of electricity [14]. Future Exascale computer systems will probably consume more than 20 MW, which will set very high requirements on power supply and cooling systems. Therefore, a large effort has to be put into research and development of more energy efficient structures and technologies in order to make possible further scaling in both capacity and performance.

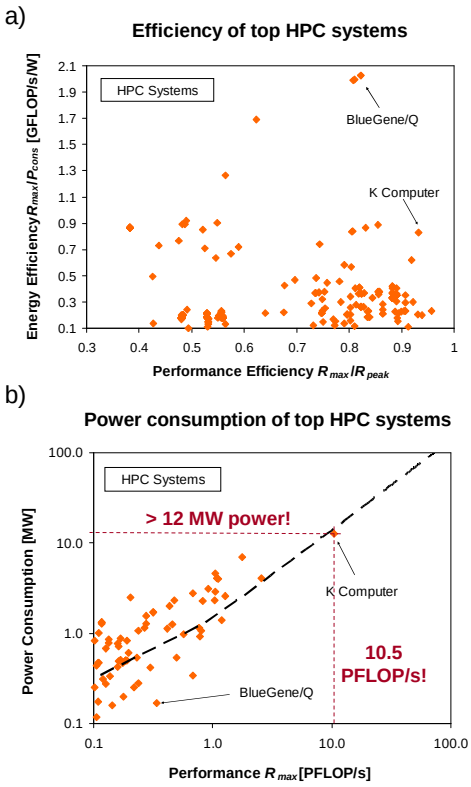


Figure 4. Recent trends in high-performance computing (HPC). Data taken from [14].

III. Methods for Improving Energy Efficiency of Communication Networks

It is widely recognized that future network elements must be able to provide high performance while featuring reduced power consumption. These high requirements can only be satisfied through a combined effort of research community, network designers, and network operators. Such an ambitious goal can be attained by developing and applying energy-efficient data transmission and processing systems, adaptive transmission links, dynamic power management and dynamically switching off or putting in sleep mode transmission links and line cards that are less utilized [15-19]. It has been shown that significant savings in energy consumed by network infrastructure can be achieved when optimizing both the network concept and the architecture of network elements with regard to energy consumption. Energy-efficient switching technologies and paradigms, power-aware routing and wavelength assignment (RWA) algorithms as well as multilevel traffic engineering and cross-layer optimization are examples of methods that can be applied at system level.

IV. Metrics for Energy Efficiency

There have been a number of metrics proposed and used to assess and compare communication systems and networks with respect to achievable reduction in energy consumption and energy efficiency. Some of mostly used metrics are listed in Table I. The most simple and intuitive metric seems to be absolute power consumption. However, absolute power consumption is not a measure

of efficiency because it says nothing about how useful the system actually is. Similarly, power consumption per subscriber, per area or even per bit of capacity does not consider the actual usage of the system. All these metrics presume a utilization of 100 %. Differently, energy consumed per bit of throughput considers the system performance and is therefore suitable for comparison purposes, but it still gives no idea of what would be theoretically possible, i.e., there is no relation to theoretical minimum energy dissipation.

TABLE I. ENERGY EFFICIENCY METRICS

Definition	Unit	Comment
Power consumption	[W]	no relation to usage at all
Power consumption per subscriber	[W/subscriber]	no relation to usage at all
Power consumption per area	[W/km^2]	no relation to usage at all
Power consumption per subscriber and area	[W/subscriber/km^2]	no relation to usage at all
Energy consumption per bit of capacity	[W/bit/sec ≡ J/sec]	no relation to usage at all
Energy consumption per bit of throughput	[W/bit/sec ≡ J/sec]	no relation to theoretical limit
EEF=10log$_{10}$[Energy/bit/($k_B T$ln2)]	[dBε]	inverse to efficiency

The recently proposed metric referred to as energy efficiency figure (EEF) [20] takes into account the theoretical minimum energy required to process one bit of information that can be calculated using the formula $k_B T$ln2 [21], where k_B is the Bolzmann constant and T is the absolute temperature. This theoretical limit can be used to compare heterogeneous systems and to indicate the theoretically possible improvements. However, EEF is, similarly to all other metrics listed in Table I, an inverse efficiency metrics. That means the higher its value the lower energy efficiency.

V. HOLISTIC APPROACH TO SUSTAINABLE NETWORKS

The overall impact of ICT on environment can be classified into direct, indirect and systemic effects [8, 22-24]. The direct effects comprise environmental impacts related directly to the life cycle of ICT infrastructure, including the production, use, recycling and disposal of ICT. The indirect impacts refer to environmental impacts that result from the use of software applications and ICT services that can lead to energy savings in other areas. Finally, the systemic impacts relate to environmental impacts that emerge from the medium or long-term changes of behavior and economic structures following from the availability of ICT applications and services and include all non-technological factors such as adjustment of individual lifestyles and rebound effects.

A study of the direct impacts should preferably include the whole lifecycle of the telecommunication network infrastructure and user devices. However, as briefly addressed in Section III, most of the studies accomplished so far concentrated on reducing the energy consumed during the operational (use) phase only [8,9,15,23,25].

Direct, indirect and systemic impacts can only be considered when applying a holistic approach such as the one illustrated in Fig. 5 [8]. The holistic approach presented here makes use of widely used thermodynamic tools to indicate and model energy and entropy flows within heterogeneous subsystems and through the boundaries between the systems, while taking into consideration peculiarities of each system. In particular, the potential of ICT applications and services can be evaluated and accessed by means of possible improvements of energy efficiency. The influence of the global communication network (the Internet) on the global energy productivity, environment and sustainable development can then be estimated by taking into account direct, indirect and systemic effects.

A. *Model of Communication Networks*

A model of communication networks representing the central part of the global system model as illustrated in Fig. 5 should take into account not only technological aspects and different realization options, but also social, economic, and demographic peculiarities as well as possible future developments. Such a model could

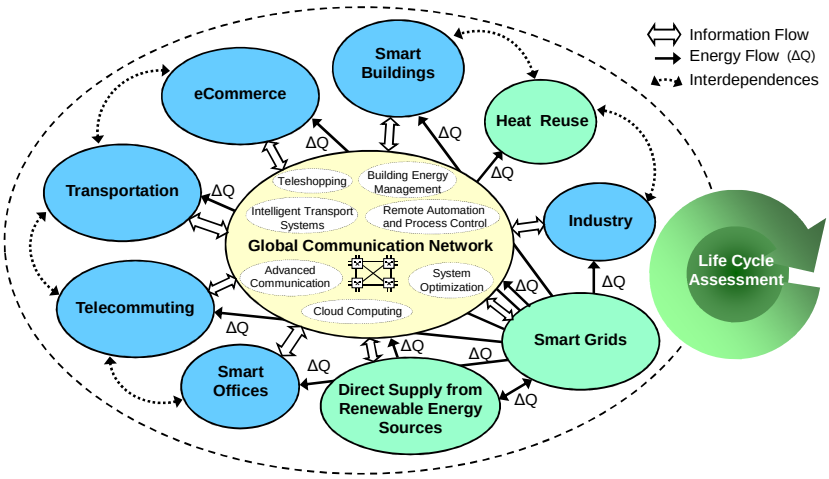
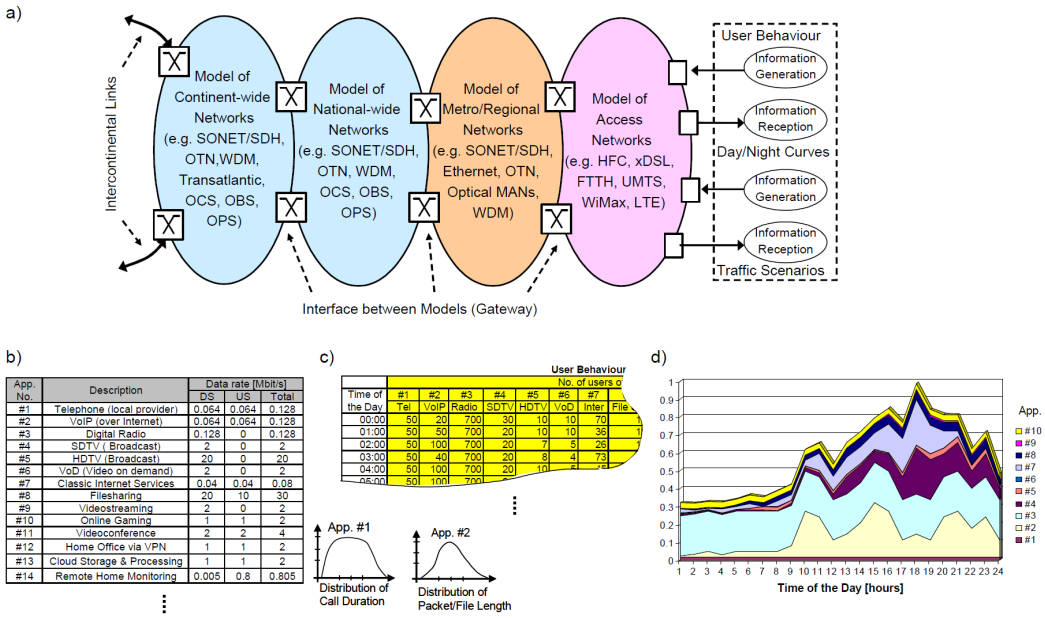


Figure 5. High-level illustration of the global system including the global communication network, applications and other subsystems that can make use of the global network infrastructure in order to improve the global energy productivity. The figure also shows information and energy flows as well as interdependencies within the global system.



App. No.	Description	Data rate [Mbit/s]		
		DS	US	Total
#1	Telephone (local provider)	0.064	0.064	0.128
#2	VoIP (over Internet)	0.064	0.064	0.128
#3	Digital Radio	0.128	0	0.128
#4	SDTV (Broadcast)	2	0	2
#5	HDTV (Broadcast)	20	0	20
#6	VoD (Video on demand)	2	0	2
#7	Classic Internet Services	0.04	0.04	0.08
#8	Filesharing	20	10	30
#9	Videostreaming	2	0	2
#10	Online Gaming	1	1	2
#11	Videoconference	2	2	4
#12	Home Office via VPN	1	1	2
#13	Cloud Storage & Processing	1	1	2
#14	Remote Home Monitoring	0.005	0.8	0.805

Time of the Day	#1 Tel	#2 VoIP	#3 Radio	#4 SDTV	#5 HDTV	#6 VoD	#7 Inter	File
00:00	50	20	700	30	10	10	70	
01:00	50	50	700	20	10	10	36	
02:00	50	100	700	20	7	5	26	
03:00	50	40	700	20	8	4	73	
04:00	50	100	700	20	10	5		

Figure 6. Hierarchical approach to model performance and energy efficiency of the global communication network; b), c) and d) hypothetical examples of applications characteristics and user behavior: b) applications' characteristics, c) user behavior respective application use and c) an example presenting application-specific traffic over the day.

comprise a number of sub-models needed to study different network areas and technologies in a detailed manner, whilst the overall picture can be obtained by joining the individual sub-models together into a global hierarchical model with a manageable complexity. For this purpose, the global communication network can be divided into different areas based on geographical, technological, and operational aspects. For each network area, a different simulation model can be developed, which considers all peculiarities of the particular area such as suitable network technologies, topologies, configuration parameters and realistic traffic scenarios (see Fig. 6).

The sub-models can be connected to each other in a hierarchical manner, where the lowest hierarchy represents the network segment close to the traffic generators (i.e., end users' equipment) and the highest hierarchical level are high-capacity long-haul transmission links that interconnect continental-wide networks (i.e., intercontinental links). Presumably, the most logical way to divide the model of the global communication network into sub-models is to follow the network area principle, namely to model separately local, metropolitan, regional, national-wide, and continent-wide networks. For each of the sub-models we can consider different network concepts and transmission technologies as depicted in Fig. 6a. Statistical data on population, ICT applications and services as well as a model of user behavior can be used to set parameters of traffic generators. An illustrative example of how to define the traffic patterns is shown in Figs. 6b, 6c and 6d, in which the main characteristics of various applications such as average data rate, maximum data rate and appropriate distribution functions as well as average usage of each application over the day are taken into account. In this way, we can study the influence of new applications and include user behavior in order to define different scenarios for possible future developments. The interface between two sub-models, i.e., between two hierarchical levels, is defined by the amount and characteristics of the aggregated traffic sent/received by the gateway interconnecting these two network areas. The following steps can be made when developing a sub-model:

- First, network nodes are modeled by taking into account technological and architectural aspects.
- Then, the network sub-models are developed under consideration of different concepts, topologies, communication protocols, and traffic scenarios.
- The total power consumption of a sub-model is calculated using realistic values for each functional block and according to the actual utilization of network elements.
- Finally, the flow of information and energy through the entire global communication network can be identified and evaluated by defining and implementing the interfaces between the sub-models.

B. Exergy-based Life Cycle Assessment

In order to build a holistic framework, the model of communication networks briefly presented in the previous section can be extended by integrating some widely applicable tools of thermodynamics. A very useful quantity that stems from the second law of thermodynamics is exergy. It can be used to clearly indicate inefficiencies of a process by locating the degradation of energy. In its essence, exergy is the energy that is available to be used, i.e., the portion of energy that can be converted into useful work. Exergy analysis has been performed in industrial ecology to indicate the potentials for improving resource utilization and minimizing environmental impacts. The higher exergy efficiency is, i.e., the lower exergy losses, the better the sustainability of the considered system or approach.

Life cycle assessment (LCA) is a technique to evaluate environmental impacts of a product or a process over its entire life cycle (i.e., raw material extraction, manufacture, transportation, operational, maintenance, and recycling phases). LCA analysis provides a system perspective across the whole life time to define the system boundaries and aid decision making for system optimization and product selection. The approach that combines the exergy concept with the LCA analysis is referred to as exergy-based LCA (E-LCA). Since exergy-based analysis is a universally applicable method to assess process efficiency, it is well suited to investigate the sustainability of heterogeneous systems [27-29]. Recently, E-LCA has also been used to assess the sustainability of ICT infrastructure and applications [30-33]. The preliminary results of these studies have indicated the importance of a sustainable technological development able to ensure both high rate of innovations and increased service lifetime of ICT devices. Even though a fast technological development can lead to both the high performance and high energy efficiency of ICT equipment, the sustainability of the overall system is strongly influenced by short service lifetime of user devices, which causes a considerable increase in the total embodied exergy and thus high environmental impacts.

VI. CONCLUSIONS

Recent and future trends in the information and communication technology (ICT) sector and in particular in technologies and applications for communication networks indicate an urgent need for a significant improvement in energy efficiency. Since ICT has become an integrated part of our business and everyday life, the complex interdependences between ICT on one side and various areas of business and society that make use of ICT on the other can only be appropriately treated using a holistic approach. Such a holistic approach for analyzing heterogeneous systems has been briefly presented in this paper. It combines thermodynamic and communication approaches in a life cycle framework in order to assess the environmental sustainability of ICT technologies as a part of the global system.

REFERENCES

[1] The Central Intelligence Agency „The World Factbook!", Available online: https://www.cia.gov/library/publications/the-world-factbook.

[2] The Climate Group, „GeSI report - Smart 2020: Enabling the low carbon economy in the information age", 2008.

[3] The Climate Group, „GeSI SMARTer 2020: The Role of ICT in Driving a Sustainable Future", December 2012.

[4] Cisco Visual Networking Index, „Global Mobile Data Traffic Forecast Update, 2013–2018", 2014.

[5] D. Evans, „The Internet of Everything", Cisco IBSG, 2012.

[6] S. Aleksic, G. Franzl, Th. Bogner, O. Mair am Tinkhof, „Energy-Efficient Internet Access", 2nd RESPONDER Workshop on Green ICT for Sustainable Consumption, Vienna, January, 2013.

[7] C. Lange D. Kosiankowski, C. Gerlach, F.-J. Westphal, A. Gladisch, „Energy consumption of telecommunicationnetworks", In ECOC 2009, Vienna, Austria, Sept. 20–24. Paper 5.5.3.

[8] S. Aleksic, „Energy-Efficient Communication Networks for Improved Global Energy Productivity", Telecommunication Systems, Vol. 54, No. 2 , 2013, pp. 183-199.

[9] S. Aleksic, „Analysis of Power Consumption in Future High-Capacity Network Nodes", IEEE/OSA Journal of Optical Communications and Networking, 2009.

[10] G. Astfalk, „Why optical data communications and why now? " Applied Physics A, Vol. 95, 2009, pp. 933-940.

[11] Cisco white paper: '„Cisco Global Cloud Index: Forecast and Methodology, 2011-2016", 2012.

[12] J. Koomey, „Worldwide electricity used in data centers", Environmental Research Letters, Vol. 3, No. 034008, 2008.

[13] J. Koomey, „Growth in Data Center Electricity use 2005 to 2010" The New York Times, Vol. 49, No. 3, 2011 pages 24.

[14] TOP500, Available online: http://www.top500.org.

[15] S. Aleksic, „Energy efficiency of electronic and optical network elements" (invited) IEEE J. Sel. Top. Quantum Electron. 2011, Vol. 17, pp. 296–308.

[16] A. Tzanakaki, K. Katrinis, T. Politi, A. Stavdas, M. Pikavet, P. van Daele, D. Simeonidou, M. J. O' Mahony, S. Aleksic, L. Wosinska, P. Monti, „Dimensioning the future pan-european optical network with energy efficiency considerations" IEEE/OSA JOCN 2011, Vol. 3, pp. 272–280.

[17] S. Aleksic „Power consumption issues in future high-performance switches and routers" (invited), Proc. of ICTON, June 22-26, 2008, Athens, Greece, pp. 194–198.

[18] W. Van Heddeghem, M. Deruyck, B. Puype, B. Lannoo, W. Joseph, D. Colle, L. Martens, P. Demeester, „Power consumption in telecommunication networks: Overview and reduction strategies", IEEE Comm. Mag., 2011, Vol. 49, pp. 62–69.

[19] M. Fiorani, M. Casoni, S. Aleksic, „Performance and power consumption analysis of a hybrid optical core node", IEEE/OSA JOCN, Vol. 3, 2011 pp. 502–513.

[20] M. C. Parker, S. D. Walker, „Roadmapping ICT: An Absolute Energy Efficiency Metric", IEEE/OSA JOCN, Vol. 3(8), 2011.

[21] R. Landauer, „Minimum energy requirements in communication" Science 1996, Vol. 272, pp. 1914 - 1918.

[22] OECF, „Greener and smarter: ICTs, the environment and climate change" Paris 2010.

[23] L. M. Hilty, V. Coroama, M. O. de Eicker, T. F. Ruddy, E. Müller, „The Role of ICT in Energy Consumption and Energy Efficiency", Scientific Survey (WP8), Project ICT-ENSURE: European ICT Environmental Sustainability Research, 2009.

[24] K. Dobernig, I. Røpke, M. Grubbe, and M. Blum, „Green ICT for Growth and Sustainability?", Background Paper to the 1st Multinational Knowledge Brokerage Event, RESPONDER, Vienna, Austria, 30 May – 1 June, 2012,.

[25] D. Simeonidou, M. Pickavet (Editors), D21.3: „Final TP report on Green Optical Networking", Project BONE, EU FP7, ICT, 2011.

[26] S. Aleksic, „Energy and Entropy Flow in Erbium Doped Fiber Amplifiers: A Thermodynamic Approach", IEEE/OSA Journal of Lightwave Communications, Vol. 30, No. 17, June 2012, pp. 1-7.

[27] T. Gutowski, J. Dahmus, A. Thiriez, M. Branham, A. Jones, „A Thermodynamic Characterization of Manufacturing Processes", Proc. of the IEEE International Symposium on Electronics and the Environment, Orlando, FL, USA, 7–10 May, 2007; pp. 1–6.

[28] A. Masini, R. U. Ayres, „An Application of Exergy Accounting to Four Basic Metal Industries", CMER, INSEAD, Fontainebleau, France, 1996, pp. 1–51.

[29] M. A. Rosen, C. A. Bulucea, „Using exergy to understand and improve the efficiency of electrical power technologies", Entropy 2009, Vol. 11, pp. 820–835.

[30] C. R. Hannemann, V. P. Carey, A. J. Shah, „Lifetime Exergy Consumption as a Sustainability Metric for Enterprise Servers", Proc. of ASME Int. Conference on Energy Sustainability, Jacksonville, FL, USA, Aug 10-14, 2008, ASME ES2008-54181.

[31] S. Aleksic, M. Safaei, „Exergy based analysis of radio access networks" (invited), Proceedings of ICEAA - IEEE APWC - EMS 2013, 2013 pp.1–8.

[32] S. Aleksic, „Energy, Entropy and Exergy in Communication Networks", Entropy, Vol. 15, No. 10, 2013, pp. 4484–4503.

[33] W. Scharnhorst, „Life Cycle Assessment of Mobile Telephone Networks with Focus on the End-of-Life Phase", Ph.D. Thesis, Lausanne, EPFL, 2006, pages 182.

PAPERS

TELECOM MARKET REGULATIONS

4G Mobile Internet, Services, Regulation and Mobile Operators in Bosnia and Herzegovina

Dragana Šerval* , Čedomir Marković*, Saša Kovačević**
* Communications Regulatory Agency/Regional sector Banja Luka, Bosnia and Herzegovina
** Communications Regulatory Agency /Monitoring department, Sarajevo Bosnia and Herzegovina
dserval@rak.ba, markovic@rak.ba, skovacevic@rak.ba

Abstract – BH mobile Internet market is mostly presented by 3G data technology. Due to limitations regarding capacity, performances, speed and QoS, new technologies such as IMT-Advanced (4G) have been developed. This article presents an overview and statistics of mobile data usage in different technologies as well as representation of different kind of applications and devices globally and by region underling situation in Bosnia and Herzegovina. Introduction of 4G technology in BH is guided by Telecommunications Policy Framework which is about to be adopted and which foresees issuing of LTE licences in 2015.

I. INTRODUCTION

4G mobile Internet was originated as an answer to users demands for increasing capacity limitations, performances QoS and speed. Technology innovations have been followed by evolution in development of services, applications and devices. Will new 4G supported services represent competition to the classic voice, sms and mms communications offered by licensed telecom operators and will BH market be able to follow introduction of new technologies with the same speed with which they enter EU market?

II. QOS OF UMTS/GSM IN BOSNIA AND HERZEGOVINA

There are three licensed operators in Bosnia and Herzegovina for provision of GSM/UMTS services, and they are: BH Telecom, Mtel and HT Mostar. Comparison of penetration of mobile telephony with EU countries shows that BH with 87.25% is below EU average which was 127% in 2012. [1]

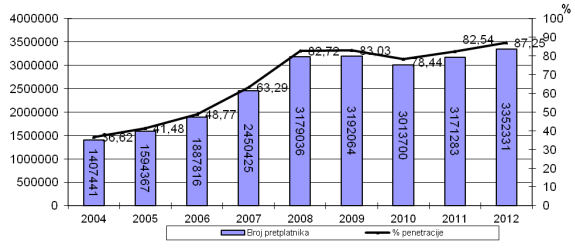


Fig. 1: Development of mobile telephony in BH 2004-2012 regarding penetration and number of subscribers

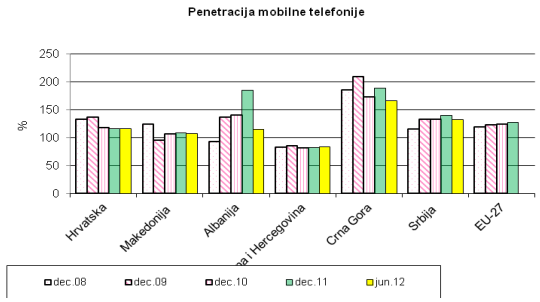


Fig 2: Penetration of mobile telephony, SEE region and EU-27 June 2012.

In regard to Article 8 of GSM License [2] issued by Communications Regulatory Agency (Agency), Decision on monitoring QoS of GSM services provides obligations in regard to methodology and kind of information which are needed for assessment of GSM QoS in BH. Mobile operators in BH are obliged to present data on QoS in line with established methodology and parameters.

Analyzing QoS of GSM/UMTS services, Agency has concluded that parameters of voice services had reached and mostly exceeded targeted values of quality (percentage of service accessibility is 96.8% which exceeds target of 95%, and call competition rate is slightly over 99%, while its target is 97%), in other words BH operators had accomplished to perform satisfactory quality of voice mobile services. [3][4]

In regard to SMS, differences between operators are more visible and values of parameters are between 90 and 99%. Also, parameter end-to-en delivery time shows that some operators need improvement in this segment. Similar is with MMS as well as data service.

Introduction of 4G network in BH market is guided by Telecommunications Policy Framework which is delivered by BH Council of Ministers. The Agency implements the Policy and new 2014-2017 Telecommunications Policy Framework is in the process of adopting. It is expected that the Policy orders issuing LTE licenses in 2015. The Agency has fulfilled its obligation to clear 61-69 channels predicted for LTE technologies. Also, introduction of digital TV is expected in June 2015. It is possible that the Agency will be instructed to make the Study for LTE license introduction in BH.

III. TECHNOLOGY EVOLUTION

In December 2010, ITU accepted only two technologies under name IMT –Advanced and those were: LTE-Advanced and Mobile WIMAX 2.0. Due to limitations in capacity, demands in utilization of radio frequencies, increasing of performances, flexibility, QoS and speed, new technologies are developed so GPRS/EDGE (2.5/2.75G) had evaluated into UMTS (3G), then HSPA (3.5G) into HSPA+ (3.75G). 3G data technology (HSPA) has imposed itself as responsible for the development of mobile Internet broadband access. Comparing data flow from 180kbps at EDGE, HSPA+ promises 42Mbps/22Mbps (down/up). With LTE (3.9G or super 3G) and LTE-Advanced (4G) mobile internet becomes even broader.

ITU statistic for 2013 states that 2.7 billion of people use Internet what is 39% of world's population. In developing countries 31% population is online and in developed 77%. Globally, Europe has the largest Internet penetration (75%) then follows America with 61%. In Africa, 16% of population uses Internet, what is half of Asia and Pacific penetration. [5]

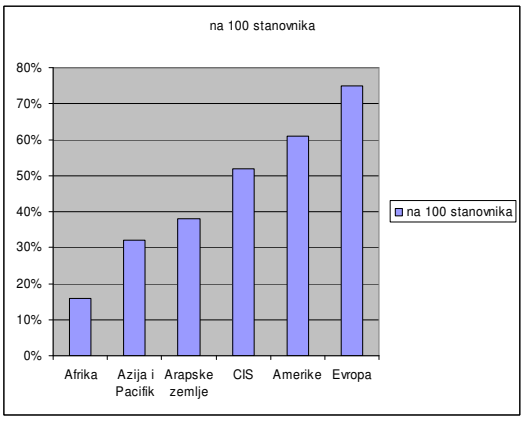


Fig. 3: Statistics by region 2013

When thinking of Internet, we always think of World Wide Web. IPV4 is dominant Internet protocol. Its successor IPV6 was originated as a result of IETF attempts from 1993 when first deficiency of address space was identified [6]. IPV6 forms base for future broadband Internet services. It is expected that new generation networks, based on IPV6 will replace those IPV4 based in order to overcome limitations in addressing, QoS, security and mobility.

Nowadays, content and online status is more important aspect of communications then ever. It is expected that future services and latest technology changes will introduce revolution in business world. Development of services will be guided by demands for new media, social networks, micropayments and real time contents.

IV. 4G MOBILE BROADBAND

ITU considers IMT-Advanced (International Mobile Telecommunications -Advanced), e.g. 4G to be the most advanced technology for global wireless broadband communications. The term was used for advanced technologies such as LTE and WIMAX and other technologies that exceed 3G potential. IMT-Advanced has a possibility to support high quality multimedia applications with huge number of services and platforms as well as improvement in performances and QoS. [7]; [8]. The term "IMT-Advanced" should be related to those systems and system components which support improvements after IMT_2000 [9]; [10].

Mobile broadband is marketing term for wireless Internet access via mobile device which experienced expansion thanks to 3G data technology (HSPA), devices with great possibilities and 2 million applications. 4G broadband Internet now represents the leader for innovations and development of devices, networks, software and technology. The greatest last year's innovations include usage of 4G possibilities, devices (smart phones, tablets) and development of over hundreds of thousands mobile applications. For the last couple of years, mobile Internet has introduced the revolution in device development and once 3G gets replaced with faster 4G, demands for mobile Internet will grow even higher.

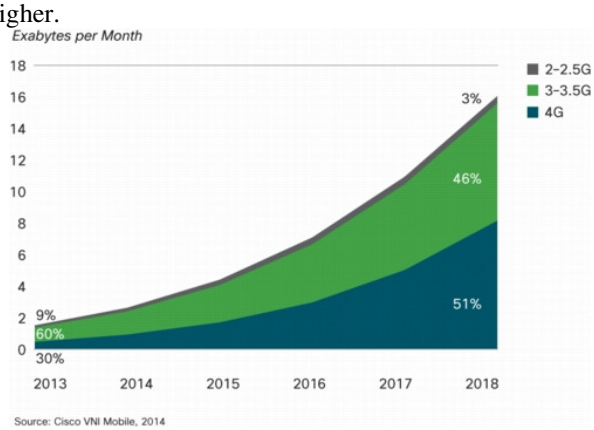


Fig. 4**:** 4G will take 51 % of total mobile Internet in 2018.

By June 2012 more then 5.6 billion consumers used GSM-HSPA, and that is ¾ of world population (7.6 billion). While 4G Internet costs are still higher then fixed, wireline broadband costs of Internet delivery continue to fall. [11]

One operator calculated that in LTE network, costs could grow up to 1€/GByte once penetration reaches 40% and consumption reaches up to 2GByte/month [12]. In Bosnia and Herzegovina that means: small market, huge investments in technology which supports competitive services such as VoIP. Yet, it is expected that during time 4G Internet demands grow up. Cisco data show exponential growth of data traffic compared to voice.

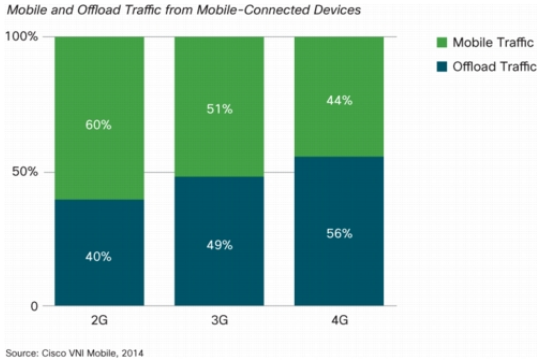


Fig. 5: Mobile data and offload traffic in 2018.

VoIP is a protocol which enables transport of voice data through wireline/wireless network that uses IP address. VoLTE is a signal protocol which enables transport of voice through 4G LTE network. Advantages of VoLTE are following: unlike VoIP, VoLTE use different channels for transport of video and voice what results in better QoS; VoLTE use IP Multimedia Subsystem which enables better synchronization of phonebook, sms and calls and also provides higher speed for data then 3G and VoIP; Average VoLTE download speed is 30 higher then VoIP, for upload 3 higher, and response time is less. Even though VoIP has surprised PSTN it has disadvantages such as noise, errors in transmission and call interruption. These disadvantages are overcome with VOLT which provides greater speed, quality, and reliability without noise and echo.

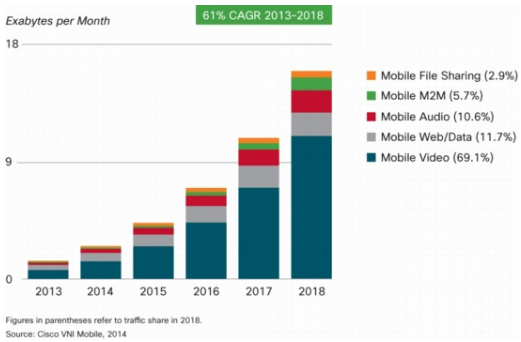


Fig. 6: Mobile video will generate 69% of mobile data traffic in 2018

Mobile Internet revolution was started in 2007 by Apple Iphone. Even though first Iphone supported just EDGE (236kb/s), that is fist phone designed to directly support web. Five generations of Iphone later, users of mobile Internet expect the same speed and experience offered by fixed Internet and desktop with high performances. There is a question: which applications are used more and also, which generate more traffic? Web traffic mostly consists of video (youtube, IPTV) and file sharing (dropbox, rapidshare, mediafire). Other traffic including web surf, VOIP and games is just 20% out of total traffic. [12]

Great deal of 4G network traffic will have multimedia interactive content. Mobile operators all over the world have experienced explosive growth of mobile Internet usage, especially after introducing *flat rate* tariffs and smart devices.

Video applications obviously precede in demands for bigger flow. In order to respond to these demands and support new generation networks, operators turn to 4G technology.

It is interesting to compare usage of PC against smartphones in 3 application categories: video, web surf and social networks. Bigger screen wins in 2 out of 3 categories because social networks users care more about online presence apart from type of device.

According to BEREC, data roaming has increased for 630% in EU from 2007 to 2012. During this period of time, roaming price was reduced and penetration of smartphones got up what explains growth in roaming data traffic. That is not the case in BH which is not in EU and regarding the fact that subscribers has high roaming tariffs, data traffic in roaming is not frequently used. It is also interesting that globally, 4G Internet is mostly used in roaming by Blackberry consumers (65%), Iphone (25%) and Android (10%). Also, data traffic mostly consists of video (30-40%), file sharing at PC, then email and audio with smartphone.

V. EVOLUTION OF SERVICES AND BUSSINESS CASE OF 4G MOBILE INTERNET

Gap between impressive traffic growth and much less growth in incomes has huge effect on mobile operators. From 2010 to 2012, data traffic incomes have risen for 8%, but traffic demands double every year.

In 2012, 4G had generated 19% more traffic then 3G and EDGE and even though 4G represents only 0.9% of mobile communications, it makes 14% of mobile data traffic.

As in fixed broadband Internet market, mobile operators should increase ARPU and find good pricing strategy in order to persuade subscribers in benefits of accepting more expensive 4G service. Comparing to other regions, EU market falls behind in traffic incomes as well as market share. According to some research, globally, US has 33%, China 17%, a EU 28% out of total market value of mobile data services [13].

In regard to amount of traffic, it is estimated that in 2012 EU market will not take even 5th place globally and Asia, North America and Pacific will generate 2/3 out of total traffic.

VI. 4G COVERAGE

By the end of 2012, approximately 96.3% of population was covered by HSPA (3G) networks in EU.

EU coverage with LTE (4G) has tripled in 2012 and took out 26.2% in Scotland and 90% in Portugal. LTE covers mostly urban area [13]. 4G is introduced in FYR of Macedonia, Spain, Ireland, Island and Netherlands. In

Bosnia and Herzegovina, one operator has tested 4G base station in 2013, in Banja Luka area.

VII. CONCLUSION

Great advantage in introducing 4G technology is represented by its flexibility, speed and QoS with negligible delay in transmission and connection. Globally, mobile data traffic of the fourth generation is being used more than the wifi, and customer base of new services based on this new technology is constantly increasing.

It is undeniable that implementation of new technologies are taking big steps forward in Bosnia and Herzegovina. There is no doubt that BH operators will follow the technology progress but market size and economy situation will certainly influence investments in new technologies as well as the speed of implementation of new services. The Agency is ready for implementation of Telecommunications Policy Framework which is about to be adopted and which schedules issuing LTE licenses in 2015.

LITERATURE

[1] Cullen International Report March 2013; http://www.cullen-international.com/asset/?location=/content/assets/research/studies/2011/11/

[2] GSM License, UMTS License; http://rak.ba/bih/index.php?uid=1267325720

[3] Communications Regulatory Agency, 2012 Telecommunications Indicators; http://www.rak.ba/bih/index.php?uid=1268383546

[4] Communications Regulatory Agency, 2012 Evaluation of telecommunication Market; http://www.rak.ba/bih/index.php?uid=1339152712

[5] ITU Fact and Figures in 2013; http://www.itu.int/en/ITU-D/Statistics/Pages/facts/default.aspx

[6] ITU-R 56; http://www.itu.int/dms_pub/itu-r/opb/res/R-RES-R.56-2007-PDF-E.pdf

[7] ITU-R M.2134 "requirements related to technical performance for IMT-Advanced radio interfaces

[8] ITU, Next Generation 4G Mobile Broadband Technologies, 2010; http://www.itu.int/itunews/manager/display.asp?lang=en&year=2008&issue=10&ipage=39&ext=html

[9] ITU-R M.1645; http://www.itu.int/rec/R-REC-M.1645

[10] RFC 2460; https://tools.ietf.org/html/rfc2460

[11] Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update 2010-2015; http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-520862.html

[12] EC services based on EITO 2012; http://epp.eurostat.ec.europa.eu/statistics_explained/index.php/Telecommunication_statistics

[13] http://www.worldtimezone.com/4g.html

Creating Regulatory Environment for Cloud Services

Suada Hadžović, Rešid Šeremet
Communications Regulatory Agency BiH, Mehmeda Spahe 1, 71 000 Sarajevo, Bosnia and Herzegovina
shadzovic@rak.ba

***Abstract* — Cloud Computing is at an early stage in Europe and the various players in the European markets can still influence how the legal framework will be designed. This paper present current initiatives of creating EU regulatory environment in order to support the development of Cloud services which can enable nations both to reap the benefits of Cloud services and mitigate some of the risks associated with them.**

***Keywords* —Regulatory, Cloud Computing, Services**

I. Introduction

While some years ago applications mainly run on servers on their own premises or dedicated data centres, now applications are outsourced to large Cloud service providers and run in a few large data centres. A simple example of Cloud Computing is Yahoo email, Gmail, Hotmail etc.

There is no universal approach for Cloud Computing, rather it is a concept covering many different services, different categories of Cloud (public, private, hybrid), and different types of user (large or small media enterprises, consumers). In Cloud Computing concept users do not see or need to know about the Cloud, they just see services [1].

Many studies indicate that Cloud Computing has substantial economic potential, therefore, EU urges Cloud Computing uptake, but users still have some reservations over security and confidentiality of information in the Cloud.

The EU market for Cloud is some years behind the US market in terms of size and maturity. The most important barriers for the growth of Cloud in Europe relate to concerns by both suppliers and users about compliance with regulatory obligations. They sometimes have to comply with different national law and that such compliance is uncertain and costly [2].

This article is structured as follows. After this introduction, Section II provides an overview of Cloud Computing definitions and types; Section III presents three main actions of EU Cloud Computing strategy; Section IV shows growth potential of Cloud Computing; Section V deals with information regarding deployment while in Section VI short conclusion is drawn.

Suada Hadžović, Regulatorna agencija za komunikacije BiH, Mehmeda Spahe 1, 71000 Sarajevo, BiH (telefon: 387-33-250600, e-mail: shadzovic@rak.ba)

Rešid Šeremet, Regulatorna agencija za komunikacije BiH, Mehmeda Spahe 1, 71000 Sarajevo, BiH (telefon: 387-33-250600, e-mail: rseremet@rak.ba)

II. Cloud Computing definitions and types

The most widely accepted Cloud Computing definition is this one by US National Institute of Standards and Technology (NIST): Cloud Computing is a model for enabling convenient on-demand network access to a shared pool of configurable Computing resources (e.g. networks, servers, storage, applications and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction [3].

Basically the Cloud is the online offering of IT as a service. According to NIST, Cloud Computing is composed of five essential characteristics, three service models based upon service provided, and four deployment models based upon location of Cloud [3].

Cloud Computing essential characteristics by NIST definition:

- On-demand self-service;
- Broad network access;
- Resource pooling;
- Rapid elasticity;
- Measured service.

Classification based upon service provided:

- *Software as a Service (SaaS)*
  Customers use applications (e.g. CRM, ERP, e-mail) from multiple client devices through a Web browser on multi-tenant and shared infrastructures without the need to manage or control the underlying resources
- *Platform as a Service (PaaS)*
  Customers use programming languages, tools and platforms to develop and deploy applications on multi-tenant and shared infrastructures with ability to control deployed applications and environments without the need to manage or control the underlying resources
- *Infrastructure as a Service (IaaS)*
  Customers use processing, storage, networks, other computing resources with ability to rapidly and elastically provision and control resources to deploy and run software and services without the need to manage or control the underlying resources. [4].

Classification based upon location of Cloud:

- *Private Cloud*
  In this Cloud both the consumer of Cloud services and the provider of those services exist within the same enterprise.
- *Public Cloud*
  In this Cloud both the consumer of Cloud services and the provider of those services exist in separate enterprises.
- *Community Cloud*
  This is form of public Cloud known also as a vertical Cloud. It is consist of public Clouds organized around a group of competing/cooperating businesses in a particular vertical market.
- *Hybrid Cloud*
  This Cloud combines multiple elements of public and private Cloud, including any combination of providers and consumers, and may also contain multiple service layers.

The above classification (most common types) is well accepted in the industry. But we can find a more granular classification on the basis of service provided, i.e [5]:

- Storage-as-a-service;
- Database-as-a-service;
- Information-as-a-service;
- Process-as-a-service;
- Application-as-a-service;
- Platform-as-a-service;
- Integration-as-a-service;
- Security-as-a-service;
- Management/Governance-as-a-service;
- Testing-as-a-service;
- Infrastructure-as-a-service.

Additionally, we can find term Everything as a Service (EaaS, Xaas, *aaS) which is a concept of Software components across the network.

Monitoring-as-a-Service (MaaS) has become increasingly popular over the last decade. It is the outsourced provisioning of security, primarily on business platforms that leverages the Internet to conduct business. Since the advent of Cloud Computing, its popularity has grown even more. Security monitoring involves protecting an enterprise or government client from cyber threats. A security team plays a crucial role in securing and maintaining the confidentiality, integrity, and availability of IT assets. The major functionality of MaaS is to monitor the working of all the three layers SaaS, PaaS and IaaS.

Cloud Computing – Primary Characteristics [4]

- On-demand self-service;
- Ubiquitous network access;
- Location independent resource pooling;
- Request driven provisioning and scheduling;
- Measured service / Pay-per-use.

Cloud Computing advantages

- Business revenue and growth;
- Agility and flexibility;
- Cost reduction;
- Innovation;
- Increased Reliability;
- Environmentally Friendly;
- Portability/Accessibility;
- Unlimited storage capacity;
- Improved performance;
- Device independence;
- Improved document format compatibility;
- Easier group collaboration.

Cloud Computing challenges

- Data integrity and security;
- New technology and incompleteness;
- Bandwidth Costs;
- App performance could suffer;
- Data might not be Cloud-worthy;
- Too big to scale.

III. EUROPEAN CLOUD COMPUTING STRATEGY

The EU Cloud Computing strategy, subtitled "Unleashing the Potential of Cloud Computing in Europe" was launched by the European Commission in September 2012. It describes the strategic plan of the European Commission aiming to enable and facilitate the adoption of Cloud services in the public and private Cloud Computing sector across the EU, outlining the most important and urgent actions. Three key actions are [2]:

- *Safe and Fair Contract Terms and Conditions*
  New model contract terms for Cloud Computing should ensure consistency and fairness in contracts for Cloud Computing services across EU. The European Commission places a number of requests on how data is handled and contemplates the model contract terms covering, for example: data preservation after the contract is terminated; data disclosure and integrity; data location and transfer; data ownership; data portability between services; and subcontracting. The model contract terms also will incorporate new mechanisms that will be introduced by the proposed EU data protection regulation.
- *Cutting through the Jungle of Standards*
  The European Commission's stated aim is introducing a new, pan-European certification schemes for Cloud Computing, including data protection. ENISA and other relevant parties are asked to assist in this process. These certification schemes focus on data protection, especially data portability, and increased transparency of Cloud service providers' security practices. Although the Commission has provided a detailed list of factors to be considered by these new certification schemes, it should be noted that participation in the schemes will be voluntary.
- *Establishing a European Cloud Partnership (ECP)*
  ECP brings together industry experts and public sector users to work on common procurement requirements for Cloud Computing in an open and fully transparent way. Part of the ECP is the *Cloud-for-Europe (C4E) initiative,* aim to help Europe's public authorities

procure Cloud products and services, so as to build trust in European Cloud Computing.

Overview of these actions is shown in Fig. 1.

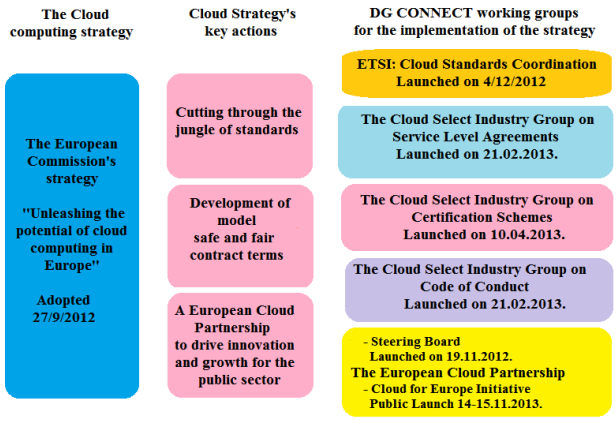


Fig.1. European Cloud Computing Strategy

IV. The Growth Potential of Cloud Computing

The European Commission contracted IDC (provider of data-driven research and analysis) to undertake research on Cloud Computing. IDC Report titled "Quantitative Estimates of the Demand for Cloud Computing in Europe and the Likely Barriers to Up-take" make comparisons between two recognized scenarios. It is estimated that first scenario named "No Intervention scenario" of Cloud adoption could generate up to €88 billion of contribution to the EU GDP in 2020. Instead of that, second scenario named "Policy-driven scenario", could generate up to €250 billion GDP in 2020. Therefore, policy driven scenario bring €162 billion over the no intervention scenario. Additionaly, IDC estimates that cumulative impacts would be even stronger, i.e. for the period 2015-2020 of some €940 billion in the "Policy-driven" scenario, compared to €357 billion in the "No Intervention" scenario [6]. These scenarios comparison is shown in Fig. 2.

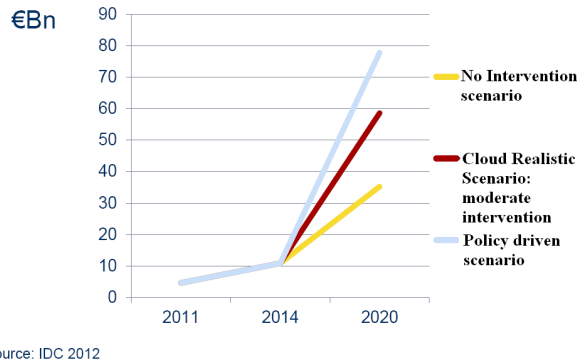


Fig. 2. EU Cloud Computing Market Growth

Considering only the potential of creation of new jobs, IDC estimates that in the "Policy-driven" scenario Cloud-related workers could exceed 3.8 million, against some 1.3 million in the "No Intervention" scenario. However, this does not take into account the jobs that would be lost or the workers displaced by Cloud-related reorganization of business processes and productivity increases.

Estimations about the creation of around 400,000 new SMEs over a period of 5 years and 3.8 million of jobs created indirectly by Cloud indicate that Cloud Computing deserves attention [7].

V. Cloud Computing Deployment

In November 2013, the European Commission has published a study on analysis of Cloud best practices and pilots for the public sector. The study provides a detailed analysis of Cloud initiatives at national level and deployments in the public sector in ten EU Member States. It identifies best practice use cases and proposes pilots for those application areas where current Cloud take-up is absent or however limited.

The study show that in the ten Member States covered by the study, the deployment of Cloud in the public sector is at a very early stage. This overview across the Member States is summarised in Table 1. [8].

	UK	IT	NL	FR	DK	DE	ES	PT	BE	AU
Privacy and security concerns	√	√	√	√	√	√		√	√	√
Regulatory issues	√	√	√		√	√			√	√
Financial issues			√		√					√
Technical issues (i.e. interoperability, techn. maturity, reliability)	√	√	√	√	√		√		√	√
Issues in SLAs (due to the immaturity in practice and market)		√	√				√	√	√	√
Changes in job content, skills and processes	√	√		√	√					√
Cooperation between administrations	√	√			√	√		√		
Cultural barriers and perceptions by users and public sector	√	√		√	√		√	√		√

Table 1. Main Cloud Computing barriers in EU countries

Based on study's analysis of the key stakeholders' needs and requirements, IDC report present recommendations for the most relevant policy actions which should be included in the European Cloud Computing Strategy to create a "Cloud friendly and proactive environment" in the EU and maximize the chances of achieving the benefits identified in the "Policy-driven" scenario.

The policy actions for removing key regulatory barriers should be the following [6]:

- Harmonising and clarifying personal data protection and privacy protection rules across the EU, protecting citizens' rights but simplifying bureaucratic requirements, no matter where the data is and developing additional instruments such as standard clauses, codes of conduct or Binding Corporate Rules for international data transfer.
- Clarifying data jurisdiction regulation and providing EU-wide guidelines about which laws apply to data stored in the EU member state or elsewhere
- Favour the identification and removal of local laws and regulation that limit the use of Cloud services

The European Union Agency for Network and Information Security (ENISA) in its annual report of November 2013 on government Cloud Computing in the

EU also states that the Cloud uptake in EU is not accelerated as expected. ENISA attributes this situation to a several issues that still need to be solved to give a boost to the public sector to adopt [9]:

- Lack of a concrete and unanimous definition of governmental Clouds;
- The Cloud uptake in national level is slow due to lack of national Cloud strategies, privacy and security constraints, lack of legal and contractual frameworks and relative immaturity of the Cloud market;
- Research needs to be promoted since many areas are still "undiscovered".

	Early adopters	Well Informed	Innovators	Hesitants
Political and legal aspects (Cloud strategy)	Cloud strategy exists	On-going implementation of a Cloud strategy	Mainly limited to the Innovation and the digitalisation of the public administration	No Cloud strategy exists
Status (Governmental Cloud)	Mainly implemented governmental Cloud or supporting initiatives	The gov-Cloud is not fully developed (in most cases at design/prototype stage)	Gov-Cloud solutions exists, but very jeopardised and not integrated into a high level national vision	No gov-Cloud exists
Type of Cloud Implementation	Private Cloud (except UK where is public Cloud)	Private Cloud (except Sweden where is community Cloud)	Public/Private Cloud	NA
Motivation for adopting Cloud	Innovation, Competitiveness, efficiency, job growth	Efficiency and cost saving, flexibility	Harmonisation of ICT, innovation in the public administration	Competitiveness and attraction of investments

Table 2. Classification of governmental Clouds in EU

ENISA study has presented the Cloud Computing status of a majority of EU countries. According to the classification of governmental Clouds across the EU, summarised in Table 2, the countries in particular groups are:

- *Early adopters:* UK, Spain and France.
- *Well - Informed*: The Netherlands, Germany, Republic of Moldova, Norway, and Ireland, Finland, and Slovak republic, Belgium, Greece, Sweden and Denmark.
- *Innovators*: Italy, Austria, Slovenia, Portugal and Turkey.
- *Hesitants*: Malta, Romania, Cyprus, and Poland.

General conclusion is that due to great diversity of Cloud solutions and the different approaches, it is difficult to reach a consensus on fundamental aspects of Cloud security such as: common Cloud contract clauses, common Cloud services model etc.

In terms of Table 2 classification, Bosnia and Herzegovina and neighbour countries belong to the *Hesitant*s group of countries which planning to implement governmental Cloud in the future to boost competitive business, but currently have no strategy or Cloud initiatives in place. But, it is sure that to the extent that the Cloud approach relies on access to communications infrastructure, then some regulation under the telecommunications framework may be guaranteed.

VI. Conclusion

Cloud Computing is a developing area and its full strengths and weakness are not yet completely researched, documented and tested. But it is sure that contains a number of issues, not all of which are unique to Cloud, that are concerns for all IT hosted services.

Additionally, despite the fact that Cloud Computing is considered as a key growth area for the ICT sector and as a key enabler of growth in ICT-using industries, it is not a solution for all consumers of IT services.

Currently, more than half of EU businesses and consumers already use some kind of Cloud services, but full adoption of the Cloud model is still far away, hindered by a wide range of bottlenecks and barriers.

Given the potential economic and social benefits of Cloud Computing, the European Commission has an interest in eliminating unnecessary obstacles to the adoption of Cloud services, while safeguarding the public interest in relevant areas such as security, consumer protection, privacy, and others.

Recent studies show clearly the potential benefits for the EU economy and industry of a proactive policy action, able to remove the main barriers to Cloud adoption.

A number of recent EU activities try to enable and facilitate the faster adoption of Cloud Computing in all sectors of the economy, helping to cut costs, boost productivity and promote the creation of growth and jobs.

Generally, the Cloud uptake in national level in EU countries, Bosnia and Herzegovina and neighbour countries is slow due to lack of national Cloud strategies, privacy and security constraints, lack of legal and contractual frameworks and relative immaturity of the Cloud market.

References

[1] Cloud Computing Hearing with Telecommunication and Web Hosting Industry, 16 November 2011, http://ec.europa.eu/information_society/activities/cloudcomputing/docs/hearingreport-telecomsv2.pdf

[2] Unleashing the Potential of Cloud Computing in Europe, September 2012, http://ec.europa.eu/digital-agenda/en/european-Cloud-Computing-strategy

[3] Recommendations of the National Institute of Standards and Technology - Cloud Computing Synopsis and Recommendations, May 2012

[4] 2011 IBM Corporation, Steve Mills: Dispelling the Vapor Around Cloud Computing, 2011 IBM Cloud forum

[5] Mohiuddin Ahmed et all: An Advanced Survey on Cloud Computing and State-of-the-art Research Issues, IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 1, January 2012, http://ijcsi.org/papers/IJCSI-9-1-1-201-207.pdf

[6] IDC 2012: "Quantitative Estimates of the Demand for Cloud Computing in Europe and the Likely Barriers to Up-take", Final report, July 2012 , http://ec.europa.eu/information_society/activities/cloudcomputing/docs/quantitative_estimates.pdf

[7] http://ec.europa.eu/digital-agenda/en/blog/whats-new-eu-Cloud-Computing-strategy

[8] Analysis of Cloud best practices and pilots for the public sector, http://ec.europa.eu/digital-agenda/en/news/analysis-Cloud-best-practices-and-pilots-public-sector

[9] ENISA Report -Good Practice Guide for securely deploying Governmental Clouds, November 2013 www.enisa.europa.eu

Emergency communication in IMS network: challenges in implementation

Ivan Krevatin*, Ivan Visković**
*Service and Platforms Development Department
**Core Network Development Department
Hrvatski Telekom
Zagreb, Croatia
ivan.krevatin@t.ht.hr, ivan.viskovic@t.ht.hr

Abstract—**Emergency communication is an important part of communication for the telecoms not only due to regulation but primary due to their responsibilities to the customers and the role they have in society. This paper describes the challenges which we are facing when implementing emergency communication in IMS network regarding the architecture and special requirements demanded by the Regulator and emergency service centers. In focus are: the architecture for customers accessing the service over fixed access and the possible implementations for customers that are accessing the service over LTE (Long Term Evolution) network.**

Keywords: emergency call handling; location based services; IMS; geolocation;

I. INTRODUCTION

Migration of legacy (PSTN) services to IMS (IP Multimedia Subsystem) domain brings several challenges regarding the implementation and replication of existing services and their behaviour. Emergency communication is one of them. Apart for its vital role that it has in the community and telecom's responsibility for its customers (both users in need and emergency centres), this field is strictly regulated by local laws and often under special observation by the Regulator. We can distinguish four main requirements that we are facing when handling emergency communication in telecom network: (1) localization of the customer, (2) selection of the nearest PSAP (Public Service Answering Point), (3) recognition and prioritization of emergency calls, and (4) alternative routing of calls in case of unavailability of dedicated PSAP. Things get more complex if you consider that PSAP and customers belong to different operators. Migration of PSAPs from PSTN to IMS network generally requires special technical solutions. In this paper we will focus on the IMS network architecture for emergency call handling and analyse the challenges when implementing the solutions for localization of customers and selection of nearest PSAP for both fixed and mobile access. It is clear that since IMS is a convergent network main parts of the solution are independent of the type of access but still the Long Term Evolution (LTE) network comes with new challenges. Voice over LTE (VoLTE) that is based purely on PS (Packet Switched) domain brings an entire new set of technical requirements to the user equipment (UE) and to the network for the support of the emergency call handling. The standardization of this topic is still open and leaves several options in implementation. In this paper we bring this topic to discussion by presenting possible solutions.

The paper is organized as follows. Section II shortly presents the requirements for emergency call handling and organization of Croatian emergency services, Section III describes the IMS network architecture focused on the elements involved in emergency call handling. Section IV describes the specifics of emergency calling from fixed access and possibilities for user localization. In Section V options for implementation of emergency call handling in LTE environment are presented. Section VI summarizes the conclusions of this paper.

II. OVERVIEW OF THE REQUIREMENTS

The existing emergency calls infrastructure has evolved together with the PSTN network following the capabilities of all actors: UE of the user in need, the network and the equipment at the PSAP. The requirements and regulation arose from that capabilities taking into account the importance and sensitivity of the subject. With the evolution of the communication, first by the introduction of mobile phones and now by moving the communication to IP which is not just IP based voice but has evolved to a real-time multimedia communication, the requirements, possibilities and expectations change. An overview of the requirements that are placed on the implementation of the IP based emergency communication can be found in [2] and [3].

Emergency service consists of enabling the user in need to reach the appropriate PSAP that is responsible for serving the area from which the user is calling. The user dials one of the emergency numbers which are dedicated for such service from the national Regulator. In Croatia this is done by HAKOM (Croatian Post and Electronic National Agency). Emergency numbers are assigned per service type for five types of services: fire brigade, medical emergency, police, sea rescue and Croatian automobile club; and one unique number for all types of emergencies (112 service). PSAP organization is mostly based on county principle (one PSAP per county) but there are services with smaller responsibility areas for each of their answering points and also national centres with only one central answering point in the country. The organization of PSAPs is not static but changes in terms of changing the responsibility area and changing the final destination number to

which the emergency call is routed. There are possibilities that PSAP changes its telecom operator but the delimitation of responsibilities and routing principles of emergency calls between operators are still not properly regulated. All this changes have to be promptly and properly handled by the operator(s) whose network is affected.

Croatian regulation regulates emergency communication with several laws which comprise following areas: availability of the service, definition and usage of service numbers, user's location, interconnection requirements, billing of the service and accessibility of the service for impaired persons. All this requirements have to be mapped into technical requirements on the telecommunication platform responsible for providing the service. In addition to this in VoLTE scenario special care must be taken in the years of coexistence of LTE and CS (circuit switched) radio coverage where single radio voice call continuity (SRVCC) is expected to ensure the consistent and seamless service and in the cases when there will be no possibilities to fall back to CS. One of the key factors in these scenarios is the transfer of the user's location information to the network which is responsible for routing the emergency call. IMS brings new challenges and possibilities like emergency registrations, handling emergency communication in roaming, emergency video and text communication, new possibilities for location delivery etc.

III. ARCHITECTURE OF THE SYSTEM

The IMS emergency call handling network architecture is depicted in Figure 1. The IMS network contains elements that are dedicated for handling emergency sessions, they are: Proxy Call Session and Control Function (P-CSCF), Emergency Call Session and Control Function (E-CSCF), Routing Determination Function (RDF), Location Retrieval Function (LRF) and Emergency Access Transfer Function (EATF). These elements are shown in the Figure 1 together with other network elements that form IMS based telecom network architecture. Emergency calls are routed based on the location of the terminal from which the service is accessed. Location is retrieved from the Location Function (LF) element in the access over e2 interface. The method for determination of the location and network element LF are different in the case the user is accessing the service over the fixed or mobile access. As per [3] one of the main architectural principles is that emergency services are independent from the access network with respect to the detection and routing of emergency sessions.

In the fixed access the location retrieval is done in the NASS (Network Attachment Subsystem) with CLF (Connectivity session Location and repository Function) [5] as key element for matching user's IP address to the physical location of the access line. In the mobile access, location of the user vary and location accuracy depends on various factors such as if it is network or terminal provided and which methods are applied. Network elements for location retrieval may be the Policy and Charging Rules Function (PCRF) or Gateway Mobile Location Centre (GMLC). Apart from routing purposes, in some cases (local regulation) it is required to transfer the location information to the PSAP in real time during the emergency session. This can be done over Le interface or in the SIP messages in case of an IP capable PSAP connected directly to IMS network.

The E-CSCF is the IMS network entity, introduced in [3], with the responsibility to route the emergency requests, query the LRF for the proper routing information/PSAP destination and retrieve location information if not included in emergency request. The E-CSCF receives emergency session establishment requests from a P-CSCF. The P-CSCF detects emergency session establishment requests, queries elements of access network for location information, prioritizes emergency sessions and forwards them to E-CSCF. It also handles emergency registrations and performs various security tasks regarding emergency and regular sessions. The LRF and RDF are key elements responsible for proper routing of emergency sessions; LRF may also contain a Location Server (LS). The E-CSCF queries LRF for routing information, information provided by the LRF to the E-CSCF includes the routing information and other parameters necessary for emergency services, which are subject to local regulations. LRF is responsible for delivering location information to PSAP. The EATF [4] is a network entity that provides IMS based mechanisms for enabling service continuity of IMS emergency sessions. It is a function in the serving (visited if roaming) IMS network, providing the procedures for IMS emergency session anchoring and PS to CS Access Transfer in cases of coexistence of LTE and CS.

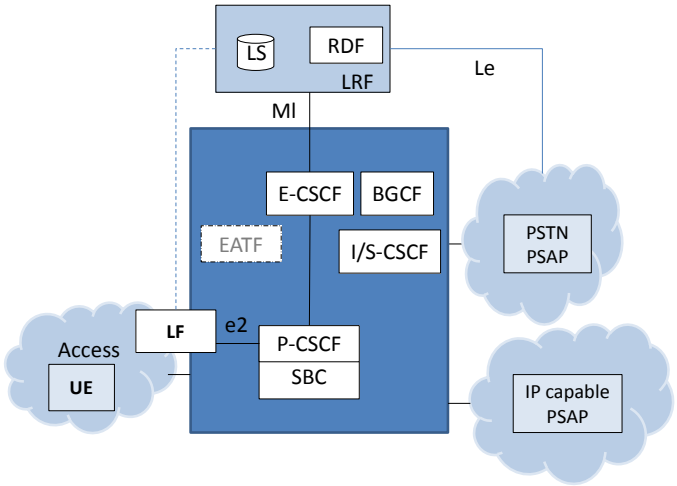


Figure 1. IMS based emergency call handling network architecture

A. *Location information*

As it can be seen from the architecture, location information is a key parameter for routing emergency calls. Standards define a number of mechanisms for providing location information for emergency support [3][6]: (*i*) it can be inserted by UE as the Geolocation header field or P-Access-Network-Info (PANI) header; (*ii*) P-CSCF can query the access network and insert it in the PANI header; (*iii*) it can be allocated as a location reference that relates to the call by LRF which can obtain the location information by querying NASS or GMLC; (*iv*) E-CSCF can query Home Subscriber Server (HSS) and insert it in the PANI header.

Location information can be transferred to PSAP in the Geolocation header field, in PANI header if allowed by configured policies or independent by SIP protocol which is more likely if PSAP is not in the IMS network. Location information is subject to privacy constraints and has to be handled accordingly.

B. Location Retrieval Function

Standards define that Location Retrieval Function (LRF) is the element that selects the PSAP address for the emergency session but don't define the mechanisms that should be implemented to support such selection. In the implementation in Hrvatski Telekom, LRF consists of the RDF and LS with following roles: RDF implements a set of rules for PSAP selection. Basic rule is to match the provided location to the PSAP's responsibility area. This matching is done in geospatial domain by first defining a set of polygons corresponding to PSAP responsibility areas and then querying for the polygon that contains the geo coordinates of the location of the user. LS has the role to provide additional location information to PSAP such as the civic address of the user and any other useful information. This information is delivered over Le interface using calling number as a key with which PSAP queries the LS. LRF itself communicates with E-CSCF over the Ml interface from where it gets the location information containing geo coordinates and reference to civic location, and the short code of the emergency service.

Challenges in implementation of such system consist in: (*i*) defining PSAP responsibility areas with polygons on digital maps; (*ii*) providing fast responses to the spatial queries under stress of the system; (*iii*) assuring proper availability of the system; (*iv*) handling border situations when coordinates of user location are on the border between polygons or outside of all polygons; (*v*) updating user information in LS; (*vi*) applying privacy policy to LS by allowing PSAP to query information only about users that have called emergency service in certain period of time and (*vii*) assuring the flexibility of the system to handle new requirements in incoming VoLTE scenario.

IV. EMERGENCY CALLS IN FIXED ACCESS

Generally, there are four types of users in fixed access that can access emergency services over IMS network: (*i*) user accessing the service over IP-MSAN emulation of POTS; (*ii*) user accessing the service over broadband (BB) access from the operator's network; (*iii*) business user accessing the service over SIP trunk; (*iv*) user accessing the service over public internet. All these users access the service over the SBC whose responsibility is to recognize the type of access an add a location parameter to SIP header i.e. the P-Access-Network-Info (PANI) header [6]. In the case when the user is accessing the service over IP-MSAN, his location is fixed to the location at the user's premises where his phone is installed. The location could be obtained dynamically based on the port on the IP-MSAN but often in implementations more users are hidden behind a single IP address. In such cases it is a real challenge to pass the port ID to the IMS network i.e. in the SIP signalization which often leads to proprietary solutions. Since the location of the IP-MSAN user is fixed, the common way to deal with this issue is to store the location in the HSS which will then be queried by E-CSCF. If this is the case then care should be taken for the situations when the same user is accessing the service over different types of access. When the user is accessing the service over BB access of the operator's network it is common that a dedicated Virtual Local Area Network (VLAN) for Voice over IP (VoIP) traffic is used and that the UE obtains its IP address from the NASS. The link between the IP address, additional network information obtained from Dynamic Host Configuration Protocol (DHCP) Option 82 and physical location information is stored in CLF which is connected to P-CSCF over e2 interface. After the setup of this mechanism it is important to keep this link between logical network and physical (user) location up to date with every new installation or changes in the network. With this mechanism in place, the user can change his location within operator's network (i.e. become nomadic) and the network will update the information about his physical location each time his UE registers for the service. Business trunking users present a real challenge in determination of the location of the user since the location of the terminals could be out of operator's control and the capabilities of terminals to provide their own location are questionable. The main task of the network for such users is to provide proper routing of emergency calls and the precision of the micro location should be agreed with the customer and properly communicated to the PSAP. The location of the users accessing the service over public internet can't be easily determined. It is important to recognize such type of access and pass this information to the PSAP. In some countries PSAP organizations dedicate call center lines handle to such calls. If the user calling from public internet is authenticated and its UE provides location information, such information can be trusted and passed to PSAP.

V. EMERGENCY CALLS IN LTE ENVIRONMENT

The 4th generation of mobile networks LTE (Long Term Evolution) is based on PS domain only and voice services solution for LTE/EPC (Evolved Packet Core) is realized in IMS network. There are two approaches for VoLTE emergency service realization in IMS network: (*i*) CS fallback based solution and (*ii*) IMS network based solution.

A. CS fallback based solution

CS fallback based solution relies on LTE standardized procedures where CS domain on 2G/3G technology is used for voice service execution. The CS fallback procedure is defined to support voice service in early LTE implementation; CS fallback [7].

If CS fallback procedure is used for emergency service in VoLTE, UE needs capability to recognize emergency service request in aim to start the procedure. Otherwise, emergency service request must be recognize by IMS network which then inform UE that CS fallback procedure need to be started. Localization of the customer and selection of the proper PSAP is done in CS domain on the way as it is realized for 2G/3G

access. The solution requires network support of CS fallback procedures and UE support of CS fallback mechanism. The solution based on CS fallback enables fast implantation of emergency service for VoLTE service in networks where CS domain already exists.

B. IMS network based solution.

IMS network based emergency service does not rely on legacy CS domain emergency call solution. However, it is required that emergency call is continuous when user moving from LTE domain to CS domain, in network where CS domain exists. Thus IMS network need to support SRVCC functionality for emergency call with EATF function for call anchoring and transferring [8]. Network support of the functionality is crucial in early LTE implementation if LTE network is not widely spread and there is a limited coverage. IMS network architecture for VoLTE emergency call is shown on Figure 2.

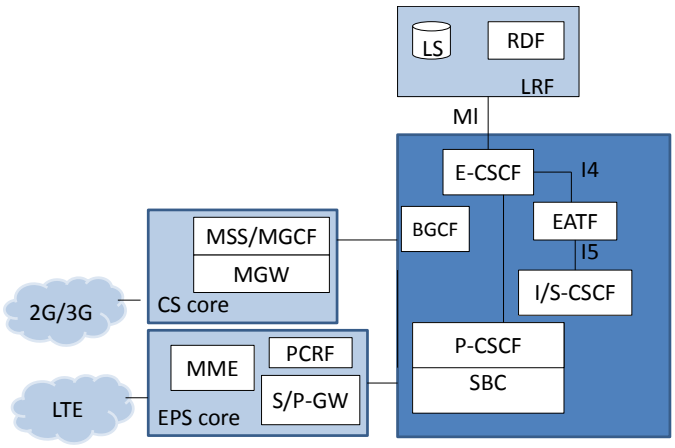


Figure 2. IMS emergency handling network architecture for VoLTE

Due to nature of mobile networks and user mobility, retrieval of proper user location information is a challenge. According to location information, IMS network discover proper destination; PSAP point and route emergency call to the PSAP point. In addition, it could be required by a local regulatory that location information must be transferred to PSAP to be used by that point. Thus, implementation of emergency service in IMS network for mobile users requires solution for user location information retrieval. There are a several type of location information which can be used; CGI (Cell Global identity), CSG Id (Closed Subscriber Group Identity), GI (Geographical Identifier), PLMN Id (Public Land Mobile Network Identifier). The most relevant and for the most network sufficient location information is Cell Id. There are a several solutions defined for location information retrieval [6][9]. In addition, according to network architecture and network elements capabilities various solutions are possible. In general, solution can rely on UE terminal capability or can be based on network capability to retrieve location information.

VI. CONCLUSIONS

In this paper we presented the IMS based emergency network architecture and the challenges we are facing when implementing emergency communication for both fixed and mobile users. Crucial thing is to localize the user to be able to perform proper routing of emergency calls and deliver the user location to PSAP. We have seen that the possibilities for the localization of the user depend on the type of access and type of the service that is provided to the user. IMS network architecture includes dedicated elements for handling emergency sessions, but the service itself has to be designed to conform to local regulations. In transition period where part of PSTN network is migrated to IMS, meaning that PSAPs and users are found in both domains, special care has to be taken in routing emergency calls between networks. Once the migration is finished it will be possible to use full potential of IMS and enable new forms of emergency communication like chat or video sharing. We can expect that this new forms of emergency communication will be included in local regulation as well.

Emergency call in LTE network can rely on legacy CS network, based on CS fallback mechanism. This solution requires small adoption of IMS network and enables fast deployment of emergency service in LTE network. However, IMS network based solution is seen as target for emergency service in LTE network. The solution requires implementation of additional functionalities in IMS network to support voice continuity between IMS network and CS network and to support retrieval of user location information.

REFERENCES

[1] GSM Association, IMS Profile for Voice and SMS, Version 7.0, March, 2013

[2] H. Schulzrinne, R. Marshall, Requirements for Emergency Context Resolution with Internet Technologies, January 2008, http://www.ietf.org/rfc/rfc5012.txt?number=5012

[3] ETSI, Universal Mobile Telecommunications System (UMTS);LTE;IP Multimedia Subsystem (IMS) emergency sessions, TS 123 167 V11.9.0, December, 2013

[4] ETSI, Digital cellular telecommunications system (Phase 2+); Universal Mobile Telecommunications System (UMTS); LTE; IP Multimedia Subsystem (IMS) Service Continuity; Stage 2, TS 123 237, V11.9.0, September, 2013

[5] ETSI, ES 282 004: "Protocols for Advanced Networking (TISPAN); NGN Functional Architecture; Network Attachment Sub-System (NASS)", V3.4.1

[6] ETSI, Digital cellular telecommunications system (Phase 2+);Universal Mobile Telecommunications System (UMTS);LTE;Internet Protocol (IP) multimedia call control protocol based on Session Initiation Protocol (SIP) and Session Description Protocol (SDP); Stage 3; TS 124 229, V12.3.0

[7] ETSI, Digital cellular telecommunications system (Phase 2+);Universal Mobile Telecommunications System (UMTS); LTE; Circuit Switched (CS) fallback in Evolved Packet System (EPS); Stage 2; TS 123 272, V11.7.0

[8] ETSI, Digital cellular telecommunications system (Phase 2+); Universal Mobile Telecommunications System (UMTS); LTE; IP Multimedia Subsystem (IMS) Service Continuity; Stage2; TS 123 237, V11.9.0

[9] 3rd Generation Partnership Project;Technical Specification Group Services and System Aspects; Study on network provided location information to the IP Multimedia Subsystem (IMS); TR 23.842, V11.0.0

Međupovezivanje i regulacija u telekomunikacijama s osvrtom na međupovezivanje i regulaciju u telekomunikacijama Bosne i Hercegovine

mr. Jasmina Gabela, doc. dr. sc. Feriz Gabela, Adnan Gabela, dipl. ecc.
Dioničko društvo BH Telecom Sarajevo
Ulica Obala Kulina broj 8, Sarajevo,
Bosna i Hercegovina
Telefon: 387 – 33 768 296; 387 – 061 507 – 530 E-mail: feriz.gabela@bhtelecom.ba

Sažetak - Sektor telekomunikacija je važan segment razvoja svake države. Kako je tržište telekomunikacija bilo u stanju zakonitog monopola, povećanjem njegove važnosti povećava se regulacija prvobitnog stanja prema stanju konkurencije pri čemu bitnu ulogu imaju regulatorna tijela. U radu je prikazana evolucija regulacije telekomunikacija u Evropskoj Uniju i analizirano međupovezivanja kao ključne mjere telekomunikacione politike. Rad daje opći prikaz regulacije u sektoru telekomunikacija u Bosni i Hercegovini, sa osvrtom na regulaciju međupovezivanja. U BiH na tržištu telekomunikacija, liberalizacija je provedena, no potrebne su daljnje regulatorne akcije nadležnih tijela kako bi se postiglo stanje konkurentskog tržišta po uzoru na razvijene države EU, kojoj BiH teži.

Summary - The telecommunications sector is an important part of development of each country. As the telecommunications market was in a state of legal monopolies, increasing its importance means increasing regulation of the original state to state competition where regulators have an important role. This paper presents the evolution of telecommunications regulation in the European Union and analysis of interconnection as a key measure of telecommunications policy. The paper gives a general overview of regulation in the telecommunications sector in Bosnia and Herzegovina, with emphasis on the regulation of interconnection. In Bosnia and Herzegovina in the telecommunications market, liberalization has been implemented, but there is a need for further regulatory action of the competent authorities in order to achieve a state of competitive markets of developed countries of the EU, which it seeks.

Ključne riječi: telekomunikacijsko tržište, regulacija, liberalizacija

Key words: telecommunication market, regulation, liberalization,

I. UVOD

Evropski sektor telekomunikacija je imao karakter jakog državnog monopola u pružanju usluga. Vlasti su u mnogim državama smatrale da je neophodno da zadrže kontrolu u sektoru telekomunikacija zbog značaja koji telekomunikacije imaju u slučaju rata ili neposredne ratne opasnosti i učešća u bruto društvenom proizvodu. Asocijacije potrošača su zastupale stav da je konkurencija najbolji način da se osiguraju povoljnije cijene usluga i unapredi kvalitet usluga u sektoru telekomunikacija. Evropska komisija je zaključila da je monopol u sektoru telekomunikacija u suprotnosti sa evropskom politikom unutrašnjeg tržišta, te da je davanje posebnih prava za upravljanje mrežama i pružanje usluga kršenje prava konkurencije.

II. EVOLUCIJA REGULATORNOG OKVIRA U EVROPSKOJ UNIJI

A. Regulatorna reforme

Evropska komisija je ključna normativna institucija odgovorna za širenje telekomunikacija u Evropi. Komisija EU je 1987. godine, objavila Zelenu knjigu o razvoju zajedničkog tržišta za telekomunikacione usluge i opremu, sa ciljem da se prilagode zahtjevima jedinstvenog tržišta EU.[1] U ovom dokumentu, Komisija je predložila da se uvede konkurencija na telekomunikacionim tržištima i postigne veći stepen harmonizacije, kako bi se iskoristile prednosti jedinstvenog evropskog tržišta. Akcioni program Komisije EU sadržavao je slijedeće:

- Uspostavljanje evropskog instituta za standardizaciju u oblasti telekomunikacija - ETSI (European Telecommunications Standards Institute);

[1] Towards a dynamic European economy: Green Paper on development of market for telecommunications services.

• Razdvajanje regulatornih i operativnih aktivnosti u državama članicama EU;
• Primjena pravila konkurencije u telekomunikacija.

Nakon objavljivanja Zelene knjige iz 1987. godine, objavljene su i:
• Zelena knjiga o satelitskim komunikacijama iz 1990. godine;
• Zelena knjiga o mobilnim komunikacijama iz 1994. godine; i
• Zelena knjiga o liberalizaciji iz 1995. godine.

Komisija je 1990. godine izdala Direktivu kojom se države članice EU obavezuju da povuku sva prava za pružanje telekomunikacionih usluga, izuzev na govornu uslugu, mobilnu telefoniju i satelitske usluge.[2] U periodu od 1994. do 1996., Komisija je serijom direktiva proširila usluge sa kojih se skidaju ekskluzivna prava i tako obuhvatila satelitske komunikacije, mreže kablovskih i mobilnih operatera i govorne usluge. Regulacija se odnosila na standardizaciju i liberalizaciju radi stvaranja jednakih tržišnih uvjeta za sve države članice EU:

(1) Standardizacija: Glavna je zadaća komisije bila usmjerena na stvaranje jedinstvenog tehničkog standarda u Evropi, što je rezultiralo uvođenjem digitalne GSM tehnologije. To je bio veliki korak naprijed, zato što su postojeći analogni sistemi bili međunacionalno inkompatibilni.

(2) Konkurencija: Druga posljedica regulacije bila je kreiranje zdravog okruženja za razvoj konkurencije. Prijelaz na digitalnu tehnologiju iskorišten je kao prilika za promjenu stanja. Propisano je da moraju postojati najmanje dva različita dobavljača GSM usluga u području 900 MHz, i određeno je da u području 1800 MHz država mora omogućiti frekvencijski pojas za još najmanje jednog operatera.

Tabelarni pregled uvođenja liberalizacije u državama EU prije zvaničnog stupanja na snagu regulatornog okvira dat je u narednoj tabeli.

Tabela 1: Proces liberalizacije u EU prestupanja

Velika B	1981 1985 1991	Liberalizacija u VAN mrežama Konkurencija u analognim uslugama Liberalizacija tržišta fiksnih veza
Švedska	1981 1991	Konkurencija na mobilnom tržištu Konkurencija na tržištu fiksne telefonije
Finska	1988 1993	Konkurencija u prenosu podataka Konkurencija na svim tržištima
Danska	1992 1993 1994 1995 1996	Liberalizacija mobilnog tržišta Konkurencija u prenosu podataka Otvaranje iznajmljenih linija Konkurencija u širokopojasnim uslugama Puna liberalizacija
Njemačka	1989 1993	Odvajanje regulatornih od operativnih funkcija Liberalizacija opreme, VAN-usluga,
Francuska	1989 1992 1999	Konkurencija u analognim mrežama Konkurencija u digitalnim uslugama Konkurencija u uslugama i infrastrukturi
Holandija	1992 1997	Liberalizacija za prenos podataka Puna liberalizacija u sektoru

[2] Commission Directive 90/388/EEC of 28 June, Official Journal L 192 (24 July 1990).

Države članice EU usvojile su 1993. godine Rezoluciju Vijeća (93/C213/01)11[3], kojom se sa 01.01.1998. godine liberalizuju sve govorne usluge, uz tranzicioni period od pet godina za manje razvijene države (Portugal, Grčka, Irska i Španija) i do dvije godine za Luksemburg. Navedeni paket direktiva predstavljao je osnovu koja je korištena za izradu Zakona o komunikacijama u Bosni i Hercegovini, koji je stupio na snagu 2003. godine.

Regulatorni proces u sektoru telekomunikacija u EU sadržan je u Rimskom Ugovoru[4] i Ugovoru iz Mastrihta[5]. Rimski ugovor koji je stupio na snagu 01.07.1987. godine, formirao je jedinstveno, umjesto zajedničkog tržišta u EU. Na osnovu Ugovora iz Mastrihta postavljeni su osnovni ciljevi evropskih integracija koji su podrazumjevali uspostavljanje unutrašnjeg tržišta, stvaranjem sistema koji će spriječiti narušavanje konkurencije i harmonizaciju nacionalnih zakonodavstava država članica, nakon čega je uslijedila liberalizacija tržišta, harmonizacija nacionalnih propisa i primjena prava konkurencije. Tržište telekomunikacija EU otvoreno je 1998. godine, po usvajanju Direktive o otvorenim mrežama[1], koja predstavlja osnovnu direktivu regulatornog okvira EU za telekomunikacije iz 1998. Evropska komisija je u okviru regulatornog paketa iz 1998. donijela i Smjernice[12] za primjenu prava konkurencije u sektoru telekomunikacija i primjeni prava konkurencije prilikom zaključivanja ugovora o pristupu mreži, kako bi se novim pružaocima usluga omogućio pristup do krajnjeg korisnika.

Nakon liberalizacije telekomunikacionog tržišta u EU, novi pružaoci usluga su dobili priliku da konkurišu uvođenjem novih servisa, pružaoci Internet usluga i usluga mobilnih mreža bilježili su porast broja korisnika, kvalitet servisa je poboljšan, a cijene su padale. Ovakav razvoj je omogućio globalnom snižavanju cijena komuniciranja, investiranju u nove servise i mreže, doprinoseći povećanju zaposlenosti.

Komisija je predložila nov regulatorni okvir za elektronske komunikacije koji je predvidio:
1) uklanjanje suvišne regulative,
2) efikasniju konkurenciju,
3) pojednostavljenje administrativnih procedura,
4) zaštitu korisnika.

Predloženim mjerma se rješava:
• Zajednički regulatorni okvir;
• Autorizacija elektronskih komunikacionih usluga;
• Pristup elektronskim komunikacionim mrežama;
• Zaštita privatnostu u elektronskom sektoru i
• Razvezivanje lokalne petlje.

[3] Council, 'Council Resolution of 22 July 1993 on the Review of the Situation in the Telecommunications Sector and the Need for Further Development in that Market,' Official Journal of the European Communities C 213 (6 August 1993).

[4] Uspostavljen jedinstvenim evropskim aktom (SEA), prvom važnom reformom Rimskog ugovora, koji je potpisan u februaru 1986., a stupio na snagu 01.07.1987. godine. Ovim Aktom je trasiran put za formiranje jedinstvenog umjesto zajedničkog tržišta.

[5] Ugovor iz Mastrihta (Mastrihtski Ugovor) o EU, je potpisan u Mastrihtu februara 1992. a stupio na snagu novembra 1993. godine.

B. *Regulatorni okvir iz 2002. godine*

Regulatorni okvir iz 2002. godine stupio je na snagu jula 2003. (osim Direktive o privatnosti i elektronskim komunikacijama koja je na snazi od 31. oktobra 2003.) i sastoji se, pored Uredbe o otvaranju lokalne petlje[14] i Odluke o radiofrekvencijskom spektru[10] i, iz Okvirne direktive[2] kojom su utvrđeni principi zaštite konkurencije i korisnika, i četiri direktive koje se bave različitim regulatornim pitanjima iz oblasti elektronskih komunikacija:
(1) Direktive o izdavanju dozvola[3] koja treba da osigura jednostavnu proceduru izdavanja dozvola za mreže i usluge primjenom režima općeg ovlaštenja;
(2) Direktive o pristupu i međusobnom povezivanju[4] koja treba da omogući pristup drugog operatera infrastrukturi pod definisanim uslovima;
(3) Direktive o univerzalnom servisu[8], kojom se utvrđuje minimalni skup usluga koji treba da bude dostupan svim korisnicima; i
(4) Direktive o zaštiti privatnosti[5].
Od posebnog značaja u pogledu stvaranja povoljnijih uslova za konkurenciju u mrežama za pristup je donošenje odredbe Parlamenta i Vijeća EU o razvezivanju lokalne petlje[6].

C. *Regulatorni okvir za elektronske komunikacije*

Prvi akt koji je Komisija usvojila u postupku revizije regulatornog paketa iz 2002. godine je Preporuka Komisije 2007/879/EC o tržištima podložnim uvođenju ex ante regulative[7], koja je korak u pravcu ukidanja regulative elektronskih komunikacija u pojedinim sektorima. Preporukom je predviđeno uklanjanje 11 tržišta sa liste relevantnih tržišta predviđenih preporukom iz 2003. godine. Osnovni ciljevi revizije regulatornog okvira bili su: jačanje nezavisnosti regulatornih tijela; smanjenje regulatorne intervencije u onim oblastima gde je postignut zadovoljavajući nivo konkurencije; brže procedure za prenosivost broja, bolju zaštitu privatnosti i sigurnosti. Usvojena je i Uredba o romingu[15] kojom su ograničeni iznosi naknada za usluge međunarodnog rominga za govorne pozive koji počinju i završavaju na teritoriji EU. Evropski parlament i Vijeće ministara usvojili su Uredbu o izmjenama i dopunama Uredbe o romingu[16], kojom je utvrđeno da period važenja Uredbe o romingu treba produžiti i nakon 2010. godine, i njenu primjenu treba proširiti na pružanje usluga SMS poruka i prenosa podataka u romingu.
Izmjenjene i dopunjene su direktive EU iz 2002., i to:
- Direktiva 2009/140/EC o izmjenama i dopunama Okvirne direktive 2002/21/EC za elektronske komunikacione mreže i usluge, Direktive 2002/19/EC o pristupu i međupovezivanju elektronskih komunikacionih mreža i Direktive 2002/20/EC o izdavanju odobrenja za mreže i usluge;
- Direktiva 2009/136/EC o izmjenama i dopunama Direktive 2002/22/EC o univerzalnim uslugama i pravima korisnika koja se odnose na elektronske komunikacione mreže i usluge, Direktive 2002/58/EC o zaštiti privatnosti u elektronskih komunikacija i Uredbe 2006/2004 o zaštiti potrošača.

Uredbom (EC) No 1211/2009, osniva se evropski regulator za elektronske komunikacije BEREC[17] (Body of European Regulators for Electronic Communications), koji ima obavezu da osigura fer konkurenciju na telekomunikacionom tržištu EU. Novo regulatorno tijelo se formira od šefova 28 nacionalnih regulatornih tijela EU i predstavnika Komisije i država kandidata za ulazak u EU (Turska i Makedonija, bez prava glasanja). Ovo tijelo okuplja predstavnike regulatornih tijela EU i pruža okvir za razmjenu iskustava među regulatorima i saradnju sa Evropskom komisijom, u cilju stvaranja jedinstvenog tržišta elektronskih komunikacija. BEREC će donositi odluke većinom glasova, a u slučajevima sporova između nacionalnih administracija, dvotrećinskom većinom. Komisija je dobila ovlaštenje da u saradnji sa BEREC-om, donosi obavezujuće odluke u situacijama kada nacionalno regulatorno tijelo kreira prepreke funkcioniranju jedinstvenog tržišta.

Regulatorni okvir iz 2009. godine uspostavlja regulatorno okruženje koje osigurava razvoj tržišta za telekomunikacione mreže i usluge u korist građana EU, te pruža visok nivo pravne sigurnosti za poticanje ulaganja u sektor i pruža dovoljnu fleksibilnost da odgovori zahtjevima dinamičnog razvoja tržišta i inovativnih tehnologija. Revidirani regulatorni okvir za elektronske komunikacije EU je implementirala 2011. godine.

D. *Međupovezivanje*

Međupovezivanje je postalo ključni instrument telekomunikacione politike sa porastom različitih mreža koje su morale da budu inter-operabilne. Pojam međupovezivanja se prema IUT definiše kao: „Komercijalni i tehnički aranžmani po kojima provajderi povezuju svoju opremu, mreže i usluge da bi omogućili korisnicima da imaju pristup mrežama korisnicima i uslugama drugih provajdera“.[9]

Postoji: paralelno, vertikalno i horizontalno međupovezivanje. Paralelno međupovezivanje je aranžman po kojem se dominantni operateri povezuju sa operaterima koji su dominantni u drugim regionima. Njihov odnos je partnerski, zajednički šire mrežu za svoje korisnike i utvrđuju zajednički cijenu. Drugi aranžman se javlja između provajdera koji nudi međumjesne telefonske razgovore i lokalnog operatera. Horizontalno međupovezivanje je aranžman gde se konkurenti za ista tržišta telekomunikacija i korisnici povezuju. Odnosi uključuju veze unutar i između entiteta i industrija:
- Postojeće i nove lokalne telefonske kompanije;

[6] Regulation (EC) No 2887/2000 of the European Parliament and of the Council of 18 December 2000, OJ L336/4, 30.12.2000.
[7] Commission recommendation of 17 December 2007 on relevant product and service markets within the electronic communications sector susceptible to ex ante regulation in accordance with Directive 2002/21/EC of the European Parliament and of the Council on a common regulatory framework for electronic communications networks and services.

- Domaći i međunarodni operateri;
- Privatne mreže;
- Kompjuterske mreže;
- Satelitska mreža i kablovska televizija.[19]

E. Uspostavljanje jedinstvenog tržišta

Početkom 2013. godine na sednici Evropskog Vijeća predviđeno je da Komisija predstavi mjere za uspostavljanje jedinstvenog tržišta informacionih i telekomunikacionih tehnologija na teritoriji EU. Ovaj prijedlog se nastavlja na glavne direktive iz 2002. godine, izmjenjene i dopunjene 2009. godine. Cilj ovog prijedloga je uspostavljanje i stvaranje jedinstvenog tržišta elektronskih komunikacija sa konačnim ciljem potpune integracije i liberalizacije. Jedinstveno tržište elektronskih komunikacija podrazumjeva da[15]:

− privredna društva koja osiguraju elektronske komunikacione mreže i usluge mogu vršiti djelatnost i pružati usluge bez obzira na mjesto nastanka ili mjesto u EU na kojem se nalaze njihovi korisnici,

− privredna društva i građani elektronskim komunikacionim uslugama mogu pristupiti sa bilo kojeg mjesta u EU, bez ograničenja ili neopravdanih dodatnih troškova.

Uspostavljanjem jedinstvenog tržišta elektronskih komunikacija podsticalo bi se ulaganje, inovacije i nadmetanje na području mreža i usluga, kao i nadmetanja u pogledu cijena. Ove mjere trebale bi promovisati ciljeve utvrđene u Digitalnoj agendi za Evropu i stvoriti jedinstveno digitalno tržište. Ovaj predlog ima za cilj da uspostavi ravnopravne uslove za sve učesnike na tržištu elektronskih komunikacija u EU kako bi se rizik ulaganja u sektoru sveo na što manju mjeru. Izazovi pri stvaranju jedinstvenog tržišta ogledaju se u:

− harmonizaciji uslova za pristup: davanjem garancija mobilnim operaterima koje se odnose na uslove dodjele i usklađenosti rokova za pristup spektru za bežični širokopojasni pristup u EU,

− otklanjanje prepreka u sistemu davanja odobrenja koja se primjenjuju na pružanje usluga tako da odobrenje dobijeno u jednoj državi članici važi u svim ostalim državama članicama,

− zaštita korisnika u EU stvaranjem jednakih uslova za sve.

I. REGULACIJA TELEKOMUNIKACIJAMA U BOSNI I HERCEGOVINI

A. Regulacija telekomunikacija

Bosna i Hercegovina je tek 2004. godine počela ozbiljnije ulaziti u sferu informacionih i komunikacionih tehnologija, u čemu inače osjetno zaostaje, ne samo za razvijenim državama Evrope i svijeta, već i za većinom tranzicijskih država.

Područje telekomunikacija u Bosni i Hercegovini uređeno je Zakonom o komunikacijama[8], kojim je uređeno područje telekomunikacija, način i uvjeti obavljanja telekomunikacijskih usluga i djelatnosti, prava i obaveze davatelja i korisnika telekomunikacijskih usluga, izgradnja, održavanje i upotreba telekomunikacijske infrastrukture i opreme, radijske opreme i telekomunikacijske terminalne opreme, upravljanje radiofrekvencijskim spektrom, te obavljanje nadzora i kontrole u telekomunikacijama. Zakon je usklađen sa starim regulatornim okvirom EU (22 direktive iz područja telekomunikacija), koji su države članice EU počele primjenjivati 1998. Postojeća zakonska regulativa ne prati novosti na telekomunikacijskom ržištu, niti prati nove izazove koje ono donosi.

U procesu usvajanja i provođenja EU propisa, a u cilju postizanja evropske perspektive za BiH, neophodno je izvršiti reformu sektora telekomunikacija, primjenjujući zakonske i podzakonske akte koje je propisala Evropska komisija. Primjena regulative se odnosi na usvajanje Zakon o elektronskim komunikacijama kao krovnog dokumenta koji regulira sektor elektronskih komunikacija u BiH, jer postojeći Zakon o komunikacijama se bazira na regulatornom okviru iz 1998. godine. Kako bi se usvojili principi novog regulatornog okvira, a primjenjujući evropske smjernice, odluke i preporuke, u BiH treba donijeti i set podzakonskih akata koji uređuju određenu oblast, a naslanjaju se na novi regulatorni okvir EU.

Novim regulatornim okvirom potrebno je ukinuti prepreke za ulazak novih sudionika na tržište kako bi se, uvođenjem slobodne konkurencije, korisnicima omogućio odabir davatelja usluge. Regulacija telekomunikacijskog tržišta treba da ide ka deregulaciji, odnosno tržište ne smije biti opterećeno mjerama regulacije. Hoće li u budućnosti biti više ili manje regulacije ovisit će prvenstveno o konkurenciji na telekomunikacijskom tržištu, gdje veća konkurencija znači i manje potrebe za regulacijom.

Liberalizirano tržište telekomunikacija omogućava privlačenje kapitala za ulaganje u telekomunikacijske mreže i uvođenje novih usluga, povećanje broja telekomunikacijskih operatora i davatelja usluga koji se međusobno natječu i time pomažu u kvaliteti usluživanja korisnika usluga. Pojavom konkurencije povećava se kvaliteta pruženih telekomunikacijskih usluga te snižava cijena za iste ili slične usluge koje korisnici telekomunikacijskih usluga očekuju.

Regulatorna agencija za komunikacije mora osigurati postizanje konkurentskih uvjeta regulacijom pristupa mrežnim resursima operatera sa značajnom tržišnom snagom što je osnova ostvarivanja konkurencije na svim telekomunikacijskim tržištima. Tržište treba biti otvoreno u segmentu usluga kao i u segmentu infrastrukture te svakako treba biti otvoreno za direktna strana ulaganja.

[8] Zakon o komunikacijama Bosne i Hercegovine, objavljen u „Službenom glasniku BiH“, broj: 31/03, usvojen 02.09.2003. godine.

B. Pristup i međupovezivanje telekomunikacionih mreža

Međupovezivanje je povezivanje javnih telekomunikacionih mreža korišteno od strane privrednih društava, s ciljem omogućavanja korisnicima jednog privrednog društva da komuniciraju s korisnicima istog ili nekog drugog privrednog društva, ili pristupaju uslugama koje pruža drugo privredno društvo. Pružaoci usluga telekomunikacija sa značajnom tržišnom snagom koji pružaju usluge iznajmljivanja linija obavezani su, na tržištu na kojem dominiraju, istaknuti minimalnu ponudu linija za iznajmljivanje s tehničkim karakteristikama u skladu s direktivama EU. Pristup mreži daje se putem priključnih tačaka koje su dostupne na tržištu ili putem priključaka na zahtjev korisnika. Sporazumi o međupovezivanju zasnivaju se na nedskriminaciji i ravnopravnom pristupu telekomunikacionim mrežama. Regulatorna agencija za komunikacije (Agencija) je 2005. godine dala saglasnosti na Referentne interkonekcijske ponude (RIP) za tri dominantna operatora, koji djeluju na tržištu telekomunikacija u Bosni i Hercegovini: BH Telecom, Telekom Srpske i HT Mostar. RIP definira uvjete pod kojima drugi licencirani operatori mogu ostvariti međupovezivanje u mrežama dominantnih operatora u Bosni i Hercegovini. Ponuda sadrži principe međupovezivanja, osnovne tehničke, komercijalne i ugovorne aspekate međupovezivanja. U svrhu osiguranja međusobnog povezivanja, osigurana je zadovoljavajuće komunikacija između krajnjih korisnika. Međupovezivanje predstavlja ključnu odrednicu za potpunu liberalizaciju tržišta telekomunikacija u Bosni i Hercegovini. Referentnom interkonekcijskom ponudom (RIP) omogućava se uvođenje tržišne konkurencije s jednakim mogućnostima ulaska novih operatora na tržište. Od 2006. godine do 2013. godine na tržištu fiksne telefonije u BiH sklopljeno je 33 interkonekcijskih ugovora između postojećih telekoma i novih operatora, dok je trenuto važećih 22. Novi operateri realizuju značajan postotak međunarodnog prometa po nižim cijenama za krajnjeg korisnika.

IV. ZAKLJUČAK

Nakon liberalizacije u telekomunikacijama, koja je započela u SAD-u i Velikoj Britaniji, u protekle tri decenije ulagani su veliki napori u EU, na stvaranju povoljnog zakonodavno-regulatornog okruženja u kome će svoj poslovni interes ostvarivati svi učesnici na tržištu pod ravnopravnim uslovima, a korisnici uživati u prednostima većeg izbora usluga uz povoljnije cijene. Regulatorni okvir elektronskih komunikacija u EU iz 2009. godine, predstavlja osnovu za održivo regulatorno okruženje usmjereno ka rješavanju preostalih regulatornih izazova i osigurava prisustvo konkurencije i podstiče investiranje. Prva od sedam inicijativa strategije "Evropa 2020" je **Digitalna agenda** iz 2010. godine, koja predstavlja Akcioni plan Komisije za jačanje evropskog prosperiteta. Agenda ističe sedam prioritetnih oblasti za akciju:

1) jačanje povjerenja na Internetu,
2) brži pristup Internetu,
3) više investicija u istraživanje i razvoj,
4) postizanje veće interoperabilnosti,
5) unapređenje digitalne pismenosti,
6) primjena informacionih i komunikacionih tehnologija u rješavanju klimatskih promjena i
7) kreiranje jedinstvenog digitalnog tržišta.

Kako bi se na najefikasniji način povećala pravna sigurnost i transparentnost i kako bi se spriječile posledice rascjepkanosti evropskog tržišta elektronskih komunikacija, Evropska komisija je predložila primjenu pomenutih mjera za koje se očekuje da stupe na snagu u 2014. godini. Procjenjuje sa da će u EU nakon primjene ovih mjera, zahvaljujući većem konkurentskom pritisku, stvoriti povoljnije okruženje za nove investicije. Ovaj prijedlog predviđa da se na veleprodajnom tržištu pristupa na nivou cijele EU i da svi proizvodi budu dostupni i standardizovani. Ova mjera će omogućiti operaterima fiksnih mreža da usluge povezivanja koje pružaju svojim korisnicima širom EU budu visokog kvaliteta i jednaki za sve. Ovo će za posledicu imati pojačano tržišno nadmetanje kroz nova ulaganja operatera u infrastrukturu i nove usluge. Jednakim pravilima koja se odnose na kvalitet usluga osigurat će se sloboda korisnika da pristupe uslugama po svom izboru na osnovu uslova ugovora u čitavoj EU, a da se njihov pristup internetu ne ograničava.

Proces liberalizacije telekomunikacionog tržišta u Bosni i Hercegovini započet je prije 10 godina. Primjenjujuci Zakon o komunikacijama, kao i pažljivo sagledavajuci iskustava drugih regulatora iz razvijenih država i prateci trendove razvoja novih tehnologija, RAK je osigurao pravni okvir za liberalizaciju i otvaranje telekomunikacionog tržišta. Danas je to ostvareno u oblasti mobilne i fiksne telefonije, Internet usluga i širokopojasnog pristupa.

U narednom periodu potrebno je prilagoditi regulativu iz oblasti telekomunikacija propisima EU, a prije svega usvojiti novi zakon o elektronskim komunikacijama usklađen sa novim regulatornim okvirom EU iz 2009. godine. Gotovo sve države okruženja su već usvojile ovaj zakon, npr. Zakon o elektroničkim komunikacijama Republike Hrvatske (stupio je na snagu 1. jula 2008. godine), Zakon o elektronskim komunikacijama Republike Srbije (usvojen maja 2010. godine) i Zakon o elektronskim komunikacijama Crne Gore (usvojen 13.08.2013. godine). U cilju povećanja stepena konkurencije sektora telekomunikacija neophodno je nastaviti aktivnosti na analizi tržišta u skladu sa Preporukom Evropske komisije o relevantnim tržištima i važećim EU regulatornim okvirom, te Pravilom o analizi tržišta elektronskih komunikacija.

LITERATURA

[1] Direktiva 97/33/EC.
[2] Direktiva 2002/21/EC o regulatornom okviru za elektronske komunikacione mreže i usluge.
[3] Direktiva 2002/20/EZ o izdavanju dozvola.
[4] Direktiva 2002/19/EZ o pristupu.
[5] Direktiva 2002/22/EZ o univerzalnom servisu.
[6] Direktiva 2002/58/EZ o obradi podataka i zaštiti privatnosti u sektoru elektronskih komunikacija.
[7] Direktiva 2009/140/EZ kojom se mijenjaju Direktive 2002/21/EZ o zajedničkom regulatornom okviru za elektronske komunikacione mreže i usluge, 2002/19/EZ o pristupu i međupovezivanju mreža elektronskih komunikacija i opreme, i 2002/20/EZ o izdavanju dozvola za pružanje mreža i usluga.
[8] Direktiva 2009/136/EZ kojom se mijenja Direktiva 2002/22/EZ o univerzalnom servisu i pravima korisnika u vezi elektronskih komunikacionih mreža i usluga. Direktiva 2002/58/EZ o obradi podataka i zaštiti privatnosti u sektoru elektronskih komunikacija i Uredba (EZ) broj 2006/2004 o saradnji između nacionalnih organa zaduženih za sprovođenje zakona o zaštiti potrošača.
[9] E. Noam, "Interconnection Practices", in: .E.Cave, S.K.Majumdar, Handbook of Telecommunications Economics, Holland, Amsterdam, 2006, pp.387-419.
[10] Odluka 676/2002/EZ o radio-frekvencijskom spektru.
[11] Proposal for a regulation of the European parliament and of the council laying down measures concerning the European single market for electronic communications and to achieve a Connected Continent, and amending Directives 2002/20/EC, 2002/21/EC and 2002/22/EC and Regulations (EC) No 1211/2009 and (EU)No 531/2012.
[12] Smjernice Komisije 2002/C 165/03 o analizi tržišta i procjeni tržišne snage.
[13] Steps towards a truly internal market for e-communications in the run-up to 2020, Ecorys, TU Delft and TNO, 2012.
[14] Uredba (EZ) 2887/2000 o pristupu lokalnoj petlji
[15] Uredba (EZ) broj 717/2007 o romingu u javnim mobilnim telefonskim mrežama u okviru EU.
[16] Uredba (EZ) broj 544/2009 o izmjenama i dopunama Uredbe (EZ) broj 717/2007 o romingu u javnim mobilnim telefonskim mrežama.
[17] Uredba (EZ) 1211/2009 Evropskog parlamenta i Vijeća Evrope od 25.11. 2009. o osnivanju evropskih regulatora za elektronske komunikacija (BEREC).
[18] Zakon o komunikacijama Bosne i Hercegovine ("Sl. glasnik BiH", br. 31/03, br. 75/06 i 32/10)).
[19] COM(2013) 147 - Regulation of the European Parliament and of the Council on measures to reduce the cost of deploying high-speed electronic communications networks.

BUSINESS MODELLING

Connecting Clouds

Ivan Vrbovčan, Tomislav Pavić, Marija Šoša Anić
Hrvatski Telekom d.d., Zagreb, Croatia
ivan.vrbovcan@t.ht.hr, tomislav.pavic@t.ht.hr, marija.sosa.anic@t.ht.hr

Abstract - Due to a declining market of telecom services, telecom providers are trying to expand their service portfolio to new markets based on cloud paradigm. To be successful with cloud services, telecom has to find offers that are interesting to their customers, that can be put on the market in short time and work with sufficient quality.

In this paper we will provide an overview of our approach for building portfolio of cloud services using locally developed and services developed by other cloud providers. In this approach all cloud services are seamlessly integrated, sold, supported and charged through local telecom channels. The overview will include description of the approach including high level architecture, main processes and cloud interface layer.

I. INTRODUCTION

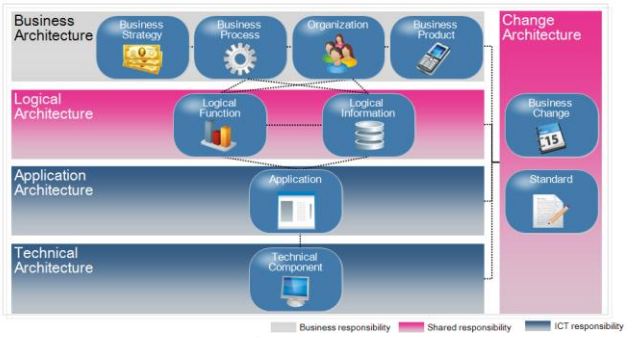
Due to a declining market of telecom services, telecom providers are trying to expand their service portfolio to new markets based on cloud paradigm. To be successful with cloud services, telecom has to find offers that are interesting to their customers, that can be put on the market in short time and work with sufficient quality.

To gain a market share, telecom providers has to build a portfolio of services attractive to their customers. As there is no a recipe on how to create a successful customer service, telecom providers are creating large number of services and hope that one in ten of services will be successful. To build services in shorter time, new business models are established in which telecom providers are partnering with external partners for cloud service delivery. Those new service delivery models can be summarized in 3 main types – integrate, distribute and co-sell [1]. As partners telecom providers are using *independent software vendors* (ISV), but also cooperating with other telecom providers from the same group operating in other countries.

Delivery of cloud services between providers requires several steps: contracting, building technical integration and operating cloud service. As building technical integration on service by service base will require too much time and effort, systematic approach that will enable mass service integration is needed. To define such approach we used industry best practices provided by organizations such as TM Forum for telecom processes and TOGAF for enterprise architecture.

II. SOLUTION ARCHITECTURE

Solution architecture represents a blueprint on how to create a solution to specific problem. Architecture is presented in several architectural views that communicate different aspect of solution to stakeholders. Architecture starts from business processes that solution has to cover, than it describes logical architecture from functional and information points, and at the end it describes how those functionalities are implemented in applications and technical infrastructure. We have adopted this approach as is described in TOGAF (The Open Group Architecture Framework) [9] and DT Architecture Meta-model [13] – see Picture 1.



Picture 1. EA Levels from DTAG EA Meta-model 1.0

In designing solution for cloud service delivery two main options were discussed:

- delivery of cloud services from central cloud through central marketplace,
- distributed delivery of cloud services in peer to peer mode between clouds without central hub.

First option requires less integration effort to bring new service to the market, but brings dependency on central cloud. Second option take into account local telecom provider needs, flexibility and consistent customer experience, but require initial high integration effort.

After evaluation second options were chosen. In this option each cloud could use service from any other cloud without need for central hub.

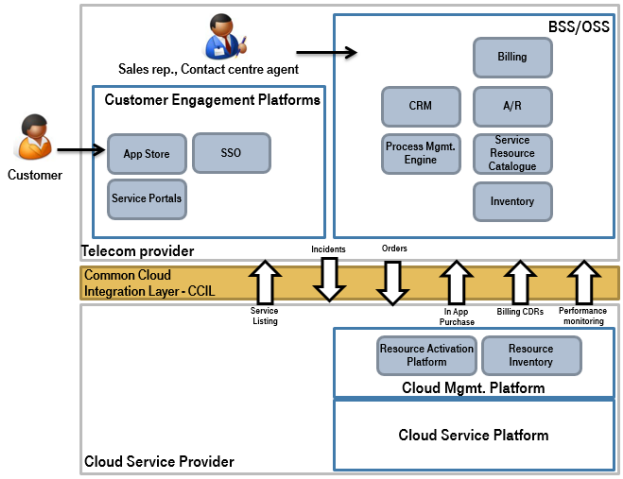
In this option all communications toward customer, customer support, providing application marketplace and customer billing remains responsibility of local provider.

Local provider finds interesting services offered by other providers of cloud services, contract service delivery and integrate it into own processes and system stacks.

In defining the solution architecture following architecture design principles were set:

- **Peer to peer cloud services providers**: application could be hosted anywhere (locally, in other telecom provider, centrally, 3rd party)
- **Easy integration into standard telecom architecture**: multiple channel support, support for standard telecom processes, one integrated invoice with telecom and local/global cloud products
- **Seamless customer experience**: customer should not notice difference between local and global cloud products
- **Easy replicable architecture**: the same architecture should be easily replicated in different environment (today telecom providers has different internal IT architecture with different systems in place)
- **Optimization of cost structure**: economy of scale, global contracts.

The idea is to connect telecom provider's clouds through standardized set of interfaces called Common Cloud Integration Layer (CCIL) – shown in brown on Picture 2. This set of interfaces will be implemented by each telecom provider. Through this set of interfaces telecom provider will be able to connect their existing BSS/OSS systems to other telecom providers in standardized way. Once initial integration is done, cloud service provider will be able to easily provide mass amount of services to other providers.



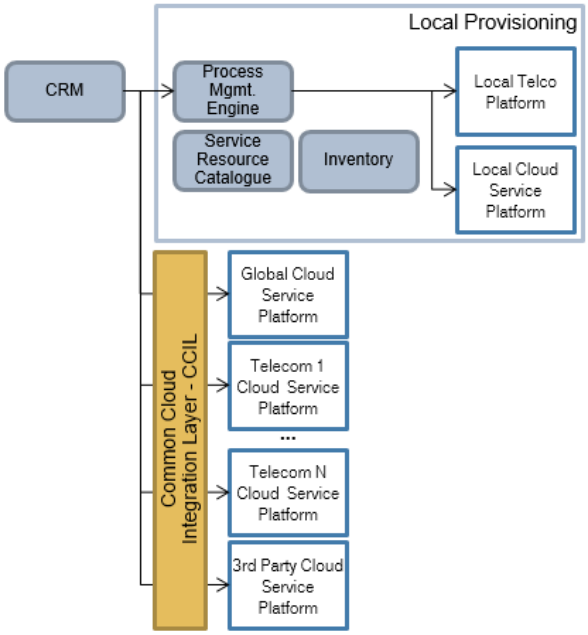
Picture 2. Common Cloud Integration Layer

Telecom provider business processes are described in standard frameworks like Frameworx from TM Forum [10] with main processes in fulfillment, assurance and billing area. Processes for cloud services has to be aligned with standard telecom processes to enable bundling with existing telecom products and using existing organization and infrastructure.

Although eventually all processes concerning cloud services could be automated, stepped approach could be defined that start with automation of fulfillment process, and in later phases automate assurance and product/service lifecycle management.

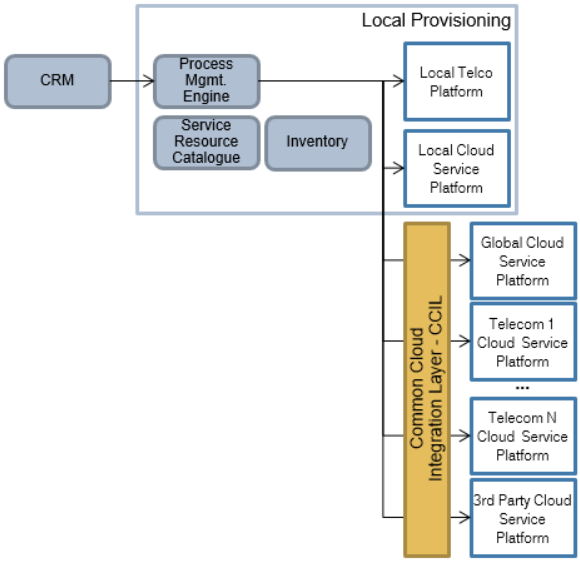
CCIL supports different architectural styles for integration with local BSS/OSS (Business Support Systems/Operating Support Systems). Common Cloud Integration Layer could be connected to local CRM as additional provisioning platform or could be connected to local OSS as an additional service platform.

Connecting CCIL to local CRM as additional provisioning platform (see Picture 3) results in faster cloud service provisioning because it bypasses the local provisioning layer. However, this approach lack service bundling and orchestration abilities that brings the local provisioning platform and requires from CRM to orchestrate order provisioning through several provisioning platforms.

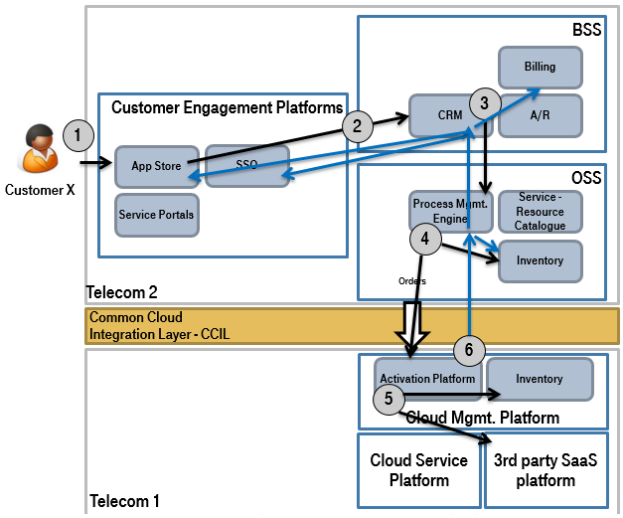


Picture 3. CCIL connected to CRM

Connecting CCIL to local OSS as additional service platform (see Picture 4), enables service bundling (for example ADSL service with cloud storage service) and using provisioning platform orchestration capabilities. However, this approach prolongs order provisioning time due to additional layer involved and also requires more configuration when a new product is introduced because product has to be configured in CRM and OSS.



Picture 4. CCIL connected to OSS



Picture 5. Use case for new cloud service fulfillment

III. PILOT IMPLEMENTATION

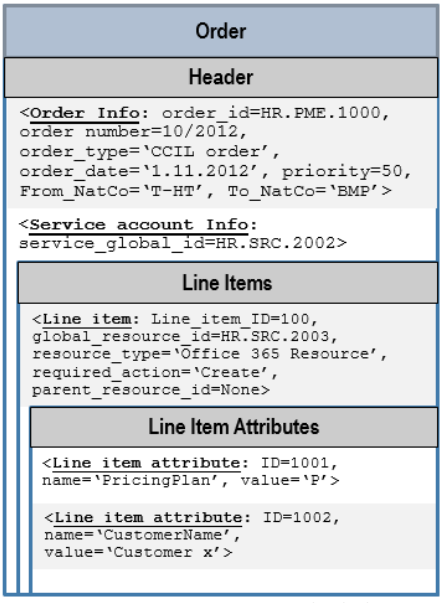
For pilot implementation of Common Cloud Integration Layer a scenario is defined in which SaaS application provided by independent software vendor to Telecom 1 is published through CCIL to Telecom 2. Only fulfilment process is covered by CCIL in the pilot implementation.

CCIL is implemented on Telecom 1 existing application marketplace and connected to Telecom 2 OSS system stack through middleware platform. To enable provisioning process, new product is configured on Telecom 2 local application marketplace, CRM and OSS systems.

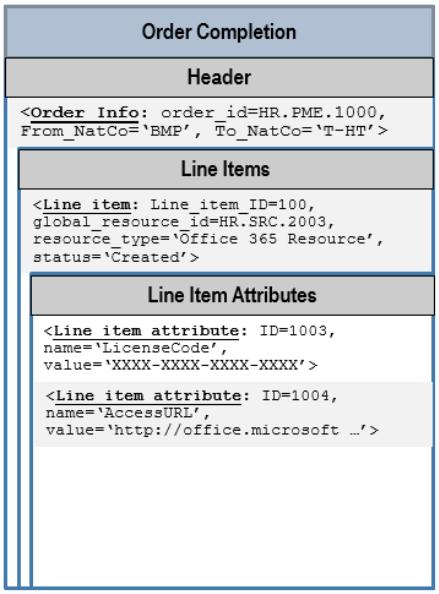
Fulfillment process for cloud service is done in the following steps (see on Picture 5):

1. Customer order SaaS application in App Store
2. App Store send product order with specific tariff option product to CRM
3. CRM decompose product order to service order and send service order to OSS Process Management Engine (OSS PME)
4. OSS PME decompose services to resources using Service/Resource Catalogue and create new SaaS application service in S/R Inventory. New resource order for creation of SaaS application resource is created and sent through CCIL to Telecom 1 Cloud Activation Platform
5. Activation platform provision SaaS application on ISV platform and store resource information in Telecom 1 cloud inventory
6. Information about provisioned resource is returned back to Telecom 2 and propagated to S/R Inventory, CRM, Billing, SSO and App Store

Interface for fulfillment process is built in such way that it is agnostic to type of cloud services ordered through this interface. Basic object sent through the interface is an order in form of XML message which an example can be seen on Pictures 6 and 7 respectively. Each order contains one or more line items with resource description, action required on the resource and resource attributes (see Picture 6). Order completion object (see Picture 7) which is sent as a response after finished provisioning of resource contains information about resource status after provisioning and resource attributes.



Picture 6. Example order message for fulfillment



Picture 7. Example message for order completion

IV. CONCLUSION

In this work we presented high level solution architecture for integrating clouds services into telecom provider environment based on Common Cloud Integration Layer. Based on this solution architecture initial pilot implementation on simplified scenario was done to prove validity of proposed models and concepts. After successful pilot implementation the work has been started on production implementation.

Although, assurance and contracting of service delivery are covered in high level through the solution architecture at this stage, they have not worked out in detail and automated, and it remains for the future phases.

LITERATURE

[1.] T. Sceales, D. Milham: B2B2X Partnering Guidebook. Morristown, NJ 07960 USA : TM Forum, 2013.
[2.] E. Troup, K. Willets, D. Milham, M. Giaccaglini: Multi-Cloud Service Management Accelerator Pack - Business Guide. Morristown, NJ : TM Forum, 2013.
[3.] TM Forum: Introduction to the Digital Services Reference Architecture. Morristown, NJ 07960 USA : TM Forum, 2013.
[4.] E. Troup, K. Willetts, D. Milham, M. Giaccaglini: Multi-Cloud Service Management Technical Guide. Morristown, NJ 07960 USA : TM Forum, 2013.
[5.] E. Troup, K. Willetts, D. Milham, M. Giaccaglini: Multi-Cloud Service Management Accelerator Pack Introduction. Morristown, NJ 07960 USA : TM Forum, 2013.
[6.] K. Willetts, D. Milham, M. Giaccaglini: Multi-Cloud Service Management Pack: Service Level Agreement (SLA) Business Blueprint. Morristown, NJ 07960 USA : TM Forum, 2013.
[7.] TM Forum. Enabling and Managing Digital Services: Positioning Interfaces. Morristown, NJ 07960 USA : TM Forum, 2013.
[8.] K. Willetts, P. Excell: Managing Partnerships - An Introductory Guide to developing successful and sustainable managed service partnerships in the digital economy. Morristown, NJ 07960 USA : TM Forum, 2012.
[9.] The Open Group: The Open Group Archtectural Framework (TOGAF), Version 8.1, . San Francisco : The Open Group, 2003.
[10.] TM Forum: Business Process Framework (eTOM), Addendum D, Process Decompostions and Descriptions, version 12.4. Morristown, NJ 07960 USA : TM Forum, 2012.
[11.] W.J. Ross, P. Weill, C.R. Robertson. Enterprise Architecture as Strategy. Boston, Massachusetts : HBR, 2006.
[12.] Reese, [2] R. J. I/T Architecture in Action. s.l. : AVRE, 2008.
[13.] Turner, D.: Enterprise Architecture Meta-Model (EAMM), DT, 2009.

Competence Transfer through Enterprise Mobile Application Development

Saša Vojvodić*, Matteo Zović*, Veselin Režić*, Hrvoje Maračić** and Mario Kušek**
* Ericsson Nikola Tesla, Zagreb, Croatia
** University of Zagreb / Faculty of Electrical Engineering and Computing, Zagreb, Croatia
{sasa.vojvodic, matteo.zovic, veselin.rezic}@ericsson.com, {hrvoje.maracic, mario.kusek}@fer.hr

Abstract - Large world corporations need corresponding information technology (IT) support as well as constant improvement of software tools, which should enable further business development and more efficient work to operational organizations. The main interest of IT support organizations is currently more and more connected with mobile IT equipment of the employees. Specific business mobile applications improve the efficiency and result in new operating possibilities.

The cooperation between Ericsson Nikola Tesla (ENT) and Faculty of Electrical Engineering and Computing (FER) of the Zagreb University, has enabled a fast development of competencies in ENT, necessary for mobile application development and quality realization of innovative solution for platforms iOS and Android. In this cooperation at the project realization the methodology of application development was defined, the corresponding competencies were developed, the system architecture was designed together with the communication of mobile application and back-end IT systems.

In the project the iterative development approach, tools for software code versioning and project control have been used thus enabling continuous insight in project progressing. At any moment it was possible to determine priorities of functional development and solution elements, as well as to realize the necessary additional transfer of knowledge.

The project has resulted in an enterprise mobile application for iOS and Android platforms, which had been implemented and tested in several countries. The complete solution enables data management by means of the portal, thus decreasing the frequent changes of users' mobile application and significantly accelerating the process of new functionalities introduction.

I. INTRODUCTION

The growing variety of devices, computing styles, user contexts and interaction paradigms, consequences of bring your own device (BYOD) programs that are significantly increasing the size of the mobile workforce is placing tremendous strain on Information Technology (IT) organizations [1]. Mobility is the megatrend in enterprise networking.

Reach mobility possibilities are demanding the new IT knowledge on creating expanded user interface models that can connect people and information in new and different ways. Mobility and BOYD create both efficiency and challenges. When employees have information and communication tools on the move and at their fingertips, enterprises can make decisions more quickly, communicate more effectively, and interface more closely with their partners and customers.

BYOD and mobility drive demand for more pervasive networks, multi-mode devices, and more sophisticated applications and user interfaces.

The worldwide tablets and smartphones shipment significantly outpaced the PC shipments numbers. In 2014, the worldwide tablet market is forecasted to grow 47 percent [2]. Sales results in Croatia in the first half of 2013 show tremendous growth: 349.000 smartphones and 172.089 tablets or 61% and 780% sales growth compared to the first half of the previous year [3]. Worldwide combined shipments of devices (PCs, tablets, ultramobiles and mobile phones) are projected to reach 2.5 billion units in 2014 (Figure 1.

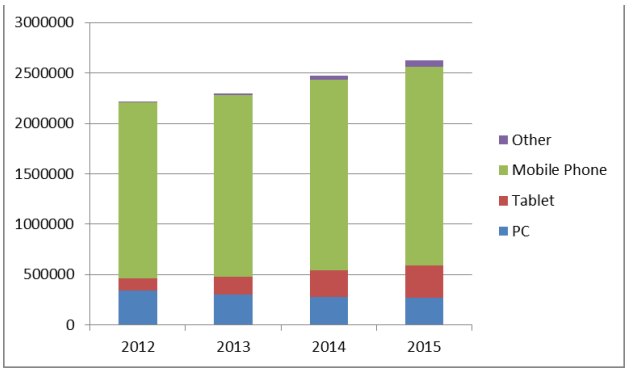


Figure 1. Worldwide Device Shipments by Segment (Thousands of Units)

Enterprise IT organizations have to act. Evolutions by means of internal transformations are needed in three different directions:

- New knowledge; focus towards mobile applications,
- New way of working; service oriented approach,
- True business partner in operational development and business excellence; savings and revenue focused innovations.

New knowledge is necessary to develop and provide enterprise mobile application that takes center stage as

mobility adoption increases. The mobile apps development lifecycle is unlike the traditional application development lifecycle, so there is a need for new management tools, and existing skills do not transfer from desktop application development to mobile app development because the patterns for building efficiency, performance, and security into mobile devices are different [4]. The new released apps should consider the application lifecycle management, application distributions, performance management, monitoring and analytics, version management, security, etc. – and this is all new for traditional enterprise IT.

To speed up the required evolution, Ericsson Nikola Tesla (ENT) decided to engage external qualified partner in know-how for knowledge transfer. The choice fell on University of Zagreb, Faculty of Electrical Engineering and Computing (FER), and cooperation in boosting mobile application development knowledge through concrete project execution has started.

II. COOPERATION

The cooperation goals were to gather information, best practices on mobile development approach and to create a new mobile application development initial team across the existing organization unit structure. Final delivered apps were supposed to be implemented on global corporate scale, so cooperation with colleagues from USA, Sweden and other test countries were established.

The knowledge management system plays a vital role in supporting the organization to increase its level of innovation by facilitating access to knowledge, encouraging the free flow of ideas, and by supporting and accelerating the learning processes within the organization. This was all achieved through this cooperation. Figure 2. describes 5-phase organizational knowledge development serving as innovation founder [5].

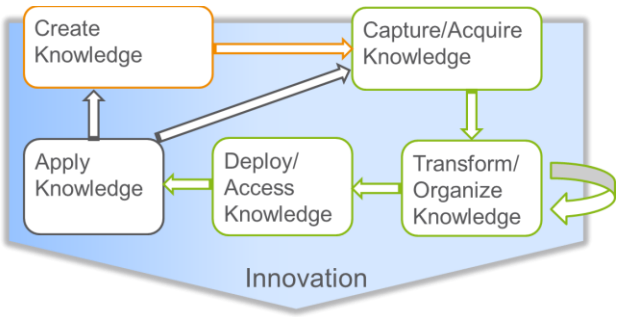


Figure 2. 5-phase organisational knowledge development

Core and non-core specific competence found inside the company and assigned to the project are the following:

- System development, software design, test and verification,
- Graphical user interfaces design,
- IP network design,
- Application and database servers design,
- Security and
- Specific application requirements competence.

In a broader perspective, the cooperation should establish overall framework to improve company competence building, mobility, administration, efficiency and productivity by deploying new innovative mobile enterprise applications. Knowledge management system is critical to get this done; hence phase approach shown in Figure 3. was established.

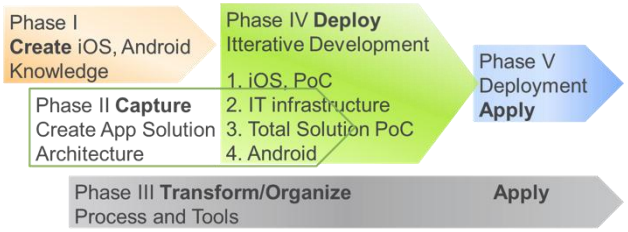


Figure 3. Cooperation phase approach

iOS [8] and Android [9] knowledge were built as starting step (Phase I) thus corresponding to Create Knowledge in 5-phase approach. Solution architecture, functional and non-functional requirements were defined and designed (Phase II) and correspond to Capture/Acquire Knowledge. Adjusting corporate ecosystem by applying tools and processes were used in Phase III in parallel with other activities. This phase is covered with Apply Knowledge from 5-phase approach. Iterative development (Phase IV) was used and each iteration was resulting with set of deployable functionalities which corresponded to Deploy/Access Knowledge in 5-phase approach. Each iteration (except the first one) was followed by deployment phase (Phase V), which corresponds to Apply Knowledge.

III. MOBILE PLATFORM

A number of different mobile platforms exist. Two of the most widely used are Android and iOS. Together, they account for more than 65% of all mobile phones [10].

The Android Open Source Project is a project initiated by the Open Handset Alliance, led and developed by Google. It is open source and manufacturer independent [11]. Different manufacturers, e.g. Samsung, HTC, Motorola, port Android to their devices, mostly by implementing drivers and other device-specific features. The operating system itself remains the same, allowing the same application to be used on all Android devices, regardless of their manufacturer. There is an Android Compatibility Program designed to ensure this. Each manufacturer needs to comply with the Android Compatibility Definition Document (CDD) and pass the Compatibility Test Suite (CTS) [12]. Android itself is written mostly in C and C++, while user interface elements are written in Java. Due to Android's use of the Dalvik Virtual Machine instead of the regular Java Virtual Machine, applications for Android are written in the Java programming language and compiled specifically for Android.

iOS, on the other hand, is a closed system, developed and used exclusively by Apple for their mobile devices, including iPhone, iPad and iPod Touch. Applications for iOS are written in ObjectiveC.

Both platforms have their respective application stores. Developers of consumer applications submit their applications to the store. The application is then reviewed for compliance with the store's terms and requirements and published. Since corporate applications can be distributed internally, there are fewer requirements since there is no store review.

New versions of both iOS and Android are published regularly. Android is currently in version 4.4, Application Programming Interface (API) level 19. When deciding which versions to support, we decided to go with version 2.1, API level 7 and newer ones. This is done because all needed features were already available in that version, there are no phones in our corporate environment with older Android versions and global use of older versions is also negligible [13]. iOS is currently in version 7, but at the time of development was in version 6. We decided to support versions 6 and 7 because all iPhones since version 3GS are capable of running those versions, all devices in our corporate environment have at least version 6, and share of these two versions is 96% [14].

Unlike iOS, which runs on only two different screen sizes on phones, Android phones have numerous different screen sizes and screen resolutions. Drivers and core services on Android phones are usually implemented by manufacturers, which means that sometimes, due to either bugs or differences in implementations, same services behave differently on different Android phones of the same operating system version. As an example, we discovered that one of our users had a phone which did not handle a certain feature properly, unlike all other manufacturers which we tried, for the same Android version. It also means that an application has to be scalable to support different screen sizes.

One of the challenges on iOS was detecting whether the user is currently in roaming. Although the iOS has this information, it is not accessible by applications, probably due to privacy concerns.

Also due to the decision to support iOS version 6 and 7, we had to switch to the new Xcode, Apple's Integrated Development Environment (IDE) for developing iOS applications. Unfortunately with the new IDE there was a change in the User Interface (UI) of the IDE and development team had to spend more time to adapt to the new version changes. In addition to the changes of the IDE, there was a change in the iOS itself. The new iOS has brought enhancements in graphical elements behavior which consequently automatically fixed graphical outstanding issues on iOS 6 that had been preventing a proper implementation of customer graphic requirements.

IV. METHODOLOGY

A. Tools

When working in a team, one of the important aspects is how to enable efficient code synchronization and sharing between team members. To facilitate this, a number of different version control systems exist, e.g. CVS, SVN, Git [15]. We decided to choose Git due to its performance and built-in support in Xcode, Apple's Integrated Development Environment (IDE) and in Eclipse, used for Android development. Online services for Git hosting are available, both paid and free, but we decided to set up our own server based on the Gitblit stack [16] considering the commercial and internal nature of the project.

Another challenge when working in a team on a project is project management. We decided to use ChiliProject, an open source, web based project management system [17]. It provides support for the complete project lifecycle, including setting up, tracking the issues and reporting work progress. It allowed us to keep track of who is working on what, and the progress of that work. It was a very important tool in tracking everyone's work and ensuring on-time deliveries.

From technical perspective, mobile development strategy with frameworks automates a large amount of the plumbing necessary to build mobile applications allowing developers to focus on the user interface, experience and integration into business processes.

On iOS we use a number of open source frameworks for different functionalities. For logging we use the Lumberjack framework [18]. It is an almost drop-in replacement for the built in logging facility i.e. uses the same syntax with a different object, but offers a variety of output devices unlike the built in facility. It enables the developer to specify one or more logging destinations, including console, debug window of the IDE, file system or a custom log processor. Single log message is then written to all these locations simultaneously.

For crash reporting in iOS, we use the QuincyKit framework [19]. It consists of the client part running on the customer's phone and the server part running on the developer's server. On every application restart after a crash the framework detects the crash and uploads all the relevant crash information to its server. This information includes the code which was currently executing at the time of the crash.

iOS already has a data persistence framework, Core Data, which enables the developer to store objects [20]. To further simplify the use of this feature, we decided to use the MagicalRecord framework [21]. It builds on top of the Core Data framework allowing clear and simple one-line fetches.

Although iOS has different kinds of progress indicators and user alerts, it is not possible to use or customize them using a public API. To address this, we used the MBProgressHUD framework which provides iOS-style progress indicators, user alerts and messages [22].

On Android we used in most cases frameworks built into the Android API.

B. Process

For the application development, we used agile approach [6] with iterative development. Daily monitoring of the progress and a scrum like framework [7] is used for

the project management. The team members were not fully assigned to this project so we defined few days in week as days for this project only. This rule was in most cases applied.

The project has been divided in four major iterations, with a cycle of two to three weeks. At the beginning of each iteration, we had a meeting to define the goals we should achieve in the current iteration. Depending on the objective of the iteration we also planned the competence development.

On a daily bases we have a short meetings analyzing current iteration tasks progress. Since team-members were on different locations, teleconference or videoconference with screen sharing has been used. Depending of gaps between existing and required knowledge to complete development tasks, additional consultations for further competence development were prepared and executed.

The above operation can be defined as scrum-like since each iteration has been divided in smaller individual tasks sorted by priorities. Each development iteration cycle can be considered as a sprint in which development team develops the application.

C. Testing

Because of the iterative application development method, short deadlines for application completions and large amount of UI element, the development was not based on the Test Driven Development (TDD) process. A unit test is created just for checking work that doesn't require UI, while other tests were performed during the development of the application and at the end of every iteration.

After completion of the integration testing of all components, we began beta testing that we conduct in several countries. User guides for all supported iPhone and Android devices were also created. These guides are divided into three parts:

- An introduction to mobile app (detailing pre-requisites),
- A guide to the app on specific mobile platform and
- A description of how to get more help, a list of procedures to perform and detail instructions how to provide feedback and get in contact with Operations and development team.

The problem with beta testing with users is how to get feedback from them and their phones. That is the reason we have integrated feedback option in both versions of application. Usage of integrated feedback, log mechanism are a preferred way of communication with development team because we get additional info about the operating system and type of device in use.

Users can send comments for improvements, malfunction reports and can send logs or crash reports. Logs and crash reports are primarily assistance for developers bug finding. Sent data logs include various information about Android phone versions of OS and phone models and skins.

Feedback is received through Representational State Transfer (REST) service and dispatched to all support team members from mail daemon running on the server.

V. BUILDING SOLUTION

Some of the important requirements for the application were as follows:

- Location independent and
- Easy to use interface.

The above requirements have been successfully implemented with the proper architecture and design solutions, and using well-known ergonomic elements.

Application is based on the Model-View-Controller (MVC) architecture. Benefits of this architecture are to have a separate Model (data, business logic), View (application screens) and Controller (input elements on screen). This model is facilitating team development, maintenance existing functionality and testing new functionality solutions.

The application design is based on a client-server communication model and the MVC architecture. Choosing this design solutions ensure that the data obtained through the server communications and stored in the model are specific for the chosen location. This design allows us to add new location, specific function for existing location or to correct data that affect the proper operation of the application without the need to change the application itself. On the server side developed App Portal enables building of location specific configuration by responsible persons through controllable access permissions.

As stated before, communication between the application and the server is based on Representational State Transfer (REST) architecture. By implementing REST architecture we have achieved faster development time, simple configurable interface and easier download or send data to the server.

A. Deployment

Delivery of the application is done in two ways.

- Enterprise application store (EAS) and
- Web page.

Enterprise application store (EAS) is a distribution channel for mobile applications where end user can download and update specific enterprise applications. It is like App Store or Google Play for enterprise.

It also provides a mechanism of remote administration and management of the client phone that allows an Enterprise more controllable and secure way to allow BYOD possibilities, and to have an efficient way to maintain company specific applications. On the other hand, if mobile device is stolen or lost, wipe out of all data on device and reset to factory settings can be executed remotely.

The second mechanism is web page distribution. Selected Ericsson corporate users for beta testing got installation instructions for iOS and Android on a web page together with mobile application installation files.

VI. CONCLUSION

Project success was measured through five different perspectives. Ordinary project perspectives: time, cost and scope plus two additional and in this project the most interesting perspectives: knowledge transfer effectiveness and impact on organization. Project valuation perspectives are shown in figure 3. With the latest iteration and final mobile apps delivery and development phase completion, the knowledge transfer effectiveness was evaluated considering factors shown in Table 1.

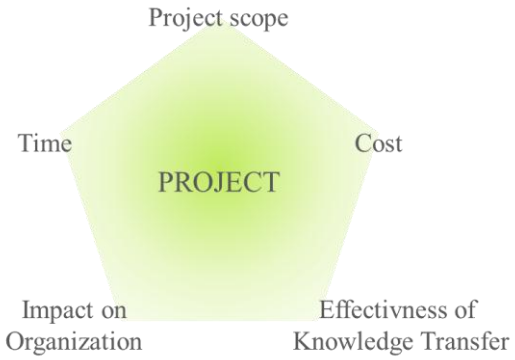


Figure 4. Project valuation perspectives

Table I. Knowledge transfer effectiveness evaluation

Mobile Apps Development – Knowledge Transfer Effectiveness Assessment of factors fulfillment in knowledge transfer	
Factor:	**Assessment:**
Cooperation readiness	Completely fulfilled
Knowledge transfer organized as a project	Completely fulfilled
Cultural differences	Not existing; no impact on knowledge transfer
Knowledge transfer complexity	In majority fulfilled(*)
Time constraints	Apps were delivered with slight delay
Budget constraints	The budget was slightly exceeded
Previous domain knowledge	There were no previous mobile apps development experience
Previous tools knowledge	Some project members were familiar with used tools
Previous methods knowledge	Some project member were familiar with scrum methodology

(*) Transfer of knowledge for this specific project scope was realized completely. Fundamentals for mobile apps development were established. However, with new and different mobile apps requirements there will be need for further competence development.

The project was successful. Joint cooperation between FER and ENT assured ENT the new IT strategic knowledge and starting capability that will be utilized as new demands from evolving corporate mobility eco-systems are coming. The whole cycle (Figure 2) for specific delivery project was closed but further cycles with continual knowledge creation, effective growth and new innovation should come.

REFERENCES

[1] Gartner Identifies the Top 10 Strategic Technology Trends for 2014, http://www.gartner.com/newsroom/id/2603623, retrieved 02.02.2014.

[2] Gartner Says Worldwide Traditional PC, Tablet, Ultramobile and Mobile Phone Shipments On Pace to Grow 7.6 Percent in 2014, http://www.gartner.com/newsroom/id/2645115, retrieved 02.02.2014

[3] GfK: Prodaja tableta u Hrvatskoj u odnosu na prošlu godinu porasla za 780 posto!, http://www.netokracija.com/gfk-prodaja-tableta-57992, retrieved 02.02.2014.

[4] Mobile app development takes centre stage as BYOD adoption increases, http://www.appstechnews.com/blog-hub/2013/may/23/mobile-app-development-takes-centre-stage-as-byod-adoption-increases/, retrieved 02.02.2014.

[5] Siniša Krajnović, „Upravljanje prijenosom znanja u istraživanju i razvoju na području informacijske i komunikacijske tehnologije “, doktorska disertacija, Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva, Zagreb, 2007.

[6] Manifesto for Agile Software Development, http://agilemanifesto.org, retrieved 08.02.2014.

[7] Ken Schwaber, Jeff Sutherland. "The Scrum Guide". Scrum.org. Retrieved 08.02.2014.

[8] iOS, http://www.apple.com/ios/, retrieved 08.02.2014.

[9] Android, http://www.android.com, retrieved 08.02.2014.

[10] Top 8 Mobile Operating Systems from Jan 2013 to Jan 2014, http://gs.statcounter.com/#mobile_os-ww-monthly-201301-201401, retrieved 04.02.2014.

[11] The Android Source Code, Android Developers, http://source.android.com/source/index.html, retrieved 04.02.2014.

[12] Android Compatibility, Android Developers, http://source.android.com/compatibility/index.html, retrieved 04.02.2014.

[13] Dashboard, Android Developers, http://developer.android.com/about/dashboards/index.html, retrieved 04.02.2014.

[14] UPDATE: iOS Version Distribution Study - 2012 through 2013, http://chitika.com/insights/2013/ios-distribution-update, retrieved 04.02.2014.

[15] Available CVS Alternatives, http://better-scm.shlomifish.org/alternatives/, retrieved 04.02.2014.

[16] Gitblit, http://www.gitblit.com, retrieved 04.02.2014.

[17] ChiliProject, https://www.chiliproject.org, retrieved 04.02.2014.

[18] CocoaLumberjack, https://github.com/CocoaLumberjack/CocoaLumberjack, retrieved 04.02.2014.

[19] QuincyKit, http://quincykit.net/, retrieved 04.02.2014.

[20] Core Data Programming Guide: Technology Overview, https://developer.apple.com/library/mac/documentation/cocoa/Conceptual/CoreData/Articles/cdTechnologyOverview.html#//apple_ref/doc/uid/TP40009296-SW1, retrieved 04.02.2014.

[21] MagicalRecord, https://github.com/magicalpanda/MagicalRecord, retrieved 04.02.2014.

[22] MBProgressHUD, https://github.com/jdg/MBProgressHUD, retrieved 04.02.2014.

An Empirical Study of Factors Affecting the Adoption of Self-Service Checkout Terminals in Slovenia

Boštjan Šumak*, Maja Pušnik* and Marjan Heričko*
* University of Maribor, Faculty of Electrical Engineering and Computer Science, Maribor, Slovenia
bostjan.sumak@um.si, maja.pusnik@um.si and marjan.hericko@um.si

Abstract - Few years ago, retailers started to employ self-service checkout terminals providing customers a choice to scan the barcodes on their own items and complete the purchase process without an interaction with retailer's staff. This study aimed to investigate factors which can have a significant impact on user's decision about how and when they will use the self-checkout unit. To understand customers' perceptions about using self-checkout terminals, the Technology Acceptance Model (TAM) was used to develop the research model. In this study, several external factors with direct or indirect impact on user's behavioral intentions are proposed: anxiety, perceived service quality, attitude toward innovation, self-efficacy and anonymousness. Empirical data was conducted using an online survey. The proposed relationships were empirically tested using the structural equation modelling approach. The implications and limitation of the study are also discussed.

I. INTRODUCTION

Merchants invest significant time and money to optimize the checkout process. The checkout process must be expedient and secure and has to offer sufficient customer service to leave the customer with a positive impression. In Slovenia in the past few years, Mercator and SPAR merchants have started implementing self-service checkout terminal (SSCT) solutions in their major stores. The need for improving in-store traffic flows, speeding up the checkout process and increasing customer service and loyalty has led merchants implement innovative solutions.

SSCT provides an alternative to the traditional checkout and enables customers to pay for purchase from a retailer without an interaction with retailer's staff. There are cases however, where the completion of the purchasing process demands attendance of the staff. For example, when buying alcohol, the purchaser has to prove the appropriate age in order to be able to buy alcohol, tobacco products, etc. Because there are cases, where the customer can't really complete the purchasing process on his or her own, these self-service terminals can also be called semi attended customer activated terminals (SACAT).

In the self-service checkout process, customers scan the barcodes of their items in order to identify items they want to purchase. The scanned items are displayed on the monitor, where customers can check for possible errors. After scanning, items are put in a special bag where these items are weighted. Customer is permitted to proceed scanning after the scanned/identified item was put into the bag and weighted and whether the observed and expected weights match [2]. The weight observed in the bagging area provides a control for ensuring that the items in the bag are items which were scanned or identified. Figure 1 shows an example of SSCT introduced in Mercator shop in Slovenia.

It is very common to see in Slovenian stores that customers don't chose to use SSCT terminals on their way out. It happens very often that customers rather step into the long line with human cashier and wait even if they have just a small amount of items in their shopping bag. Although SSCT makes possible to avoid waiting in long lines, customers don't choose this option. Such observations that can be seen every day inspired this study, which aimed to investigate the factors that can affect customer's decision when he or she will use SSCT



Figure 1. IBM's self-checkout terminal in Mercator retail shop (Slovenia) [1]

in order to complete the purchase process.

The paper is organized as follows: in the next section, theoretical backgrounds about acceptance theories are given. In section three, the research model and a list of research hypotheses are proposed. In section four, the research methodology is described, where the development of a measurement instrument is explained and the data collection and statistical analysis processes are described. In section five, data analysis and results are given. Last section concludes the paper including discussion about the implications and limitations of the study.

II. THEORETICAL BACKGROUNDS

Studies, dealing with identification of factors and assessment of their impact on user's decision whether he or she will accept and use a certain technology, are usually based on a technology acceptance theory. Table 1 provides the list of most common technology acceptance theories being used in existing acceptance research studies. In technology acceptance studies researchers are mostly interested in the structural relationships among constructs in the research model that help explain an individual's acceptance of a technology. The empirical data is usually statistically analyzed using structural equation modeling (SEM).

Davis [3] proposed TAM to explain the potential user's behavioral intentions of using a technological innovation. TAM is based on the theory of reasoned action – TRA [4], which is a psychological theory that can be used to explain behavior. The Motivational Model (MM) is another psychological theory that is often being employed in studies dealing with the factors that have an impact on the end user's motivation in the use of an information technology.

TABLE I. ACCEPTANCE THEORIES

Theory	***Factors***
Theory of Reasoned Action (TRA)	attitude toward behavior, subjective norm, behavioral intentions, actual use
Technology Acceptance Model (TAM)	perceived usefulness, perceived ease of use, attitude toward use, behavioral intentions, use
Motivational Model (MM)	intrinsic motivation, extrinsic motivation, behavioral intentions
Theory of Planned Behaviour (TPB)	attitude toward behavior, subjective norm, perceived behavioral control, behavioral intentions, actual use
Model of PC Utilization (MPCU)	facilitating conditions, social factors, perceived consequences, long-term consequences, behavioral intentions, actual use
Innovation Diffusion Theory (IDT)	relative advantage, ease of use, image, visibility, compatibility, results demonstrability, voluntariness of use, behavioral intentions
Social Cognitive Theory (SCT)	computer self-efficacy, perceived outcomes, affect, anxiety, use
Unified Theory of Acceptance and Use of Technology (UTAUT)	performance expectancy, social influence, effort expectancy, facilitating conditions, behavioral intentions, actual use

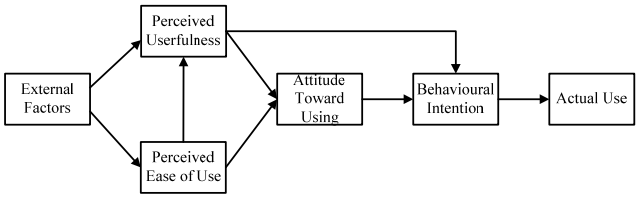


Figure 2. Technology Acceptance Model

TAM is one of the most widely used theoretical models in IT/IS acceptance research because of its understandability and simplicity [5] and is applicable to different information technologies. TAM involves two main independent factors which have a direct or indirect influence on users' behavioral intentions (see Figure 2):

1. *Perceived usefulness* is the degree to which a person believes that using a particular system would enhance his or her job performance.
2. *Perceived ease of use* stands for the degree to which a person believes that using a particular system would be free of effort.

The TAM model has one dependent variable – behavioral intention, which is the degree to which a person has formulated conscious plans to perform or not perform some specified future behavior.

Several existing systematic literature reviews and meta-analysis [5]–[7] show that TAM is the most common theory being used as a ground theory for designing the research in the field of acceptance and use of technology. This is why the TAM model was chosen as a ground theory in this study.

Research about acceptance and use of an information technology can be divided into two types of studies: (1) an empirical validation of the theory (e.g. the TAM theory) in the context of a specific information technology, and (2) an extension of the theoretical model (e.g. the TAM model) with user specific factors. The main objective of this study was not only to evaluate the TAM research model in the context of acceptance and use of self-service checkout terminals, but also to identify external factors, which may have a significant impact on TAM related factors.

III. RESEARCH MODEL AND HYPOTHESES

In this study, we proposed extending the TAM research model with following external factors: (1) anxiety, (2) perceived service quality, (3) attitude toward innovation, (4) self-efficacy and (5) anonymousness. Figure 3 presents the extended TAM research model with hypothesized relationships between constructs. According to the research model, following hypotheses were proposed:

H1: Anxiety will have a significant (negative) impact on user's intention to use SSCT.

H2: Perceived Service Quality will have a significant influence on user's perception about usefulness of SSCT.

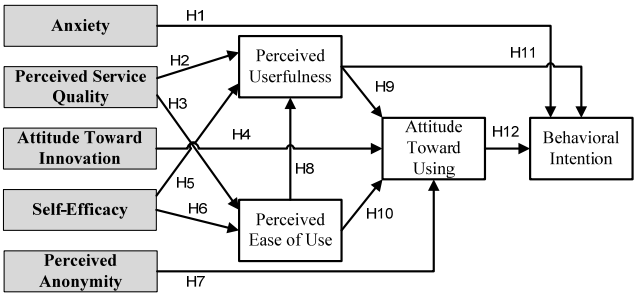


Figure 3. The research model

H3: Perceived Service Quality will have a significant impact on user's perception about ease of use of SSCT.

H4: Attitude toward innovation will have significant influence on user's attitude toward using SSCT.

H5: Self-efficacy will have significant influence on user's perception about usefulness of SSCT.

H6: Self-efficacy will have significant influence on user's perception about ease of use of SSCT.

There are cases, when a customer could be too embarrassed to buy certain products. Customer could be afraid that the cashier will for example mock or condemn or have some negative comprehension during the scanning of his or her items. Traditional checkout process involves the face-to-face interaction between the customer and the cashier and can cause a low level of anonymity. Therefore, one of the reasons for which customers would choose SSCT is because they seek for anonymity [8]. We therefore propose following hypothesis:

H7: Perceived anonymity will have significant influence on user's attitude toward using SSCT.

Hypotheses that follow were proposed according to the original TAM theoretical model and are adapted to the context of usage of SSCT.

H8: Perceived ease of use will have a significant influence on user's perception about usefulness of SSCT.

H9: Perceived usefulness will have a significant influence on user's attitude toward using SSCT.

H10: Perceived ease of use will have a significant impact on user's attitude toward using SSCT.

H11: Perceived usefulness will have a significant influence on SSCT usage intention.

Attitude toward using technology is an individual's overall affective reaction to using a system. It stands for an individual's liking, enjoyment, joy and pleasure associated with technology use. The attitude toward using a system has been shown as a determinant of behavioral intention in different studies. We therefore suggest the following hypothesis:

H12: Attitudes toward use is related to behavioral intention.

IV. RESEARCH METHODOLGY

Quantitative research in the form of an online questionnaire based survey was performed to test the hypotheses, stated in previous section. The questions were organized into two groups: (1) demographic questions about the respondents' gender, age, frequency of use of SSCT, the purchase value, for which users choose to use SSCT, rather than internet experience, Moodle experience, voluntariness, etc. (see Table 2 for the characteristics list); and (2) measures for the TAM constructs and external factors.

The TAM constructs and external factors were adapted into the context of use of SSCT. The measuring items were assessed by a seven-point Likert-type scale from "strongly agree" to "strongly disagree", where the respondents gave the extent to which they agreed or disagreed with each statement of the constructs.

To reduce measurement error, a pre-test of the questionnaire was performed in order to improve the content of the measuring items. In the pre-test, colleagues at the faculty were asked to examine the questionnaire for meaningfulness, relevance and clarity. With remarks and propositions, the measurement items were improved.

After the pilot test, the request form for participation in the online survey was sent to random students and staff at the faculty. They were also asked to forward the request to their friends. 212 online surveys were started, of which 170 were successfully finished and 42 returned incomplete. Thus, the usable response rate was 80%.

TABLE II. SAMPLE CHARACTERISTICS

Demographic characteristics	***Freq.***	***%***
Gender		
Male	134	78.8
Female	36	21.2
Age		
less than 16 years	1	0.6
16 – 20 years	31	18.2
21 – 25 years	79	46.5
26 – 30 years	24	14.1
31 – 35 years	18	10.6
36 – 40 years	5	2.9
41 – 45 years	2	1.2
46 – 50 years	1	.6
more than 50 years	9	5.3
Frequency of use		
allways, when shopping	14	8.2
almost always when shopping	55	32.4
halftime, when shopping	39	22.9
sometimes	56	32.9
never	6	3.5
Using SSCT when value of the purchase is		
10€ or less	52	30.6
up to 50€	49	28.8
up to 100€	2	1.2
the purchase value is not important	67	39.4
Using SSCT when buying up to		
3 items	7	4.1
5 items	26	15.3
8 items	35	20.6
10 items	40	23.5
number of items is not important	62	36.5
Using also other self-service terminals (self-service check-in at airport, etc.)		
Yes	106	62.4
No	64	37.6

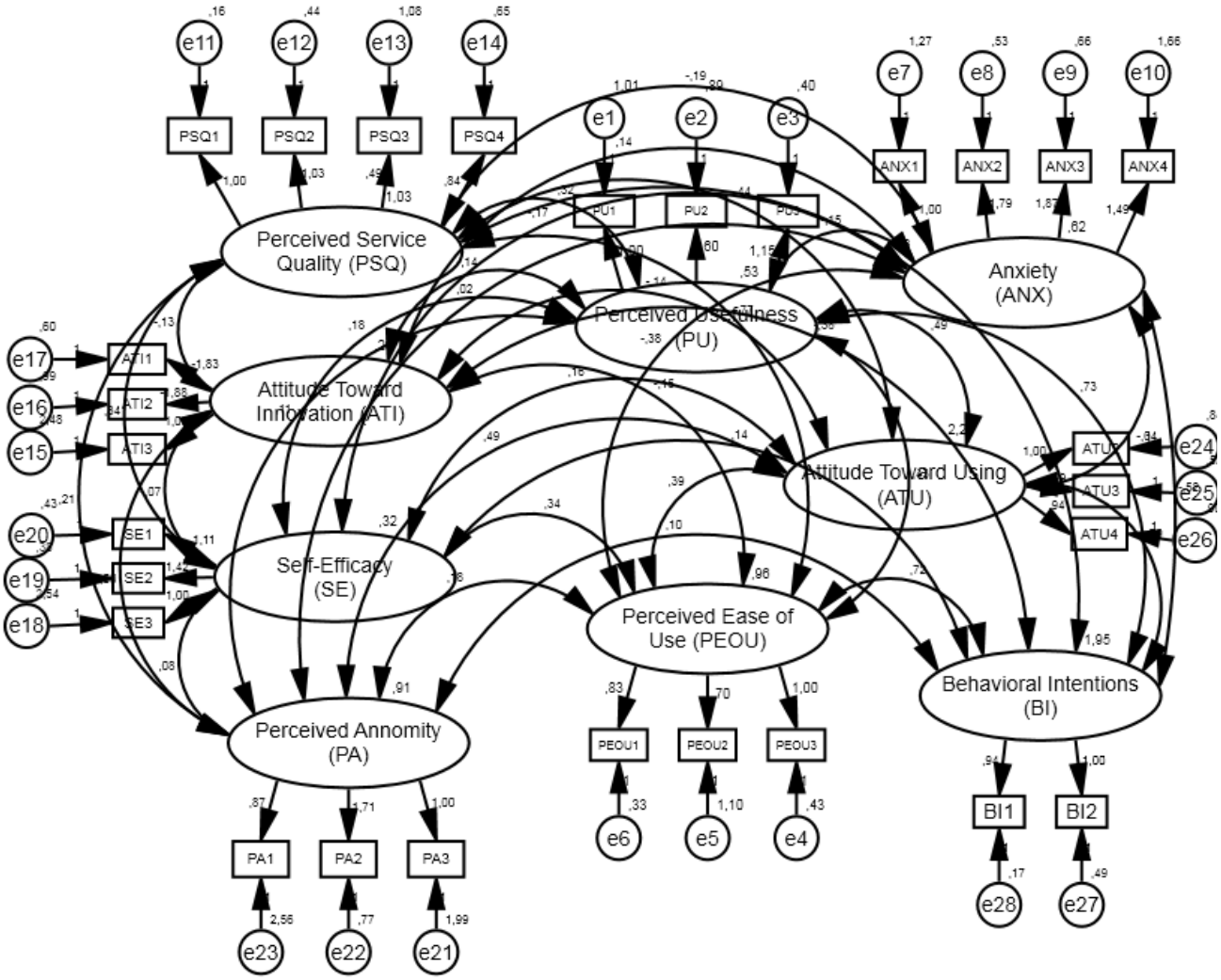


Figure 4. The measurement model

To describe the main features of an average participant in this study, descriptive statistics was used on the respondents' characteristics data. To understand user's perceptions about using SSCT, the extended TAM research model and hypothesized relationships between constructs were empirically tested using the structural equation modelling (SEM) approach. SEM was used to test the fit of the proposed theoretical model with empirical data. The measurement model (see Figure 4) was estimated using confirmatory factor analysis to test whether the proposed constructs possessed sufficient validation and reliability. After the final measurement model passed the goodness-of-fit tests, the structural part of the research model was estimated using SEM on the structural model. The structural model (see Figure 5) was also tested for a data fit with appropriate goodness-of-fit indices. A statistical analysis was performed using the SPSS 22 statistical package together with AMOS 22 software.

TABLE III. HYPOTHESIS TESTING RESULTS

Hypothesis	*Effect*	*Path coefficient*	*Remarks*
H1	ANX→BI	-0.327^{***}	Supported
H2	PSQ→PU	-0.058^{NS}	Not Supported
H3	PSQ→PEOU	0.581^{***}	Supported
H4	ATI→ATU	-0.022^{NS}	Not Supported
H5	SE→PU	-0.028^{NS}	Not Supported
H6	SE→PEOU	0.316^{**}	Supported
H7	PA→ATU	0.298^{***}	Supported
H8	PEOU→PU	0.651^{***}	Supported
H9	PU→ATU	0.396^{***}	Supported
H10	PEOU→ATU	-0.013^{NS}	Not Supported
H11	PU→BI	0.574^{***}	Supported
H12	ATU→BI	0.123^{NS}	Not Supported

V. DATA ANALYSIS AND RESULTS

The characteristics of the respondents are presented in Table 2. The typical respondent is a 21-25 year old male that uses SSCT sometimes or halftime when shopping. The respondent choice to use SSCT is not related to the number of items he is buying. The purchase value is also not the main factor upon which the purchaser would chose to self-checkout. The respondent is also using other self-service terminals (e.g. self-service check-in terminals at the airport).

The results of the final structural model are shown in Figure 6 and the summary of hypothesis testing is presented in Table 3. The analytical results show that anxiety has a negative impact (β=-0.327, $p<0.001$) and perceived usefulness a strong positive impact (β=0.574, $p<0.001$) on user's intention to use SSCT. These results provide support for hypotheses H1 and H11. However,

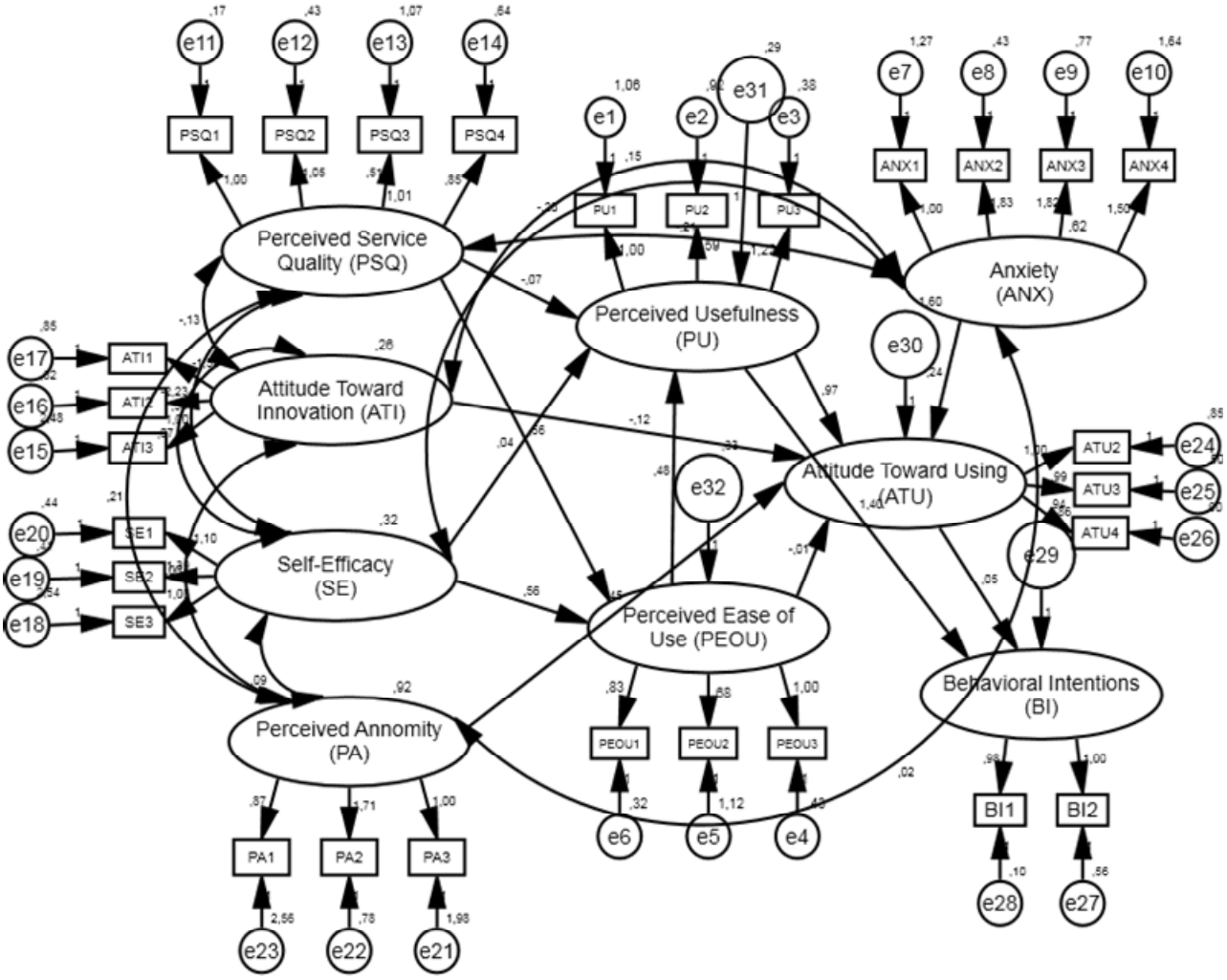


Figure 5. The structural model

there was no significant relationship found between attitude toward using SSCT and user's intention to use SSCT. Therefore, the hypothesis H12 was rejected.

Perceived ease of use of SSCT has a significant impact on perceived usefulness (β=0. 651, p<0.001), therefore the hypothesis H8 was accepted. However, proposed external factors perceived service quality and self-efficacy didn't have a significant impact on user's perceptions about usefulness of SSCT. Therefore hypotheses H2 and H5 were rejected.

A strong significant relationship was found between perceived service quality (β=0.581, p<0.001) and perceived ease of use. Self-efficacy also had a significant impact on user's perceptions about ease of use of SSCT (β=0.316, p<0.05). Therefore, hypotheses H3 and H6 were accepted. User's attitude toward using SSCT was also affected by perceived anonymity (β=0.298, p<0.001) and perceived usefulness (β=0.396, p<0.001). Thus, hypotheses H7 and H9 were supported. However, there was no statistically evidence regarding the impact of attitude toward innovations and perceived ease of use on user's attitude toward using SSCT. This means that in this study, the analysis did not support hypotheses H4 and H10. Therefore, hypotheses H4 and H10 were rejected.

VI. CONCLUSION

The main purpose of this study was to identify external factors which may have a significant influence on user's decision whether he or she will use the SSCT to finish the purchase. According to the literature review, this study is one of the few attempts to investigate user's perceptions about acceptance and use of an SSCT.

The results of this study showed that perceived usefulness and perceived anonymity have had a significant direct effect on attitudes toward using SSCT. However, perceived ease of use and attitude toward innovation don't influence customer's attitude toward using SSCT. There was no prof for significant influence of attitude toward using on customer's behavioral intention. However, customer's behavioral intention is predicted by perceived usefulness and anxiety, where for the latter factor a strong negative relationship was found. User's perceptions about ease of use of SSCT are influenced by self-efficacy and perceived service quality, where perceived service quality was the strongest indicator of perceived ease of use. Perceived usefulness was determined by user's perceptions about ease of use of SSCT.

The implications of the results of this study are important especially to retailers who employ SSCTs in

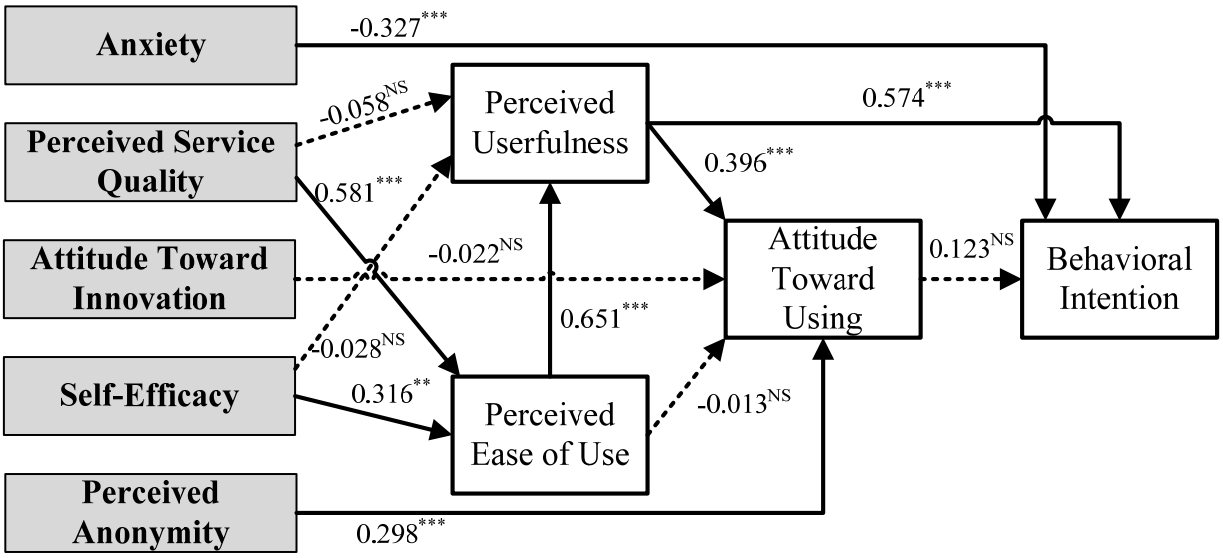


Figure 6. The research model results

their stores. Customers don't use SSCT because it is easy to use, but rather because it is useful and because it provides a certain level of anonymity when finishing the purchase process. It is also important that SSCT providers assure sufficient quality level in SSCT checkout application, because customers will have a better understanding of the application and will thus find it more useful. The customer must also have a strong belief that he or she is self-sufficient and doesn't need assistance in the checkout process when using SSCT.

As in all empirical research, this study has some limitations that need to be identified and discussed. First, the sample is limited to Slovenian customers that use self-service checkout terminals in major stores from Mercator and SPAR. Although the results from this study are useful for describing the characteristics of a large population of customers, the generalizations of the results are limited.

REFERENCES

[1] "IBM's self-checkout machine in a Mercator retail shop (Slovenia)." [Online]. Available: http://commons.wikimedia.org/wiki/File:Mercator_-_samopostre%C5%BEna_blagajna.jpg. [Accessed: 09-Feb-2014].

[2] D. Miletic, "A new way to shop — check it out for yourself - National - theage.com.au," *http://www.theage.com.au/news/national/a-new-way-to-shop-151-check-it-out-for-yourself/2008/04/21/1208742852723.html*, 22-Apr-2008. [Online]. Available: http://www.theage.com.au/news/national/a-new-way-to-shop-151-check-it-out-for-yourself/2008/04/21/1208742852723.html?page=fullpage. [Accessed: 09-Feb-2014].

[3] F. D. Davis, "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology," *MIS Q.*, vol. 13, no. 3, pp. 319–340, Sep. 1989.

[4] M. Fishbein and I. Ajzen, *Belief, attitude, intention, and behavior: an introduction to theory and research*. Addison-Wesley Pub, 1975.

[5] W. King and J. He, "A meta-analysis of the technology acceptance model," *Inf. Manage.*, vol. 43, no. 6, pp. 740–755, Sep. 2006.

[6] J. Schepers and M. Wetzels, "A meta-analysis of the technology acceptance model: Investigating subjective norm and moderation effects," *Inf. Manage.*, vol. 44, no. 1, pp. 90–103, Jan. 2007.

[7] S. Y. Yousafzai, G. R. Foxall, and J. G. Pallister, "Technology acceptance: a meta-analysis of the TAM: Part 2," *J. Model. Manag.*, vol. 2, no. 3, pp. 281–304, 2007.

[8] L. S. K. Leung and M. J. Matanda, "The impact of basic human needs on the use of retailing self-service technologies: A study of self-determination theory," *J. Retail. Consum. Serv.*, vol. 20, no. 6, pp. 549–559, Nov. 2013.

Cash Register Lines Optimization System Using RFID Technology

D.Budić, Ž. Martinović, D. Šimunić
Faculty of Electrical Engineering and Computing / Department of Wireless Communications
Unska 3, 10000 Zagreb, Croatia
E-mail: dora.budic@fer.hr, zeljko.martinovic@fer.hr, dina.simunic@fer.hr

By doing business the old way, retail chains don't use all the potential of developed technologies in order to increase their efficiency. Cash register lines optimization system uses RFID (Radio Frequency Identification) technology to follow the number of shopping carts and baskets in front of cash register. Developed system consists of RFID reader and passive RFID tags with assigned ID. System works on ultra-high frequency (UHF, 868 MHz) with long range (up to 15m) which enables development of protocols using simultaneous communication between reader and tags. One reader can collect data from 8 antennas at the same time. Concept of following shopping carts and baskets using RFID is based on backscatter coupling which enables powering RFID tags with energy from magnetic field of the reader. System uses SQLite3 database and Python 2.7 programming language. Web application is written using JavaScript jQuery framework and HTML/CSS Twitter Bootstrap framework. Also, for sending SMS, mail2sms technology is used. JSON via HTTP protocol is used for communication between reader and server and between server and web application. In order to inform the store manager, SMS is generated. Web application is used for tracking the whole system. Statistics integrated in web application enable store managers to organize employees in advance considering observations from past periods.

I. INTRODUCTION

More and more retail chains today have the need for automated monitoring of the number of customers in cash register lines in real time. Previously developed systems have proved to be insufficiently effective because they involved additional and unpleasant interaction between customers and employees or monitoring in the way that threatens the privacy of customers. To avoid such problems, it was necessary to design different solution which would implement new technologies.

There are many different jobs that employees can do: working in bakery department, cash register, placing the products on the shelves etc. Store managers have the task to distribute employees in the most efficient way. The decision about distribution can be made on the basis of real-time status analysis, statistical analysis and combination of the two methods. Real-time distribution means that the store manager has the data based on the real-time status of the store, and statistical analysis can assume what the distribution should be based on the previous observations. Combination of the two methods provides better precision because it considers previous observations and real-time status analysis.

System has the task to collect and analyze data that can be used for concluding how many customers wait in the cash register lines. In case where system concludes that there are more customers that it can be handled in optionally given time frame, it is necessary to warn the store manager that conditions are met where it is needed to redistribute workers in the way that more people work on cash registers. Also, store manager is warned when too few customers are in the lines so some cash registers should be closed and extra workers redistributed on other jobs. All collected data is permanently stored in the database. In every moment it can be used to offer insight into real-time status and statistical analysis status [1].

Technology that makes development of system like this possible is RFID.

II. SYSTEM CONCEPT

RFID is the wireless non-contact use of radio-frequency electromagnetic fields to transfer data, for the purposes of automatically identifying and tracking tags attached to objects. Every RFID system consists of several basic elements:

- Reader



Figure 1: RFID reader Motorola FX9500

- Antenna and tags

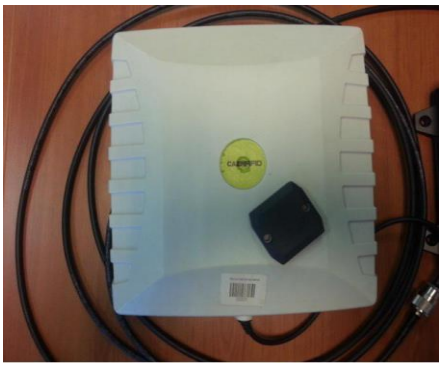


Figure 2: RFID antenna and tag

RFID systems differ in operating frequency, coupling, range, memory type and battery presence in the tag.

Operating frequency is the one that reader uses for impulsion. The different transmission frequencies are classified into the three basic ranges:

- Low frequency (LF) from 30 to 300 kHz
- High frequency / radio frequency (HF/RF) from 3 to 30 MHz
- Ultra-high frequency (UHF) from 300 MHz to 3 GHz / microwave > 3 GHz

A further subdivision of RFID systems according to range allows us to differentiate between close-coupling (0-1 cm), remote-coupling (0-1 m) and long-range (>1 m) systems.

Close-coupling systems are coupled using both electric and magnetic fields and does not rely upon the radiation of fields because of closeness of the tag. Close-coupling systems are primarily used in applications that are subject to strict security requirements, but do not require a large range.

Almost all remote coupled systems are based upon an inductive (magnetic) coupling between reader and transponder. These systems are therefore also known as inductive radio systems.

All long-range systems operate using electromagnetic waves in the UHF and microwave range. The vast majority of such systems are also known as backscatter systems due to their physical operating principle. In addition, there are also long-range systems using surface acoustic wave transponders in the microwave range. Ranges of 15 m and above can be achieved using active (battery-supported) backscatter transponders.

Tags can be passive or active considering the battery presence. Passive tags don't have any power source (battery). Electromagnetic field of the reader provides all the energy necessary for tag to work via tag antenna. This means that the emitted energy of the reader is used to transfer data from reader to tag and from tag to reader. Active tags have their own power source (battery or solar cell) which is used for powering the chip. This means that electromagnetic field of the reader isn't necessary for powering like it is necessary with the passive tags. This feature enables using the reader with weaker field to transfer data. Tags also vary in memory capacity. 1-bit tags have only one bit of memory that stores 1 if the tag is in the field and 0 if tag isn't in the field. Other tag type is the one with ID stored in the memory by the factory. Third type is the one that has enough memory for storing ID and extra data (by the user). Tags are produced in plastic and glass housing in shape of coin, sticker, capsule, little box and card [2].
RFID uses the low-end of the electromagnetic spectrum and signal strenght is limited

The RFID UHF signal strenght is defined under ISO/IEC 18000-6:2004standards. Signal strenght can be in range mW up to several mW and therefor can not represent danger for human health.

III. SYSTEM ARCHITECTURE

The idea is to put readers and antennas on cash registers and tags on shopping carts and baskets. When a tag on the shopping cart or basket is read as present for a certain period of time in front of cash register, system recognizes that the customer is standing in that line. This way it can be determined how long is the line in front of cash register and in case that it is too long, system sends SMS to store manager that more cash registers should be open. Also, if the lines are too short, store manager gets information that some case registers should be closed.

We wanted to make the system as simple as possible to modify for additional demands and to integrate in existing store system. Because of that, the whole system is divided into three separate entities:

- *Pinkie* – part of the system in charge of reading ID from tags
- *Brain* – part of the system that stores collected data into database and processes them to become data useful to store manager
- *Protocol* – part of the system that enables communication between Pinkie and Brain parts

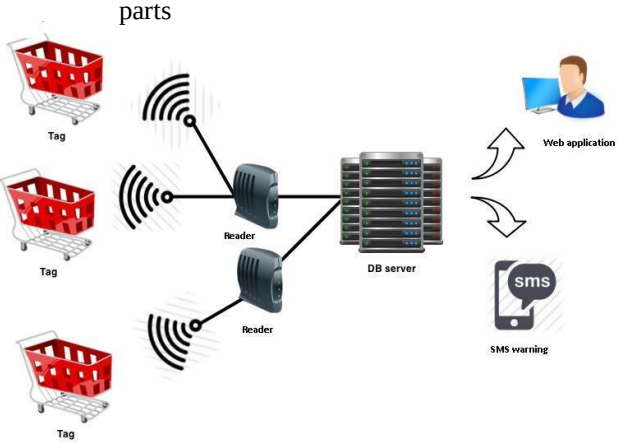


Figure 3: Shema of the system

A. Pinkie

Pinkie presents part of the system in charge of RFID technology. When a cash register is opened, antenna reading is automatically turned on in front of that cash register. Every reading first comes to the reader that has embedded software. That software enables data manipulation using modified built-in functions that are required for correct data reading.

Reader can be configured to send signals to the tags at specified time intervals. Constant reading ensures greater accuracy, but creates data redundancy. For this reason, reader functions can be modified in the way that the rest of the system gets an update only if there is a change in the system. Eg., a particular tag enters the antenna field so the reader reads its ID and sends the information to the rest of the system because the change in the system occurred. Since we do not want to read data every second of the time that tag remains in the field of antenna, a modified function sends an alert only if the tag gets out of the field of the antenna. Additional functionality is reading multiple tags at the same time and ignoring tag readings that are in the antenna field too short.

For the purpose of testing the developed software solution for optimization of cash register lines using RFID technology, we used Motorola FX9500 fixed RFID reader (868 MHz) that supports .NET and C as API support [3]. Also, passive tags in thicker plastic housing that have ID memory have been chosen (due to better work with metal surfaces). Used antenna is circularly polarized (dimension 24.5x23.5 cm).

B. Brain

Brain is in charge of storing, processing and displaying of received data. Data sent by Pinkie is stored in the database. Used database is Open Source SQLite3. To make collected data visible to store managers, we developed web application that can run on mobile phone, computer and tablet. Web application enables a graphical representation of the real-time status and status in selected period of time (so the statistical analysis can be performed). Also, web application (whose interface has automatic update every few minutes) sends SMS to store manager in the need for redistribution of employees. Data processing is written in Python 2.7 programming language. Web application is created using JavaScript jQuery and HTML/CSS Twitter Bootstrap framework that gives functionality like buttons or drop-down menu to the application. Sending of SMS is implemented using mail2sms technology.

C. Protocol

To make parts of the system communicate with each other, it is necessary to implement communication protocol. For this system we used sending of JSON (JavaScript Object Notation) through HTTP protocol. JSON is an open standard format that uses human-readable text to transmit data objects consisting of attribute-value pairs. It is used primarily to transmit data between a server and web application.

HTTP protocol is an application protocol that defines format (JSON in this case) and the way of message transportation between client and server.

D. Results and additional functionalities

Figures 4, 5 and 6 show results of developed system:

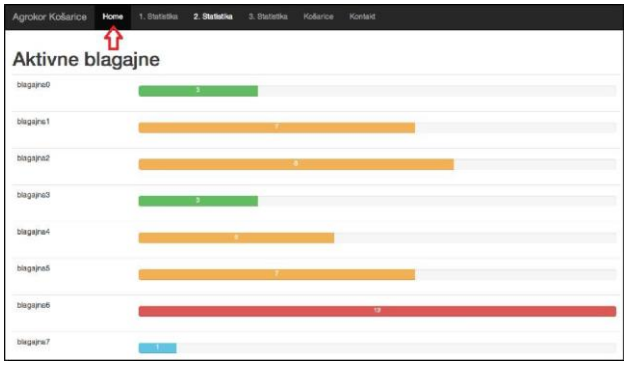


Figure 4: Web application

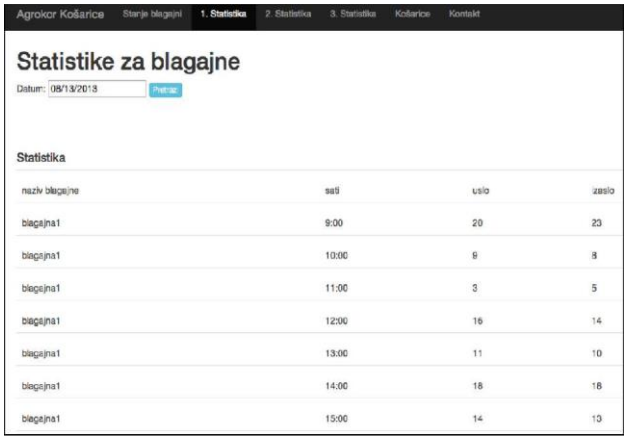


Figure 5: Cash register statistics



Figure 6: Real-time cash register state

Additional implemented functionality is basket counter. In front of every cash register stands a place where used baskets are collected. System offers automated monitoring of number of baskets to warn a store manager that collected baskets have to be removed. Functionality is developed the same way as the rest of the system.

IV. CONCLUSION

Cash register lines optimization system using RFID offers automated monitoring of number of customers in cash register lines without uncomfortable interaction between customers and employees of the store and also without threatening customers privacy. The whole system is designed to be easily upgradable with many additional features and integrated with already existing solutions in stores. All the newest technologies in the field of RFID wireless communications and also web applications have been used.

For further development, we recommend using active tags to reach an optimal range of 12 m. Also, for better communication between tag and the reader, we recommend using narrow radiation pattern antenna.

LITERATURE

[1] D. Budić, „Cash register lines optimisation – Documentation“, unpublished

[2] K.Finkenzeller, „RFID Handbook“, 3rd ed., John Wiley & Sons, Ltd., 2010., pp. 13-28.

[3] “Product Spec Sheet – Motorola FX9500 Fixed RFID Reader“

[4] J.P.Curty, M. Declercq, C.Dehollain, N.Joehl, „Design and optimization of passive UHF RFID systems“, Springer 2007, New York,SAD.

Pozicija BH Telecoma d.d. Sarajevo prema Adižesovoj teoriji životnih ciklusa preduzeća

Adnan Gabela*, Lejla Pamuk ** i Jasmin Gabela***
* BH Telecom d.d. Sarajevo, Bosna i Hercegovina
** BH Telecom d.d. Sarajevo, Bosna i Hercegovina
***USAID, Sarajevo, Bosna i Hercegovina

I. UVOD

Jedan od vodećih eksperata iz oblasti menadžmenta i ekonomije **dr. Isak Adižes** svoju teoriju "Životnog ciklusa preduzeća" i principe kako upravljati promjenama u kompaniji razvija i sprovodi posljednjih trideset godina. Njegova teorija i principi daju objašnjenje zašto kompanije rastu, stare i umiru, i šta se povodom toga može uraditi. On opisuje i analizira uobičajen put kojim kompanije prolaze dok rastu i optimalan put kojim bi trebalo da idu, da bi izbjegle probleme karakteristične za rast i starenje.[1]

Kao i kod živih organizama, rast i starenje privrednih društava se manifestuje u međusobnom odnosu dva faktora – **fleksibilnosti** i mogućnosti **kontrolisanja** procesa. Kad su organizacije mlade, one su veoma fleksibilne, ali se ne mogu uvijek kontrolirati. Kako organizacije stare, odnos se mijenja. Mogućnost kontrole raste, a fleksibilnost opada. Veličina i vrijeme nisu uzroci rasta i starenja, pa ne možemo reći da su velike kompanije sa dugom tradicijom po pravilu stare, a male kompanije bez tradicije mlade. Postoje stogodišnje mylade kompanije i petogodišnje stare (treba li ovo?).

Nijedan sistem - ljudski, organizacioni, tehnički - ne može istovremeno biti i maksimalno **fleksibilan** i maksimalno sposoban da održava kontrolu nad procesima. Stadijum kada organizacija ima istovremeno najveću moguću **fleksibilnost** i sposobnost **kontrole** nazivamo **„top forma“** i to je stanje koje bi svaka organizacija trebala nastojati postići. U svakoj fazi organizacija je suočena sa normalnim i patološkim problemima, karakterističnim za tu fazu. Organizacija **„top formu“** može postići tipičnim putem - put rasta i starenja - popraćen svim standardnim problemima u pojedinim fazama ili optimalan putem – bržim i putem bez problema karakterističnih za tipični put.

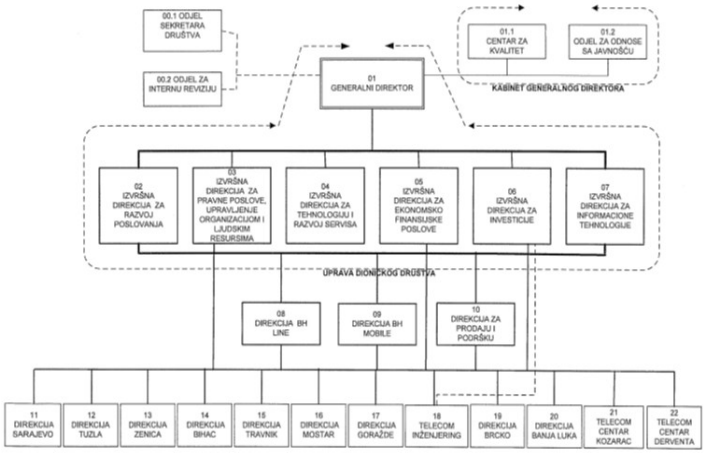
[1]

II. ORGANIZACIJA BH TELECOMA

A. *Osnovni podaci i organizacija BH Telecoma*

BH Telecom d.d. Sarajevo će ostati vodeći telekom operator u Bosni i Hercegovini koji će stalnim razvojem novih usluga i kvalitetom istih odgovoriti na zahtjeve tržišta, te udovoljiti potrebama korisnika i više od njihovog očekivanja.

Organizaciona shema BH Telecoma



B. *Pozicija BH Telecoma prema teoriji životnih ciklusa privrednog društva*

Prema navedenoj teoriji, stanje BH Telecoma je **„aristokratija“**, **fleksibilnost** 44/100, **kontrolisanost** 83/100 (graf na stranici ispod). Ovu fazu karakteriše relativno nizak stepen **fleksibilnosti**, a relativno visok stepen **kontrolisanosti** kompanije. Karakteristike ove faze životnog ciklusa, koje se konkretno mogu primjetiti u BH Telecomu su:

- Sve manja očekivanja za rast;
- Pažnja se usmjeravaja na dostignuća iz prošlosti;
- Sumnjičavost na promjene;

- Nagrađuju se oni koji rade ono što im se kaže;
- Novac se troši na kontrolne sisteme, nagrade i uređenje prostorija;
- Veća briga o tome kako se radi nego o tome šta i zašto se radi;
- Zapošljavaju se pojedinci koji brinu o vitalnosti kompanije, ali radni moto je: "ne talasaj";
- Zanemarljivo malo unutrašnjih inovacija – Kupuju se tuđa rješenja da bi se došlo do novih proizvoda, tržišta i ideja;
- Mnogo gotovine, potencijalna meta za nove vlasnike;
- Komuniciranje – pismena forma, niko otvoreno ne govori o problemima;
- Zaposleni žele što manje sukoba i što manje promjena;
- Promjene se svode na najmanju moguću mjeru;
- Opada proces reagovanja na kratkoročne potrebe, jer se zanemaruju dugoroče mogućnosti;
- Organizaciona klima je ustajala;
- Ne priznaje se stvarnost;
- Gubi se tržište i sve je teže održati konkurentnost sa svojim proizvodima i uslugama;
- „Dividende se moraju raspodjeliti";
- U organizaciji preovladava osjećaj sveopšteg zadovoljstva.

C. *Opis trenutnog stanja i obrazloženje zašto se BH Telecom nalazi u fazi **aristokratije***

Postoji više razloga za navedenu situaciju, a jedan od najizraženijih je prevelika administracija i jako duboka linija odlučivanja i na horizontalnom i na vertikalnom nivou. Osim toga BH Telecom je kao javno preduzeće dužan poslovati u skladu sa Zakonom o javnim nabavkama BiH, čime se ionako veliki i komplikovani način nabavke i odlučivanja dodatno usložnjava. Samim time se BH Telecom stavlja u nepovoljniji položaj u odnosu na konkurente koji ne podliježu zakonu. BH Telecom je predmet konstatnih kontrola poslovanja od strane Regulatorne agencije za komunkacije (RAK), Uprave za indiraktno oporezivanje, Ombdusmana, Federalne inspekcije, eksterne revizije, interne revizije, internog audita i drugih. Zbog svoje specifičnosti i činjenice da je gotovo neprekidno predmet kontrola po raznim osnovama BH Telecom je ograničen da djeluje brzo i dužan je pratiti gotovo beskonačan broj pravila i procedura. Prateći trendove u svijetu RAK konstantno nastoji da tržište telekomunikacija učini konkurentnijim, pružajući alternativnim operatorima povoljnije uslove u odnosu na dominante, što dodatno otežava poslovanje BH Telecoma. Ipak i pored brojnih navedenih kontrola i inspekcija, jako je teško u organizaciji od skoro 3.500 zaposlenih izmjeriti produktivnost zaposlenika i izvršiti adekvatnu evaluaciju složenosti radnog mjesta, zbog čega se javljaju problemi u motivisanosti, manjak produktivnosti, te činjenica da neki zaposlenici odrađuju mnogo, a neki vrlo malo posla.

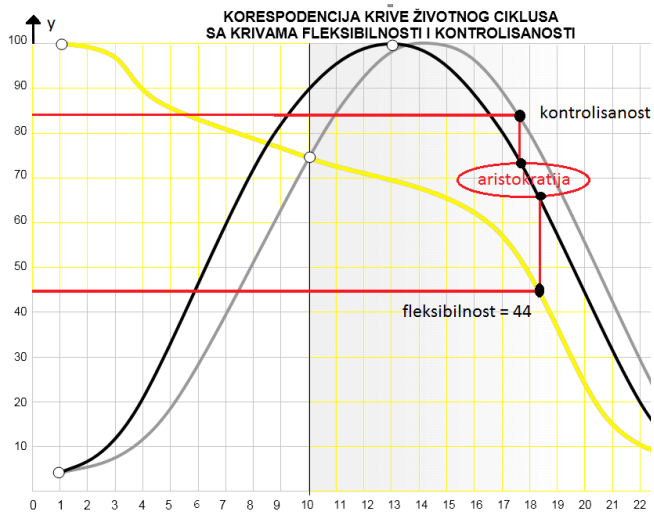
Drugi razlog za sporost u odlučivanju i provođenju odluka je sama fizička odvojenost organizacionih djelova koji moraju da sarađuju u velikom broju projekata. Trenutnom organizacijom ID za razvoj poslovanja, FD za prodaju i podršku i ID za informacione tehnologije se nalaze u jednoj zgradi, ID za ekonomsko finansijske poslove i ID za pravne poslove, upravljanje organizaciom i ljudskim resursima u drugoj, a ID za tehnologiju i razvoj servisa i ID za investicije u trećoj. Zbog činjenice da svi dokumenti u BH Telecomu cirkulišu u hard copy (papirnatoj) formi, sam protok dokumenata između različitih organizacionih jedinica oduzima jako puno vremena. Kada se na sve to doda i nejasan raspored odgovornosti i ovlaštenja između direkcije, te generalne klime bježanja od istih, očito je da zbog svojih internih problema, tačnije prevelikog i tromog administrativnog aparata organizacija ne djeluje na optimalnom nivou produktivnosti.

Kao primjer navedene problema sa prevelikom adiminstracijom prikazujemo put fakture za proviziju posrednicima koji vrše prodaju eDopuna BH Telecoma. Od momenta prijema fakture na protokolu, faktura prolazi 27 koraka u različitim organizacionim jedinicama do momenta plaćanja:

Faktura se prima na protokolu
Faktura se šalje generalnom direktoru na signaturu
Faktura se vraća na protokol
Saradnik/operator iz Službe za platni promet vrši prijem i parkiranje fakture
Saradnik iz Službe za kontrolu prihoda vrši preuzimanje fakture i provjeru ispravnosti
Faxom se faktura šalje u F Direkciju za prodaju i podršku
Saradnik iz Službe za UP pri FD PiP preuzima fakturu i sačinjava obrazac 6.
Obrazac 6. potpisuje šef Službe za upravljanje plasmanom RK
Obrazac 6. potpisuje rukovodilac Sektora za prodaju
Obrazac 6. potpisuje direktor Direkcije za prodaju i podršku
Obrazac 6. se šalje redovnom poštom prema ID za EFP
Obrazac 6. se prima na protokolu
Obrazac 6. ide na signaturu kod izvršnog direktora ID za EFP
Obrazac 6. ide na signaturu kod rukovodioca Sektora za finansijske poslove
Obrazac 6. ide na signaturu kod šefa Službe za kontrolu prihoda
Saradnik iz Službe za kontrolu prihoda vrši prijem obrasca
Saradnik iz Službe za kontrolu prihoda šalje fakturu na plaćanje
Šef Službe za kontrolu prihoda potpisuje fakturu da je saglasan da se šalje na plaćanje
Saradnik/oprator iz Službe za platni promet vrši prijem i likvidaturu fakture
Fakturu, da je saglasan da ide na plaćanje, potpisuje šef Službe za platni promet
Fakturu, da je saglasan da ide na plaćanje, potpisuje rukovodilac Sektora za finansijske poslove
Fakturu, da je saglasan da ide na plaćanje, potpisuje izvršni direktor ID za EFP
Faktura se vraća saradniku/operatoru iz Službe za platni promet
Original fakture se šalje prema Sektoru za ekonomske poslove, knjigovodstvo i kontroling
Saradnik/operator iz Službe za računovodstvo preuzima fakturu
Saradnik/operator iz Službe za računovodstvo vrši uknjiženje fakture u SAP
Saradnik/oprator iz Službe za platni promet vrši plaćanje fakture

Osim gore pomenutih ograničenja, telekomunikacioni operateri generalno su suočeni sa novim konkurentima, kao naprimer sve većim rastom, i masovnom upotrebom alternativnih načina komunikacije i prenosa podataka (Facebook, Twitter, Viber, WhatsApp, Skype), te nametnutim smanjenjima cijena saobraćaja. Iz gore navedenih razloga je očita nemogućnost brzog djelovanja na tržištu i prilagođavanja promjenama, čime kompanije

ubrzano gube efikasnost i efektivnost, te ubrzano stare po **Adižesu**.



D. *Znakovi da kompanija ide u korak sa vremenom i trendovima*

Uprkos gore navedenom očiti su pokazatelji da kompanija prati trendove oblasti TK industrije.

Godine 2009., BH Telecom je pružao osnovne telekomunkacione usluge, uz mogućnost kupovine telefona na rate za pretplatnike. Jedini način naplate tih računa je bio u JP BH Pošta i bankama.

Samo u zadnje četiri godine, kompanija je uvela ili je u fazi uvođenja sljedećih projekata: naplata računa u vlastitim Telecom centrima, putem POS terminala, SMSom, online, dostava računa online. Pored prodaje telefona na rate, u istom periodu je omogućena ili će uskoro biti prodaja laptopa, tableta, televizora, igračkih konzola i dodataka za sve navedene uređaje. Trenutno je pokrenut projekat i za razvoj m2m rješenja sa partnerima.

Pokrenut je proces bolje saradnje sa bankama, čime je omogućeno plaćanje računa u prostorijama BH Telecoma putem POS aparata, putem SMSa i za neke usluge i online (planirano je da se omogući i plaćanje svih računa onlina). Ovim inovacijama BH Telecom je značajno smanjio troškove vezane za naplatu računa i omogućio veći komoditet svojim pretplatnicima (mogućnost preferiranog izbora naplate, eliminisanje čekanja u redovima i sl.)

BH Telecom je omogućio i dostavu računa korisnicima elektronskim putem (e-mailom), što se radi za sada samo na zahtjev korisnika. Dostava računa ovim putem značajno smanjuje troškove dostave, ali zbog nesprovođenja Zakona o elektronskom potpisu, BH Telecom nema mogućnost prelaska dostave svih računa na ovaj način, nego to može provesti isključivo na zahtjev korisnika.

U toku je i projekat izdavanja jedinstvenog računa za sve usluge BH Telecoma, čime bi se još više smanjili troškovi BH Telecoma, a povećalo zadovoljstvo korisnika.

BH Telecom nastoji iskoristiti svoju široku prodajnu mrežu i veliki broj korisnika te u saradnji sa partnerima plasirati prodaju tk i srodnih uređaja. BH Telecom veliki broj aparata nastoji plasirati komisionom prodajom, čime se koriste kanali distribucije partnera, njegova skladišta, a BH Telecom unosi široku prodajnu mrežu, veliki broj korisnika i dobre mehanizme fakturisanja i naplate.

Trenutno je u fazi projekat za razvoj m2m (Machine to machine) rješenja sa partnerima u oblastima: javni prevoz, zdravstvo, nadzor vozila, nadzor ljudi, sigurnost i metering javnih usluga. Ovaj projekat predstavlja ulazak BH Telecoma u potpuno nove oblasti, gdje sa svojim specifičnim karakteristikama postoji mogućnost za zauzimanje novog segmenta tržišta. Implementacijom navedenih m2m rješenja, te planiranim razvojem nfc rješenja (Near Field Communication) BH Telecom prati trendove u oblasti telekomunikacione industrije i pozicionira se kao lider u domenu pružanja novih usluga.

E. *Prijedlozi i rješenja za povratak kompanije u stanje **„top forme“***

- Rad u timovima sa posebnim naglaskom na nove tehnologije

Najveći dio navedenih projekata je uveden od strane stručnih timova, koji su imenovani od svojih organizacionih jedinica i imali su zadatak da osmisle, razviju i implementiraju konkretno rješenje. Članovi tima su na osnovu svoje kompetentnosti dobili određen stepen autonomije za svoj rad i zajedničko djelovanje u cilju ostvarenja zadatka. Svaki član je definisao i sproveo zadatke ispred svoje organizacione jedinice, ali kao dio tima imao je ovlasti da u nekim segmentima djeluje i van svoje organizacione jedinice u cilju implementacije zadatka.

U skladu sa **Adižesovom** metodologijom, rad u komplementarnim PAEI timovima i predstavlja najbolje rješenje za djelovanje i u konkretnoj situaciji predstavlja jedan od preduslova za povratak kompanije u **„top formu“**. Budući da BH Telecom djeluje u oblastima koje su konstatno podložne promjenama i inovacijama, potrebno je povećati stepen **fleksibilnosti**, a smanjiti stepen **kontrolisanosti**, kako bi se kompanije pozicionirala i održala u stadiju **„top forme“**. [2]

Rad ovih timova svakako treba podsticati, ali je jako bitno izabrati članove koji imaju znanje, volju za radom i sticanjem novih znanja i spremnost na rad u timu. U idealnim uslovima članovi tima bi trebali da formiraju komplementaran PAEI tim, ali zbog činjenice da svi aspekti poslovanja moraju biti prisutni, kao i nastojanja da timovi nemaju pretjeran broj članova (zbog gubitka produktivnosti), komplementarnost ne može uvijek nužno biti ispunjena. Po prirodi svog posla, predstavnici ID za

[2]

RP osmišljavaju uslugu i daju koncept (E), tehnika (BH Line, BH Mobile) i prodaja (FD za PiP) pružaju uslugu (P), a ID za EFP, ID za PPUOiLJR i ID za IT sistematizuju i institucionališu uslugu (A), a Uprava BH Telecoma u saradnji sa ID za RP integriše tim (I).

- Pojednostavljenje administrativnog aprata i eliminisanje internih ograničenja

BH Telecom je u prvoj polovini 2014. godine planirao preseljenje u novu zgradu, gdje bi se na istom mjestu nalazile organizacione jedinice koje direktno određuju strategiju, čime bi se značajno izvršila ušteda na vremenu. Još jedan dobar način uštede na vremenu je uvođenje novog sistema protokola, kojim bi se svi dokumenti zavodili elektronski u jedinstven registar, a dalje cirkulisali unutar organizacije elektronskim putem.

- Povećati motivaciju zaposlenih

Zbog prirode svoje djelatnosti, BH Telecom je primoran konstantno kupovati i poboljšavati svoju infrastrukturu i softwarsku podršku. Definitivno postoji opravdanost kupivine velikog broja platformi, stanica, licenci, programa i aplikacija, ali postoje i brojna rješenja koja se trenutno kupuju od strane softwareskih kuća, iako BH Telecom zapošljava veliki broj informatičara i TK inžinjera koji su u stanju razviti određene aplikacije. Tako je naprimjer BH Telecom kao jedan od prvih u Bosni i Hercegovini kupio knjigovodstveni program SAP, koji je dokazan kao najbolji u svijetu i opravdanost takve kupovine ne može biti dovedena u pitanje. S druge strane kupljena su brojna rješenja za aplikacije koje su mogle biti razvijene interno. Razvoj programa i mogućnost poboljšanja kompanije iznutra, može se ostvariti uvid u znanje i sposobnosti svojih vlastitih kadrova čime bi se značajno poboljšala motivisanost zaposlenih.[3]

Osim toga primjetna je činjenica da u većini timova za razvoj novih usluga i tehnologija učestvuju skoro isti ljudi. Članstvo u ovim timovima, samo po sebi dokazuje kompetentnost i sposobnost, te budi motivisanost zaposlenika, ali budući da se radi o krucijalnim projektima BH Telecoma, potrebno je izvršiti detaljniju selekciju članova po projektima i program motivacije i nagrađivanja u ovisnosti o uspjehu projekta.

- Vršiti konstatne obuke zaposlenika koji su u direktnom kontaktu sa kupcima

U svakoj kompaniji, posebna pažnja se treba posvetiti zaposlenim koji ostvarju direktne kontakte sa kupcima. U BH Telecom bi se to posebno odnosilo na prodajno osoblje i podršku korisnicima (call centri, monteri). S obzirom da se često slika o organizaciji i percepcija o proizvodi i kvaliteti kreira putem odnosa sa osobljem, potrebno je konsatno provoditi posebne treninge, radionice i seminare za navedene zaposlenike.

- Poboljšanje „last mile“ infrastrukture

Bakarne parice, na osnovu kojih maloprodajni pretplatnici ostvaruju povezanost sa tk infrastrukturom su u Bosni i Hercegovini često u jako lošem stanju, najčešće zbog starosti i istrošenosti. Sve brzine koje BH Telecom garantuje (po osnovu ADSL, VOIP, IPTV...) se odnose na brzinu do pristupnih tačaka, i u glavnom dijelu infrastrukture te brzine su na zadovoljavajućem nivou. Problem se međutim javlja u „last mile“ infrastrukturi, koja vodi do pretplatnikovog doma.[4] Ovaj dio infrastrukture nije u vlasništvu BH Telecoma, nego krajnjih pretplatnika (vlasnika zgrada i sl.), ali s obzirom da kvalitet „last mile“ infrastrukture direktno utiče na kvalitet usluga potrebno je učiniti više napora i iznaći rješenja za popravak ovog dijela infrastrukture za sve pretplatnike.

III. Zaključak

Na osnovu navedenog, zaključuje se da postoji mogućnost povratka kompanije u stanje **„top forme”**, ali da je za realizaciju tog cilja neophodan angažman zaposlenih, kako na strateškom tako i na operativnom nivou. S obzirom na specifičnosti industrije telekomunikacija i veliki broj inovacija i promjena u ovom području, sljedećih nekoliko godina bi moglo biti ključno za razvoj i održivost kompanije. Povećanjem **fleksibilnosti**, a smanjivanjem **kontrolisanosti**, kompanija bi uspjela da se pozicionira na željeno mjesto, uz pretpostavku da se provođenje procesa provodi uz visok nivo posvećenosti i kompetentnosti, što je optimlano realizovati kroz rad komplementarnih timova, te poštivanje ostalih tačaka navedenih ovim radom.

Literatura

[1] Adizes, I. Corporate Lifecycles: How and Why Corporations Grow and Die and What To Do About It. Englewood Cliffs, NJ: Prentice Hall, 361 pp. (1988).

[2] Adizes, I. The Ideal Executive: Why You Cannot Be One and What to Do About It. Santa Barbara, CA: The Adizes Institute Publications (2004).

[3] Crandall, W. Robert , Competition and Chaos: U.S. Telecommunications Since the 1996 Telecom ACT (Paperback) , (2005).

[4] Gertner, Jon The Idea Factory: Bell Labs and the Great Age of American Innovation (Hardcover) (2012.)

[5] Crawford, Sudan, Captive Audience: The Telecom Industry and Monopoly Power in the New Gilded Age (Hardcover) (2013.)

[6] Sesia, Stefania, LTE, The UMTS Long Term Evolution: From Theory to Practice (Hardcover) (2009.)

[7] www.bhtelecom.ba

3

4

Unaprjeđenje upravljanja resursima uz pomoć sustava za podršku

Darko Markulin, M.Sc., Kristina Musa, M.Sc.
Odjel za usluge i poslovanje s korisnicima
HT d.d.
Kupska 2, 10000 Zagreb, Hrvatska
Telefon: (385) 1-4912 721 Fax: (385) 1-4917 777
E-mail: darko.markulin@t.ht.hr, kristina.musa@t.ht.hr

Sažetak - Potrebne akcije na strani pristupne mreže telekomunikacijskog operatora tijekom realizacije usluga u velikoj mjeri utječu na ukupne troškove operatora potrebne za realizaciju usluge te isto tako imaju veliki utjecaj na ispunjavanje postavljenih ciljeva u vidu KPI-ova (ključnih pokazatelja uspješnosti). U velikoj mjeri akcije na strani pristupne mreže izvršavaju zaposlenici operatora ili vanjskih izvođača. Upravo zbog toga postupak automatske dodjele zadataka za realizaciju ovih akcija može znatno doprinijeti uštedama i efikasnijoj realizaciji usluga.

U sklopu automatske dodjele zadataka potrebno je osigurati optimalan put zaposlenika koji izvršavaju akcije, osigurati zadovoljavanje definiranih poslovnih uvjeta s vanjskim izvođačima, osigurati realizaciju usluga unutar zadanih KPI-ova te osigurati realizaciju usluga u vremenskim rokovima dogovorenim s korisnicima.

U radu su prikazani osnovni principi i postavke koje treba zadovoljiti automatska dodjela zadataka u sklopu dodjele zadataka na području pristupne mreže koje se primjenjuje u HT-u. Ne ulazi se u opis samog algoritma dodjele zadataka.

Ključne riječi: automatsko dodjeljavanje zadataka, poslovni procesi, usluga, resursi, sustavi podrške

I. UVOD

Dvije su osnovne komponente koje u znatnoj mjeri utječu na uspješnu realizaciju telekomunikacijske usluge i optimalne troškove. To su tijek poslova (workflow) i dodjela zadataka (workforce).

Tijek poslova definira potrebne akcije vezane uz konfiguraciju mrežnih elemenata i definiranje potrebnih zadataka za akcije koje akteri poslovnog procesa trebaju poduzeti u cilju realizacije usluga. Ispravnim i optimalnim kreiranjem tijeka akcija osigurava se uspješna realizacija usluga što je vrlo bitno kod složene realizacije više usluga ili poslovnih procesa. Veliki utjecaj na uspješnost realizacije i smanjenje troškova u ovoj komponenti poslovnog procesa ima automatizacija akcija vezanih uz konfiguraciju mrežnih elemenata i sustava za podršku [1], [2].

S druge strane određene akcije tijekom poslovnog procesa ne mogu se automatizirati i njih moraju izvršavati akteri poslovnog procesa. Primjeri za to su ožičenje na razdjelniku na kojem se pristupna mreža spaja s mrežnim elementima telekomunikacijske mreže te konfiguracija usluge na strani korisnika. Za realizaciju tih usluga akter poslovnog procesa treba utrošiti određeni dio svog radnog vremena na putovanje i realizaciju usluge. Ovo znatno utječe na ukupne troškove realizacije usluge. S druge strane potrebno je paziti na vrijeme realizacije usluge. Različiti zadaci imaju i različitu važnost u odnosu na realizaciju usluge. Pravovremena realizacija najvažnijih zadataka mogu također znatno utjecati na ukupne troškove telekomunikacijskog operatora vezano uz procese realizacije usluga. Primjer za to su realizacije reguliranih usluga gdje kašnjenje u realizaciji može uzrokovati znatne troškove kroz različite penale koje operator zbog kašnjenja mora plaćati.

Stoga se zbog svega navedenog nameće potreba za automatskom dodjelom zadataka kojom će se osigurati optimalna dodjela zadataka. Na taj se način žele smanjiti troškovi realizacije usluga, povećati efikasnost aktera poslovnog procesa i povećati zadovoljstvo korisnika. Sve to zajedno doprinosi poboljšanju poslovnog procesa realizacije telekomunikacijskih usluga.

U nastavku rada opisani su osnovni principi i postavke automatske dodjele zadataka čijom se primjenom u poslovnim procesima realizacije usluga u HT-u želi znatno doprinijeti poboljšanju realizacije usluga i smanjenju ukupnih troškova realizacije. Opis i definicija samog algoritma dodjele zadataka na temelju kojeg se stvoreni zadaci dodjeljuju na aktere poslovnog procesa ne razmatra se u ovom radu. U radu se ne razmatra niti način prikazivanja dodijeljenih zadataka. Ipak, važno je napomenuti da mogućnost pregleda i dodjele zadataka kroz kartu područja na kojem se dodjeljuju zadaci dodatno doprinosi uspješnoj dodjeli zadataka i njezinom praćenju.

II. OSNOVNI PRINCIPI AUTOMATSKE DODJELE (DISPEČIRANJA) ZADATAKA

Prilikom specifikacije automatske dodjele zadataka nameće se potreba za definiranjem početnih uvjeta koje je potrebno zadovoljiti prilikom dodjele zadataka. Ti uvjeti ovise o poslovnim pravilima koja vrijede kod operatora za kojeg se specificira automatska dodjela zadataka i zapravo znače dodatne parametre koje je potrebno uzeti u obzir prilikom automatske dodjele zadataka. Kod primjene automatske dodjele zadataka u pristupnoj mreži HT-a potrebno je osigurati:

A. Dodjela zadataka na vanjske izvođače

Prilikom dodjele zadataka na vanjskog izvođača potrebno je više pravila. Neka od tih pravila su:

- Na određenim područjima svi se zadaci dodjeljuju na vanjskog izvođača (npr. otoci),
- Na ostalim područjima definiran je postotak zadataka koji se dnevno dodjeljuje na vanjske izvođače,
- Potrebno je zadovoljiti uvjete dogovorenih poslovnih pravila s vanjskim izvođačima koje se tiču predviđenog vremenskog roka za pojedini vrstu realizacije. Prilikom automatske dodjele zadataka potrebno je voditi računa o količini naloga s određenom vrstom realizacije jer se ona razlikuju po iznosu plaćanja, pa je tu moguć velik utjecaj na ukupne troškove realizacije usluge,
- Definira se određeni postotak naloga koji se dodjeljuje na vanjskog izvođača, a ostatak se dodjeljuje na interne izvođače operatora,
- Određeni se zadaci odmah dodjeljuju na vanjskog izvođača bez utjecaja kriterija koje odnosi algoritam. Primjer za to je usluga koju s korisnikom ugovara vanjski izvođač,
- Ukoliko realizacija zadatka zahtjeva ulazak u prostorije operatora, a vanjski izvođač nema dozvolu samostalnog ulaska u nju zadatak se ne može automatski dodijeliti na vanjskog izvođača.

B. Utjecaj željenog vremena

U slučaju kada je potrebno uslugu realizirati u određenom vremenskom periodu (period od nekoliko sati ili jedan dan) vrlo je bitno zadatak dodijeliti točno u vremenskom periodu kada ga je potrebno odraditi. Takve zadatke je potrebno dodjeljivati s visokim prioritetom i točno unutar vremenskog razdoblja kada ga je potrebno odraditi.

C. Utjecaj dogovorenog vremena realizacije s korisnikom

Kod operatora koji koriste mogućnost dogovora realizacije uključenja ili otklanjanja smetnje s korisnikom vrlo je važno to i osigurati kroz automatsko dodjeljivanje. Kao i kod zadataka koje je potrebno odraditi u željeno vrijeme zadacima za čiju je realizaciju dogovoreno vrijeme potrebno je dodijeliti visoki prioritet dodjele, a u isto vrijeme potrebno je voditi računa da se ti zadaci dodjele točno u periodu kada je i dogovorena realizacija.

D. Utjecaj definiranog KPI-a

Vrlo bitan faktor koji utječe na uspješnost automatskog dodjeljivanja je realizacija zadataka unutar definiranih rokova kroz KPI. Tim je zadacima potrebno postaviti takve prioritete koji mu osiguravaju dodjelu na vrijeme. Vrlo je bitno istaknuti da akteri poslovnog procesa odrađuju i zadatke koji su preduvjeti za konačnu realizaciju usluge. Primjere je ožičenje na razdjelniku koji je preduvjet za realizaciju usluge, ali i kojeg je potrebno realizirati u vrijeme koje omogućava dodjelu konačnog zadataka za realizaciju usluge u predviđenom roku. Ovo je bitno za zadatke s dogovorenom realizacijom usluge s korisnikom, zadatke sa željenim vremenom realizacije usluge i zadatke s definiranim KPI-om. To algoritam automatske dodjele zadataka treba osigurati.

E. Utjecaj na dodjelu slijedećeg zadatka

Moguće je definiranje više različitih faktora koji utječu na automatsku dodjelu zadataka. To mogu biti lokacija zadnjeg izvršenog zadatka, pozicija aktera poslovnog procesa te prioritet odrade zadatka.

F. Mogućnost da se zadatak ne realizira, a automatska dodjela zadataka se može nastaviti

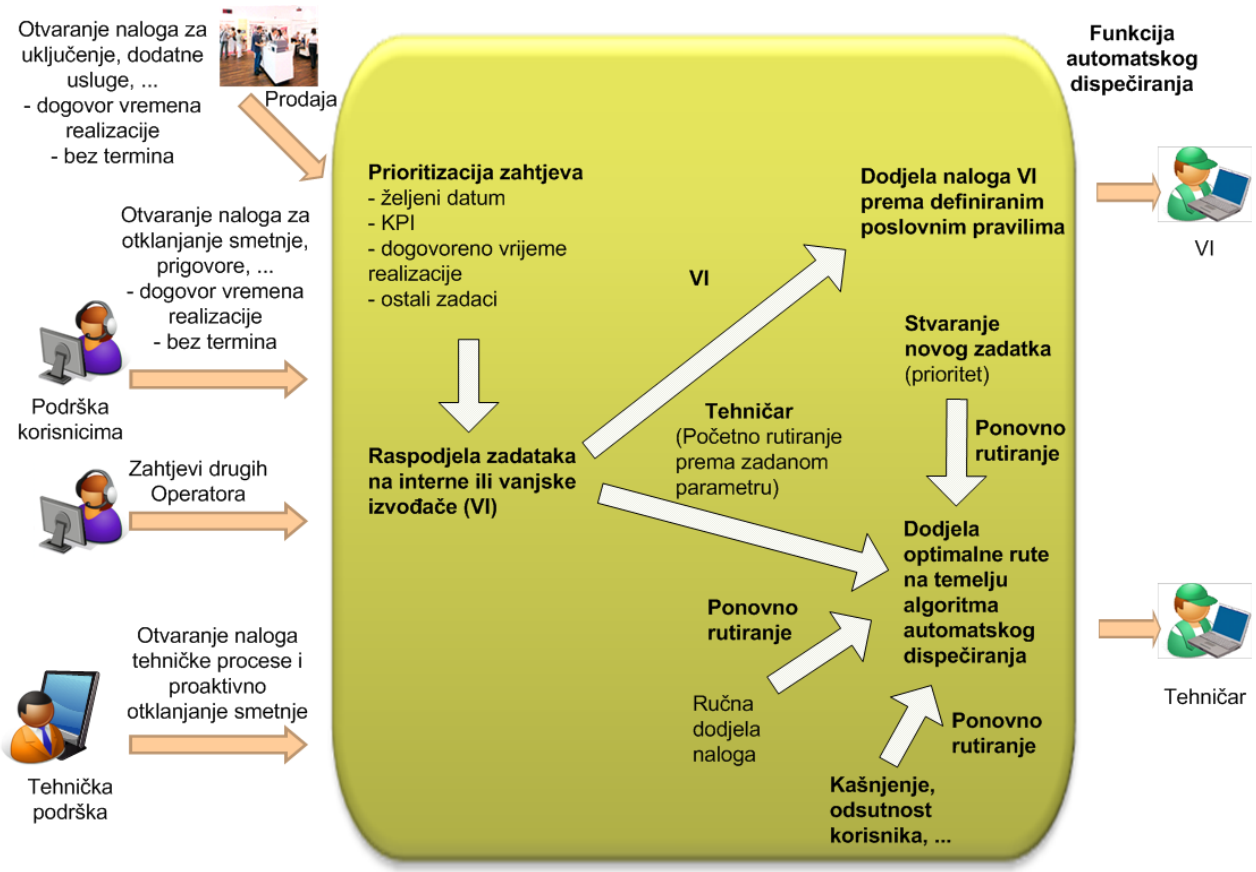
Postoje situacija kada akter poslovnog procesa ne može do kraja realizirati zadatak odmah, već nakon nekog vremena. U tim situacijama potrebno je omogućiti automatsku dodjelu zadataka bez daljnjeg utjecaja ovog zadatka. Taj će se zadatak odraditi kroz neki od ostalih predviđenih postupaka dodjele zadataka.

G. Dodjelu zadataka mimo utjecaja algoritma automatske dodjele zadataka

Ukoliko je potrebno hitno odraditi neki zadatak bez utjecaja algoritma automatskog dodjeljivanja potrebno je omogućiti dodjeljivanje takvih zadataka. Nakon realizacije takvog zadatka uzima se njegova lokacija u proračun slijedećeg zadatka koji se dodjeljuje kroz automatsku dodjelu.

Kako telekomunikacijski operatori koriste vrlo veliki broj poslovnih procesa kao posljedica vrlo složenog djelovanja vrlo je teško uklopiti sve poslovne procese u postupak automatskog dodjeljivanja i predvidjeti sve moguće situacije prilikom realizacije telekomunikacijske usluge. Stoga je potrebno ostaviti mogućnost ručnog dodjeljivanja zadataka čija realizacije direktno ne utječe u rad algoritma automatskog dodjeljivanja. No posredno ovo može značajno utjecati na efikasnost algoritma. Stoga je tu potrebno osigurati kontrolu u vidu mogućnosti ručnog dodjeljivanja zadataka koji se nalaze unutar unaprijed definirane udaljenosti od zadnjeg realiziranog zadatka kroz automatsko dodjeljivanje. Također se može definirati maksimalni broj zadataka koji se može dodijeliti kroz ručno dodjeljivanje.

Vrlo važna karakteristika automatskog dodjeljivanja zadataka je mogućnost definiranja različitog utjecaja određenih parametara na rad algoritma dodjele. Definiranjem težinskih vrijednosti mogu de definirati različiti utjecaji na rad algoritma s obzirom na udaljenost između lokacija zadnjeg i slijedećeg realiziranog zadatka, pridodanog prioriteta zadatku s obzirom na njegovu važnost te početne pozicije internih izvođača s koje započinje svoj radni dan. Na slici 1 prikazan je osnovni princip rada primijenjene automatske dodjele zadataka.



Slika 1. Prikaz principa primijenjene automatske dodjele zadataka

Može se izdvojiti nekoliko osnovnih koraka u postupku automatske dodjele zadataka.

A. *Kreiranje naloga na osnovu kojih se pokreću poslovni procesi*

Vrsta i način otvaranja naloga utječe na način i važnost dodjele zadataka unutar poslovnog procesa koji se primjenjuje na području telekomunikacijske pristupne mreže. Ukoliko je to nalog koji se otvara na zahtjev drugih operatora izvršavanje zadataka u sklopu odgovarajućih poslovnih procesa najvećeg je prioriteta. Vrlo velikog prioriteta su nalozi u sklopu kojih je s korisnikom dogovoreno vrijeme realizacije. Najmanjeg prioriteta su nalozi koji ne pripadaju jednoj od nabrojanih skupina te za koje nije definiran KPI.

B. *Dodjela zadataka vanjskim izvođačima*

Kako vanjski izvođači svoje zadatke trebaju dobiti na vrijeme koje je osigurano kroz dogovorena poslovna pravila automatska dodjela zadataka zasniva se na nešto drugačijim principima od dodjele zadataka internim izvođačima. Automatska dodjela zadataka vanjskim izvođačima pokreće se odmah nakon kreiranja zadatka koji se dodjeljuje prema prethodno nabrojanim principima.

C. *Dodjela zadataka internim izvođačima*

Konačno, automatska dodjela zadataka pokreće se na način da se zadovolje principi kojima se osigurava realizacija usluge na optimalan i efikasan način. To uključuje realizaciju u traženo ili dogovoreno vrijeme, zadovoljavanje postavljenih KPI-ova te optimalan put tehničara tijekom svog radnog vremena.

S obzirom na različite karakteristike područja na kojima djeluju interni izvođači i poslovne jedinice unutar kojih djeluju potrebno je omogućiti različita poslovna pravila koja vrijede unutar tih poslovnih područja. Tako se može definirati da se na određene djelatnike mogu dodjeljivati određeni zadaci s obzirom na područje djelovanja, vrstu procesa, vrstu zadatka i doba dana. Isto tako interni izvođač treba imati odgovarajuće znanje za odrađivanje određenog zadatka.

Kako su principi automatske dodjela zadataka vrlo složeni posljedica su vrlo složene forme kroz koje se uvode poslovna pravila i način djelovanja algoritma automatske dodjele. Njihovo popunjavanje i ažuriranje zahtjeva vrlo dobro poznavanje poslovnih procesa i načina rada prilikom realizacije telekomunikacijskih usluga.

III. ZAKLJUČAK

Dodjela zadataka akterima poslovnog procesa vrlo je složen postupak jer je istovremeno potrebno zadovoljiti više čimbenika koji utječu na uspješnu i efikasnu dodjelu zadataka, a samim time i uspješnu realizaciju usluga. Uspješan rad algoritma automatske dodjele zadataka nije sam po sebi dovoljan ukoliko se pri tome ne zadovolje svi postavljeni preduvjeti koji uvelike utječu na postupak automatske dodjele zadataka. Preduvjeti se znatno mogu razlikovati između operatora i stoga je prilikom njihove definicije potrebno dobro poznavanje poslovnih procesa i poslovnih pravila koji se koriste kod operatora koji će koristiti automatsku dodjelu zadataka.

Mehanizam automatske dodjele zadataka vrlo je složen i stoga je definicija svih potrebnih parametara za njegovu uspješnu primjenu vrlo složen postupak. I tu je potrebno vrlo dobro poznavanje svih mogućih prilika i događaja prilikom dodjele zadataka akterima poslovnog procesa.

Primjena automatske dodjele zadataka na području pristupne mreže u HT-u na samom je začetku i njezin daljnji tijek pokazat će u kojoj mjeri postavljeni principi i rad automatske dodjele zadataka pridonose efikasnosti rada aktera poslovnih procesa i koliko pridonose uštedi u sklopu ukupnih troškova realizacije uključenja telekomunikacijskih usluga i ostalih poslovnih procesa koji se primjenjuju u HT-u.

Iako su princip rada i funkcionalnosti automatske dodjele zadataka prilagođeni uvjetima i načinu rada na području pristupne mreže u HT-u, planiraju se dobiveni rezultati usporediti i s iskustvima ostalih telekomunikacijskih operatera te na taj način definirati daljnja moguća poboljšanja rada automatske dodjele zadataka.

LITERATURA

[1] D. Markulin, K. Musa, „Proposal for business process modelling and development in condition of complex telecommunication market", Mipro12, Opatija, 2012. , str. 549-553.

[2] D. Markulin, K. Musa, „Realizacija složenih procesa uključenja kroz sustave za podršku", Mipro09, Opatija, 2009, srt. 177-181.

Predprodajna reklamna aplikacija telekomunikacijskih usluga

Jasna Mulabegović, Halil Šabanović, Adisa Arslanagić
BH TELECOM d.d.
Obala Kulina Bana 8 71000 Sarajevo
E-mail: halil.sabanovic@bhtelecom.ba, jasna.mulabegovic@bhtelecom.ba, adisa.arslanagic@bhtelecom.ba

Sažetak - Prodajne mogućnosti koje telekom operatori nude svojim korisnicima nužno moraju biti interaktivne prirode kako bi korisnik mogao birati opcije koje odgovaraju njihovim željama i potrebama. Uzimajući u obzir prodajna iskustva koja se temelje na prodajnim rezultatima ostvarenim kroz tradicionalne kanale prodaje jasno je da korisnički zahtjevi idu ka opcijama prodajne ponude koja nudi mogućnost odabira telekomunikacijskih usluga najprije putem e-prodajnog kanala. Kreiranje prodajnog pristupnog kanala za ovu namjenu temeljenu na modelu „Predprodajne reklamne aplikacije“ je jedan od mogućih načina unaprijeđenja prodajnog procesa sa ciljem omogućavanja ostvarenja individualnih zahtjeva korisnika. Cilj ovog rada je pokazati mogućnosti integracije i distribucije korisničkih znanja u predprodajnu reklamnu aplikaciju telekomunikacijskih usluga sa svrhom primjene individualnih korisničkih zahtjeva kako bi se proizvela dodatna vrijednost u smislu pozitivnih korisničkih iskustava tokom procesa prodaje te kako bi se temeljem ovakvog pristupa prodajnom procesu ostvarili pozitivni efekti u povećanju prihoda.

I. UVOD

U skladu sa prodajnom strategijom telekom operatora koja se temelji na konceptu zadržavanja postojeće baze korisnika te potom njenog uvećavanja, osim dobro osmišljene ponude telekomunikacijskih usluga temeljenih na različitim tehnologijama neophodno je omogućiti korisnicima jednostavan pristup takvoj ponudi te odabir paketa usluga sa odgovarajućim opcijama prema individualnim korisničkim zahtjevima i potrebama.

Široka paleta usluga koje telekom operatori nude je osnov na kojem je moguće temeljiti cilj zadržavanja korisnika te eventualno dodatnog rasta, a mogućnost jasnog i jednostavnog odabira odgovarajuće usluge za korisnika je važan aspekt na temelju kojeg daje svoje povjerenje telekom operatoru. U ovom smislu od presudne važnosti su profili prodajnih kanala i njihova efikasnost u predprodajnoj podršci potencijalnim korisnicima usluga. S obzirom na kompleksnost i različitost opcija ukupne ponude usluga jednog savremenog telekom operatora, za uspješnu prodaju usluga zaposlenicima u prodaji je neophodna adekvatna pomoć u vidu aplikacije ili nekog drugog informatičkog alata (baza znanja, pretraživač, savjetnik i sl.).

II. CRM I ISKUSTVO KORISNIKA

Primjena CRM sustava u upravljanju iskustvom korisnika prema istraživanjima provedenim u različitim industrijama u kojima je ovakav pristup korišten, pokazala se vrlo ovisnom o okolnostima pod kojima se takav sustav koristi [1]. Uspješnost primjene CRM sustava u velikoj mjeri je u ovisnosti od stupnja integracije znanja temeljenih na informacijama prikupljenim od strane samih korisnika unutar određenog tržišta te potom integriranja takvih znanja u pojedine razine CRM sustava [2]. Svrha primjene CRM sustava je povećanje zadovoljstva korisnika kod isporuke željene usluge pri čemu fokus telekom operatora postaje sam kupac, te tako iskustva kupaca u procesu prodaje i isporuke usluga postaju relevantni za poslovanje. Kvalitet usluga koje telekom operator nudi, potom komunikacija u prodaji i isporuci usluga i na koncu cijena usluge bi trebali odgovarati ili premašiti očekivanja kupaca te za rezultat imati pozitivan utjecaj na lojalnost korisnika.

Djelotvorni i učinkoviti komunikacijski kanali u kontaktima „licem u lice“, putem telefona, interneta ili putem drugih interaktivnih medija omogućavaju telekom operatoru bolju dostupnost korisnicima te na taj način stvaraju povoljne okolnosti za razumijevanje njihovih potreba i želja. Logički slijed aktivnosti koji korisnicima olakšavaju izbor optimalne telekomunikacijske usluge, stvara preduvjete za pozitivna korisnička iskustva u kupovini i korištenju usluga telekom operatora. U tom smislu definiranje i kreiranje e-predprodajne reklamne aplikacije kao logičkog koraka nakon analiziranja korisničkih iskustava te integriranje ovog rješenja u potencijalni CRM sustav i WEB portal je dodatna mogućnost podizanja kvaliteta brige za korisnička iskustva od strane telekom operatora.

CRM procesi trebaju promjene na način da telekom operatori mogu prepoznati individualne korisnike te ih predstaviti u vidu informacija: ko su ti korisnici i što žele. Telekom operatori trebaju biti u poziciji stvaranja dodatnih vrijednosti u predprodajnom i prodajnom

procesu na način da prilagođavaju ponude svojih usluga korisničkim zahtjevima. Na koncu takav pristup se ogleda i u razvoju interaktivnih prodajnih kanala korisnik-telekom operator-korisnik.

III. E- PRODAJNI KANAL

A. *Razlozi kreiranja predprodajne reklamne aplikacije*

Telekom operatori koji imaju mogućnost korisnicima ponuditi različite vrste usluga ili iste vrste usluga ali preko različitih tehnologija, što je vezano i sa različitim cjenovnim pragovima za takve usluge, primjenom predprodajnih reklamnih aktivnosti olakšavaju korisniku optimalan izbor seta usluga u skladu sa njegovim trenutnim potrebama i mogućnostima. Cilj kreiranja i definiranja predprodajne reklamne aplikacije je da stvori poslovnu vrijednost kroz povećanje zadovoljstva korisnika čime je stvorena pretpostavka korisničke lojalnosti svom operatoru.

B. *Principi primjene predprodajne reklamne aplikacije*

Osnovi principi predprodajne reklamne aplikacije su zasnovani na grupama upita koje korisnici popunjavaju i na koncu ovog slijeda upita u ovisnosti od odabranih odgovora korisniku se nudi set usluga prema njegovim zahtjevima. Ovakav pristup prodaji usluga predstavlja u izvjesnom smislu evoluciju od pristupa upravljanja telekomunikacijskim uslugama kao proizvodom (Product Managament) prema upravljanju korisnikovim iskustvom (Costumer Experience Managament).

Izbor potrebne telekomunikacijske usluge, pri čemu je taj odabir u skladu sa korisnikovim trenutnim potrebama i mogućnostima, iz veoma široke ponude telekomunikacijskih usluga, korisnicima stvara nedoumice oko toga što odabrati i često predstavlja cilj do kojeg je korisnicima prilično komplicirano doći. Ključni uvjeti za uspješan proces prodaje sa stajališta prve linije prodaje i isporuke usluga korisnicima su:

- Visoka razina obučenosti i kompetentnosti prodajnog osoblja da korisniku preporuči optimalno rješenje za njegove potrebe, pri čemu su te kompetencije i znanja usklađena sa dinamikom isporuke novih rješenja telekomunikacijskih usluga
- Velika dinamika kreiranja novih telekomunikacijskih usluga u ponudi telekom operatora, praćena efikasnom dinamikom same isporuke novih usluga korisnicima
- Visoka prodajna iskorištenost usluga kao posljedica usklađenosti termina isporuke usluga sa prodajnim akcijama
- Brza isporuka usluga usklađena sa korisničkim očekivanjima.

Telekom operatori koji svoju razvojnu strategiju baziraju na pružanju usluga korisnicima po različitim tehnologijama imaju za potrebu napraviti kvalitetnu sponu između korisnikovih zahtjeva i tehnoloških rješenja telekomunikacijskih usluga iz svoje ponude. Ovakvi telekom operatori često jednu vrstu usluge mogu pružiti preko različitih tehničkih rješenja, koji kao svoje komponente sadrže različite cjenovne rangove.

Cilj ovog rada je prikazati principe i metodologiju moguće praktične primjene „Predprodajne reklamne aplikacije“ sa ciljem da se što jednostavnije omogući korisniku uspješan odabir optimalne telekomunikacijske usluge. Dodatni cilj je istražiti i analizirati druge mogućnosti aplikacije kao što je postizanje pozitivnog iskustva kod korisnika i zaposlenika u prodaji usluga, te procijeniti moguće efekte primjene ovog tipa predprodajne reklamne aplikacije na proces prodaje usluga kod telekom operatora.

IV. PRINCIPI I METODOLOGIJA PRIMJENE PREDPRODAJNE REKLAMNE APLIKACIJE

Efikasna predprodajna reklamna aplikacija treba ispuniti sljedeće pretpostavke:

- WEB bazirana aplikacija lako dostupna neovisno o lokaciji korisnika aplikacije;
- Aplikacija daje mogućnost izbora optimalnog seta usluga korisniku kroz jednostavan izbor lako razumljivih parametara;
- Aplikacija daje mogućnost pristupa korisniku svim dodatnim informacijama o svakom paketu usluga na jednostavan i razumljiv način, čime je korisniku omogućen lakši izbor optimalne usluge prema korisničkim potrebama i zahtjevima;
- Aplikacija omogućava prikupljanje vrijednih informacija telekom operatoru o korisničkim željama, potrebama i sugestijama na aktualnu ponudu usluga.

Rješenje „Predprodajne reklamne aplikacije“ je implementirano u testno okruženje kod telekom operatora gdje su analizirani efekti primjene aplikacije. Razvoj aplikacije i njena integracija u testno okruženje kod telekom operatora je trajao dva mjeseca. Proces integracije rješenja bio je prilagođen postojećem informacijskom sistemu te bitnih prepreka za okončanje integracije nije bilo. Za praktičnu implementaciju predprodajne reklamne aplikacije izabrana je grupa lako razumljivih parametara za definiranje optimalnog seta usluga:

- način plaćanja (putem računa, bez računa)
- mobilnost (na jednoj lokaciji, u pokretu)
- kategorija korisnika (fizičko lice, pravno lice)
- uključeni servisi (govor, prijenos podataka, IP TV, pristup internetu).

Parametar „način plaćanja“ definira profil usluga sa dvije mogućnosti, postpaid profil usluga- plaćanje putem računa ili prepaid profil usluga- bez računa (plaćanje unaprijed npr. e-vaucher). Parametar „mobilnost“ određuje tehnologiju, „u pokretu“- mobilna mreža ili „na jednoj lokaciji“- fiksna mreža. Parametar „kategorija korisnika“ određuje dvije grupe usluga namjenjene za fizička ili pravna lica, te parametar „uključeni servisi“

(govor, prijenos podataka, IPTV i pristup internetu) bliže određuju pakete usluga.

A. Primjer scenarija rada sa aplikacijom

Korisnik bi nakon pristupa aplikaciji putem web portala birao uslugu npr. pristupa internetu. Pri tome bi odabrao parametre:

- Način plaćanja- putem računa
- Mobilnost- na jednoj lokaciji
- Kategorija korisnika- fizičko lice
- Uključeni servisi- Govorni pozivi- Ne, Prenos podataka- Da, IP TV- Ne, Pristup internetu- Da.

Nakon završenog odabira parametara korisnik ima pristup ponudi paketa usluga prema zadanim kriterijima:

- 1. BH Line CLL (10 KM/mjesec)
- 2. BH Line L.Net (15 KM/mjesec)
- 3. BH Line WLL (18 KM/mjesec)
- 4. BH Line ADSL (20 KM/mjesec).

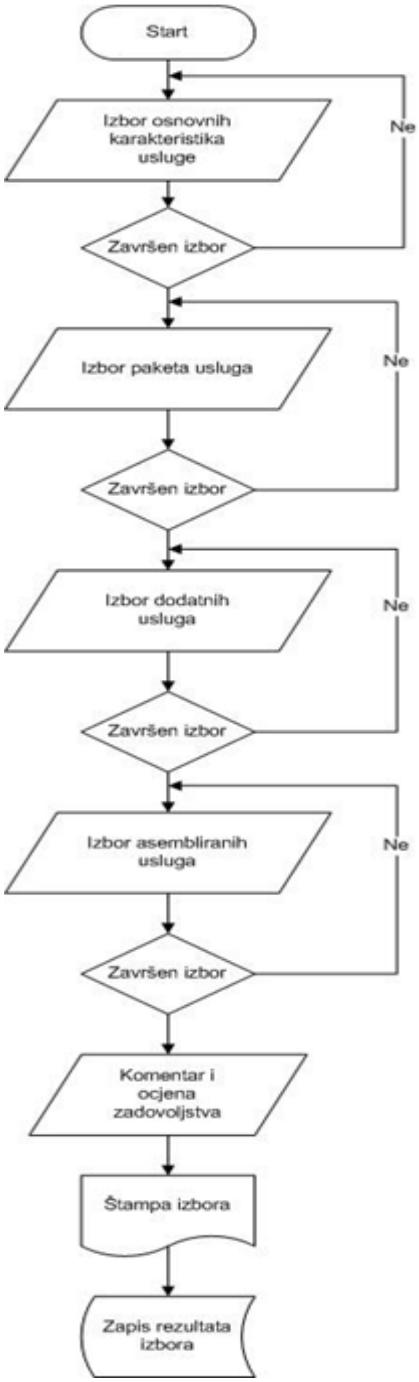
Nakon osnovnog seta upita korisnik ima mogućnost izbora jednog od četiri ponuđena paketa usluga. Drugi set upita dodatno daje mogućnost korisniku odabrati opciju dodatnih usluga kao i uređaja iz ponude telekom operatora koja mu najviše odgovara.

Uvid u detaljni opis svake od ponuđenih usluga korisniku daje mogućnost da odabere uslugu po svom izboru npr. korisnik može odabrati „L.Net“ uslugu koja je korisniku na raspolaganju i isporučiva od strane telekom operatora a zadovoljava korisnikove zahtjeve. Predprodajna reklamna aplikacija korisniku omogućava i dodatni uvid u cjenovnik za dati paket usluga kao i uvid u ponudu dodatnih usluga predviđenih za odabrani paket. Kao još jednu mogućnost osim izbora dodatnih usluga na odabrani paket, korisnik ima mogućnost odabira uređaja iz liste asembliranih uređaja za odabrani paket usluga, za ovaj primjer ponuđena je opcija prodaje USB modema za „L.Net“ paket.

Dodatna vrijednost korištenja predprodajne reklamne aplikacije se svakako nalazi u dostupnosti ovog tipa aplikacije samim korisnicima putem web pristupa ovakvom tipu prodajnog kanala. Takvim pristupom korisnik ima mogućnost odvojiti vremena koliko je potrebno da bi analizirao i uspoređivao ponuđene opcije usluga prema traženim zahtjevima i na koncu samostalno odlučivao o izboru optimalne usluge.

V. LOGIKA PREDPRODAJNE APLIKACIJE

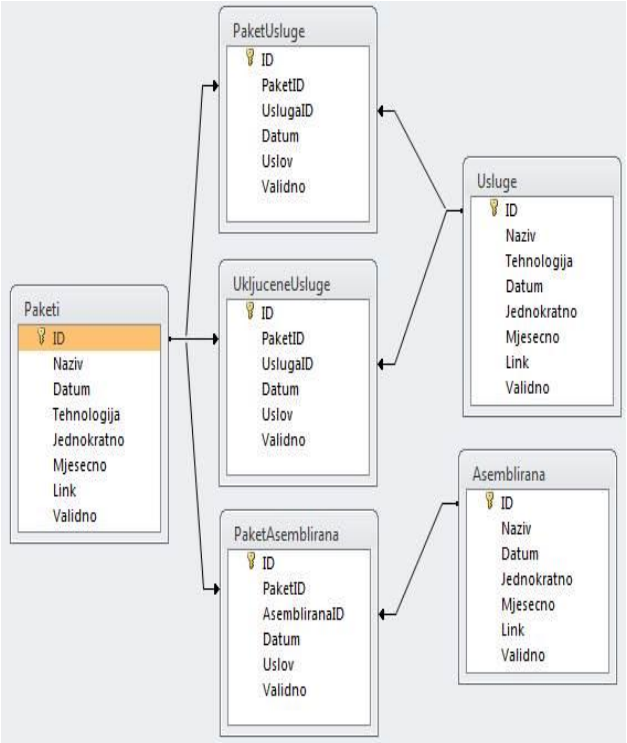
Istraživanja su pokazala da jedan od nedostataka koji dovodi u pitanje uspješnost CRM procesa jeste sam način prikupljanja i obrade podataka o klijentima [4]. CRM procesi, tehnologije i CRM informacije se nužno razvijaju u uskoj uzajamnoj sprezi. Kompanije kojima je kupac fokus poslovanja moraju preispitivati svoje procese, posebno u području prikupljanja podataka o klijentima, njihovoj obradi kao i analizi takvih podataka . Primjena efikasne metrike u korištenju prikupljenih podataka o korisnicima utječe na povećanje zadovoljstva i lojalnosti kod korisnika te omogućuje smanjenje troškova i povećanje prihoda kod telekom operatora.



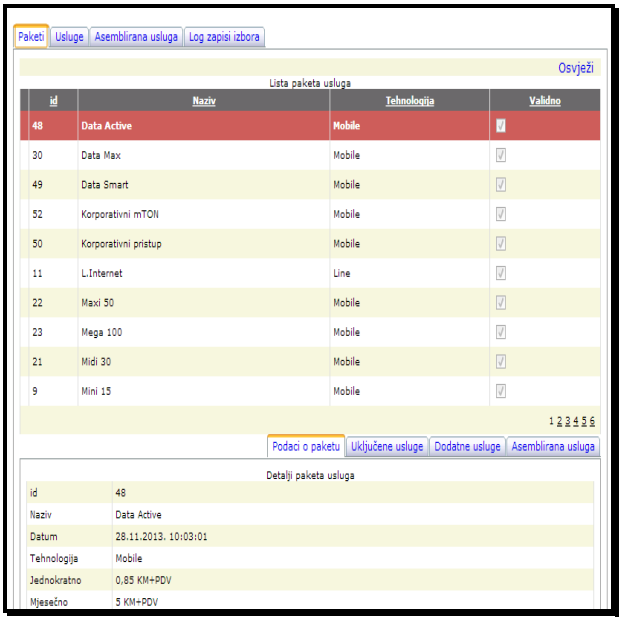
Slika 1. Logička struktura predprodajne-reklamne aplikacije

Logička struktura predprodajne reklamne aplikacije (Slika 1. i Slika 2.) treba osigurati prikaz obimne ponude usluga telekom operatora. Ujedno se sa pravilno definiranim relacijama među podacima može pojednostaviti algoritam same aplikacije, čime se ubrzava njen rad i pojednostavljuje održavanje.

Za održavanje podataka o ponudi usluga koristi se administratorska aplikacija (Slika 3.) putem koje se ažuriraju izmjene, dopune i nove usluge, te uređuju relacije između paketa usluga i samih usluga.

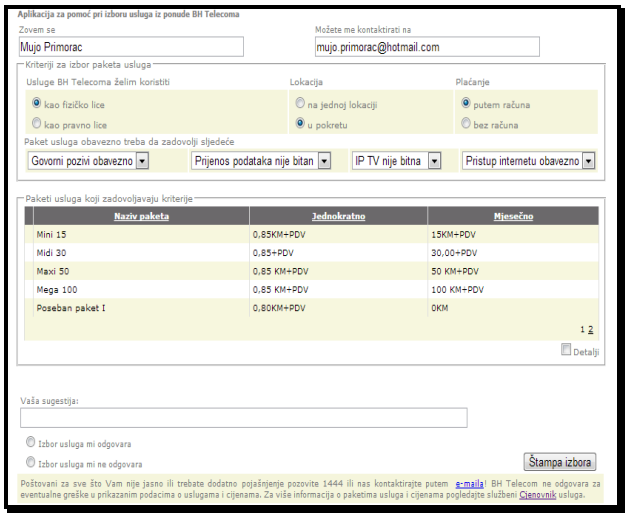


Slika 2. Struktura podataka predprodajne reklamne aplikacije

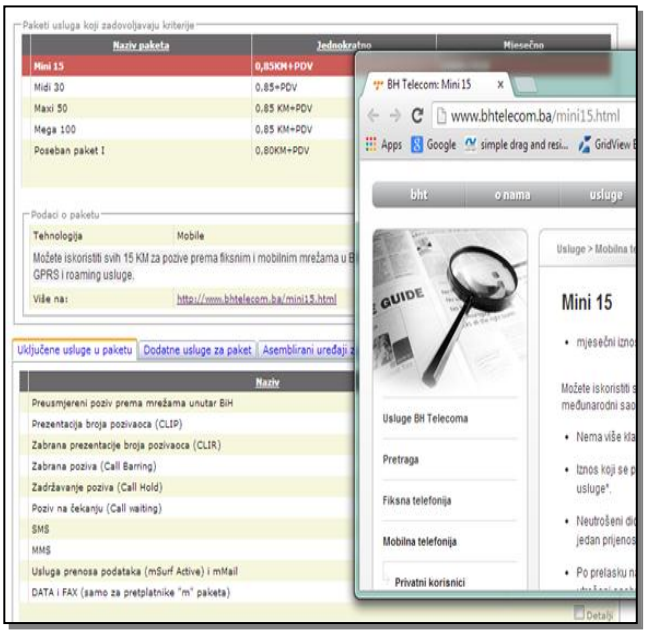


Slika 3. Administratorski dio predprodajne reklamne aplikacije

Korisnička aplikacija (Slika 4.) svojim sučeljem prilagođena je upotrebi korisnicima sa „prosječnim“ tehničkim znanjima na način da je dovoljno imati relativno malo znanja o telekomunikacijskim tehnologijama. Fokus aplikacijske logike je na sužavanju ponude paketa usluga prema postavljenim zahtjevima od strane korisnika te omogućavanju korisniku izbora optimalnog paketa usluga koji mu odgovara prema cijeni korištenja i njegovim realnim potrebama u trenutku korištenja aplikacije. Na osnovu odabira opcija jasno definiranih paketa usluga moguće je imati potpun uvid u ponudu paketa usluga za pojedine tipove usluga.



Slika 4. Korisnički dio predprodajne reklamne aplikacije

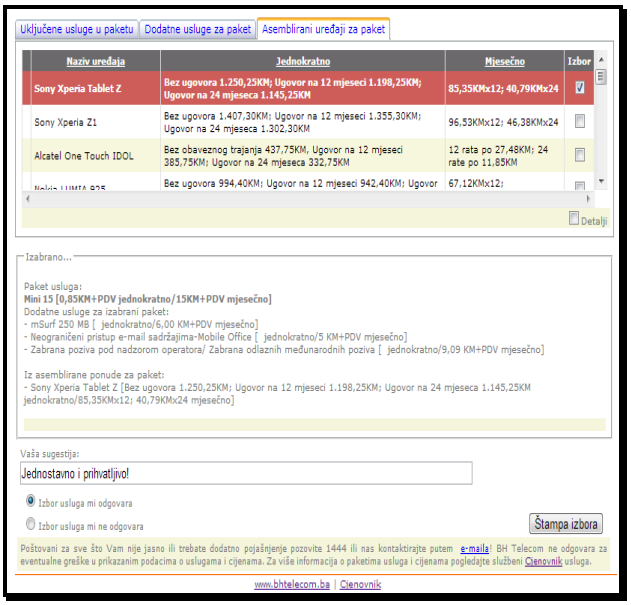


Slika br. 5 Izbor seta usluga na osnovu dodatnih informacija

Za svaku pojedinu uslugu na raspolaganju je skup osnovnih podataka i link na druge web stranice koje sadrže dodatne podatke o uslugama, za one korisnike aplikacije koji se ne mogu lako odlučiti (Slika 5.).

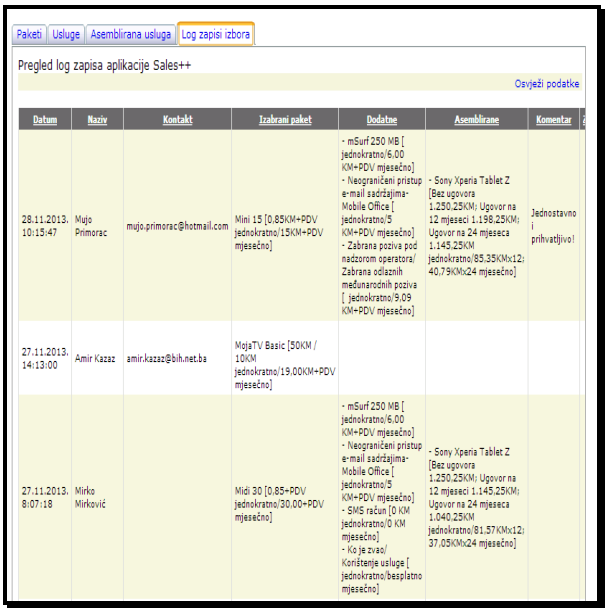
Korisnička aplikacija obezbjeđuje povratnu vezu sa telekom operatorom na način da korisnik može dodati svoju sugestiju/ preporuku/ prigovor na ponudu usluga, te može izraziti svoje zadovoljstvo izborom usluga (Slika 6.). Aplikacija nudi mogućnost izjašnjavanja korisnika o zadovoljstvu izborom usluga (Izbor usluga (ne)

odgovara), odnosno štampom stranice (Štampa izbora) zapisuje podatke o izabranim uslugama zajedno sa povratnim podacima korisnika (Slika 7.).



Slika 6. Podaci o izabranim uslugama i povratne informacije korisnika

Prikupljeni podaci iz predprodajne reklamne aplikacije mogu biti veoma korisni za analizu potreba korisnika, njihovo viđenje ponude usluga, te kao takvi mogu poslužiti i za proaktivnu ponudu na osnovu iskazanog interesovanja korisnika.



Slika 7. Podaci o korisnicima, izboru usluga i povratne informacije korisnika aplikacije

Dakle, moguće je zaključiti da prodajni kanal temeljen na „predprodajnoj reklamnoj aplikaciji“ telekomunikacijskih usluga korisnicima nudi mogućnosti odabira odgovarajuće usluge pri čemu korisnik u interakciji sa opcijama aplikacije dobija potpuniji uvid u ukupnu ponudu telekom operatora, sa svim opcijama usluga, opis tehničkih karakteristika i cijene korištenja. Istovremeno telekom operator dobija na jednostavan način vrlo vrijedne informacije o interesovanjima korisnika i eventualnim nedostacima svoje ponude.

Primjenom rješenja „Predprodajne reklamne aplikacije telekomunikacijskih usluga“ ponuda telekom operatora postaje posve transparentna za korisnike čime odabir željene usluge postaje jednostavan. Bitno se povećava dostupnost za odabir širokog spektra dodatnih telekomunikacijskih usluga sa kojima korisnici vrlo često nisu bili upoznati zbog činjenice da su takve usluge vrlo brojne. Dodatno, korisnicima je omogućen jednostavan odabir uređaja za kupovinu kao vrlo važan aspekt korisničkih zahtjeva, što prema prodajnim iskustvima kod telekom operatora povećava zadovoljstvo korisnika. Za telekom operatora se primjenom ovakvog rješenja otvara prostor brže i efikasnije prodaje, povećanja prodajnih rezultata, kreiranja različitih prodajnih akcija kojima bi se efikasno moglo upravljati te na taj način ostvarivati bolje prodajne rezultate. Poseban efekat ovakvog rješenja je moguć u segmentu povećanja prodajnih rezultata za nove usluge koje su na ovaj način dostupne za kupovinu odmah, pri čemu je za telekom operatora iznimno važno smanjenje vremena upoznavanja korisnika sa novom uslugom.

VI. ZAKLJUČAK

Strateško opredjeljenje kod telekom operatora koje se ogleda u poticanju sposobnosti za razvoj poslovne prakse u kojoj se kupac stavlja u fokus poslovnih procesa, obavezno podrazumjeva i to da se informacije o mišljenju korisnika prikupljaju, analiziraju i prevode u bolju vrijednost i pozitivno iskustvo za kupca. Predstavljeni model predprodajne reklamne aplikacije pokazao je samo neke od mogućnosti putem kojih su dostupna aktualna znanja o korisničkim navikama i potrebama temeljem kojih je omogućeno definiranje razvojne strategije kao i redefiniranje poslovnih procesa prema aktuelnim okolnostima. Prilagođenost ponude jedne kompanije stvarnim potrebama kupca posebno na tržištu kakvo je telekomunikacijsko tržište, čine telekom operatore sposobnijim zadržati postojećeg kupca ali i privući novog. Krajnji rezultat uspješne ponude su povećanje kupovne orijentacije korisnika i bolje upravljanje iskustvom kupca. CRM kao složeni sustav za upoznavanje korisnika i upravljanje njihovim iskustvom ne može se kupiti kao ad-hoc rješenje. Ovakav sustav se gradi od brojnih komponenti kao što je i opisana predprodajna reklamna aplikacija, a u takve komponente se postepeno ugrađuje nakupljeno iskustvo zaposlenika kompanije, te prikupljena znanja o korisnicima kao i najbolje prakse drugih uspješnih kompanija. Široka paleta ponude usluga telekom operatora je preduvjet za uspješan rad na dinamičnom telekomunikacijskom tržištu. Izbor usluge ili seta usluga a posebice optimalan izbor prema potrebama i zahtjevima korisnika je trenutno veoma zahtjevan posao, kako za prodajno osoblje tako i za potencijalne korisnike usluga. Iz tog razloga cilj prikazanog rješenja predprodajne reklamne aplikacije za telekomunikacijske usluge predstavlja mogući iskorak ka olakšavanju i pojednostavljenju izbora optimalne usluge prema zahtjevima korisnika telekom operatora.

LITERATURA

[1] Khalid Rababah, Haslina Mohd and Huda Ibrahim, Principles and Guidelines for Developing Implementation Methodologyof Customer Relationship Management (CRM) Systems, 2012

[2] NRG Working Paper Series,A Study into the Foundations of CRM Success, Ed Peelen, Kees van Montfort Rob Beltman, Arnoud Klerkx, no. 06-09, March 2006

[3] Customer Relationship Management: Implementation Process Perspective, Alok Mishra, Deepti Mishra, Department of Computer Engineering Atilim University, Ankara, Turkey, 2009

[4] Application and Model of CRM Strategy Map: an Empirical Study in China Qingliang MENG, Nongji ZOU, Guangming ZHANG, 2008

INTERNET & WEB APPLICATIONS

Multi-user web application for content management on multi-touch devices

Maja Cvijanović*, Petra Kirin*, Davor Pleša*, Josip Soko*, Sara Stamenov*, Ante Cikojević*, Stjepan Kelemen* and Ante Karin*

* University of Zagreb, Faculty of Electrical Engineering and Computing, Zagreb, Croatia
maja.cvijanovic@fer.hr, petra.kirin@fer.hr, davor.plesa@fer.hr, josip.soko@fer.hr, sara.stamenov@fer.hr, ante.cikojevic@fer.hr, stjepan.kelemen@fer.hr, ante.karin@fer.hr

Abstract - Nowadays, using touch applications is almost ubiquitous. It is appealing to different age groups, including very young generations who have just started their education. This paper describes development of a product that implements existing web applications into environment that enables multiple users to simultaneously interact with multi-touch device. Web application is developed for children to help them solve various tasks encountered for the first time. This paper analyzes web application requirements and implementation of required functionalities. The main requirement is to provide multi-user environment. Another important requirement is the ability to adjust game representation to child's position at the table. This is why the web application enables users to drag, rotate and resize each instance of the game. This paper also describes how the above mentioned content management is implemented. The other possible ways of implementation will also be mentioned, even though they are not used in the final product. Considering the fact that some children might have mobility difficulties, both gesture and button controls are implemented and can be used, depending on the preferences of a single child. In addition, this paper describes limitations, problems and solutions that have arisen during the web application development process.

I. INTRODUCTION

Nowadays, multi-touch devices are well accepted in the global population. With multi-touch foil padded on, almost any device with display can perform as interactive multi-touch device. Touch screen technology is increasingly popular in all age groups, including the youngest that are just starting their education. Usage of multi-touch devices brings a whole new dimension in solving mathematical tasks, helping children to learn faster by combining education and fun.

This paper describes the development of web application which enables simultaneous use of different types of web applications (e.g. educational games) in a multi-touch and multi-user environment. It initially includes two educational games, Mathematical Carousel and Mathematical Playroom, developed as part of ICT-AAC project at University of Zagreb, Faculty of Electrical Engineering and Computing.

Developed web application is primarily intended for preschool children to help them in coping with various tasks which they face for the first time. For this reason the application is primarily intended for use on large multi-touch tables, but it can also be used on other appropriate multi-touch devices.

One of the main goals in development of this web application was ensuring stable multi-user environment. Therefore, each child can use a web application instance of their choice at the same time and in any part of the same web browser. Because of planned use with multi-touch tables, each instance should be independently adjustable in size, position and rotation according to needs of each child and their position at the table. Therefore, controls for dragging, rotation and resizing should be implemented. Taking into account the fact that some of the children might have motoric difficulties, it is important to support different ways of managing instances – using both simple button controls and multi-touch gestures. Thus each child can choose which way of managing their instance they prefer.

Next section gives an overview of related work in using multi-touch tables and multi-touch and multi-user gesture detection. The third section describes application design and structure, implemented button controls and gesture detection. Article ends with conclusion and references sections.

II. RELATED WORK

Recently, there has been a significant amount of both software and hardware oriented work related to multi-touch devices and applications. In this section, some of the work that relates closely to this research will be described.

Paper [1] describes a set of interaction techniques which enables the usage of gestures in interactive multiuser direct-touch tabletop environments. It provides a variety of interesting functionalities including drag, rotate and resize gestures. Second paper [2] describes an adaptation of a single-user desktop application DigiQuilt into DigiTile, a multi-user multi-touch tabletop application. An example which includes both hardware and software development is presented in the third paper [3]. The work designs and implements a smart dining table which fits the requirements of the open-plan kitchen living room. Presented solution includes interactive entertainment games.

As part of this research, a number of multi-touch gesture detection libraries, offering certain capabilities

necessary for development of the application, were encountered.

The most interesting among observed was Hammer.js, which is a JavaScript library for multi-touch gestures. Hammer.js supports a variety of one-finger and two-finger gestures, namely drag, rotate, tap, double-tap and pinch zoom (resize). Library can be effectively used for detection and control over user gestures on multi-touch devices which was suitable for the purpose of this research. However, although it provides most of the needed functionalities, Hammer.js does not provide multi-user support which, for us, was of upmost importance. Therefore, it was used only for development of demo versions of the application which were mostly used for testing purposes, but not for the development of the final version.

Another library considered was SwinxyTouch which is a JavaScript library for multi-user gesture detection. SwinxyTouch provides a set of cross platform and cross device touch events. However, library is a commercial software and therefore wasn't appropriate for the purposes of this research.

Related work mentioned in this section represents a collection of interesting solutions which, when combined, support all of functionalities needed for development of the core application. However, none of the solutions fully fit the requirements mainly because they aren't web-oriented or exhibit a lack of multiuser support. The goal of this project was to develop an application which would satisfy both requirements mentioned above. Although this project shares a lot of qualities with presented examples of related work, it stands out by implementing those qualities into web-oriented environment which offers a new platform for various multi-touch and multi-user intended applications, as well as a new and simple access point for end-users.

III. SYSTEM DESIGN

A. *Application structure*

Developed application is implementation of existing web applications into the multi-user framework that supports multi-touch. Multi-user framework gives possibility to run multiple instances of web applications (e.g. educational games) and every user can interact with their instance independently and simultaneously, on the same multi-touch device. Every instance runs in its own separate container. Except the games that are available in the current solution version, developed application is very easy to upgrade by adding new web applications.

The user interface is presented in Figure 1.



Figure 1. User interface

Figure 2 presents the application structure with its key HTML elements mapped with corresponding JavaScript script on which particular HTML element relies on.

Application consists of a toolbar and main container. There are three controls within toolbar: "New game" button for starting a new game, "Settings" button and "Info" button for review of user manual. New instance containers for web applications are instantiated in the main application container. Every instance container has two sub-elements with controls for managing game instance and sub-element in which the game is loaded.

When loading the new instance in the browser window, it is necessary to set initial size of the instance

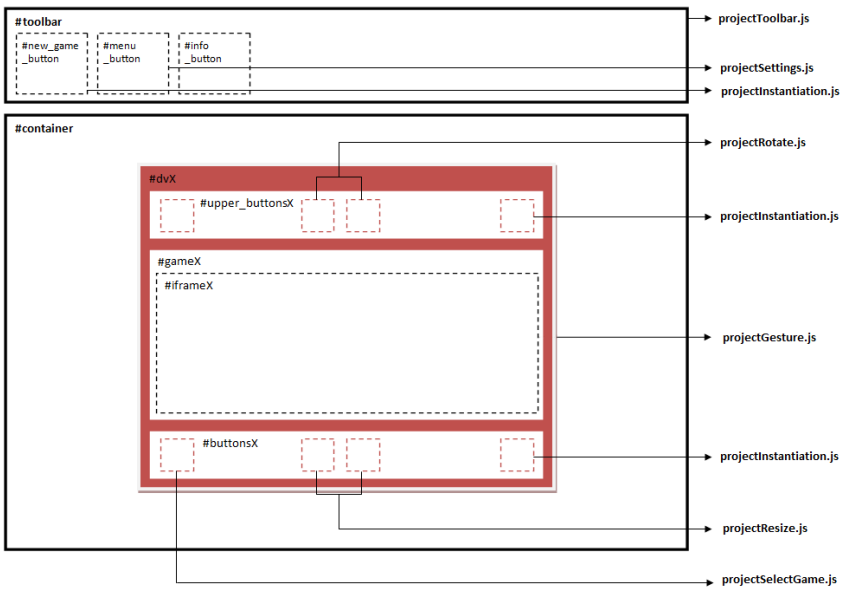


Figure 2. Application structure

and its position in the browser window. Initial size depends on the size of the screen – or more accurate, on the size of the browser window. This enables the scalability of the application on different devices, from mobile phones to large-sized screens. Besides the initial size, there is also the minimal and the maximal size to be set, dependently on the browser window size as well. Every new game is initially positioned in the middle of the screen.

B. *Application controls*

The most important part of the developed application is the user's ability to adapt instance containers according to their preferences by managing size, position, and viewing angle. The main goal was to create the user interface that suits the needs and preferences of users with different psycho-physical qualities, age groups and technology knowledge. Users can adjust instance containers either by well-known multi touch gestures or by using simple button controls. Every instance can be controlled independently and users can combine these two control methods.

By using button controls, as seen in Figure 2, users can change the size of their game instance, up scaling or down scaling the original container by clicking the buttons. Also, there is the ability to reposition (drag) the container anywhere inside the browser window. Finally, users can adapt every container to their viewing angle by changing the angle of the container using the buttons. Every button click rotates the container by 45 degrees in clockwise and counterclockwise directions, respectively.

C. *Multi-touch and multi-user gesture detection*

Another way to manage instance containers is by using multi-touch gestures. As noted in the previous section, there are gesture detection libraries available, but they don't fulfill all the necessary requirements – most importantly, multi-user support. For this reason, during this project, a custom gesture detection algorithm was built. It focuses mainly on detecting one-finger and two-finger gestures (i.e. drag, rotate, and resize). Detection of these gestures is separately and simultaneously available for each instance container.

Implemented algorithm uses a set of low-level web browser events described in [4], a W3C Recommendation. The prerequisite for activating basic touch events is the existence of one or more touch points (e.g. fingers) on multi-touch surface. Algorithm uses data stored in Touch objects, namely the target HTML element and coordinates of touch point within the browser window. Implementation also uses the TouchEvent interface and associated events (i.e. *touchmove* and *touchend*).

Touchmove event is used as the main event for detecting drag and transform (i.e. rotate and resize) gestures over a generic container. When such an event is detected, the algorithm counts the number of touch points over each instance container, and thus determines whether a drag or a transform gesture was performed. If a single touch point was recorded over the instance container, event is processed as a drag gesture, and if two touch points were recorded, it is processed as a transform gesture.

Simultaneously, a list of previous recorded touch points over each instance container is maintained. This enables the calculations necessary for determining the scaling factors and rotation angles that are to be applied on target instance container in case of performing a transform gesture. On the other hand, if a *touchend* event – which signals the release of a touch point, and thus the termination of a multi-touch gesture – is detected, the list of previous recorded touch points for the specific instance container is to be revised. The specific instance which lost a touch point is determined by comparing *touches* and *changedTouches* list provided by the TouchEvent interface.

After the multi-touch gesture is successfully detected and all the necessary parameters are dynamically calculated, same logic is applied as when using button controls.

D. *Implementation*

Web application has been implemented in HTML, CSS and JavaScript. JavaScript code is used primarily for detecting touches and dynamically adjusting CSS of target instances in order to control their shape, size, and position. Having reusability in mind, the functionality of JavaScript library jQuery, which is rich in essential functions on which this application relies on, was used.

IV. CONLUSION

Developed application enables multi-user usage of web applications using multi-touch devices. It is primarily developed for bigger multi-touch devices such as multi-touch tables. Multiple users can simultaneously use different instances of applications on a single multi-touch device.

Solution includes two different applications that users are able to choose from, however the nature of solution enables it to easily add numerous different applications to choose from. Users can use gestures and buttons to manipulate their instance of application. One of the biggest problems that we faced during the development of the application was the lack of libraries for simultaneous multi-user gesture detection.

Further improvements can be made by improving the gesture detection algorithm, namely making it more precise. Also, it is possible to implement an automatic positioning of application instances depending on number of instances that are used. Automatic positioning will automatically calculate the best position for each instance so there will be no need for users to manually manipulate the instances as much.

In the future, application can be used mostly in kindergartens and different educational institutions, since it can offer a fun experience while learning.

REFERENCES

[1] Shen, C. et al., "Informing the Design of Direct-Touch Tabletops, Computer Graphics and Applications", IEEE, 2006, vol. 26, Issue 5, pp. 36 – 46.

[2] Rick, J., Rogers, Y., "From DigiQuilt to DigiTile: Adapting Educational Technology to a Multi-Touch Table", 3rd IEEE International Workshop on Horizontal Interactive Human Computer Systems, 2008.

[3] Ling-Erl Cheng, Chao-Hsuing Tseng, Chun-Lin Lu, "Design and Implementation of the Smart Dining Table", International Conference on Consumer Electronics, Communications and Networks (CECNet), 2011.

[4] Schepers D., Moon S., Brubeck M., Barstow A., "Touch Events – W3C Recommendation 10 October 2013", 2013.

Front end optimization methods and their effect

Eng. Ilian Iliev
Web Software Developer
Sofia, Bulgaria
ilian@i-n-i.org
http://ilian.i-n-i.org

Assoc. Prof. Georgi Petrov Dimitrov, PhD
University of Library Studies and Information Technologies
Sofia, Bulgaria
geo.p.dimitrov@gmail.com

Abstract - With the evolution of the internet the loading speed of the web pages has become a vital factor for their success. A few hundred milliseconds difference in the response time are enough for visitors to choose one page instead of its competitors. This paper describes the increasing demand for faster loading webpages and the need of methods for optimization of the web code in order to increase their speed. It describes the process of loading webpages in browser and lists several methods for front-end performance optimization. The understanding of the loading process allows the correct identification of the parts that can be affected and improved. It also helps for understanding the effect of these methods and the cases when they can be applied. Although the total effect of the optimization depends on the specific webpage the appliance of these rules does not only reduce the page loading speed but it also affects the total bandwidth and load of the web server.

Keywords: front-end, optimization, web, performance, speed, optimization methods

1. INTRODUCTION

In the last few years we are witnessing the increasing usage of internet sites and web based applications. They have become an essential part of our lives and daily schedule. Unnoticed by the regular users together with the increased number of websites there is a huge increase of the number of resources used by these webpages. The increased load on the webservers, forces the hosting companies to add new webservers, to handle it, which leads to increased expenses for the business. The increased number of resources not only adds load to the servers hosting the websites and applications but also to the users. In addition to this the loading speed of the websites has become a major factor determining the user experience and one of the criteria used by search engines for positioning in the search results. The loading time and bandwidth for the websites can be decreased by applying different front-end optimization methods. As a side effect these methods decrease the load of the web servers as a result of the decreased bandwidth and number of established connections.

Section two presented the change of webpage size covering a period of 17 years together with the current state of the web pages. The sections analyzes the effect of the constant growth of the page's size on both the server and client side and marks the basic areas that optimization should target.

Section three starts with description of the webpage loading process followed by presentation of several different methods for front-end optimization of web pages with explanation of their target and effect.

Sections four contains the data of an experiment measuring the webpage size and loading time before and after the appliance of optimization methods. It also shows the difference of the server response time before and after the optimization.

2. HISTORY AND CHANGE OF THE WEBPAGE SIZE

History of the web page size

The evolution of the computer technologies in the last few years resulted in rapid increase of the internet connection speed and also the size of the web pages. According to a research published by websiteoptimization.com in November, 2012 the average size of the webpages has more than tripled for the period 2008 - 2012[1].

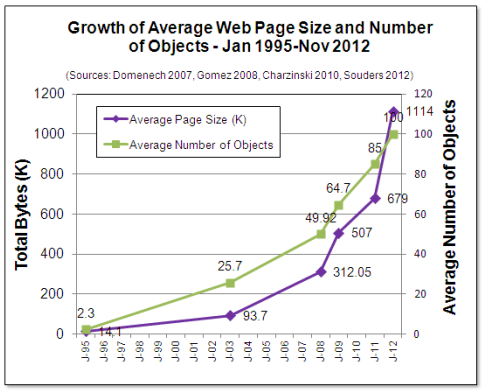


Fig.1 Average webpage size growth 1995 - 2012

As shown on fig.1 together with the growth of the average size there is also a significant growth in the average number of resources per page. From 2.3 up to 100 - this number has increased almost 50 times.

According to HTTP Archive for the period between November 2012 and November 2013 the average webpage size has raised from 1.25MB to 1.7MB, resulting in 40% growth for the last year[2].

Together with the growth of the page size there is also a constant increase of the end user bandwidth. A research made by Jakob Nielsen that covers the period from 1983 to 2013 shows a 50% growth of the connection speed per year[3]. The increase of the connection speed also raises the demand of the customers for faster loading and highly available websites.

The result of page size growth

The growth of the webpages has 2 major effects – on one side it increases the loading time for the end client and on other it increases the load of the server. The second of which is caused by the higher number of resources per page and the increased bandwidth.

For example if the total size of a webpage including its resources is 1.5MB, the total number of concurrent connections that a server can handle is 100 and the average connections speed is 1.5MB/s, which means that for 1 minute the server can serve the page up to 6000 times. This example ignores the fact that modern web browsers can open multiple simultaneous connections to the server to fetch page resources which adds additional load because of the process of establishing and closing connections.

The higher number of concurrent connections and their duration affects directly the total number of clients that the website host can serve. In order to lower the server load and increase the maximum number of users that the server can handle is necessary to either decrease the bandwidth or the number of connections per page.

3. WEBPAGE LOADING PROCESS AND FRONT-END OPTIMIZATION

Webpage loading process

The understanding of the webpage loading process is essential for the understanding of the effect of the front-end optimization methods.

The loading of the website starts by entering an URL in the address bar of the browser or clicking on a link. Then the browser parses the URL and determines the correct protocol, mainly HTTP or HTTPS, the domain, port resource address and request arguments if provided. Once the browser has determined the domain it makes a DNS lookup in order to find the IP address of the server hosting the page. Then a network connection is established between the server and the browser and the browser requests the specified page. The server generates a response and returns it to the browser. The response can be either a static resource or it can be generated by an application.

After the browser has received the response and if it is an HTML document it starts parsing it. In case of incorrect syntax the browser suppresses the error and tries to generate a structure as close as possible to the markup. Then a Document Object Model (DOM) is created. It is a hierarchical structure where every tag and attribute is represented as a separate node. Once the DOM is completely loaded the browser starts to load the additional resources specified in the code – styles, images and scripts.

After the load of each style resource the browser parses it and starts to apply the defined styles to the nodes in the DOM tree. The styles do not amend the DOM tree and as a result the browser can continue loading other resources simultaneously with their processing.

Due to the fact the scripts may cause changes in the DOM tree, during their loading, parsing and execution all other processes are halted. An exception here is the fact that some browsers as Mozilla and Safari continue loading other resources during the execution of the script using separate threads. In this case the DOM manipulations are still placed only in the main thread of the browser. Once the DOM tree is generated and the base resources are loaded and the browser starts rendering the webpage.

The front-end optimization methods primary target the process of page resources loading, their order and the usage of caching techniques.

The effect of front-end optimization

According to Steve Souders' "Performance Golden Rule" average of 80-90% of the user response time, while loading a webpage is spent in the front-end[4]. This time is spent mainly downloading resources and processing them. It is additionally increased by the fact that most browsers pause all other processes while downloading and executing script resources. Based on this fact Souders recommends to the developers to start optimizing the front-end parts of the websites and applications before focusing on back-end optimization[5].

The main goal of the frontend optimization methods is to reduce the loading time by lowering the generated traffic; respectively the time spent downloading resources. Another part of the optimization process is to reduce the number of resources required by the page, again in order to optimize the loading time for the end client. The second main goal is to allow the browser to load the page and its resources progressively and display a rendered page as fast as possible leaving the user with the feeling for quickly loading webpage. Here are described some of these methods and how they affect the loading process and the server performance.

CSS/JavaScript minification – this is the process of removing obsolete characters like blank lines and comments from the code in order to lower its size. The process of minifying JavaScript also includes a code obfuscation that replaces the names of the variables and methods with shorter ones. This process is also known as compression and can combine different scripts or styles into a single resource. As a result not only the size of the resources is lowered but also their number is decreased which leads to lower number of connections opened per page load.

Using Expires and Cache-Control headers – these headers are used to instruct the browsers whether a specific resource can be cached or not. The Cache-Control header was introduced in HTTP/1.1 in order to overload some of the limitations of the Expires header, for example while Expires value is specific date the Cache-Control max-age directive can set a time interval since the request has been served e.g. "X days after the resource has been served". Although the Cache-Control header is actually a replacement of the Expires one, it is recommended to use both of them together in order to ensure compatibility with browsers not supporting HTTP/1.1.

The main benefit of the use of Expires and Cache-Control headers is that once the resource has been cached the browser will make no more requests to fetch it from the server until the resource is considered an expired. This way once the page is loaded and the resources are cached they would not be requested again if the page is reloaded or during navigation inside the website.

Using GZip compression – this method is controversial because on one side it reduces the

bandwidth but on other the process of compressing the resources adds an additional load to the server. However, combining this method with proper caching techniques leads to lower number of requests i.e. lower number of executions of the compression algorithm. This method has good effect on text based resources like HTML documents, styles and scripts but has poor effect on images and other documents. In version 1.1 of the HTTP protocol the capabilities of the Accept-Encoding header are have been extended in order to allow the browsers to send additional data to the web servers regarding the supported compression methods and their preference order.

Using a Content Delivery Network – with this method the task of serving static resources is delegated to different network of servers. The Content Delivery Network (CDN) is a large geographically distributed network of servers aimed to serve content to end-users with high speed and availability. The distribution is made in order to serve the user request from the closest and fastest location. Such networks are widely used in the modern web sites to serve static content, downloadable resources, media and other.

In the CDN networks the servers are split into two groups – edge and origin servers. The end-user communicates with the edge servers and asks them for the needed resources. If the edge server does not possess the required resource it makes a request to an origin server for it. Once the resource is retrieved from the origin server the edge serves it to the user and also stores a copy of it for future uses. This way the end-users can retrieve the static resources of the webpages much faster because of the closer location of the server.

When using a CDN the application server should handle only the request to dynamic pages, e.g. those for which data calculating and preparation is required. In this case, the server hosting the application should not static resources.

Using only external JavaScript and CSS files is another method that can reduce the traffic generated to load the page. If measured directly, loading a page with internal JavaScript and CSS files is faster than using external ones, which is caused by the reduced number of requests.

However this way the end-user cannot benefit from the caching capabilities of the browser because the styles and scripts will be served as part of the HTML code for each page. This way, the traffic generated when navigating inside the website is increased and so is the loading time of the webpages. Vice versa putting all scripts and styles into external resources allows the browser to cache them and this way reduce the traffic and loading time for the page.

All methods presented this far are related to the bandwidth and number of requests made in order to completely load a webpage. The following methods affect the process of progressive loading of the page and the moment when the browser starts rendering the content.

Placing the styles at the top of the page – this way the styles are among the first resources that the browser starts to load after the HTML document has been processed. As a result the declared styles are applied to the DOM tree before the loading and execution of other resources and the browser can start rendering the webpage while waiting for other resources to be loaded. This is the reason why the style tags should be placed as higher as possible in the "head" section of the HTML document.

Placing the styles at the bottom of the page is another method for speeding up the loading of resource and progressive rendering. As described above during the loading and execution of the script resources the browsers halt all other processes. This also affects parallel downloads. Although browsers can download multiple resources simultaneously, most of them stop the downloading process when executing scripts, even if the resources are coming from different hosts. To avoid this blockage the script tags should be placed at the bottom of the page, usually before the closing body-tag.

Removing the duplicate scripts is another essential part of the front-end optimization. Having duplicate scripts has two major drawbacks. First of them is the increased loading time as a result of the repeated parsing and execution of the script. The second drawback is caused by the fact that frequently duplicate scripts lead to unexpected behavior in the front-end and may also lead to errors breaking the whole page.

In modern website quite often some visual elements like images, menus or even sidebars are removed when displaying the page in mobile devices. These changes are focused in two directions – lowering the bandwidth and the loading time due to the lower number of resources to load and also providing better experience for users by better utilization of the smaller screen. However such techniques are usually part of the design and user experience so they are only mentioned in this paper without being described thoroughly.

4. MEASURING THE EFFECT

The total effect of these optimization methods strongly depends on the specifics of the web page – the number of resources that can be cached and that are reused between the different pages of the site, the percent of bouncing users i.e. the ones that visit only a single page and then leave the website and others. However, testing a sample HTML template can show the general effect of the frontend optimization on both the server and the client.

In this case the following optimization methods are applied:

- CSS/JavaScript minification and compression
- Use of Expires and Cache-control headers
- Use of only external styles and scripts
- Placing styles at the top
- Placing scripts at the bottom

As shown on table 1 after the optimization the size of the webpage is reduced by 13.5% for initial load and up to 99.7% while reloading. The cause of the small reduce of the initial size of the pages is the heavy usage of images that represent a significant part of the webpage size and cannot be compressed.

The much lower traffic when reloading the page is due to the heavy use of caching for the resources on the webpage. In this case all resources except the HTML page are cached, and the page itself is served after GZip compression.

TABLE I. FRONT END OPTIMIZATION EFFECT

	First Load		**Reload**	
	Size(KB)	***Loading time(s)***	***Size (KB)***	***Loading time(s)***
Before	790.5	1.82	790.5	1.82
After	683.5	1.28	1.1	1.24
Difference (%)	13,5	29,7	99,7	31,9

The loading time of the page is reduced by 29.7% for the initial load and by 31.9% for the consecutive ones. This decrease shows the effect of the reduced size of the page and mainly the reduced number of requests made for loading external resources. The small difference in the loading time between the initial and the consecutive load results from the fact that the major part of the loading time of the page is spent while processing and executing scripts and not downloading resources.

The response time of the webserver before and after the optimization can be easily measured with the Apache JMeter tool. The test settings used for the test are as follows:

- Test duration: 600 seconds
- Number of threads: 50
- Ramp-up period: 50 seconds

The number of threads means the total number of concurrent users of the webpage and the ramp-up period is the time for starting all threads. The usage of the ramp-up period is for avoiding the extreme load of the webserver at the beginning of the test.

The test case is also set to imitate browser behavior i.e. to use the caching mechanisms used by the browsers and to reuse the connections with "keep-alive". The results of the test are shown in table 2.

TABLE II. SERVER RESPONSE TIME

	Number of samples	Response time (ms)		
		Average	*Minimal*	*Maximal*
1	10041	33	9	10091
2	10049	0	0	159

Row 1 of the table shows the data for the page before optimization and row 2 after the optimization process. The 0 values for the response time after the optimization are result of using caching techniques for the static resources. This way only the main page is loaded on consecutive loads and the time for its retrieval from the server is close to zero due to the fact that the testing tool and the server reside on the same machine.

In production environments, users are rarely in the same network with the hosting servers, so the time required to download the HTML page will increase, but still it will be much lower if the static resources are cached.

Although for sites and applications in production caching all of the static resources required by the webpages is not always possible the test results still show a valid example of the effect of the optimization. The much lower response time is a strong indicator of the reduced server load and proves the two-side effect of the front end optimization.

5. CONCLUSION

Undoubtedly in the last two decades the web has evolved from a small system for information exchange between scientists to a network connecting billion of users. In the context of this evolution the loading speed of the web pages has become a major factor for the success of websites and applications.

The methods presented in this paper cover the basic and most effective steps for front-end optimization. Their effect is mostly based on the reduced traffic as a result of the usage of compression and resource caching.

As stated above caching all resources is rarely possible but still a major improvement of the bandwidth when navigating between the pages of the website. This, combined with the correct placement of styles and scripts, allows the browser to start rendering the page earlier than for non-optimized pages, giving the user a faster response and progressive loading which improves the user experience and the satisfaction from the usage of the site. In addition to the effect on the browser the server load is reduced due to the lower number of connections and traffic.

The wide usage of mobile devices not only opens additional opportunities for bandwidth and loading speed optimization but also requires it due to the lower connection speed and more expensive bandwidth for mobile traffic.

To summarize these methods improve both the response time of the webpage and the loading speed and also reduce the web server load. Having in mind the fact that the number of internet users is constantly increasing applying these techniques will allow website owners to serve more users without changes of the existing hosting equipment.

The reduced number of requests lowers the load of the server processors so the main load remains on the storage. The reduced expenses for bandwidth and processing power can be invested in faster storage – SSD disks and/or more RAM for even faster resource delivery.

Finally, on the business side this improvement means lowering the expenses for servers and bandwidth, or keeping them the same while increasing the number of users. Even in the case of using a CDN which may add an additional cost to the expenses, the result of increased user satisfaction, respectively conversion rates will diminish the increased expenses.

We are yet to see the new opportunities for optimization that technologies as SPDY[6] and the second version of the HTTP protocol – HTTP 2.0[7] will provide in the area of webpage optimization but the path is clearly set. The effort is focused in the creation of better transport protocols, providing lower latency, better compression and reduced page load time.

REFERENCES

1. Websiteoptimization.com, http://www.websiteoptimization.com/speed/tweak/average-web-page/
2. HTTP Archive : http://httparchive.org/compare.php
3. Nielsen, Jakob, http://www.nngroup.com/articles/law-of-bandwidth/
4. Souders, Steve, High Performance Websites, O'Reilly Media 2007
5. Souders, Steve, The Performance Golden Rule, http://www.stevesouders.com/blog/2012/02/10/the-performance-golden-rule/
6. SPDY: An experimental protocol for a faster web http://dev.chromium.org/spdy/spdy-whitepaper
7. Hypertext Transfer Protocol Bis - https://datatracker.ietf.org/wg/httpbis/charter/

Comparative analysis of Google and Bing SEO on leading Croatian News portals

Ivan Britvić, Jurica Đurić, Dalibor Bužić
Visoka škola za informacijske tehnologije, Klaićeva 7, Zagreb, Croatia
ivan.britvic@vsite.hr, jurica.duric@vsite.hr, dalibor.buzic@vsite.hr

Search engine optimization is a important task in professional web site design. Google dominates the search engine market but there are others, namely Bing, Yahoo, etc. How well are web sites optimized for other search engines and is it worth a trouble? What are the main differences between search engine algorithms?

This research paper tend to answer these questions by comparing Google and Bing. Research includes example study of Croatian news portals, analyzing their SEO parameters and results they get on a Google and its primary contender, Microsoft Bing.

I. INTRODUCTION

Enormous quantity of content on World Wide Web is not well organized. A web search engine is complex software which enables user to easily find required information on Web. Search engine basically acts as answering machine. Delivering relevant results to user's query is it's critical success factor – user who found information with one search engine will most likely use it again.

Search engines like Google, Bing, Yandex, etc. have developed different algorithms to rank results and present them to a user. Ranking algorithms are business secret, so knowledge about them are in fact assumptions based on literature. Google algorithm is complex, and it includes more than 200 elements which influence on relevance (and consequently position in result) of web pages which content is closest to user's queries [1]. First position on Google's search engine result pages (SERP) brings 32,5% while second position brings 17,6% of all clicks from SERP [1][2]. That means, if 40.000 queries in Croatia for term 'sport' are submitted in one month, first ranked result could receive about 13.000 of all clicks from users. So, natural tendency for website content publishers is to have their web page as higher as possible on SERP. Better positioning means more visitors to their website.

Based upon [3][4][5], search engine optimization (SEO) can be considered as continuous process which helps web pages to be better positioned on SERP. One class of SEO techniques can be applied directly on particular web page (so it is called on-page SEO). The second class is performed out of web page (it is called off-page SEO). But first of all, it is necessary to emphasize high importance of quality web page content. Quality content means unique, informative and/or interesting content from visitor's perspective. According to [6][7][8][9] quality content could hold visitor for longer time on website, and sometimes user will share that content with his friends on social networks – and that is exactly a part of off-page SEO. Part of off-page optimization can be made by user – with no cost. On the other hand, poor content can decrease visitor's interest, so he can leave from website quickly.. Search engines learn from user experience, and if many visitors leave website from first page they entered within short time period, that page could fall down on SERP. Creating better content decreases efforts in SEO. If more publishers create equally good content, better SEO could help in better positioning.

II. SEO PHASES

Based upon [10][11][12][13] four main phases in SEO process can be differentiated: search term analysis, on-page SEO, off-page SEO and, finally, measuring results. If measuring results shows weak success, returning to previous phases with redefined tactics is necessary.

A. Search term analysis

Starting point in SEO is developing search terms strategy. Search term is word (or words) which user types into search box. Special online services known as keyword tools help analyst to find out appropriate synonyms, variations and phrases for specific keyword. Some of them provide search volume estimation on monthly basis. Important part of analysis is competitor identification. Trying to get top ranking with one generic single-word term can be unreachable goal for new website if many established competitors exists. With terms compound of two, three or even four words it can be easier to get in top ten ranking in SERP (having in mind that longer search terms have smaller search volume).

B. On-page SEO

Some of most important on-page optimization elements are [4][5][11]: title and description tags, headings, link structure, keywords density and consistency. Results of search term analysis (in further text: keywords) will be applied in on-page elements. Title tag is web page descriptor which is, besides in browser, shown to user in SERP. It is desirable to create eye-catching, not too long (up to 80 characters) title tag which contains exact match or variation of user's query. Meta description tag often (but not always) shows up in SERP, so effort to create concise description on page content can be rewarded by gaining more visitors. HTML header tags contribute to readability of longer text. Using clean URLs (URLs easily read by humans, not containing query strings or parameters) makes search engine easier to crawl pages on website. URL containing important keywords is descriptive to user who can see it in SERP and in shared content on social networks.

Preferred keywords should appear in title, headings, description, URL and, of course, in page content. Myth is that keywords need to be an exact match, or there is an ideal keyword density in page. It just should be used in a reasonably manner [14]. High keyword density could have negative impact on ranking.

Page load time is also one of important on-page SEO element which influences on user experience. Smaller load time is better, and can be achieved by using caching method, gzip compression, minifying CSS, HTML and JavaScript, prioritizing visible content etc.

C. *Off-page SEO*

Search engines count link to a page from another website as vote for that page [3]. More external links (or backlinks) does not automatically mean better ranking. Key is value of backlink. Search engines have ways to determine values of links. Backlinks from reliable or popular sites have greater value [6]. More value for sport page has link from sport shoes related website than from jewellery related website. Search engines prefer links from more authoritative and trusting sites because such links are harder to get.

Recently anchor text (visible text in hyperlink) was highly appreciated by search engines, but now things have changed [15].SEO masters used same anchor texts excessively in links from other websites, but it was indicator to search engine that such links could be made by professionals [16]. There is an opinion in SEO community that search engines now like more imperfectly structured links (variations in anchor text, using 'click here' in anchor text, links without anchor text) [17]. That could mean non-professional user put link to web page, and such user usually does not know anything about anchor text. Mix of two above mentioned types of inbound links is considered as the best for SEO at this moment [18].

Social media links are unavoidable part of off-page SEO. Social media activity helps search engines to determine page relevancy. Higher social network activity increases authority of website and it also directly brings visitors to website.

D. *Measuring results*

Performance measurement activities are necessary in efficient SEO. Keyword rank monitoring, traffic tracking, link tracking and cost of SEO (time and money) are inputs for adaptation, improvement and planning a new SEO cycle.

III. BING VS. GOOGLE SEO

Search engines publish guidelines to webmasters for optimizing web pages. Some of general Bing recommendations are [19]:

- clear, deep, easy to find, rich and engaging content which provides visitors information they need
- inbound links as signal of popularity
- social sharing as impact on organical rank (unpaid position on SERP)
- technical part such as using sitemaps, robots.txt files, fast page load time, redirects and canonical tags
- using title, description and header tags

Some of Google guidelines are [20]:

- create useful, information-rich site with clear hierarchy and text links
- include words users would type to find a page in page content
- create pages primarily for users, not for search engines
- use sitemaps, robots.txt, optimize load times
- reasonable number of links on a page

Search engines emphasize avoiding techniques which try to deceive users and search engine (known as black hat SEO) such as:

- cloaking (showing one content to search engine and another content to visitor),
- link buying, link spamming, participating in link schemas
- duplicate content
- meta refresh redirects
- automatically generated content
- hidden text and links
- sending automated queries to search engine.

Differences between Google and Bing SEO ranking factors which make search engines unique are not easy to determine. Comparing various resources on Web, some findings however can be obtained [21][22][23] as shown in Table 1.

Table 1. Differences between Bing and Google SEO ranking factors

	Bing	Google
Caching	caches first 100 KB of most pages	has no problem with larger pages
Indexing	discards pages with no ranking authority	indexes every file
Keywords in title	heavier emphasize on keywords, doesn't consider position of keyword in the title of page	lighter emphasize on keywords, considers keyword in the title of page as ranking factor
Natural links	less emphasize	more emphasize
Flash	more friendly	less friendly
Fresh content	less important	more important
Exact matching keywords	necessary	not necessary thanks to better understanding of synonyms
Domain age	more important	less important

Google is dominant search engine. According to [24] Google is used by 97.49%, Bing by 0.95% and Yahoo by 0.48% users in Croatia in 2013. Most of Croatian SEO efforts are aimed toward Google. But Bing US market share is 18.2% [25]. Since US market determines worldwide trends it is clear that Bing is not irrelevant.

IV. SEO ANALYSIS

In order to determine how are Croatian sites optimized for search engines and what results they achieve on two major search engines we selected five leading news portals in Croatia: www.index.hr, www.jutarnji.hr, www.vecernji.hr, www.net.hr and www.tportal.hr. List is kept only five sites short in order not to lose focus by analysing too much similar sites. Criteria for selection was site popularity, but also a differences in site designs and SEO approach. News portals are selected because they should rank rather high in search results in Croatian language.

Basic SEO parameters that are taken in account are: Woorank grade, Google PageRank, number of indexed pages, backlinks, Facebook likes, Twitter backlinks and load time. Tools used in all analysis are Woorank and Webrank.

A. *Woorank grade*

Woorank grade is dynamic grade on a 100 point scale that represents internet marketing effectiveness at given time. Many parameters are considered in Woorank grade, so it's a relevant global grade for SEO.

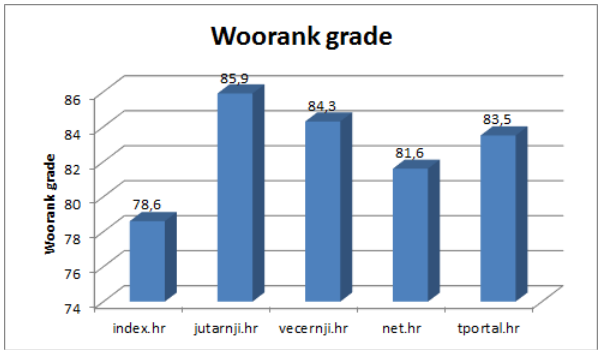


Figure 1. Woorank grade

Jutarnji.hr had the highest score and index.hr the lowest, but that grade is dynamic so results may vary in future.

B. *Google PageRank*

PageRank is an algorithm used by Google Search to rank the relevance of websites in their search engine results. According to Google: PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. The underlying assumption is that more important websites are likely to receive more links from other websites [26]. PageRank goes from 1 to 10, 1 is the worst and 10 is the best PageRank.

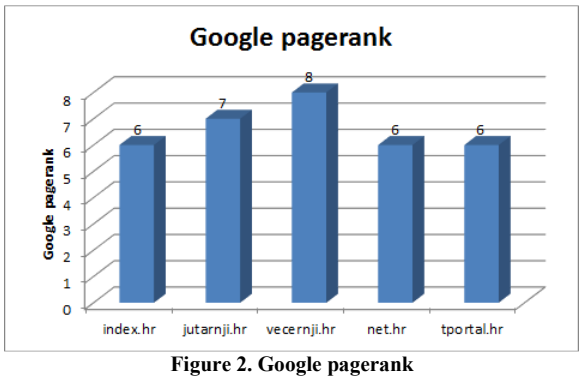


Figure 2. Google pagerank

The highest PageRank has vecernji.hr, followed by jutarnji.hr, remaining three news portals are equally leveled.

C. *Indexed pages*

Every search engine has its crawler engine that scans all web pages on single site and index them. Number of pages changes regularly, but it is very important factor in search engines ranking. Generally more is better. Here are the results for our portals:

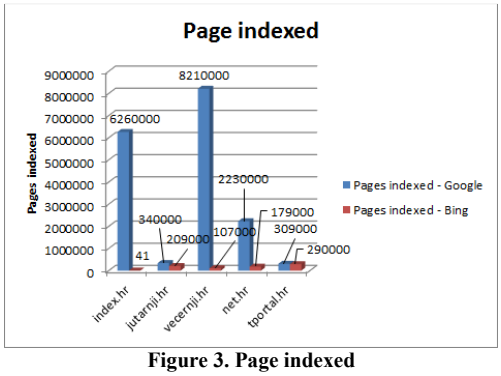


Figure 3. Page indexed

D. *Backlinks*

Backlinks, also known as incoming links, inbound links, inlinks and inward links, are incoming links to a website or web page. In basic link terminology, a backlink is any link received by a web node (web page, directory, website, or top level domain) from another web node [27].

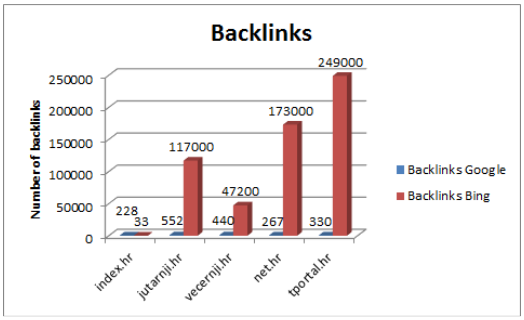


Figure 4. Backlinks

There is noticeable difference, Google has far less backlinks then Bing. It is consequence of very strict polices of Google to remove all links that are considered unimportant. This way Google ignores all marginal or

fraudulent links. Bing is at this point less advanced and can be easily misled by fake links.

E. Social

Social networks mentioning (likes, shares, etc.) have a huge influence in SEO. Two major social networks Facebook and Twitter are considered in this analysis.

I. Facebook

Importance of Facebook in todays web cannot be overstated. But here Bing has advantage, as Facebook data is pretty much inaccessible for Google which is one of the greatest problems that Google faces in order to keep market dominance. Here are the Facebook likes for our sites:

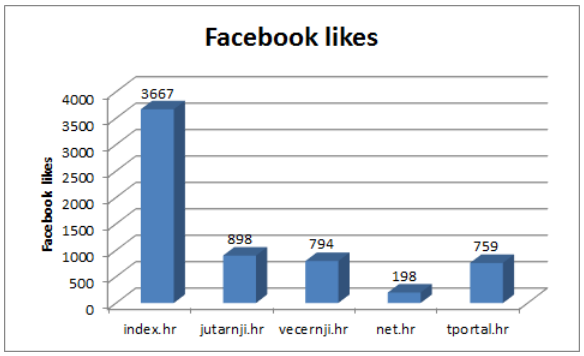


Figure 5. Facebook likes

It is clear that index.hr dominates in this segment which is clear indication of better integration of this site with social media sites.

II. Twitter

Twitter data are accessible as paid service to both of search engines and is considered as important factor in search results.

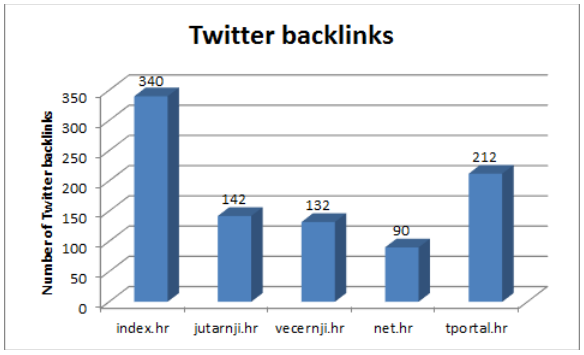


Figure 6. Twitter backlinks

Twitter backlinks are evenly distributed but index.hr has slight lead even here.

F. Load time

Load time describes amount of time needed to present requested page to the end user. Higher load times often indicate not just more multimedia content but poorly designed sites. Naturally, sites that have inadequate structure, unclean design and HTML and JavaScript errors score noticeable lower in search results. It is general opinion that Google put much more accent on this parameter.

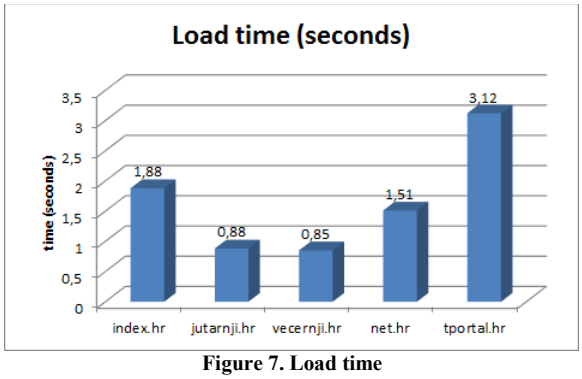


Figure 7. Load time

Tportal.hr has highest load time which should indirectly result in worse search rankings.

V. Search Engine Results Analysis

Search results ranking for selected news portals is tested for typical daily news search phrases in Croatian. Having in mind sites relevance, they all should rank rather high in observed searches, definitely in top one hundred. However, this wasn't always the case in performed analysis. Also there are sometimes significant differences in Google and Bing rankings. Achieved results are shown in following table (0 indicates site is not listed in top hundred results):

Table 2. Search phrase analysis

Search Phrase	index.hr		vecernji.hr		jutarnji.hr		net.hr		tportal.hr	
	Google	Bing	Google	Bing	Google	Bing	Google	Bing	Google	Bing
novosti	92	0	77	0	5	7	0	86	0	0
vijesti	1	8	4	51	2	10	7	46	8	76
najnovije vijesti	1	1	3	7	4	12	6	9	34	59
vijesti iz hrvatske	1	3	6	4	2	2	4	17	9	0
novosti iz inozemstva	75	0	9	20	4	45	0	18	85	0
hrvatska	18	41	37	0	0	0	47	0	12	51
svijet	8	0	3	2	4	8	25	32	6	51
sportski rezultati	53	0	0	0	15	0	2	8	50	0
vremenska prognoza	21	95	23	0	17	0	23	32	19	83
ljepota i zdravlje	2	33	57	0	42	0	16	18	0	0
dnevni horoskop	0	0	11	36	5	10	6	57	2	43

VI. CONCLUSION

Presented results clearly show that targeted sites usually rank lower than they should in search results. Search term "novosti" is common also for Serbian, Bosnian and similar languages so ranking here is a bit lower but still, "not in first 100", is not acceptable result for any of selected sites. One factor, not initially considered, emerged as truly important: clean URL's (uniform resource locators) [26]. Well defined URL's that include search phrase words were ranked high, even when they

originated for not so relevant sites. Also, in search results from Table 2, URL structure that didn't comply to the clean URL standards resulted in lower rankings.

When we compare Google and Bing results, it is clear that selected sites are usually ranked far worse on Bing. This is partially proof of superiority of Google search engine but also a strong indicator that sites didn't took Bing in consideration, maybe even didn't created XML site maps for Bing. It should be possible to greatly improve search results with dedicated Bing SEO effort. Having in mind recent iOS Bing integration, it is clear that although Bing is distant second in market share, it cannot simply be ignored in SEO process. Generally, Bing SEO should bring more visitors and enhance profitability for specific news sites.

Altogether, we can assess that SEO is done only partially on selected sites and have potential of further improvement. At this point, all observed sites are losing potential visitors due to unsolved SEO issues. There is definitely potential for improvement in SEO process for researched Croatian news portals.

REFERENCES

[1] https://www.google.com/intl/en/insidesearch/howsearchworks/algorithms.html

[2] The Value of Google Result Positioning, http://chitika.com/google-positioning-value

[3] R. Kumar and S. Saini, "A study on SEO monitoring system based on corporate website development", International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.1, No.2, June 2011 pp.42–49.

[4] N. Yalçın and U. Köse, "What is search engine optimization: SEO?", Procedia - Social and Behavioral Sciences, 9, 2010, pp 487-493.

[5] O. Reinhard and D. Johnson, "Search Engine Optimization for the Website of a South African Organization as an Internet Marketing Strategy", Lecture Notes in Engineering and Computer Science , 2012.

[6] C. Kerdvibulvech and K. Impaiboon, "A new method for web development using search engine optimization", International Journal of Computer Science and Business Informatics, Vol.3, No.1, July 2013, pp. 1–21.

[7] S. Frost, "7 Ways to Improve Your On-Page SEO and Enhance Your User Experience", 2014, http://www.rankranger.com/blog/7-ways-improve-onpage-seo

[8] D. Maynard "Why Bounce Rate is Your Most Important SEO Metric", 2013, http://www.ostraining.com/blog/marketing/bounce-rate/

[9] 6 SEO trends in 2013 - quality of content and social signals are on focus, http://itprism.com/blog/107-seo-trends-content-marketing

[10] Phases of SEO Life Cycle, http://www.gadgetcage.org/phases-of-seo-life-cycle/9504/

[11] X. Xiaojie, F. Yuan and W. Jian, "The Basic Principle and Applications of the Search Engine Optimization", Proceedings of the 2012 International Conference of Modern Computer Science and Applications Advances in Intelligent Systems and Computing, Vol. 191, 2013, pp 63-69.

[12] Establishing Your International SEO Strategy: How to Start Your International Web Presence, 2013, http://www.seerinteractive.com/blog/international-seo-strategy-guide

[13] Our Strategy: Not Really Having One…, http://www.sociallyinfused.com/services/ser/strategy/

[14] S. Duk, D. Bjelobrk, and M. Čarapina, "SEO in e-commerce: balancing between white and black hat methods", MIPRO 2013, May 2013, pp. 443–448

[15] Anchor text strategy, 2013, http://www.redmudmedia.com/2013/05/03/seo-anchor-text-strategy/#axzz2wtF5Mb8C

[16] K. Strong, "5 Reasons Why Exact Match Anchor Text is Bad", http://blog.act-on.com/2013/07/5-reasons-why-exact-match-anchor-text-emat-is-bad/

[17] S. Anderson, "Free Link Building Tips For 2014", http://www.hobo-web.co.uk/link-building-strategy-for-beginners-a-month-of-free-tips/

[18] N. Patel and B. Dean, "The Advanced Guide to Link Building", http://www.quicksprout.com/the-advanced-guide-to-link-building-chapter-1/

[19] Bing webmaster guidelines, http://www.bing.com/webmaster/help/webmaster-guidelines-30fba23a

[20] Webmaster guidelines, https://support.google.com/webmasters/answer/35769

[21] SEO ranking factors – rank correlation 2013 for Bing USA, http://www.searchmetrics.com/en/white-paper/ranking-factors-bing/

[22] J. Sherrod, "SEO for Bing – Google and Bing indexing differences", http://www.searchdiscovery.com/blog/seo-for-bing-google-and-bing-indexing-differences/

[23] Bing optimization, http://www.webconfs.com/bing-optimization-article-25.php

[24] StatCounter global stats http://gs.statcounter.com/#all-search_engine-HR-monthly-201301-201312-bar

[25] M. McGee, "Bing Ends 2013 With All-Time High In US Market Share, But Google Also Up [comScore]", http://searchengineland.com/bing-ends-2013-with-all-time-high-in-us-market-share-but-google-also-up-comscore-181876

[26] http://en.wikipedia.org/wiki/PageRank

[27] http://en.wikipedia.org/wiki/Backlink

[28] http://en.wikipedia.org/wiki/Clean_URL

CHALLENGING ICTs

State of the art in IPTV

Biljana Veselinovska
Innovation Dooel
1000 Skopje, Macedonia
biljana.veselinovska@innovation.com.mk

Marjan Gusev
Ss. Cyril and Methodius University
1000 Skopje, Macedonia
marjan.gushev@finki.ukim.mk

Toni Janevski
Ss. Cyril and Methodius University
1000 Skopje, Macedonia
tonij@etf.ukim.edu.mk

***Abstract*—IPTV is an essential new technology in ICT, which changes the perspective of the way end users consume TV services, but also serves as a platform for new ideas and concepts in other areas, such as medicine, science and industry. It is considered as a technology that brings significant contribution in different scientific areas, as well as medicine and industry, besides the development of information and communication technology itself. This paper gives an overview of recent state of the art research and technology development, appropriate architectures, design issues, upgrades and hot topic features. The analysis of recent relevant IPTV research papers and the comparison of the novelties they bring to the IPTV technology itself has led to the conclusion about the existence of three main approaches to state of the art IPTV research: IPTV Channel Control and Management, P2P IPTV systems and Technology Assisted Living. These approaches are further observed and discussed as an attempt to prove the importance of IPTV as a technology.**

***Index Terms*—IPTV; Full-duplex EAS; P2P TV; UMB Smart-TV**

I. Introduction

The ubiquitous growth of broadband Internet connections, characterized with enhanced data transmission rates and minimized overall losses, arises many different opportunities for exploiting the advantages of the IPTV technology. Since the future in ICT largely relies on the advancement of IPTV, its overall quality must be significantly improved. Thus, numerous researches are being undertaken in order to discover the ways in which IPTV can be enhanced not only as a technology, but also as a service platform used in different areas (economy, politics, industry etc.). The motivation for this paper is identifying and classifying the opportunities for improving the IPTV technology so that human society can take a step forward and fully benefit from this essential service platform.

Mainly, this paper elaborates on the existence of three different approaches to IPTV technology research: system and technology functionality upgrade (contextualized IPTV monitoring, IPTV channel control algorithm, reducing energy consumption in IPTV networks, IPTV channel control algorithm, reducing channel zapping time by creating adjacent multicast groups, the mesh-pull P2P streaming approach), performance analysis (Measuring Bit-torrent and large-scale P2P IPTV systems) and creating or improving technology-assisted living (Full-duplex emergency IPTV alert system, An IPTV platform for Integrating Multimodal Technologies, IPTV in health telemonitoring as well as education).

In Section II we explain the definition of IPTV as a term, the most popular interactive IPTV applications, as well as an overview of IPTV system architecture. An overview of recent related research papers is given in Section III. We believe that they represent state of the art exploitation of IPTV as a technology. Section IV explains the differences between some of today's most popular STB implementations. Finally, the discussion and conclusions presented in Section V will elaborate on future intentions and trends in the explained area of interest.

II. Background

A. Defining IPTV

According to O' Driscoll [1], IPTV is a technology that aims at securely delivering high quality broadcast television and/or on-demand video and audio content over a broadband network. Being based on the Internet Protocol, IPTV has a number of features which include:

- *Support for interactive TV*. The two-way capabilities of IPTV systems allow service providers to deliver interactive TV applications (standard live TV, HDTV, interactive games, high speed Internet browsing).
- *Time shifting*. A mechanism for recording and storing IPTV content for later viewing, along with a digital video recorder
- *Personalization*. The bidirectional communications with the IPTV system allow users to decide what they want to watch and when they want to watch it.
- *Low bandwidth requirements*. Service providers only stream the channel that the end user has requested instead of delivering every channel to every end user, thus conserving more bandwidth.
- *Accessible on multiple devices*. Consumers can use not only televisions, but also their PCs and mobile devices to access the IPTV content.

It is also important to state that IPTV and Internet TV differ in terms of their approaches in delivering IP based video, even though they share the same core base of technologies. Their main differences are shown in Table I [1].

B. Interactive IPTV applications

Although the key IPTV services are broadcast digital TV and Video on Demand (VoD), IP based technologies also allow service providers to offer a range of advanced interactive TV (iTV) applications. O' Driscoll [1] states that iTV changes

TABLE I: Differences between Internet TV and IPTV [1]

Features	Internet TV	IPTV
Platform for delivering video services	Leveraging the public Internet	Secure dedicated private networks managed by the SP
Geographical reach	No geographical limitations	Located in fixed areas
Network infrastructure ownership	QoS is not guarantied	QoS is guarantied
Access medium	STB device	PC
Costs	Mainly free	monthly subscription
Content generation methodologies	User generated content	Traditional shows and movies

the TV set from being the source of one-way, passive entertainment to two-way active entertainment, infotainment and communication. iTV applications can be embedded, network centric or server centric. These applications are listed as follows.

By bringing web-like interactivity to the traditional TV viewing experience, IPTV allows the end users to view hundreds of channels, simultaneously record TV material and access a range of interactive TV services by offering at least the following interactive IPTV applications:

- *EPG*. The Electronic Program Guide informs the subscribers about the available program channels. Subscribers can also personalize their EPGs to meet their own needs.
- *On-demand video content*. Subscribers access video content on their own schedule by using their IPTVCD.
- *IPTV browsing*. The IPTV STB browser adjusts existing Web pages for presentation on a TV.
- *IPTV e-mail*. Subscribers access their e-mail accounts by using an advanced digital IPTVCD with a keyboard.
- *Digital video recording*. Recording and storage of video content, made possible by adding a hard diet to the IPTVCD.
- *Walled garden portal*. Emulating a Web portal, whose interface is branded with the logo of the IPTV host.
- *Instant IPTV messaging*. Integration of IM with a number of commercially available IPTV networking platforms.
- *IPTV commerce*. Delivery of electronic commerce applications on a TV.
- *On-screen caller ID*. A combination of video content and telephony services.
- *Targeted interactive TV advertising*. Advertisers as well as service providers generate additional revenue by inserting adverts into their IP based video channels.
- *Gaming on demand*. A result of the combination of the TV medium and the two- way capabilities of broadband.
- *Parental control*. Parents managing the type of content viewed by their children.
- *IP based emergency alert systems*. Notifying residents of emergency situations (severe weather storms and terrorist attacks). A proposed system for emergency alerts [2] will be explained in the next section.
- *IPTV program related interaction*. Presentation of additional information on the screen while the viewer is watching the video content.

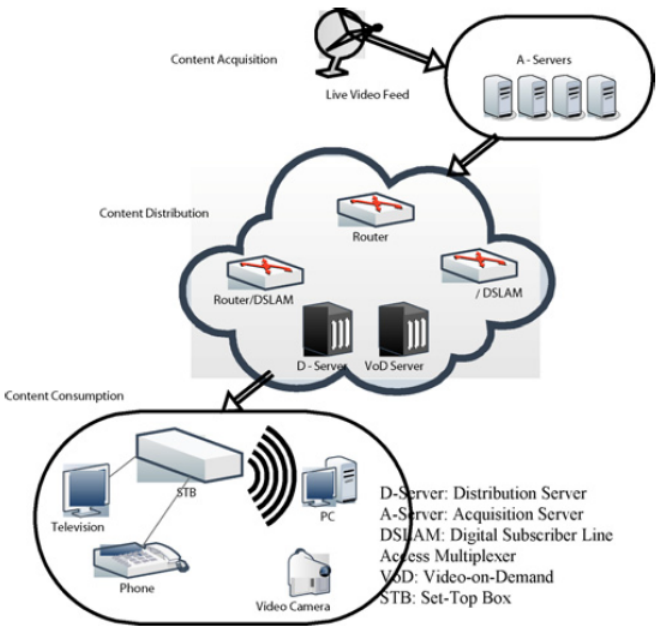


Fig. 1: A basic IPTV system [3]

C. IPTV Architecture

Leaving any technological differences aside, a typical IPTV infrastructure is comprised of three conceptual domains: content acquisition, content distribution, and content consumption [3], as shown in Fig 1.

The comprising entities of each domain are explained as follows.

- Acquisition servers (A-servers): They encode video and add DRM metadata.
- Distribution servers (D-servers): They provide caching and QoS control.
- VoD creators and servers: They retain a library of encoded VoD content to provide VoD services.
- IP routers: They route IP packets and provide fast reroute in case of routing failures.
- Residential gateways: IP routers for bundled services at home.
- STBs: A STB is a device on the customer side that interfaces with the user terminal (e.g., TV, PC, laptop, and others) with a DSL or cable wiring. [3]. As technology evolves, this device has become more than just a means for the typical user to receive their IPTV provider's predefined subscription content. That is, major ICT companies have created their own STB products with many advantages over the 'standard' ones. The following

section gives a list of some of today's hottest STB brands. Furthermore, [4] discusses the option of delivering IPTV services in the absence of an STB device.

III. Overview of Research Topics

In this section we will present an overview of recent research papers which aim at fully exploitation of the IPTV potential in various domains or at improving its quality as a service.

A. Optimizing the bandwidth

Optimizing the channel bandwidth, improving the QoS/QoE as well as energy savings are the goals of IPTV channel control. Some research on this topic will be presented as follows. An effective channel control algorithm is described in [5]. This research paper focuses on effectively designing the IPTV service system in terms of optimal management of the channels' bandwidth [5]. Joo et al. state that despite the STB device immediately displaying the selected channel, it still receives all of the channels regardless of whether they are used or not. However, due to the lack of bandwidth, the IPTV service cannot transmit all the channels at the same time.

Furthermore, the channel zapping time term as an essential parameter [5]. Channel zapping time is defined as the time difference (i.e. delay) between the moment a channel becomes available at the STB, and the one when the user watches the display of that selected channel on their IPTV watching device. This occurrence exists due to the computational overhead at the router when the number of broadcasting channels increases as a result of the IP multicast technology which is used to reduce the duplicate data transmission over the network. Thus, the proposed algorithm concurrently considers both the broadcasting channel distribution state, as well as the video encoding structure as control variables to effectively guarantee channel zapping time.

The same goal is being achieved in [6]. In this research, Cho et al. intend to improve the channel zapping time by making adjacent multicast groups of a current channel come to the home gateway in advance. Therefore, multicast streams for the adjacent channels of the one that is currently watched are always available to the home gateway. When the STB requests join to an adjacent channel, the home gateway can forward the corresponding multicast stream immediately, thus reducing the channel zapping time.

The similar goal of optimizing the channel bandwidth is being achieved with the OptiBand FP7 project [7] which we have previously observed in [8]. These two researches differ in the part of the IPTV network which they intend to improve. [7] focuses on improving the the access line rates (predominately ADSL) for the delivery of multiple HD streams over a single ADSL line, and by that enabling multiple HD channels per household, while [5] aims at effectively guaranteeing channel zapping time by an algorithm that takes both broadcasting channels' distribution state as well as the video encoding structure.

Despite bandwidth savings, according to Ramos et al. reducing the energy consumption can be done by a dynamic power consumption model scheme that pre-joins only a selection of TV channels instead of multicasting all TV channels at all times to everywhere, also aimed at achieving a very small number of TV channel switching requests. As stated above, the final result is comprised of bandwidth, energy and cost savings [9].

Finally, we will mention the opportunity for contextualized IPTV monitoring executed by the design of a scalable system for monitoring the state of an IPTV network, which will not only capture and analyze network traffic at key network points, but also provide granularity associated with hundreds of subscribers [10].

The presented solution [10] is a loosely coupled system with a server side and a large number of dedicated software agents running on distributed IPTV STBs across the IPTV network. The server side is responsible for receiving, storing and processing messages in real time, while each IPTV STB acts as an IPTV quality sensor node by capturing quality and telemetry related data, which then sends to the server side of the system.

B. P2P systems

One of the ways of effective bandwidth usage and providing an IPTV service with service guarantees, aimed towards the users' Quality-of-Experience (QoE) is by implementing the P2P (peer-to-peer) method of content delivery. This helps the Internet backbone and access networks which are constantly being overwhelmed by the ever-increasing IPTV video traffic. Also, according to Hei et al. [11], the most successful IPTV deployments have employed mesh-pull P2P streaming architectures. A lot of research exists about the structure and measurements of IPTV P2P systems.

[12] explains the mesh-pull P2P streaming approach. Hei et al. state that the operation of the mesh-pull system rely on the bandwidth contribution from each peer, even though they may have different upload capacity due to different types of access links. Furthermore, [11] was aimed towards developing a large-scale P2P IPTV system. The global view of user behaviors includes modifying the peer tracking methodology, which was followed by investigating the evolution of participating users, as well as the connections and traffic characteristics. The final results were achieving shorter star-up delay, higher rate streaming, smaller peer lags and better NAT traversal [11].

On the other hand, Biernacki [13] describes the implementation of an OMNeT++ based P2P TV simulation which enables to observe how the internal structure and the rules describing the interactions between the comprising elements influence the macroscopic behavior of the system. The goal is creating a common simulation framework for rapid P2P TV systems. Typically, the four main components are: a source (broadcaster), peers, superpeers and a tracker. The results describe how the upload to download ratio, the number of maximum allowed connections, and the number of sources and super- peers' influence on the system performance [13].

Similarly, as we discussed in [8], the NAPA-WINE project [14] aims at providing suggestions for P2P developments and Internet providers regarding the design of systems that minimize the impact on the underlying transport network, and the cost actions that can be taken in order to better exploit the network bandwidth This was achieved by creating a software library with the studied algorithms, as well as releasing a P2P-TV application.

Furthermore, one way to reduce startup delays which occur in channel switchings due to bandwidth limitations and the employment of buffers and overlay structures in today's P2P IPTV systems is shown in [15]. Manzato et al. propose four novel schemes for providing fast channel switching that reduce the occurrence of latency: Linear, Fast, Slow and Immediate schemes. The idea behind the proposed schemes is that users do not need to receive streams in full quality to verify the content being broadcast on a channel, instead they can receive streams at their minimum quality in order to do so [15].

C. IPTV and Internet TV in Technology-Assisted Living

As we already mentioned, IPTV as a technology provides endless opportunities for implementing different technology-assisted living systems. We will discuss this opportunity as follows.

Thang et al. [2] explore the possibility of using IPTV in the users' homes for disaster scenarios. The focus of their research is a two-way emergency alert system (EAS) with very high availability. The main goal of the system is to maintain a high availability of service, which can be achieved by applying the following solutions: redundancy of devices, multiple means of alert notification, as well as scalability and prioritization of data. The architecture of the system contains four conceptual parts- gateways, in-house gateways, outside networks and a database [2].

Mlakar et al. [16] present UMB-SmartTV, a novel platform that combines traditional IPTV services with multimodal-based non-IPTV services (automatic speech synthesis, speech recognition ambience control). It is important to note that any service within IPTV in this platform can be operated by using speech. The core of this platform is IMS, which is combined with a distributed DATA architecture in order to merge IPTV services and other non-IPTV services into a novel and highly modular solution. The first block, named Content core, represents an application server for content production and presentation. The second block is named IMS core. In the context of IPTV system, IMS represents multimedia service platform architecture. The third block is named Multimodal services. It actually represents the novel contribution of this platform- a distributed system intended for providing non-web speech-based multimodal services (automatic speech recognition, automatic text-to-speech synthesis, virtual assistants, home automation, personalization, etc.). Finally, the fourth block is named Environment controller, which controls several devices in the household (either by using TV remote, keyboard, mobile device, or speech)[16].

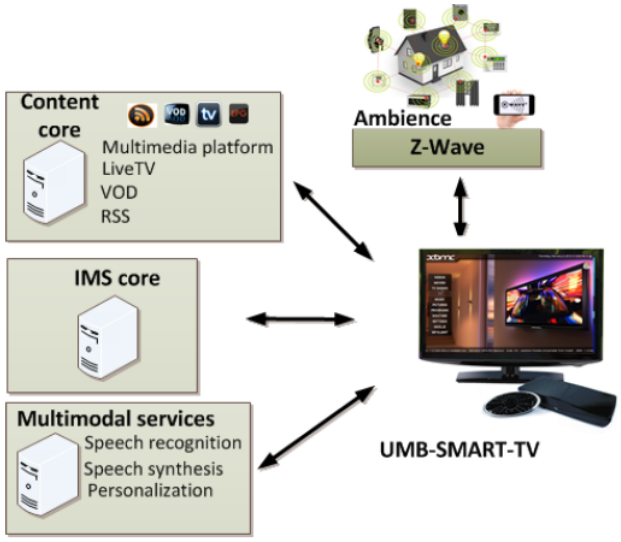


Fig. 2: The UMB-SmartTV environment [16]

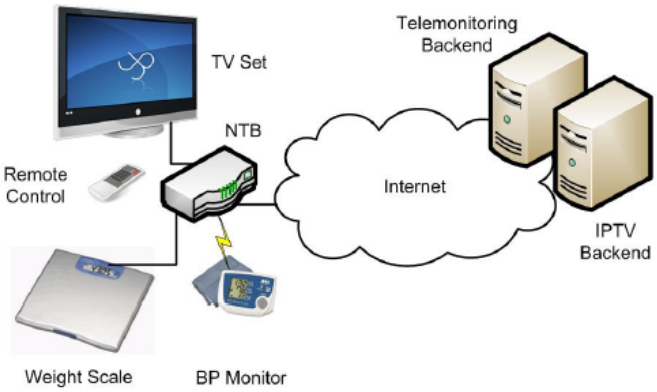


Fig. 3: Health monitoring system architecture [17]

An example of exploiting the benefits of IPTV as a technology in designing personalized health care systems is given in [17]. This research focuses on the development of a personal health telemonitoring application integrated into the IPTV home communication environment. The application enables monitoring of blood pressure and body weight, with rather common sensors and weight scales that have built-in communication interfaces, thus taking medicine to a whole new level. The application also supports on-line medical interviews. The user interface is manageable by a simple TV remote-control. The proposed system architecture is given in 3 [17].

Another example of using the IPTV technology for enhancing advancement in the field of medicine is the Vital Mind project [18], [8]. Here, instead of improving the diagnostics (remote measurement of blood pressure and weight), the goal is increasing the lifespan of the elderly by enabling them to actively participate in mind fitness exercises while watching TV by using a combination of cognitive psychology, the TV medium and advanced interactive ICT. This is done by

TABLE II: IPTV research papers comparison

Research	State of the art contribution
Functionality Upgrades	
[10]	scalable monitoring system which captures and analyzes network traffic, provides granularity (hundreds of subscribers) and pinpoints the source of errors
[5]	a channel control algorithm which effectively guarantees channel zapping time
[9]	bandwidth, energy and cost savings by creating a dynamic power consumption model scheme
[15]	Reducing latency in channel switchings within P2P IPTV systems with four novel fast channel switching schemes
[6]	improving the channel zapping time by making adjacent multicast groups of a current channel
[12]	increasing bandwidth savings with the mesh-pull P2P streaming approach
System Performance Analysis	
[13]	common simulation framework for rapid P2P TV systems
[11]	achieving shorter star-up delay, higher rate streaming, smaller peer lags and better NAT traversal in large-scale P2P mesh-pull IPTV systems
Technology-Assisted Living	
[2]	a highly available two-way EAS with redundancy of devices, multiple means of alert notification, scalability and prioritization of data
[16]	improving user's interactivity by a distributed system of non-web speech-based multimodal services (automatic speech recognition, text-to-speech synthesis, virtual assistants, etc.)
[17]	blood pressure, body weight and on-line medical interviews provided by a personal health telemonitoring application integrated into the IPTV service at home
[19]	IPTV in tertiary education, continuing professional development (CPD), and personally tailored education services

new methods for users control, production of Authoring and Production tools for IDTV STB and the development of usage of USB Flash-Device-UFD as an alternative to a broadcast delivery system [8].

Furthermore, IPTV easily finds its role in education. This is discussed in [19]. The purpose of this article is to provide examples of the IPTV technology being used in a variety of situations across the educational spectrum, including tertiary education, continuing professional development (CPD), and personally tailored education services in the home for the wider community. IPTV's efficacy in an educational setting has been proved by introducing 'Uni TV' at the The Melbourne Dental School because of its many video-content creation avenues [19].

D. Comparison of related research papers

In this subsection we give a comparison of the researches we analyzed. We have observed three different approaches to IPTV technology research: system and technology functionality upgrade, system performance analysis, and creating or improving technology- assisted living. This is presented in Table II.

IV. Comparison of different STB implemenations

This section gives an overview of some of the contemporary STB devices coming from different vendors, such as Apple TV, Google Chromecast, HBO Roku, Netgear NeoTV, Sony Bravia Smart Stick, which are yet to be proved as efficient enough to replace the existing 'standard' STBs that only deliver the service provider's predefined content to which a user has previously subscribed. It is important to note that STBs can be used both in IPTV and Internet TV, since their role is decoding the video downstream. However, given the main difference between these two technologies being the QoS support, STBs are mainly used in IPTV, while PCs, smartphones or laptops are usually used to access Internet TV.

Apple TV is Apple's popular streaming device runs on 512MB memory, 8 GB NAND Flash for Cache, and requires a Wi-Fi (802.11a/b/g/n) or 10/100 Ethernet connectivity. Its operating system is consisted of Apple TV Software 6.0 (based on iOS 7) [20]. It allows users to access more than two dozen channels, including free channels [21]. Supported Internet media services [20] include: iTunes Store, iCloud or Flickr photos, Netflix and Hulu streaming, sport channels including NBA, MLB, WatchESPN, than HBO Go, Sky News, Disney channels, Weather etc.

Google Chromecast is a digital media player that plays audio/video content on a high-definition television by streaming it via Wi-Fi from the Internet or local network. Users select the media to play on their television from the Google Chrome web browser on a personal computer or from a supported app on their mobile device [22]. Although it is by far the cheapest streaming video device, it has the weakest lineup of viewing channels.

Nevertheless, third-party monthly service is attempting to deliver more channels. Furthermore, ChromeCast requires users to select videos with a smartphone or tablet, via Android operating device. intended only for users with Android phones [21].

Roku is the streaming device sold by the Roku consumer electronic company offering a handful of upgrades and can stream more than 1,000 channels, including HBO Go. With low prices, Roku attempts to offer a set-top box for everyone. It is also one of the few companies that integrates a USB port for hard drive extensions, a headphone jack and a motion-controlled remote for gaming [21].

Netgear offers a series of set-top boxes: NeoTV, NeoTV Pro and NeoTV Max with the ability to stream a wide array of channels. The actual number of streaming channels is roughly 100 including Hulu, Netflix, YouTube, CinemaNow and Vudu.

Bravia Smart Stick is Sony's latest USB connected streaming device. Designed as the company's next Google TV device, it only works with Sony TVs. Hulu and other Bravia apps will be available even though they are not currently supported by the Google TV format. WIth its price being a bit higher than Apple TV's and Chromecast's, this device might become the least popular one [21].

In terms of having the aforementioned IPTV features in Section II, these STB devices have support for interactive

TV, time shifting, personalization of available TV channels, low bandwidth requirements and are accessible on multiple devices. However, this access is limited by the operating system. For example, Chromecast can only be accessed by devices that work with Android. Unfortunately, when it comes to IPTV applications, not all of those listed above are available to the users, such as IPTV e-mail, walled garden portal, instant IPTV messaging, a combination of video content and telephony services, gaming on demand as well as IP based emergency alert systems. Thus, the vendors of STBs should consider the implementation of some (if not all) of these missing IPTV applications which the technology itself can host, so that they can offer the users a more attractive product and increase their popularity on the market.

A. IPTV in the abscence of a STB

Recent research in the area has led to a new opportunity for delivering IPTV services, which excludes the usage of an STB device. This is made possible by the implementation of DVB-CI plus standard, whose 1.4 version was introduced in September 2013. As stated in [4], this DVB Common Interface (DVB-CI) plus standard is a result of resolving a security loophole of the firstly DVB-CI standard presented in 1997. It described the interface between a Conditional Access Module (CAiptvM) and the host, which is typically a TV set. When the IPTV customer inserts the SmarDTV CI Plus 1.4 CAM into an IP-connected TV set, an automatic discovery process starts to find the location and type of the available IPTV Channels, The application is launched and the first IPTV channel is displayed. The QoE has proven to be the same as in TV consumtion from other sources.

This standard has benefits for both consumers and IPTV service providers. Consumers would no longer have to make space for a STB and cables in their homes, and would save on their electricity bill, while the providers would save on equipment handling and cost by using a standards-based solution [4].

V. Conclusion and Future Work

The analysis of the IPTV state of the art has lead us to a conclusion that IPTV is an new emerging technology that represents a possible direction of future ICT ubiquitous and pervasive computing solutions. We have presented basic IPTV features and interactive applications, as well as the analysis of the impact of current research.

While in [8] we give an overview of recent research and technology development of appropriate architectures and design issues, as well as features and functionalities concerning IPTV as a technology in FP7 projects, this paper mainly focuses on the advancements in IPTV technology made in scientific research papers.

The fact that IPTV can easily be provided at alternative venues such as libraries, health centres, and other public places, such as the relatively inexpensive STB makes IPTV a useful technology in education, medicine and other scientific areas via sophisticated entertainment.

Having learned that IPTV can rise many opportunities for overall ICT growth, and technology assisted living, our future work will be aimed towards implementing the functionalities in new concepts and application areas enabling game oriented sophisticated concept of realizing continuous education, systems that support better living, etc.

References

[1] Gerard O'Driscoll, *Next Generation IPTV Services and Technologies*. John Wiley and Sons, 2008.

[2] T. C. Thang, A. T. Pham, Z. Cheng, and N. P. Ngoc, "Towards a full-duplex emergency alert system based on IPTV platform," in *Awareness Science and Technology (iCAST), 2011 3rd International Conference on*. IEEE, 2011, pp. 536–539.

[3] S. Zeadally, H. Moustafa, and F. Siddiqui, "Internet protocol television (iptv): Architecture, trends, and challenges," *Systems Journal, IEEE*, vol. 5, no. 4, pp. 518–527, 2011.

[4] Justus. (2013) Dvb-ci plus 1.4 demo, iptv without a set-top box, presented at ibc. [Online]. Available: http://www.iptvseminar.com/dvb-ci-plus-1-4-demo-iptv-without-a-set-top-box-seen-at-ibc/

[5] H. Joo, H. Song, D.-B. Lee, and I. Lee, "An effective iptv channel control algorithm considering channel zapping time and network utilization," *Broadcasting, IEEE Transactions on*, vol. 54, no. 2, pp. 208–216, 2008.

[6] C. Cho, I. Han, Y. Jun, and H. Lee, "Improvement of channel zapping time in iptv services using the adjacent groups join-leave method," in *The 6th International Conference on Advanced Communication Technology*, vol. 2, 2004, pp. 971–975.

[7] OptiBand. (2010). [Online]. Available: http://www.optiband-project.eu/

[8] B. Veselinovska, M. Gusev, and T. Janevski, "Overview of current trends in IPTV related FP7 projects," in *Information, Communication and Energy Systems and Technologies (ICEST) 2013 48th International Scientific Conference on*, 2013, pp. 111–114.

[9] F. Ramos, R. J. Gibbens, F. Song, P. Rodriguez, J. Crowcroft, and I. H. White, "Reducing energy consumption in IPTV networks by selective pre-joining of channels," in *Proceedings of the first ACM SIGCOMM workshop on Green networking*. ACM, 2010, pp. 47–52.

[10] U. Sedlar, M. Volk, J. Sterle, A. Kos, and R. Sernec, "Contextualized monitoring and root cause discovery in IPTV systems using data visualization," *Network, IEEE*, vol. 26, no. 6, pp. 40–46, 2012.

[11] X. Hei, C. Liang, J. Liang, Y. Liu, and K. W. Ross, "A measurement study of a large-scale p2p iptv system," *Multimedia, IEEE Transactions on*, vol. 9, no. 8, pp. 1672–1687, 2007.

[12] X. Hei, Y. Liu, and K. W. Ross, "Iptv over p2p streaming networks: the mesh-pull approach," *Communications Magazine, IEEE*, vol. 46, no. 2, pp. 86–92, 2008.

[13] A. Biernacki, "Simulating Performance of a BitTorrent-Based P2P TV System," in *Computer Networks*. Springer, 2011, pp. 448–458.

[14] NAPA-WINE. (2013). [Online]. Available: http://www.napa-wine.eu/cgi-bin/twiki/view/Public

[15] D. A. Manzato and N. L. d. Fonseca, "Providing fast channel switching in p2p iptv systems," *Selected Areas in Communications, IEEE Journal on*, vol. 31, no. 9, pp. 326–337, 2013.

[16] I. Mlakar, Z. Kačič, D. Zimšek, and M. Rojc, "A novel ims based iptv platform for integrating multimodal technologies," in *Proceedings of the 2013 International Conference on Systems, Control and Informatics*, 2013, pp. 364–370.

[17] M. Pustišek, E. Stojmenova, L. Zebec, and D. Kervina, "Bringing health telemonitoring into iptv based ami environment," *Previous Publications of the AMEA Association*, 2013.

[18] Vital Mind. (2011). [Online]. Available: http://www.vitalmind-project.eu/

[19] N. Parkes, "The role of iptv in education," *NSW*, vol. 2013, pp. 10–23, 2013.

[20] Apple tv. [Online]. Available: http://en.wikipedia.org/wiki/Apple_TV

[21] L. Bedigian. (2013) Apple tv vs. roku vs. chromecast vs. neotv vs. bravia smart stick. [Online]. Available: http://www.benzinga.com/tech/13/10/3969572/apple-tv-vs-roku-vs-chromecast-vs-neotv-vs-bravia-smart-stick

[22] Chromecast. [Online]. Available: http://en.wikipedia.org/wiki/Chromecast

Kinect-based Presenter Tracking Prototype for Videoconferencing

Davor Gadanac[1], Mićo Dujak[1], Damir Tomić [2], Domagoj Jerčić[2]
[1]Ericsson Nikola Tesla d. d., Krapinska 45, 10000 Zagreb, Croatia
[2]Faculty of Electrical Engineering and Computing, University of Zagreb, Unska 3, 0000 Zagreb, Croatia
E-mails: {davor.godanac, mico.dujak}@ericsson.com; damir.tomic@fer.hr; domagoj.jercic@fer.com

ABSTRACT: Video conferencing is a quickly growing market. Suppliers are enhancing their services in order to get more functionality beyond the traditional video services. Such add-on functionality is the presenter tracking system. This solution can operate with any videoconferencing system or other video capture device. It allows complete room freedom and adds movement flexibility utilizing smooth continuous tracking.
In this paper, a kinect-based presenter tracking system for video conferencing systems is introduced. The prototype is validated with Ericsson's Visual Communication system. Ericsson Visual Communication (VisualCom) offers an operator hosted, enriched video communication experience to enterprises and professionals, enabling the operators to evolve their communications. VisualCom was originally made for holding video conferences in smaller rooms. Presenter Tracking System expands VisualCom's capabilities by adding a moving HD camera (also known as a pan tilt zoom camera) that focuses a single presenter, and can be controlled both automatically and manually.

KEY WORDS

1. Kinect, 2. Videoconferencing, 3. Presenter tracking, 4. Visual communication

I. INTRODUCTION

As video conferencing services become more focused on supporting user activity, a growing need for presenter's location and activity is recognized. To support such videoconferencing systems, a tracking application and algorithms are required.

In this paper, we present a Presenter Tracking System which consists of Kinect Sensor, which is used for tracking the presenter. Kinect is a cost-efficient device which contains multiple sensors (skeletal tracking, voice recognition, infrared, stereoscopic), as well as a powerful API for tracking human movement. In this project Kinect for Windows was used, although the application can be run with Kinect for Xbox, but with lower performance. Note that the Xbox version does not support all features, such as the near-mode tracking. Other part of the solution is HD PTZ camera which is used to follow presenter movements around the predefined space.

II. PRESENTER TRACKING SYSTEMS

Generally, tracking is the observing of people or objects on the move and supplying a timely ordered sequence of respective location data to a model e.g. capable to serve for depicting the motion on a display capability. In virtual space technology, a tracking system is generally a system capable of rendering virtual space to a human observer while tracking the observer's body coordinates. For instance, in dynamic virtual auditory space simulations, a real-time head tracker provides feedback to the central processor, allowing for selection of appropriate head-related transfer functions at the estimated current position of the observer relative to the environment [1].

As stated earlier, PTS is a solution which consists of Kinect Sensor, which is used for tracking the presenter, and PTZ camera. The Kinect sensor provides head coordinates of the presenter to the application, that turns them into appropriate camera pan and tilt values (Fig 1).

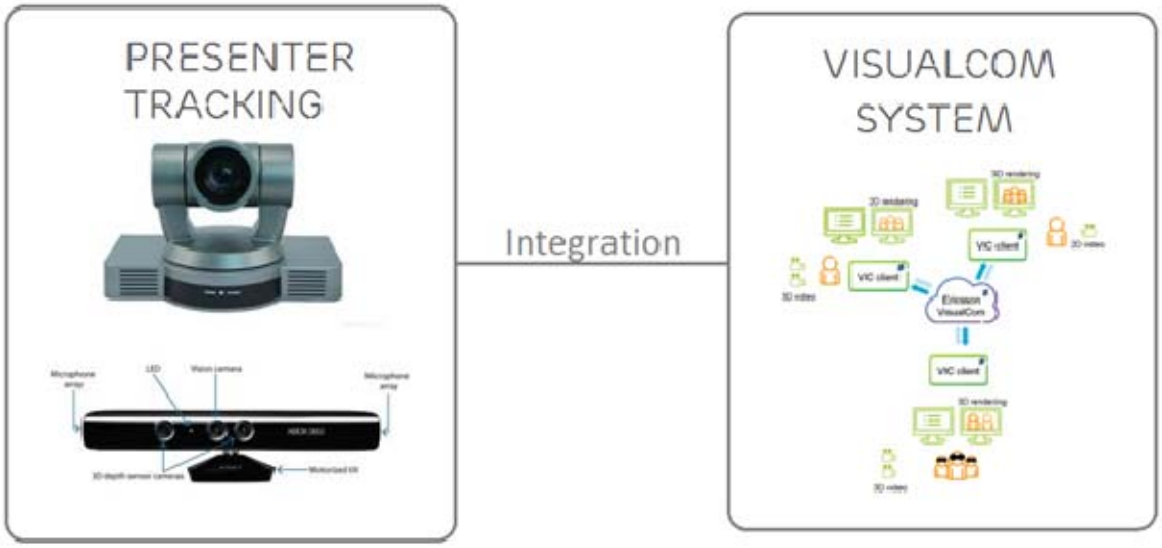


Fig 1 Kinect-based Presenter Tracking System solution overview

III. ERICSSON VISUAL COMMUNICATION (VISUALCOM)

The Ericsson Visual Communication solution is aimed at allowing operators to provide their business and consumer customers with easy to use, interconnected, and interoperable high quality video conferencing experiences globally. It offers an operator hosted, enriched video communication experience to enterprises and professionals, enabling the operators to evolve their communications offering in the quickly growing video conferencing market (Fig 3). VisualCom solution is an implementation based on the Ericsson IMS System with enhancements for high quality and high definition video communication. The solution brings together multiple nodes that employ standardized entities and interfaces [2].

Ericsson Visual Communication shows how high quality video communication can be realized with Ericsson and optional third party products. It is a complete network solution, bringing together multiple nodes that are based on standardized entities and interfaces.

Standard VisualCom Solution is based on Ericsson VisualCom Client HW, VisualCom Client SW, consumer electronics and standard VisualCom stand. VisualCom was originally made for holding video conferences in smaller rooms (Fig 2).



Fig 2 Visualcom Hardware Client for end-users

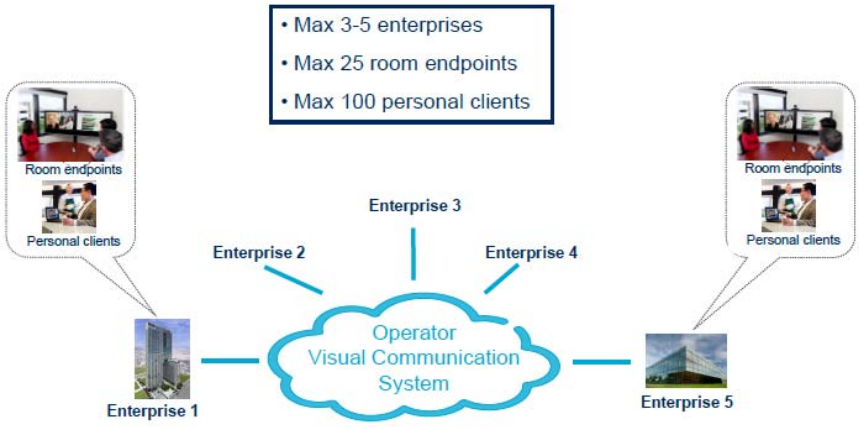


Fig 3 Ericsson Visual Communication Service

IV. SYSTEM ARCHITECTURE

Presenter Tracking System expands VisualCom's capabilities by adding a moving HD camera (also known as a pan tilt zoom camera) that focuses a single presenter, and can be controlled both automatically and manually. Manual control can be useful for focusing the camera vision on the product being currently presented or even, people in the audience. The HD camera can use up to 18x zoom, making it possible to install the camera far away from the presenter. Visualcom combined with the presenter tracking system adds support for much larger rooms, with 16 to 100 people inside, as well as stages. Visualcom & PTS can be used for educational purposes, promoting eLearning.

Presenter tracking system is a solution that uses a PTZ camera to follow a presenter over a medium sized area (3.5x4m). It is compatible with the VisualCom video conferencing system, allowing the user to walk around on stage, giving his presentation to the audience located in the room, while the camera is automatically following his movement. A moving camera adds a much larger viewing angle for the remotely connected audience and the automatic zoom function also gives a more enjoyable video conferencing experience.

The presented prototype consists of following components:

1. Kinect Senzor

A Kinect sensor (also called a Kinect) is a physical device that contains cameras, a microphone array, and an accelerometer as well as a software pipeline that processes color, depth, and skeleton data. The physical device is displayed on the picture bellow.



Figure 4 Kinect sensor

2. Kinect for windows SDK

The Kinect for Windows SDK provides a sophisticated software library and tools to help developers use the rich form of Kinect-based natural input, which senses and reacts to real-world events [5]. The Kinect and the software library interact with the developed application.

The needed components of SDK are following

a) Kinect hardware - The hardware components, including the Kinect sensor and the USB hub through which the Kinect sensor is connected to the computer.
b) Kinect drivers - The Windows drivers for the Kinect, which are installed as part of the SDK setup process as described in this document. The Kinect drivers support:
   - The Kinect microphone array as a kernel-mode audio device that you can access through the standard audio APIs in Windows.
   - Audio and video streaming controls for streaming audio and video (color, depth, and skeleton).
   - Device enumeration functions that enable an application to use more than one Kinect.
c) Audio and Video Components
   - Kinect natural user interface for skeleton tracking, audio, and color and depth imaging
d) DirectX Media Object (DMO) for microphone array beam forming and audio source localization.
e) Windows 7 standard APIs - The audio, speech, and media APIs in Windows 7, as described in the Windows 7 SDK and the Microsoft Speech SDK. These APIs are also available to desktop applications in Windows 8.

3. PTZ camera

For this project, PTZ Camera KT-HD60 was used. It is a high resolution, 18x zoom video capturing device. It outputs the HD signal through a HDMI out. Camera acquires commands via serial port and supports multiple pan tilt zoom camera protocols, such as Pelco D/P, and VISCA. In this solution, we have used VISCA protocol to communicate with the camera. Note that the camera doesn't follow Pelco D/P standards closely, so some commands will not function when using Pelco protocols. To use the camera in the upside down mode, you must flip the jumper that is located on the bottom side of the camera. Camera commands such as moving the motor left and right are automatically flipped, meaning any software adjustments are not necessary [3].

V Concept Validation

As mentioned before In this project we have used Kinect for Windows, although the application can be run with Kinect for Xbox, but with lower performance. Other part of the solution is HD PTZ camera which is used to follow presenter movements around the predefined space. the concept validation consisted of following goals:

1. Integration of Kinect with PTZ camera,
2. Developing Graphical User Interface (GUI),
3. Concept validation on existing videoconferencing system.

The integration was done in C#, .NET framework 4.0 and the Kinect SDK. Concept validation was done on Ericsson Visual Communication System. Presenter tracking system was tested using the Ericsson Visual Communication Hardware Client in small and large rooms. The system is very mobile, and can be quickly reconfigured to work in different rooms. The benchmarks proved that Presenter tracking system is a useful upgrade for the Ericsson Communication System.

User interface consists of two parts. The first part is image data received from the Kinect sensor, it represents the visual field of the Kinect sensor, and its application is for the presenter to know if he is locked down by the sensor. The second part of the interface are camera and Kinect controls and Kinect and camera data, such as 3D coordinates, presenter velocity, camera pan and tilt angle (See Figure 5). Kinect has limited tilting capabilities [6]. The user can change the sensor tilt angle by moving the tilt slider. User interface also provides a way to take over manual control of the camera. This function can be used to focus the camera vision onto a product that is being presented. Manual controls are started by pressing the Tracking ON button. While the manual controls are active, automatic tracking will be disabled, and vice versa. Camera movements are controlled with keypad arrow buttons (up, down, left, right). Camera zoom is controlled with keypad + for zooming in, and keypad – is for zooming out. It is also possible to force stop zoom by pressing the * key. Above the Tracking ON/OFF button there are 2 sliders.

In this project, customized Fzbin Kinect Gesture library [4] was used for Gesture implementation. System possibilities such like adding new gestures, using gestures with skeleton data and presenter gestures (lock-on and unlock gestures) were developed and implemented.

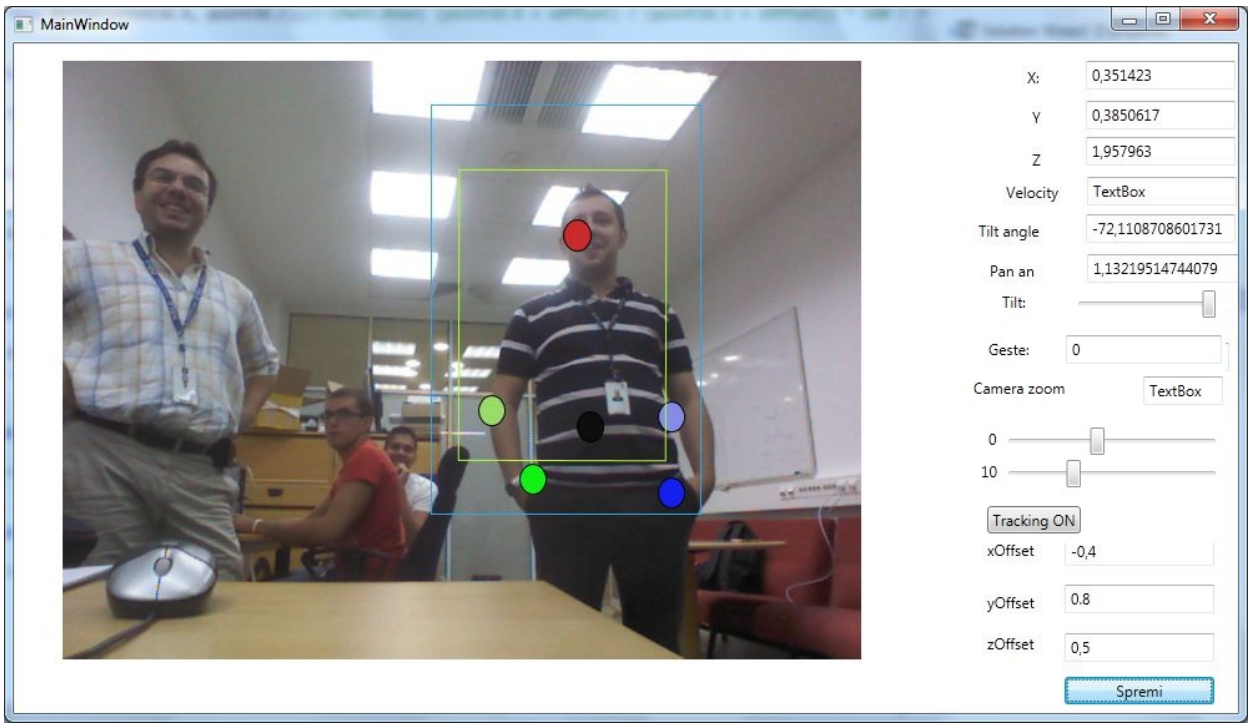


Figure 5 Hardware and Software Interaction with an Application

VI. Future work and development

Presenter Tracking system is visually stunning solution for halls and big rooms. Solution can be expanded in several different ways. Currently, the biggest problem for PTS is space for the presenter to move. One of the possible solutions for this problem is adding multiple Kinect sensors to cover more space. Future development can also include implementing new presenter gestures, such as swipe left and swipe right which can be used for controlling the presentation. For example, swipe left could mean next slide, while swipe right can mean previous slide. Current solution has gesture support system, so adding new gestures can be described as a task of medium difficulty.

Main direction for future development should be integration with VisualCom user client interface. This way, no additional processes would be needed, and PTS could be automatically started with the video conference. This could be described as a task of medium difficulty. PTS could also be expanded by adding an option for a second presenter, and move the camera depending on which of the presenters is currently speaking. Kinect API already has voice direction recognition, but the logic of camera movement depending on which of the presenter is speaking would be a problem. This could be described as a task of high difficulty.

Another suggested feature includes tracking members of the audience with an additional PTZ camera. Using Kinect's sound sensors, it is possible to determine the approximate direction the voice is coming from and then pinpoint it using facial recognition algorithms to determine which person is opening its mouth. The person in the audience who is talking would also appear on the screen, showing the dialogue between the presenter and the person in the audience, rather than just presenter talking to an unknown voice. This would work especially well for recording university classes, where a lot of questions are asked during the lecture. It's worth noting that many famous universities such as Stanford, UCLA are starting to offer online lectures.

VI. CONCLUSION AND FUTURE WORK

Audio and video communications are a current world trend. This trend produces a certain market, not only for end users customers, but for large companies as customers. VisualCom is one of the products that can help Ericsson conquer the market of HD visual and audio communication. It targets larger companies as end users. It is mainly designed for small conference rooms, but to expand the market this project was created and implemented.

Presenter Tracking System helps expand quality of

service for VisualCom serving as a plugin for video conferences and presentations in bigger rooms. This way, Ericsson can expand its offer to companies. In a sense, presenter is interacting with the camera using gestures, while the Kinect serves as a user-interface.
Implementation of Presenter Tracking System is complete; however some of the components in the solution could be replaced with higher quality products. For example, PTZ camera has a very slow auto focus, unstandardized values for VISCA protocol, low quality image stabilization at higher speeds. Regardless, PTS is a solid solution for economic expansion of Ericsson product portfolio.

VII. REFERENCES

[1] S. Arseneaz, J. R. Cooperstock: Presenter Tracking in a Classroom Environment, Industrial Electronics Society, 1999. IECON '99 Proceedings. The 25th Annual Conference of the IEEE

[2] Ericsson Internal materials

[3] HD Camera & VISCA documentation: http://www.ccrp.usp.br/cultura/EQUIPAMENTOS/MANUAIS/08CAMERA.

[4] Kinect Fizbin library: http://www.exceptontuesdays.com/gestures-with-microsoft-kinect-for-windows-sdk-v1-5/ (accessed 12.09.2013.)

[5] Kinect Documentation: http://www.microsoft.com/en-us/kinectforwindows/develop/resources.aspx (accessed 12.09.2013.)

[6] Kinect latest API changes, http://robrelyea.wordpress.com/2012/02/01/k4w-details-of-api-changes-from-beta2-to-v1-managed/

Iterative Receiver for Qc-LDPC Coded Underwater Acoustic Communication Systems

Liang Zhao*
* Xi'an Research Institute of Navigation Technology, Xi'an, P.R.China
xdyjs204@hotmail.com

Abstract - Quasi Cyclic-Low Density Parity Check (QC-LDPC) codes are easy to construct and provide the considerable coding gain, which is suitable for underwater acoustic communication (UWAC). Single-carrier (SC) transmission with frequency-domain equalization (FDE) is today recognized as an attractive alternative to orthogonal frequency-division multiplexing (OFDM) for communication application with the inter-symbol interference (ISI) caused by multi-path propagation, especially in shallow water channel. In this paper, the turbo theory is applied on the QC-LDPC codes and minimum mean square error (MMSE) decision feedback equalizer (DFE) to design iterative data processing for underwater acoustic communication system. In the proposed iterative structure, the MMSE based FD-DFE and QC-LDPC decoder exchange soft information through an iterative manner so that the performance of underwater acoustic communication system can be improved greatly. Based on sound speed profiles (SSP) measured in the lake and finite-element ray tracking method, the shallow water channel is constructed to verify the validity of the proposed system structure.

I. INTRODUCTION

The major cause of performance degrade in underwater acoustic communication system is the inter-symbol interference (ISI) introduced by multi-path propagation [1]. Multi-channel adaptive decision feedback equalizer (ADFE), as in [2-4], has been certified as effective approach to eliminate the ISI for the underwater acoustic multi-path channel. Meanwhile, channel coding is an effective technology to ensure reliable data transmission. Low-density parity-check (LDPC) code, firstly presented by Gallager [5], is the one most close to the Shannon limit of a class of error-correcting codes. From the point of view of modern coding theory, Mackay [6-7] proved that LDPC codes are capable of performance close to the Shannon limit. LDPC code has become the hot point in current error-correcting code field. Quasi Cyclic-Low Density Parity Check (QC-LDPC) codes [8] exhibit excellent decoding performance, simple encoding structure, lower encoding complexity and easily to implement with hardware.

For channels with severe delay spread, such as underwater acoustic channel, frequency domain equalization (FDE) is computationally simpler than corresponding time domain equalization because an efficient FFT operation is involved. A single carrier (SC) system with FDE (SC-FDE) has essentially the same performance and low complexity as an orthogonal frequency division multiplexing (OFDM) system when combined with FFT processes and the use of a cyclic prefix (CP) or zero padding (ZP) [9-10].

But, single technique, such as equalization or coding technique, is difficult to obtain satisfied data transmission because of the complexity of UWA channel, especially in shallow water channel. So, in this paper, an iterative receiver is developed combining MMSE based FD-DFE and Qc-LDPC decoder. In the proposed iterative receiver, the Qc-LDPC decoder and FD-DFE iteratively exchange soft information. And thus, the performance of equalizer can be improved utilizing decoding gain provided by decoder such that the performance of communication system is improved.

II. SYSTEM DSCRIPTION

For MPSK, let's define $m = \log_2 M$, where, M is number of phases. Irregular Qc-LDPC code is characterized by a $M \times N$ parity-check matrix $H(M, N)$ with variant column weight and variant row weight. The code rate can be calculated by $R = 1 - M / N$. In this paper, irregular Qc-LDPC codes with code rate *R=1/2* and *2/3*, are adopted in the proposed communication system.

In receiver, the received signal is firstly transformed into baseband by multiplying the local carrier frequency. And then, iterative data processing is performed. The structure of iterative receiver is shown in Fig.1. The concrete algorithm will be given in Section III.

III. ITIRATIVE RECEIVER

The block diagram of the proposed iterative receiver is illustrated in Fig.1. In its *l*th iteration, the frequency domain received vector feeds to the FD-DFE. This produces the vector feeding to a Qc-LDPC decoder producing *a posteriori probability* (APPs) of data, which are exploited to generate the extrinsic information. This is then processed to produce a soft estimate of the ISI in the received signal samples, and feedback to the feedback filter of FD-DFE for the next iteration.

In this section, we first derive FD-DFE algorithm. And then, an iterative receiver based on equalization and decoder is developed. As shown in Fig.1, the space diversity technology is adopted.

A. Fast Encoding of Qc-LDPC

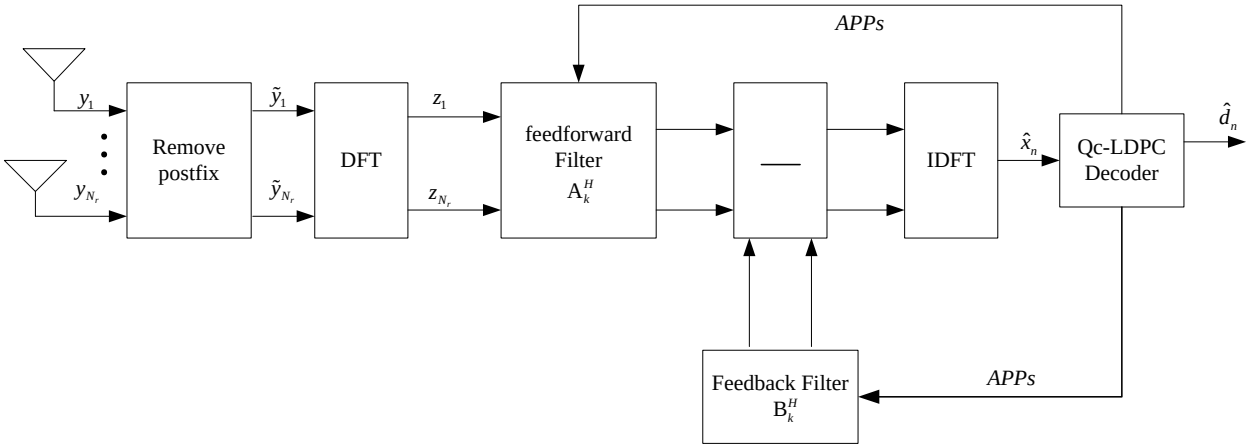


Fig.1 Block diagram of the proposed iterative receiver

Parity-check matrix H*(M,N)* can be divided into six sub-matrixes, that is

$$H = \begin{bmatrix} A & B & T \\ C & D & E \end{bmatrix} \tag{1}$$

where, the size of matrixes A, B and T are $(m-z)\times k$, $(m-z)\times z$ and $(m-z)\times(m-z)$, the size of matrixes C, D and E are $z\times k$, $z\times z$ and $z\times(m-z)$.

The encoded symbols can be divided into two parts: information bits S and parity bits $P=[P_1;P_2]$. The length of P_1 and P_2 are z and $m-z$ respectively, which can be calculated as follows:

$$P_1^T = \left(ET^{-1}A+C\right)S^T \tag{2}$$

$$P_2^T = T^{-1}\left(AS^T + BP_1^T\right) \tag{3}$$

So, the outputs of Qc-LDPC encoder are $[S;P_1;P_2]$.

B. MMSE Based Frequency Domain Decision Feedback Equalization (FD-DFE)

After performing *K* point discrete Fourier transform (DFT), the received symbols in frequency domain (FD) $\mathrm{Z}=\left[Z_0,Z_1,...,Z_{K-1}\right]^T$ are produced

$$\mathrm{Z}_k = \mathrm{H}_k S_k + \mathrm{V}_k \tag{4}$$

Where, $H_k(f)=H(k/KT)=\sum_{l=0}^{L-1}\hbar_l e^{-j2\pi k\tau_l/KT}$ is the frequency response of $h(n)$, and V_k is complex additive white Gaussian noise (AWGN) in FD. S_k is the *k*th component of the vector $\mathrm{S}_K=\left[S_0,S_1,...,S_{K-1}\right]^T$,which is the DFT of x_K, that is $\mathrm{S}_K=F_K\mathrm{x}_\mathrm{K}$, where, F_K is the *K*-point DFT matrix with the $(p+1,q+1)^{st}$ entry as $\left(1/\sqrt{K}\right)e^{-j2\pi pq/K}$.

The corresponding error in frequency domain (FD) is

$$\mathrm{E}_{DFE} = S_k - \hat{S}_k = S_k - A_k^H\mathrm{Z}_k + B_k^H\tilde{S}_{k'} \tag{5}$$

Appling MMSE criteria on eq.(5), and after mathematic derivation, we can get the optimal forward filter and feedback filter as

$$A_k = \mathrm{cov}(S_k,S_k)(\sigma_v^2 I + H_k\,\mathrm{cov}(S_k,S_k)H_k^H)^{-1}H_k^H \tag{6}$$

$$B_k = A_k H_k^H E(\tilde{S}_{k'},S_k^H) - E(\tilde{S}_{k'},S_k^H) \tag{7}$$

where σ_v^2 is the noise variance.

In iterative FD-DFE, the soft symbol estimations at last iteration $\tilde{S}_{k'}$, are used as the input of feedback filter. Thus, the $E(\tilde{S}_{k'},S_k^H)$ is equal to the correlation between transmitted and detected data sequences.

In order to estimate the parameters in eqs. (6) and (7), we introduce the mean, power, and variance of S_k in FD:

$$E(S_k) = \frac{1}{K}\sum_{i=0}^{K-1}E(x_k)\times\exp\left(-j\frac{2\pi ki}{K}\right) \tag{8}$$

$$E\left(|S_k|^2\right) = \frac{1}{K}\sum_{i=0}^{K-1}E\left(|\hat{x}_k|^2\right)\times\exp\left(-j\frac{2\pi ki}{K}\right) \tag{9}$$

$$\sigma_S^2 \triangleq \mathrm{cov}(S_k,S_k) = E\left(|S_k|^2\right) - |E(S_k)|^2 \tag{10}$$

The correlation $E(\tilde{S}_{k'},S_k^H)$ in eq.(7) can be expressed as

$$\begin{aligned}\gamma_{\tilde{S},S} &\triangleq E(\tilde{S}_{k'},S_k^H)\\ &= \frac{1}{K}\sum_{i=0}^{K-1}E(\tilde{x}_i,x_i^H)\exp\left[-j\frac{2\pi(k-k')i}{k}\right]\\ &= \frac{1}{K}\sum_{i=0}^{K-1}\rho_s\delta(k-k')\exp\left[-j\frac{2\pi(k-k')i}{k}\right]\end{aligned} \tag{11}$$

Where, $E(\tilde{x}_i,x_i^H)$ is the correlation in TD. $\tilde{x}_i$ is the estimation of the x_i in TD at the last iteration. So, in eq.(11), the ρ_s describes the reliability in the quality of the estimate $\tilde{x}_i$.

If $\tilde{x}_i$ is an excellent estimate of x_i, then $\rho_s \approx 1$, and nearly all of the interference is subtracted from the received symbols.

During the course of iterative processing, the APP $P(\hat{x}_k = \alpha_i)$ delivered by Qc-LDPC decoder is used to estimate the transmitted symbols. So, we can use it to approach the ρ_s :

$$\rho_s \cong \max_{i \in \mathbb{S}} (P(\hat{x}_k = \alpha_i)) \tag{12}$$

which means $\tilde{x}_k$ equal to x_k with the maximum probability.

Thus, using eq.(12), eq.(11) can be simplified to

$$\gamma_{\tilde{s},s} \triangleq E(\tilde{S}_{k'}, S_k^H) \cong \frac{1}{K} \sum_{k=0}^{K-1} \max_{i \in \mathbb{S}} (P(\hat{x}_k = \alpha_i)) \tag{13}$$

So, the optimal forward filter and feedback filter can rewrite as

$$A_k = \sigma_s^2 (\sigma_v^2 I + H_k \sigma_s^2 H_k^H)^{-1} H_k^H \tag{14}$$

$$B_k = A_k \hat{H}_k^H \gamma_{\tilde{s},s} - \gamma_{\tilde{s},s} \tag{15}$$

After IDFT, we can get TD transmitted symbol $\hat{x}_k$, which will be send to the decoder to generate the final bits decision $\hat{d}_k$ and the APPs of transmitted symbols for next iteration. As mentioned above, the output vector after IDFT is applied to a Qc-LDPC decoder producing the APPs of data, which are exploited to generate the extrinsic information. This is then processed to produce a soft estimate of the ISI in the received signal samples, and to adjust the iterative FD-DFE for the next iteration.

In this paper, Qc-LDPC decoder adopts min-sum algorithm (MSA) [11] as the decoding algorithm, which has the linearly complexity.

IV. Simulation Results

In this section, based on sound speed profile (SSP) measured in the lake and finite-element ray tracking method, underwater acoustic channel is constructed to evaluate the performance of the proposed iterative receiver.

A. Channel Model and System Parameters

The sound speed profile (SSP) measured in the lake is shown in Fig.2. The water with mixed gradient is about 33m depth. Based on the SSP, the shallow water channel with multipath propagation is constructed using ray tracking method.

Before data symbols are transmitted, the preamble symbol is transmitted to detect the channel impulse response (CIR). The LFM signal is used as preamble in our system. In Fig.3, we can see that there are two receiving elements are deployed (Ch1 and Ch2).

The system parameters of computer simulation are shown in Table 1. The transmitted signal with QPSK modulation is adopted.

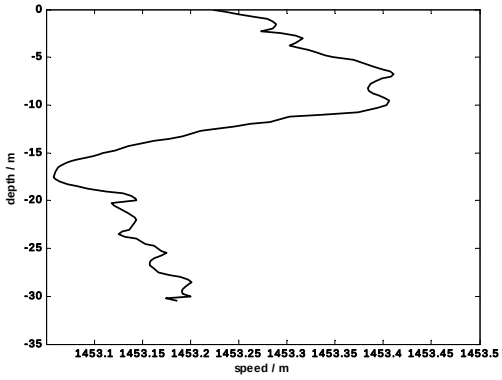


Fig.2 Sound speed profile (SSP)

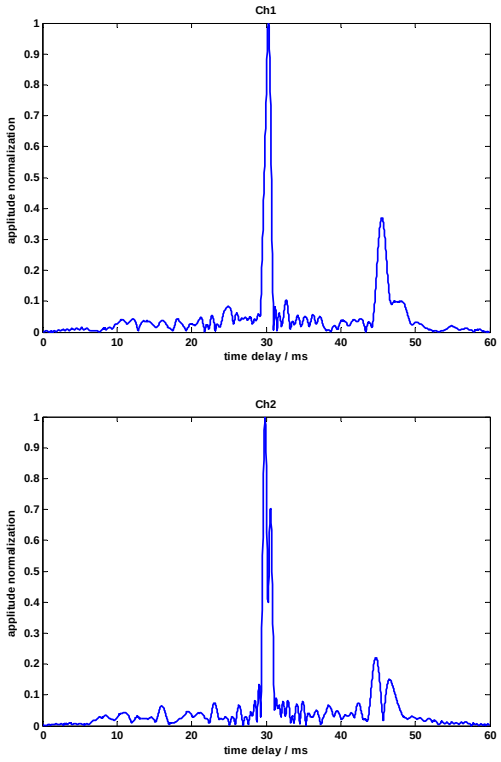


Fig.3 Channel impulse response (Ch1 and Ch2)

TABLE I. System Parameters

Parameters Name	***Values***
Carrier frequency	10 *KHz*
Symbol rate	5 *Ksps*
Sample rate	40 *KHz*
Transmission distance	2000 *m*
DFT length (*K*)	600
Code rate	1/2 ; 2/3
Receiving elements	2

B. *Performance of Iterative Receiver*

Fig.4 and Fig.5 show the bit error rate (BER) curves of the proposed iterative receiver using two receiving antennas with variant numbers of iterative equalization and decoding (marked with "IED" in Fig.4 and 5) and iterative Qc-LDPC decoding (marked with "Qc-LDPC decoding" in Fig.4 and 5). We can see that the proposed iterative receiver can sufficient utilize the decoding gain provided by decoder to improve the system performance. Thus, the data transmission with lower BER can be obtained. What's more, more Qc-LDPC iterations can obtain better performance under the condition of the same IED iterations.

Comparing Fig.4 with Fig.5, we can find that the system performance will reduce as the code rate rose. That is because there are more parity bits for lower code rate such that the more error will be corrected.

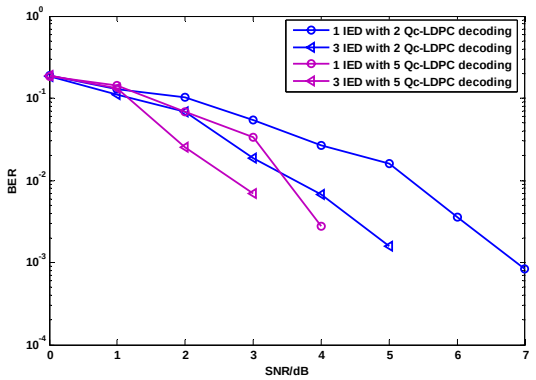


Fig.4 The performance of the iterative receiver with 1/2 code rate

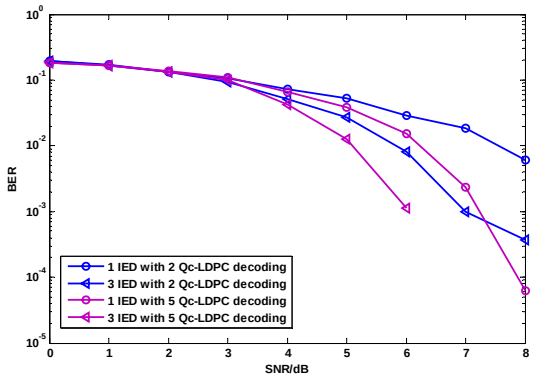


Fig.5 The performance of the iterative receiver with 2/3 code rate

V. CONLUSION

In this paper, according to the characteristics of underwater acoustic channel, an iterative receiver based on frequency domain decision feedback equalization (FD-DFE) and Qc-LDPC decoder is investigated to overcome Inter-symbol interference (ISI) caused by multi-path propagation. Based on SSP measured in the lake and finite-element ray tracking model, the shallow water channel is constructed to evaluate the performance of the proposed iterative receiver. The results suggest that the proposed iterative receiver can significantly improve the performance utilizing decoding gain provided by decoder and obtain better data transmission. Moreover, the proposed FD-DFE algorithm is developed for multi-channel receiver. So, it can easily extend to MIMO system.

REFERENCES

[1] R.S.H. Istepanion, M. Stojanovic. Underwater Acoustic Digital Signal Processing and Communication Systems. Kluwer Academic Publisher, London, UK, 2002.

[2] D.B. Kilfoyle, A.B. Baggeroer, "The state of the art in the underwater acoustic communications," IEEE J. Ocean. Eng., vol. 25, No. 1, 2000, pp.4-27.

[3] M. Stojanovic, L. Frieitag. "Wideband Underwater CDMA: Adaptive Multichannel Receiver design," Proceedings of Oceans 2005 Conference, vol.2, 2005, pp.1508-1513.

[4] L. Zhao, W.Q. Zhu, M. Zhu. "Adaptive Equalization Algorithms for Underwater Acoustic Coherent Communication System," J. Electron. Inf. Technol., vol.30, No.3, 2008, pp.648-651.

[5] R.G. Gallager. Low-Density Parity-Check Codes. Cambridge :MIT Press,1963.

[6] D.J.C. Mackay, R.M. Neal. "Near Shannon limit performance of low density parity check codes," Electronic Letters, vol.32, No.8, 1996, pp.1645-1646.

[7] D.J.C. Mackay, "Good error-correcting codes based on very sparse matrices," IEEE Trans.Inform.Theory, vol.45, No.2, 1999, pp.399-431.

[8] M.P.C. Fossorier. "Quasi cyclic low-density parity-check codes from circulant permutation matrices," IEEE Transactions on Information Theory, vol.50, No.8, 2004, pp.1788-1793.

[9] A. Czylwik. "Comparison between adaptive OFDM and single carriermodulation with frequency domain equalization," Proceedings of 47th Vehicular Technology Conference−VTC, vol.2, 1997, pp.865–869.

[10] F. Pancaldi, G.M. Vitetta, R. Kalbasi, N. Al-Dhahir, M. Uysal, and H. Mheidat. "Single-carrier frequency domain equalization," IEEE Signal Process. Mag., vol.25, No.5, 2008, pp.37–56.

[11] T. Richardsen, R. Urbanke. "The capacity of low-density parity check codes under message passing decoding," IEEE Trans. On Info. Theory, vol.47, No.2, 2001, pp.599-618.

SOFTWARE PLATFORM FOR WIRELESS INDOOR LOCALIZATION PLANNING BASED ON THE STRENGTH OF RECEIVED RADIO SIGNAL

Igor Vitas, Bojan Lukovac, Dina Šimunić

Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia

ABSTRACT

Abstract – Radiofrequency positioning problem within indoor space becomes significant with respect to the specific area for which the radiolocation is required. The area of interest may be an apartment, shop or storage area of any shape and size. The specificity of the space defines various requirements on the radiolocation system and the need for software that would facilitate the planning of such a system is obvious. Using well-known empirical and deterministic models for propagation of electromagnetic waves in enclosed spaces and precise analysis of specific closed areas, localization based on the RSS (Received Signal Strength) method should give good positioning results in any enclosed space. Software solution that was developed in this work allows you to determine the optimal position of the receiver on the walls of any space that is selected as an area in which implementation of a system of RSS indoor localization is required. The optimal position of the receiver provides precise and reliable localization within the indoor enclosure. The software can be used to optimize the indoor localization of people or objects, with the requirement that special radio frequency tags or radio frequency transmitters are mounted on them. The software supports the ability to import an arbitrary layout of indoor space over which optimization will be carried out from AutoCAD or a CAD tool.

Keywords **–** radiofrequency indoor positioning, software tool, received radio signal strength

1. INTRODUCTION

Wireless positioning has been an issue of high research interest in the last two decades [1]. Numerous systems for wireless positioning are already in use, most of them are global systems used for positioning of people or objects (e.g.. cars, trains, planes) on the surface of the Earth. The most popular system used for such a purpose is definitely GPS – the Global Positioning System. Positioning of people and objects is an efficiently solved problem in open space areas (Earth surface), because satellite positioning systems are advanced systems built for the specific application of positioning. A significant problem with that kind of system is that they do not work in areas that cannot be covered with the radio signal from a satellite, such as indoor spaces, tunnels, Earth surface areas with high forest density and so on. The strength of the radio signal required for normal work of the GPS system is usually too weak in indoor environments, so specific indoor positioning systems are being developed[2][3][4].

Similarly to outdoor positioning systems, indoor positioning systems also work on three fundamental principles of radiolocation:

1. Received radio signal strength
2. Time of flight of the radio signal
    a) Time of arrival
    b) Time difference of arrival
    c) Phase of arrival
3. Angle of arrival of the radio signal

The indoor propagation environment is much more complicated than an outdoor propagation environment like free space. It is, for example, hard to precisely predict a relationship between distance and received radio signal strength because of the unpredictability of the indoor propagation environment. Indoor propagation is very space-dependent because the number of indoor spaces that can be used as a propagation scenario is infinite. Also, indoor spaces can contain very different kinds of materials that effect propagation such as a walls, furniture or any kind of in-house equipment (e.g. electrical equipment with metal housing).

Propagation in indoor environments can be described easily with a modified Friis equation (1). Indoor path loss (or path gain) for a given distance d will be a random variable that will vary around some expected value.

$$PG_{dB} = 10n \log\left(\frac{\lambda}{4\pi d}\right) + X_\sigma \quad (1)$$

X_σ is a random variable with standard deviation. Parameters for the varible X_σ regarding various indoor spaces can be seen in Table 1.

In the case where equation (1) is not used for the calculation of path loss in indoor space, some empirical equations that determine path loss in specific indoor space can be used. Most popular empirical equations and models used for determination of path loss in indoor space are COST 231 Multi Wall, COST Average Wall and Motley-Keenan model [5].

Table 1 Comparison of standard deviation of random variable X_σ and propagation exponent for various indoor spaces and frequencies.

Propagation space	Frequency (MHz)	Propagation exponent	Variable deviation (σ)
Supermarket	914	2.2	8.7
Office, frequency 1	1500	3.0	7.0
Office, frequency 2	900	2.4	9.6
Factory	1900	2.6	14.1

Performance of radio technologies in indoor environments is restricted, so many systems for indoor positioning are hybrid systems that use radio technologies coupled with other technologies such as infrared, ultrasound or optical systems [6][7][8] . Another very popular option is also a combination of radio technologies coupled with electronic accelerometers that can be used for determination of motion. Motion detection with use of an electronic accelerometer provides good results in various indoor localization systems.

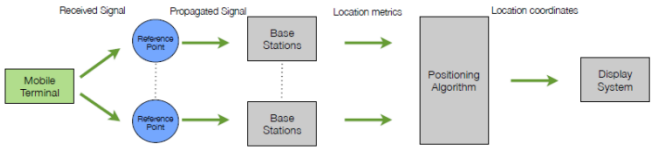


Figure 1 Architecture of typical wireless positioning system.

Because all commercial wireless technologies that are used in indoor space such as Bluetooth, WLAN (IEEE 802.11), ZigBee (IEEE 802.15.4) have the possibility to measure the received signal strength of an incoming radio signal [9], the easiest way to implement an indoor positioning system is to measure the received signal strength from multiple access points on one mobile node. In that case, the mobile node needs to calculate its own position using signal strength measurement and known location (coordinates) of all access points in indoor space [10]. The minimal number of access points needed for this unilateral measurement procedure is three. The process of finding the position of a mobile node in a network that is made of three wireless access points and one mobile node is called trilateration. Trilateration is calculated using equations (2), (3) and (4).

$$(x_1 - x)^2 + (y_1 - y)^2 + (z_1 - y)^2 = D_1^2 \quad (2)$$
$$(x_2 - x)^2 + (y_2 - y)^2 + (z_1 - y)^2 = D_2^2 \quad (3)$$
$$(x_3 - x)^2 + (y_3 - y)^2 + (y_1 - y)^2 = D_3^2 \quad (4)$$

$D_1, D_2, D_3 \rightarrow distance\ between\ access\ point\ and\ mobile\ node$

$(x_n, y_n) \rightarrow\ access\ points\ coordinates$

$(x, y) \rightarrow mobile\ node\ coordinates$

The localization algorithm that is used in a received signal strenght localization system needs to minimize function (5) quickly and accurately. Minimizing the equation F, the system can calculate position coordinates of a mobile node in indoor space.

$$F = \sum_{i=1}^{M} (\sqrt{(x_i - x_e)^2 + (y_i - y_e)^2} - D_i)^2 \quad (5)$$

$M \rightarrow number\ of\ system\ access\ ponts$

For systems that work with a received signal strength localization algorithm, it is very important to make a good prediction of path loss in the envirnonment in which the system is located. The path loss model will be used by the positionig system to determine values D_i in equation (5) and because of that, it will directly affect positioning system precision.

Choosing the correct path loss model for a specific indoor space scenario is not an easy task. The propagation exponent and deviation shown in Table 1 will theoretically be differenet for every indoor space and they will highly depend on the current situation in that space, such as the type of the building walls, presence of people, furniture, position of wireless access points on walls of the space and so on. The most accurate way to choose a path loss model for specific indoor space is the method of radiofingerprintig. Radiofingerptinting can use a finite number of measurements inside a specific indoor space (e.g. measurement of path loss between a raster of points in the space and every acces point in the space) to determine the average propagation exponent and variable deviation for that space. With the calculated average propagation exponent and variable deviation found by radiofingerprinting, it is possible to most accurately calculate values of D_i , and mobile node position. If we consider that the number of radiofingerprinting measurements for a specific indoor space is N, we can calculate propagation exponent for every point of measurement i in indoor space as shown in equation (6).

$$prop_exponent(i) = \sqrt{G_T G_R \frac{P_T}{P_{R(i)}} \frac{\lambda^2}{16\pi^2}} \quad (7)$$

Average propagation exponent that is used to describe the entire indoor space is then calculated as (7).

$$prop_exponent_space = \frac{1}{N} \sum_{i=1}^{N} prop_exponent(i) \quad (7)$$

$N \rightarrow number\ of\ measurement\ points\ in\ indoor\ space$

2. SOFTWARE PLATFORM

The goal of our developed software is to optimize a specific indoor space in terms of indoor radiofrequency-based positioning. An imported AutoCAD file(.dwg format) represents an architectural visualisation of indoor space (e.g. an apartment, house, warehouse, factory, shopping center). Processing is done over the AutoCAD file to produce a geometric model of the indoor space over which calculations will be taken. The software uses propagation models of

electromagnetic waves that can accurately simulate propagation conditions in any indoor space that is imported from the AutoCAD tool to our software. With those propagation models, the software can suggest cooridantes of ponts in the indoor space in which installation of wireless access points is required to cover the entire area with sufficient radio signal strength for normal operating of a specific wireless system (e.g. WLAN).

Positions of wireless access points are choosen with two conditions:

1. Optimal coverage of the entire indoor space with sufficient strength of radio signal – first access point is set in the geometric middle of the indoor space, every next access point is placed in the point with the minimal level of radio signal strength and so on, until every surface is covered with sufficient signal for at least three access points. In this example,the chosen threshold for determing minimal possible allowed radio signal strength in any point of the space is -89 dBm (IEEE 802.11 system with 8% PER and 11 Mbps transmission speed).

2. Positioning of access points in space so they can be efficiently used in the trilateration algorithm for indoor localization using received radio signal strenght – in the ideal case, trilateration would be most precise if transmitters of the system would form an equilateral triangle as shown in figure 2. A special algorithm is used to place access points of the system (wireless transmitters) in space so the three access points form three vertices of a geometrical shape that is as close as possible to an equilateral triangle.

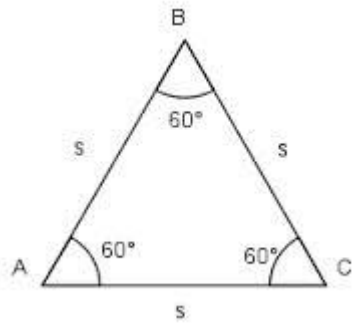


Figure 2 Equilateral triangle with wireless transmitters A, B and C in its vertices.

Virtual radiofingerprinting of the indoor space is also calculated in a resolution that the user can input (ideally it would be $\frac{\lambda}{10}$ meters for radio signal with a wavelenght of λ). Figure 3 shows the indoor space (apartment) that is imported to our software with the purpose of optimization, the selected resolution in this example is 0.5 meters. Higher resolution of space also means a higher computation time for the software to determine indoor space optimization.

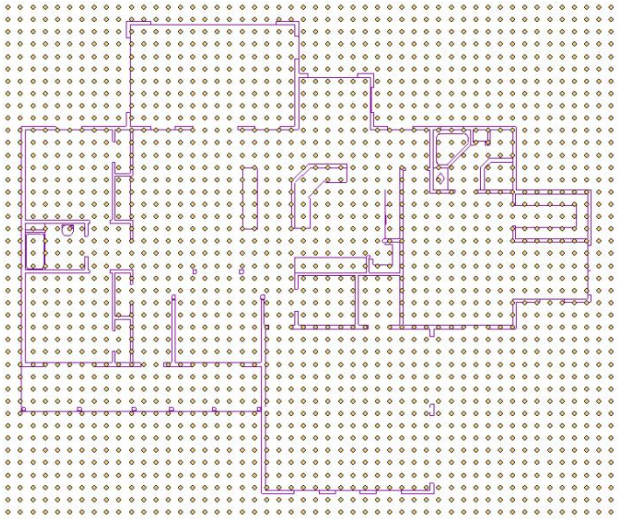


Figure 3 Flat with 0.5 meter resolution of calculation domain.

When radiofingerprinting of the whole space is completed and when the average propagation exponent is calculated with equation (7), that information is used to make the indoor localization system path loss model. This model will be adapted to fit the specific indoor space that is imported to the software so that the localization with that model in a real system that would be implemented in actual indoor space would be as accurate as possible.

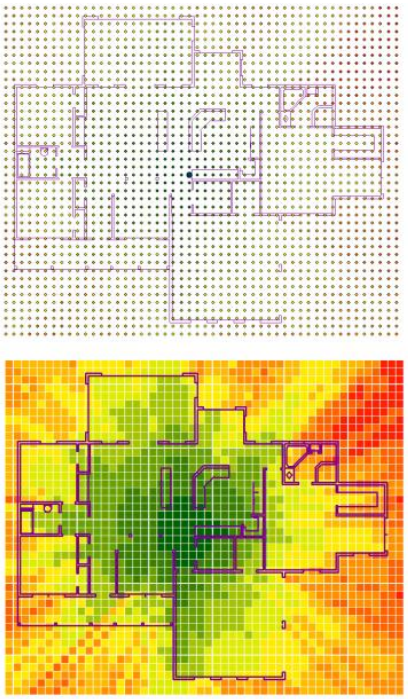


Figure 4 Radio coverage of the indoor space with first transmitter in the middle of the apartment (blue point on the first picture).

Figure 3 shows the placement of the first transmitter in the geometric middle of the flat. The second and third transmitter are set on the positions determined by the software regarding conditions 1 and 2. The placement of the second transmitter is shown on figure 4 and the placement of the third

transmitter with respective radio coverage of the flat is shown in figure 5.

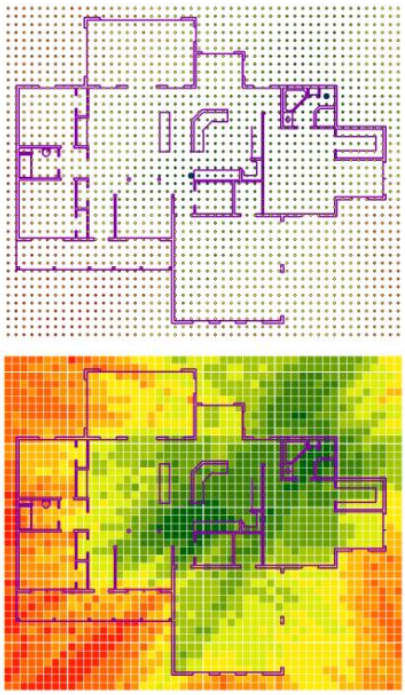


Figure 5 Radio coverage of the indoor space with first two transmitters in the apartment (two blue points on the first picture).

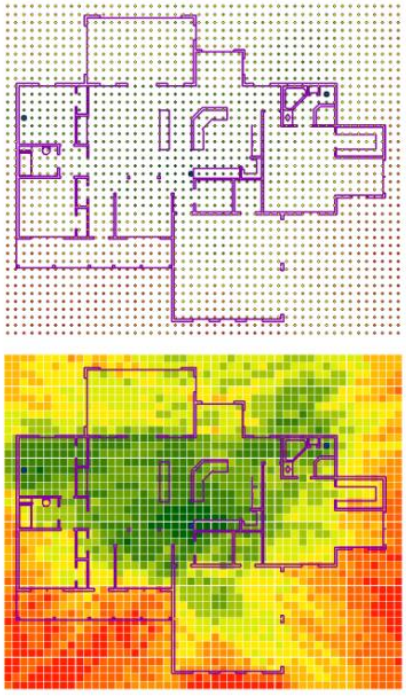


Figure 6 Radio coverage of the indoor space with three transmitters in the apartment (three blue points on the first picture).

In the used example for indoor the layout showed in figure 3, the software calculated $prop_exponent_space$ parameter to be 2.7568. The transmitter power used in this example was 20 dBm which is the typical strength of a WLAN 802.11 transmitter. The propagation model that was used to simulate real world propagation was the COST Average Wall model (8).

$$L(dB) = L_0 + 20\log d + k_w L_w \quad (8)$$

$k_w \rightarrow number\ of\ walls$
$L_w \rightarrow attenuation\ of\ average\ wall$

When software is over with calculations for specific space that is being optimized, all informations for indoor positioning system implementation are ready. Mobile node that is being tracked by three access points set by the software needs to pass initialization before it is put in use. Initialization consists of memorizing the indoor space map, coordinates of three access points and the $prop_exponent_space$ parameter. When the initialization of mobile tracked node is completed, the node can be moved freely in space and trilateration can be done on the mobile node device for the application of determing mobile node position.

Figure 7 shows scale that can be used to read radio signal strength (dBm) for every radio coverage map example in this work.

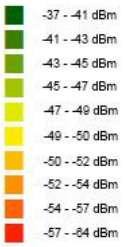


Figure 7 Radio signal strength values.

3. POSSIBLE ERRORS IN MEASUREMENT OF RADIO SIGNAL STRENGTH WITH COMMERICAL WIRELESS TECHNOLOGIES

Received signal strength indicator (RSSI) in some wireless technologies, (for e.g. WLAN, Bluetooth, ZigBee) is primarily used to set the output power of the transmitter with respect to the estimated distance of the second wireless module. Indicator of the received radio signal strength usually does not show its measured radio signal power in dBm, it actually shows a voltage which would ideally be increased linearly with the logarithmic change of the radio power at the entrance of radio signal strength measuremnt circuit. The output of the signal strength indicator is in linear scale, but the input is shown in a logarithmic scale because of the large variation in the strength of the received radio signal which may occur in the radiofrequency receiver circuit of wireless access point.

Implementation of the wireless localization system with some wireless technology (e.g. Bluetooth, WLAN, ZigBee)

technology depends largely on the implementation of the RSSI circuit. Transmission characteristics of the RSSI circuit [9] that is presented in Figure 9 shows typical transmission characteristic of the RSSI circuit implemented in CMOS technology, the error in the RSSI measurement of this circuit for the worst case is around 7.5 dB. Design of a typical RSSI microelectronic circuit which is implemented in CMOS technology and incorporated into commerical chips is shown in Figure 8 [9]. The circuit consists of a cascade of limiting amplifiers, full wave current rectifiers and low-pass filters.

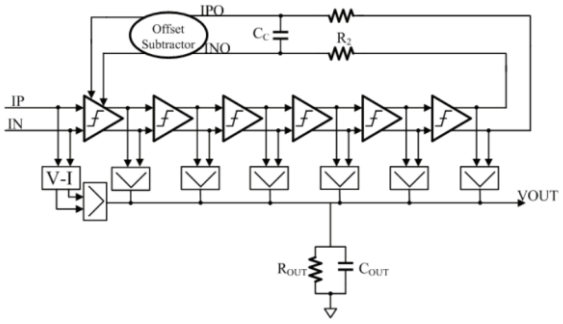


Figure 8 Typical microelectronic design of RSSI circuit [11].

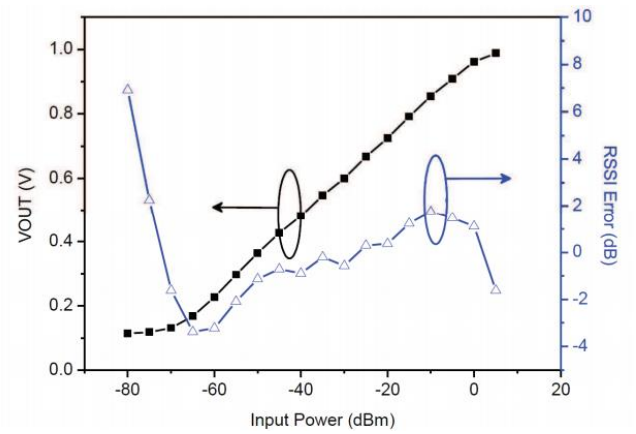


Figure 9 Transfer characteristic of the RSSI circuit [11].

RSSI circuits that are implemented, for example, in commercial Bluetooth devices are not equal because of their non-ideal manufacturing process. In Figure 10 deviation of RSSI circuit transfer characteristic for three Bluetooth devices can be seen. All three compared Bluetooth devices from Figure 10 were bought as the components with equal technical characteristics.

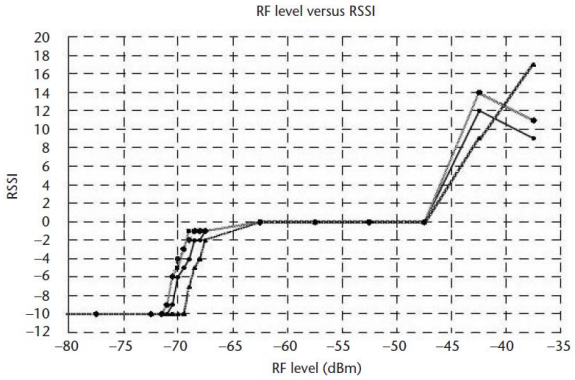


Figure 10 Example of RSSI characteristic for three randomly chosen Bluetooth devices.

4. CONCLUSION

This work presents developed software that is used for the optimization of various indoor spaces in terms of radiofrequency positioning and radio coverage. This software can be used to implement radiolocation services in indoor spaces with cheap, "off the shelf" wireless technologies that possess the ability to measure received radio signal strength. To make this software more accurate and reliable, a more complex 3D propagation model such as an indoor ray-tracing model needs to be implemented as a model that represents real world propagation in a particular environment. Errors that are introduced in the measurement of radio signal strength by various wireless technologies such as WLAN, Bluetooth and ZigBee will also be introduced in the next version of our software.

REFERENCES

[1] Alan Bensky, „Wireless Positioning Technologies and Applications“, 2008 ARTECH HOUSE, INC. 685 Canton Street, Norwood, MA 02062

[2] Morrison, J.D., „IEEE 802.11 Wireless Local Area Network Security Through Location Authentication“, Thesis from Naval Postgraduate School, Monterey, CA, September 2002.

[3] Hallberg,J.,M.Nilsson,and K.Synnes,“Positioning with Bluetooth“,10th International Conference on Telecommunications ICT 2003, Tahiti, French Polonesia, pp. 954-958 vol.2, February 23-March 1, 2003.

[4] Feldmann, S.,et al., „An Indoor Bluetooth-Based Positioning System: Concept, Implementation and Experimental Evaluation“, International Conference on Wireless Networks, page 109-113. CSREA Press, *(*2003)

[5] C. Seroido, L. Coutinho, L. Reigoto, J. Matias, A. Correia, P. Mestre, „A Lightweight Indoor Localization Model based on Motley-Keenan and COST“, Proceedings of the World Congress on Engineering 2012, July 4-6, 2012, London, U.K.

[6] J. Hightower, R. Want, and G. Borriello, “SpotON: An indoor 3D location sensing technology based on RF signal strength,” Univ. Washington, Seattle, Tech. Rep. UW CSE 2000–02-02, Feb. 2000.

[7] “Software and Hardware For High-Precision Indoor Localization And Navigation of Mobile Devices”, University of Florida, patent number UF #14439, March 2003.

[8] Tilch,S.Mautz,R., “Current investigations at the ETH Zurich in optical indoor positioning”, IEEE Positioning Navigation and Communication (WPNC) conference, pp. 174-178, 2010.

[9] Wang, K.,et al., „A CMOS Low-Power Receiving Signal Strength Indicator using Weak-Inversion Limiting Amplifiers“, 2011 International Conference of Electron Devices and Solid State Cicuits (EDSSC), pp. 1-2, 17-18 Nov. 2011.

[10] Bandara, U., et al., „Design and Implementation of a Bluetooth Signal Strength Based Location Sensing System“, IEEE Radio and Wireless Conference RAWCON 2004,pp. 319-322, Atlanta, Georgia, September 19-22, 2004.

Management in M2M networks

V. Cackovic, Ž. Popovic
Ericsson Nikola Tesla d.d., Zagreb, Croatia
vanesa.cackovic@ericsson.com
zeljko.popovic@ericsson.com

Abstract - The vision of future networks usually depicts huge amount of different type of devices connected through heterogeneous networks. All those devices do not act alone, but many devices act together to achieve some goal. Although distributed processing is critical to achieve the mass deployment of systems for Machine-to-Machine (M2M) communication, centralized decision making and centralized management of billons of devices will become the key challenge for achieving the global mass market.

Unlike management of traditional telecom network, M2M networks have some characteristics specific to M2M communication that need to be considered while designing efficient M2M network management. Examples of those characteristics would be networks including sleeping devices, low power lossy area networks, heterogeneous networks and device intelligence. Certain characteristics like mobility, time sensitivity of data and data volume differ from application to application but also need to be considered since they put requirements for efficient end to end management.

The paper will focus on identifying the minimum requirements needed for M2M network management, propose mechanisms for efficient management in M2M networks as well as describe the relevant standardization activities for network management in M2M systems.

I. INTRODUCTION

The market for Machine-to-Machine (M2M) communication services is not new, but is rapidly developing, with a predicted yearly growth of more than 30%. Challenges in the space include a multitude of business models, ecosystems and a huge diversity in services, protocols and devices. Ten predictions for M2M and Internet of Things (IoT) in 2014 identified by global researchers and analysts [1] on M2M are:

1. Dawning realization of the evolution from M2M to IoT. There has been a lot of buzz around IoT in 2013. This year is when that becomes a reality. The old "stove-pipe" M2M application approach is giving way to a more integrated approach to application development which draws from a wider range of data sources (including M2M connected devices of different types and also corporate and other IT systems) and stitches these together with more sophisticated applications. This evolution presents massive opportunities but also threatens to sideline any company in the M2M ecosystem that does not evolve.

2. More traffic and more average revenue per user (ARPU). The old days when M2M was only about low volume low ARPU sensor networks is passing. There will be more high bandwidth applications, generating higher ARPU.

3. Fears over 2G re-farming will continue to interfere with the growth of M2M. Buyers of M2M are struggling with how to cope with mobile network operators (MNO) plans for 2G switch off and the impact it will have on their technology choices. This will inhibit growth in some geographies in some sectors in 2014.

4. Low power wide area networks will see much more widespread deployment. There will be much more substantial deployments of low power wide area radio access technologies such as SigFox and Weightless. These networks are very good fit for many M2M applications and a number of announcements of network deployments are expected during 2014.

5. Customers will move away from roaming-based services, towards localized connectivity using remote provisioning (and mobile network operators/service providers will resolve associated commercial issues). In December 2013 the GSM Association published its specifications for Over-The-Air (OTA) provisioning of M2M SIMs. This is on the back of pre-standard solutions available from vendors such as Giesecke&Devrient [2] and Gemalto [3]. Customers and operators alike will recognize the benefits of multi-national (i.e. relationships with facilities-based operators) in each territory, rather than relying on roaming SIMs. The service providers and operators offering these localization options will resolve the commercial arrangements necessary to let it happen.

6. Those selling to the M2M market will find their focus, and smaller operators will find their feet. Today almost everyone in the M2M value chain is attempting to address all customers: MNOs, service providers, module vendors, systems integrators, and everyone in between, are trying to sell to the same set of M2M to customers. As a result they are spreading themselves too thinly trying to be all things to all potential customers. There will be a settling down in the value chain in 2014 as players find their position and the customers to which they should most appropriately be addressing themselves. Smaller, national MNOs will begin to assert their position in the M2M market as they come to better understand their optimum strategy and key market segments.

7. The big data analytics segment will rise...and fall. The value of data is in having a monopoly on the right kind of data or in being able to use it in some clever way. But the market will come to realize that there isn't

much value in the actual analytic part of big data analytics.

8. More merges and acquisitions are going to happen. The M2M industry is evolving very rapidly and with changing dynamics we would expect more consolidation, as all major players gear up to become full service providers. The consolidation will cut across all sectors, including device vendors, service providers and software/cloud platforms.

9. The ratio of products revenue to services revenue will sort the big players from smaller ones. With huge growth expected across all of M2M, it will be those players best able to productize their offerings that will win through. The cost of service provision tracks service revenues, only products scale allowing products-focused companies to outcompete services-focused companies in the same domain.

10. Connected homes finally arrive. They have been promising so much for so long, but 2014 looks like it will be the year when it actually gains some momentum.

To summarize, the world of M2M communications is continuing to grow steadily as predicted few years ago. Operators and service providers that build M2M solutions in a smart and efficient way, solutions that are based on relevant standards [4], [5], [6], [7], [8] will be better in the game of getting the M2M market lead than ones building vertical isolated solutions.

Unfortunately there is still a lot of misunderstanding regarding the complexity and impact on existing networks when talking about end-to-end M2M solutions. Similar situation happened in the past with IMS (IP Multimedia Subsystem) when IMS was firstly perceived as one single platform that will easily enable launching "killer applications". The fact is that IMS solutions similar to M2M solutions consists of several nodes responsible for different functionalities, that integration with supporting systems like charging, provisioning and similar is not trivial and that holistic approach which takes into account clients, network, services and so on is needed when designing network.

Reducing complexity and enabling cost efficient management of devices by enabling one horizontal device management solution instead of a multitude of vertical management solutions will be one of the key components of M2M communication solutions and it will be explained in more details in the following chapters.

II. Overview of M2M Architecture

M2M system is a combination of various heterogeneous electronic, communication, and software technologies. The general architecture of M2M networks such as those being specified in ETSI TC (technical committee) [5] M2M is shown in Figure 1 below.

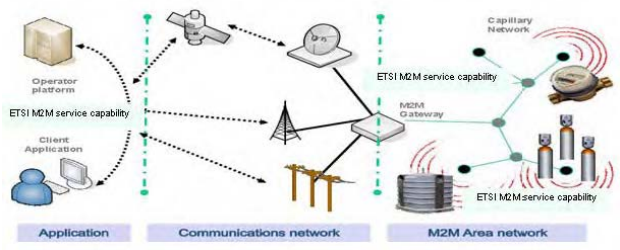


Figure 1 ETSI M2M architecture

In reference to this architecture, M2M devices (intelligent and communication enabled) form an M2M area network; this could be from a small-scale home environment to a bigger environment such as company. The M2M area network is connected to the communication network such as cellular base stations through the M2M gateway. Through the communication networks, they are connected to the M2M management server on the M2M service platform and subsequently reaching the M2M applications on the other side of the M2M management server.

Data in M2M systems travel from an area network to a communication core network. When data leaves the area network and arrives to communication core network, its first destination is usually a gateway. The gateway is a device which decides the communication protocols and converts the incoming data into legacy system format. A middleware layer that routes data and converts the format of the data could also be part of the communication core network. Middleware layer could also perform network management tasks such as logging, notification, auto-configurations, etc. Some parts of the business logics could also be applied in these middleware's and gateways.

Communication network can be of any type including wireless LAN, telephone lines, Ethernet, satellite or cellular networks. Finally, the data gets integrated into a business process at the M2M application layer such as smart metering, fleet management, eHealth etc. The data flow explained above is not always exactly the same; it varies from application to application. Some components will merge with others as time passes and M2M systems mature.

III. Standards for M2M Network Management

Since standards for M2M communications are written relatively recently, M2M network management is also a new area of investigation. In this section, the main standards related to M2M network management will be explained.

ETSI defined in M2M Functional architecture document [5] the overall end to end M2M functional architecture, including descriptions of the functional entities, the related reference points and the service capabilities; M2M identifiers, information model; procedures for bootstrapping, security, and management; charging; and implementation guidance. Besides its other functionalities, the Service capability layer (see Figure1) supports remote management of gateways, devices and

area network, including performance management, configuration management, fault management and software and firmware upgrading management [5]. In the functional architecture management resources are described as MgmtObj. The resources are handled with the RESTful style of data exchange. The four basic methods in the RESTful architecture are Create, Retrieve, Update and Delete.

ETSI "Machine-to-Machine communications (M2M); BBF TR-069 compatible Management Objects for ETSI M2M" The document [11] defines the corresponding BBF-TR-069 compatible Data Model. In its TR-069 the BroadBand Forum (BBF) specifies a protocol for communication between a CPE (Customer Premises Equipment) and an ACS (Auto-Configuration Server). ETSI M2M defines a Device Management function that allows to remotely configure the M2M Devices by using already existing Device Management protocols (e.g. BBF-TR-069, OMA-DM, etc.). Due to the RESTful nature of the interface with the M2M Applications, some ETSI Management Objects are specified and modelled as REST resources in [5]. Some of these ETSI resources may be mapped on already existing Data Models / Management Objects as defined by BBF or/and OMA. The corresponding mapping rules are specified in [6]. Some other resources do not correspond to any existing Data Models/Management Objects as defined by BBF or/and OMA.

Similar to the previous technical specification, ETSI published also ETSI TS 103 092 for OMA DM compatible Management Objects [12].

Open Mobile Alliance (OMA) on the other side, sharing the same idea as ETSI, has responded to market demand in the M2M area, understanding that a common set of standards for managing light weight and low capability devices on a variety of networks is not just a nice option, but a mandatory approach. Considering the billions of devices expected to be connected to networks in homes, as well as smart metering and medical devices to name a few, the ability to monitor, provision and manage these billions of connections becomes absolutely essential. Facing the new challenges offered by the market, the OMA DM Working Group with its expertise in remote management addressed the M2M environment with OMA Lightweight M2M protocol, which focuses on constrained cellular and sensor network M2M devices.

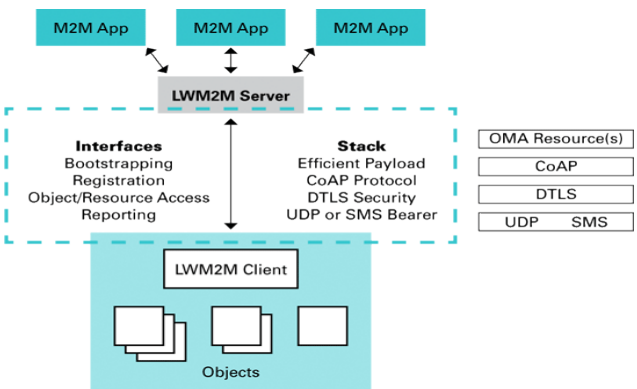


Figure 2 OMA Lightweight M2M protocol

The OMA lightweight specifies simple management interfaces based on IETF standards (CoAP and DTLS with UDP and SMS bindings), with plain text, TLV and JSON serialization for data payload and an Object Model based on manageable Resources (accessible via URI) and Instances concepts (borrowed by OMA DM technology).

OMA LightweightM2M 1.0 is targeted for release in 3Q 2013.

The Telecommunications Industry Association (TIA) Smart Device Communications Engineering Committee TR-50 is responsible for the development and maintenance of access agnostic interface standards for monitoring and bi-directional communication of events and information between machine-to-machine (M2M) systems and smart devices, applications or networks. These standards development efforts pertain to but are not limited to the functional areas as noted: Reference Architecture, Informational Models and Standard Objects, Protocol Aspects, Software Aspects, Conformance and Testing, and Security. TR-50 is developing an M2M Smart Device Communications framework that can operate over different underlying transport networks (wireless, wired, etc.) and can be adapted to a given transport network by means of an adaptation/convergence layer. The TR-50 framework will make its functionality available to applications through a well-defined Application Programming Interface (API) that is agnostic to the vertical application domain (eHealth, Smart Grid, Industrial Automation, etc.). TR-50 works with international standards organizations to ensure end-to-end functionality and interoperability, to avoid overlap or duplication of work, and to foster collaboration and coordination among organizations addressing various components of Smart Device Communication Systems [13].

IP for Smart Objects (IPSO) supports IP for smart objects. This alliance was formed to promote IEEE and IETF standardization bodies. Standards established by the 6lowpan, Roll and CoRE working groups under IETF can provide complete M2M solutions. 6lowpan provides an adaption layer for IPv6 over IEEE 802.15.4. Their technique includes packet fragmentation, reassembling, header compression and optimized neighbor discovery for Low power and Lossy Networks (L2Ns). The Roll working group defines the routing solutions for L2Ns. The first draft of the Roll group's IPv6 Routing Protocol for L2N (RPL) is almost stable. The CoRE working group specifies a framework to monitor and control smart objects intended to run on L2N. They will define a RESTFul style of API and specify the proxy handling mapping between CoAP and REST APIs. CoAP will also support cashing for sleepy devices [14].

Since the M2M industry is fragmented and spread across many different sectors (telecommunications, energy, health, automotive and so on) the current situation is that industry verticals are solving their M2M needs on their own depending on dominant technologies in their sector. Many organizations are working on standardization. These efforts need both leadership and consolidation. There is an opportunity for telecoms companies to take the lead here and convene the world's

experts in the field to work together on this important issue.

IV. Requirements for M2M Network Management

In this section we will elaborate the network management functionalities required for M2M networks and applications.

1) Fault Management (FM) is the set of functions that detect, isolate, and correct malfunctions in the network. FM also compensates for environmental changes, and include maintaining and examining error logs, accepting and acting on error detection notifications, tracing and identifying faults, carrying out sequences of diagnostics tests, correcting faults, reporting error conditions, and localizing and tracing faults by examining and manipulating database information. FM is essential part of all types of M2M networks and applications. Fault management could be periodic, on demand or event driven and M2M network management shall support reporting of faults with specific parameters (i.e. time, date and fault data). Apart from general event driven faults, accidental, environmental and theft related context information can also be triggered as an event in fault management of M2M.

2) Configuration Management (CM) - For a device to be operational several parts need to be configured. This includes the following functionalities:

- Periodic inform - Configuration for how often and in which timeframes a device must contact the management system
- Management server - Configuration for which management system to contact and what authentication method to use
- Connection request - Configuration for which username and password the management system must use when sending connection requests to a specific device
- If managing devices behind a NAT gateway the STUN server information must be provisioned.

In addition to the things needed for device management the service parameters needed for the different services to work is of course possible to configure. This can be e.g. parameters for VoIP service as specified in TR-104 or presence service as specified in presence MO.

3) Mobility Management - is critical requirement in M2M applications with high mobility (such as fleet management or good tracking). Mobility management will include location awareness, navigation support, vertical-horizontal handoff and roaming facilities depending on the application needs.

4) Service Assurance - The service assurance function shall support the service provider in the following use cases:

- Assurance/trouble shooting - Performing tests on individual devices in case of problems reported
- Activation - Performing tests on individual devices after installation at the customer premises in order to make sure that the offered bandwidth etc. is fulfilled.
- SLA tests - proactive and continuous tests towards several devices in order to make sure that agreed SLA is fulfilled

5) Account management – Charging scheme for M2M service usage can be prepaid or postpaid. M2M account management shall provide support for both type of charging schemes.

6) Security - Hackers and malware are finding tempting targets in connected devices across all vertical markets, from consumer electronics and automobiles to medical equipment and military hardware. One of the key elements for a successful deployment and adoption of M2M systems is to ensure trustworthiness, authenticity, and privacy of data being communicated.

V. Management in M2M Solutions – Implementation Considerations

To achieve the mass market, M2M solutions will need to operate in very lean and effective ways. It is therefore essential that management and support functions are well thought out and integrated from the beginning.

Furthermore, M2M solutions cover a broad spectrum of solution spaces that might have quite different business requirements. In the M2M solution space an “off-the-shelf” solution is seldom good enough. There are, however, many management and support functions that fit well across solutions. From experience around 80% of the functionality is common between verticals and thus existing functionality can be reused which allows cost efficiencies to be reached, the remaining 20% of the functionality is unique per vertical.

It has been shown from the implementation experience that horizontal platforms for managing the end devices are the best option. Example of the management platform would look like shown in Figure 3:

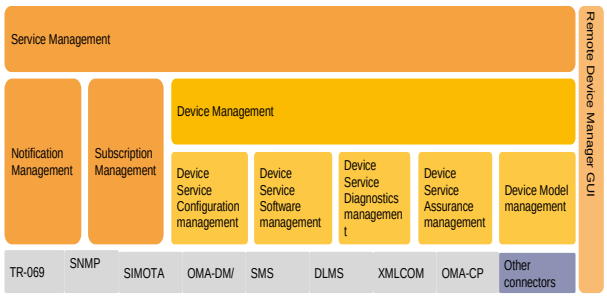


Figure 3 Internal Management Platform Structure

The Management provides a graphical user interface for administration and customer care. The Remote Device Manager GUI uses the services interface exposed by each component. These services are also exposed as external web services interfaces and can be used for integration towards other systems and end user portals.

- Every component owns its data and is storing persistent data in the resource layer (database layer)
- Every component exposes a public web services API which can be used to integrate with external systems
- The protocol connectors are handling the sessions with the managed devices and SIM cards. The protocol connectors are abstracting the differences in the protocols to the upper layers. Examples of connectors shown in the picture are currently implemented protocol connectors are:
  - TR-069 connector
  - OMA-DM connector
  - SNMP connector
  - XMLCOM connector
  - OMA-CP connector
  - Proprietary SMS connector
  - SIMOTA connector
  - DLMS connector

In this way, the platform is modular and can be expanded with additional connector for supporting required technologies. The common functionalities like configuration management or diagnostics reside atop of specific connectors and can be reused regardless of underlying technologies. Benefits of this approach are in simplified integration, creating a one to many (in place of a costly many-to-many) interface to support multiple devices and technologies with a single northbound API that supports all device management functions.

Common functionalities support all device-related use cases for implementation of the complete M2M system, ie. Smart metering application:

- Device Configuration Management – ensure correct settings; continuously update; lock or reboot; apply automatic firmware/software updates
- Network Provisioning – provide device information; initiate provisioning upon device events
- Device Lifecycle Management: device installation, substitution, software update, configuration, removal, termination, (re)activation
- Data & Event Collection – scheduled or on-demand retrieval of device data; cleanse device data; publish device data for multiple uses
- Diagnostics – verify and resolve device software and configuration settings
- Fault Management – detect device alarms and forward to external systems
- Monitoring – collect and forward vital device KPIs; support quality of service, optimize campaigns and maintain SLAs

The Management platform will enable reuse of the device data it collects. In this way, the operator, for example, can use this to improve processes for supply, market planning and customer relationships.

With fewer, shorter and more efficient customer care calls, the overall service experience becomes more positive, leading to happier customers and increased profits for the service.

VI. CONCLUSION

When deploying a new service that is aimed for thousands or even millions of consumers, the operator and service provider face a number of challenges. Many of these are related to equipment handling, and the time and skill needed to cope with manufacturing, installation and configuration. In addition there are challenges once the equipment is in service; to keep the software up to date at all times, assure the equipment is functional.

There is a logistic challenge to manufacture M2M devices that will be deployed globally. The software version may differ from market to market and the data connectivity e.g. access points will differ from the selected service provider.

One answer to this challenge is late configuration of an M2M device. By integration with the mobile network it is possible for Management platform to act upon a network attach of an M2M device, and automatically initiate configuration of the device (e.g. by setting the APN) and/or by installing the latest firmware and software onto the device. This enables a full automated process to bring the device into service.

Depending on the area an M2M device may have a very long lifetime. During this lifetime the software has a need to be updated (with bug fixes or even software enabling new services). Management platform is a tool to enable remote update of the device firmware and software.

If a device is deemed as malfunctioning, e.g. noticed by the M2M Enterprise by gaps in its data collection, or via the Management platform monitoring feature then analyses of what might be wrong is desired. Sending out a service technician is of course costly and thus remote troubleshooting is desired. Some M2M devices do support remote trouble shooting and diagnostics and thus Management platform can be the tool to perform these tasks, thus saving time and money in the process.

M2M devices may also collect and store service KPI's which Management platform can collect using the monitoring functionality as well as to forward data to backend systems. This way the quality of the M2M service can efficiently be assured which is a huge value for the M2M service operations.

In summary, device management is a very important and essential part in order to secure cost efficiency in managing the vast amount of devices soon deployed.

Rolling out M2M services can be complex, as different industries can have diverse requirements and deployment characteristics. The described M2M device management platform is oriented around ETSI M2M standards, which

provides a simplified functional architecture fir addressing the M2M needs of multiple industries. This is key for operators and service providers that want to extract value from a single platform serving multiple customers on millions of endpoints concurrently.

REFERENCES

[1] https://machinaresearch.com/

[2] http://www.gi-de.com/en/about_g_d/company/company.jsp

[3] http://www.gi-de.com/en/about_g_d/company/company.jsp

[4] ETSI TS 102 689 V2.1.1, Machine-to-Machine communications (M2M); M2M service requirements

[5] ETSI TS 102 690 V2.0.13, Machine-to-Machine communications (M2M); Functional architecture

[6] ETSI TS 102 921 V2.0.9, Machine-to-Machine communications (M2M); mIa, dIa and mId interfaces

[7] ETSI TR 102 966, V0.9.1, Machine to Machine Communications (M2M); Interworking between the M2M Architecture and M2M Area Network technologies

[8] http://www.onem2m.org

[9] Roy Thomas Fielding, "architectural styles and the design of networkbased software architectures", http://www.ics.uci.edu/~fielding/pubs/dissertation/top.htm, 2000.

[10] BBF TR-069: "CPE WAN Management Protocol, Issue 1", Amendment 4, July 2011, Protocol version 1.3.

[11] ETSI TS 103 093, V 1.1.1, Machine-to-Machine communications (M2M); BBF TR-069 compatible Management Objects for ETSI M2M

[12] ETSI TS 103 092 V2.1.1, Machine-to-Machine communications (M2M); OMA DM compatible Management Objects for ETSI M2M

[13] http://www.tiaonline.org/m2m

[14] http://www.ipso-alliance.org/technical-information/ipso-guidelines

[15] „Service Enablement", http://www.ericsson.com/ourportfolio/telecom-operators/service-enablement

[16] Ericsson White Paper, "More than 50 billion connected devices, 284 23-3149 Uen"

[17] IOT – i The Internet of Things Initiative, „D1.3 Analysis of Existing IoT Strategic Research Directions and Priorities", http://www.iot- i.eu/public/public-deliverables/

[18] P. Russel, ETSI M2M Architecture Introduction, "A brief overview for Release 1 and Release 2", ETSI M2M Workshop, 24.-25. listopad 2012., http://www.etsi.org/news-events/news/609-2012-m2mworkshop

Next-Generation Utilities based on M2M Communications

Mićo Dujak[1], Zdenko Kljaić[1], Matej Šerbec [2], Hrvoje Jelačić[2]
[1]Ericsson Nikola Tesla d. d., Krapinska 45, 10000 Zagreb, Croatia
[2]Faculty of Electrical Engineering and Computing, University of Zagreb, Unska 3, 0000 Zagreb, Croatia
E-mails: { mico.dujak, zdenko.kljaic}@ericsson.com; matej.serbec@fer.hr; hrvoje.jelacic@fer.hr

ABSTRACT: Utilities today are being challenged as never before. We are witnessing the creation of Next Generation Utilities, which differ from the old utilities in their massive use of extremely sophisticated ICT equipment and processes. In particular the introduction of Smart Grid Communications and Smart Metering requires the ability to deal with ICT projects of a scale and complexity that are beyond the experience and resources of most utilities. There is a need for utilities to make use of partners with advanced ICT skills regarding deployment of systems based on machine-to-machine communication and development of new services like electric vehicle charging.

In this paper, a mobile electric vehicle charging station is being introduced. As plug-in hybrid electric vehicles and battery electric vehicle ownership is expanding, there is a growing need for widely distributed publicly accessible charging stations, some of which support faster charging at higher voltages and currents than those available from domestic supplies. Many charging stations are on-street facilities provided by electric utility companies. Mobile charging stations have been recently introduced. Presented proof-of-concept provides a complete solution for remote control of electric vehicle charging stations, for both customers and owners using web clients.

KEY WORDS

1. Utilities, 2. M2M communication, 3. Electric Vehicle Charging

I. INTRODUCTION

As an electric vehicle charging (EVC) station, also called EV charging station, electric recharging point, charging point and EVSE (Electric Vehicle Supply Equipment), is an element in an infrastructure that supplies electric energy for the recharging of plug-in electric vehicles, including all-electric cars, neighborhood electric vehicles and plug-in hybrids.
Some charging stations support faster charging at higher voltages and currents than those available from domestic supplies. Many charging stations are on-street facilities provided by electric utility companies. Mobile charging stations have been recently introduced. Some of these special charging stations provide one or range of heavy duty or special connectors and/or charging without a physical connection using parking places equipped with inductive charging mats.



Figure 1 Ericsson Electric Vehicle (EVC) charging Station [4]

Machine to Machine (M2M) communication is communication between two or more entities that do not necessarily need any direct human intervention. M2M services intend to automate decision and communication processes. M2M uses a device (such as charging station) to capture an event (such as start of charging, voltage, etc.), which is transferred through a network to an application, that translates the captured event into meaningful information [1].
There are many benefits regarding integrating EVC system with M2M Data Management (M2M DM) platforms like Ericsson MDM DM. Ericsson M2M DM is capable to receive the data measurements on the southbound interface and store the data but also to provide data visualization capabilities on northbound interface. M2M DM is excellent with dealing and processing thousands of requests. Security is another big advantage. With intelligent M2M data management system, customers and available devices can always be tracked [2] [3].

II. STANDARDS FOR ELECTRIC VEHICLE CHARGING SYSTEMS

T in SAE (Society of Automotive Engineers) International, 240 volt AC charging is known as level 2 charging, and 500 volt DC high-current charging is known as DC Fast Charge. Owners can install level 2 charging station at home, while businesses and local government provide level 2 and DC Fast Charge public charging stations that supply electricity for a fee or free [5]. The International Electro technical Commission (IEC) modes definition:

- **Slow** - slow charging from a regular electrical socket (1- or 3-phase)
- **Semi-fast** - slow charging from a regular socket equipped with some EV specific protection arrangement (e.g., the Park & Charge or the PARVE systems)
- **Fast** - slow or fast charging using a specific EV multi-pin socket with control and protection functions

There are three connection cases:

A. Case A is any charger connected to the mains (the mains supply cable is usually attached to the charger) - usually associated with modes 1 or 2

B. Case B is an on-board vehicle charger with a mains supply cable which can be detached from both the supply and the vehicle - usually mode 3

C. Case C is a dedicated charging station with DC supply to the vehicle. The mains supply cable may be permanently attached to the charge-station such as in mode 4.

III. SYSTEM ARCHITECTURE

User connects to EVC web application from any device with internet browser. The browser is used to access the end-user service portal. Service portal is an web application with EVC application. Web application communicates with server that delivers all of the functionality displayed in pictures below. Once web server receives request for charging to start, it assigns logged in user in M2M to selected gateway (Charging station) and sends confirmation SMS message. Every 5 seconds EVC charging station sends its status to M2M southbound interface. Web server pulls data from northbound and shows them on user's device as shown on Figure 2.

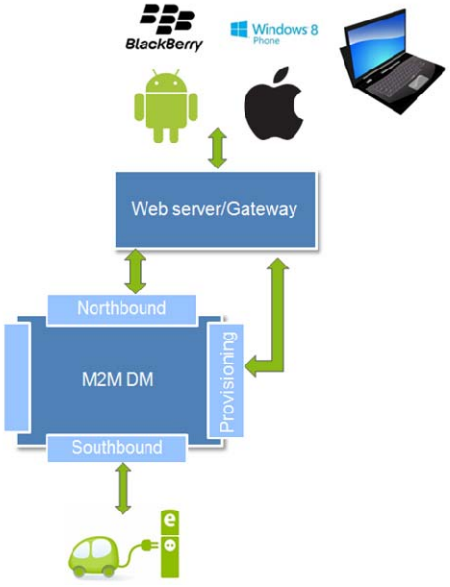


Figure 2 Electric Vehicle Charging – generic system architecture

Sequence diagram on Figure 3 shows successful charging scenario:

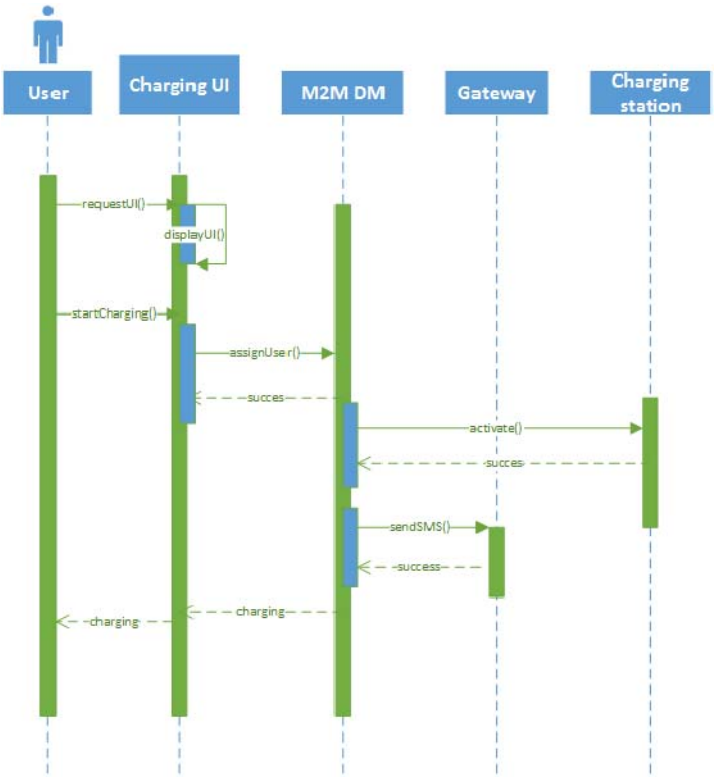


Figure 3 Sequence diagram for EVC using M2M DM

- User sends „requestUI()“ message to Charging UI to display user interface.
- Charging UI executes „displayUI()“ message and displays the user interface to user.
- Users sends „startCharging()“ message to ChargingUI.
- ChargingUI sends „assignUser()“ message to M2M DM node.
- M2M DM assigns current user to gateway instance and sends confirmation message „success“ to Charging UI.
- Presenter M2M DM then sends „activate()“ message to the Charging station to start the charging.
- Charging station turns the light bulb on and responds with „success“ message back to M2M DM node.
- M2M DM sends „sendSMS()“ message to SMS gateway. Gateway sends SMS message to designated cellphone number and responds with „success“ message back to M2M DM node.
- M2M DM responds with „charging“ message back to Charging UI which forwards the message back to User.

This is a critical point of the system because M2M DM plays the role of the data collector, while devices (Charging station) play the role of data providers. In this form of distributed architecture, the data providers connect to the data collector and post the measured data from the devices sensors.

As Gateway shown in Figure 3, Ozeki SMS gateway was used. SMS Gateway enables sending the applications to send/receive SMS messages to mobile devices the battery status information. Also in order to reserve, start and stop the charging process the gateway was needed and used.

IV. CONCEPT VALIDATION

As part of our project, we have developed a service portal, which communicates with the M2M through northbound interface and through provisioning interface. The communication is being done with standard HTTP protocol. The northbound interface is used for querying the M2M and for receiving the data provided by the M2M. All sensors are connected to M2M through the southbound interface. This interface is used for inputting the data from the sensor devices and it is based on REST principles. The protocol for communication is also standard HTTP in combination with DIGEST authentication. Important thing to mention is that the communication is done both ways, from the charging station (EVC) to M2M DM, and from M2M DM to EVC.

The validated concept has the following hardware and software solution components:

- Ericsson Vehicle Charging station
- ZyXEL power line communication device
- M2M DM for capturing and processing the events
- EVC end-user and administrator web application.

End-user web application has following pages and functionalities:

- Welcome page
- User registration page
- Car battery status page
- Charging station map and reservation page
- Start & Stop charging page
- Charging history page

Administrator web application has following pages and functionalies:

- Sign in page
- Overview of available charging points
- Adding / removing charging points
- Overview of utilization by individual charging points

In welcome screen user has option to choose application language. Also they have a choice to sign in if they have EVC account or register for a new one as show on Figure 4.

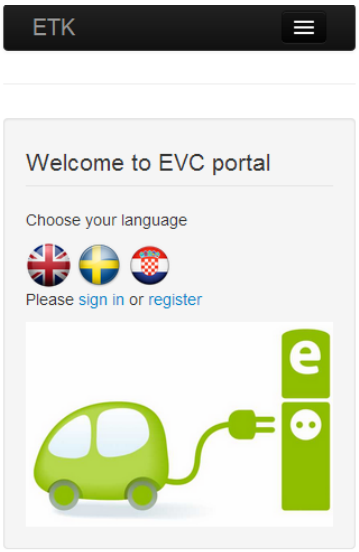


Figure 4 End-user welcome screen

User has to fill out some personal information. Username, email and password are necessary fields while others are optional (on demand that can change in future). Term of Services is yet to be determined and filled out. If all entered information is correct user is going to see success message, otherwise error message with list of invalid fields is going to show up.

Under locations tab user can see maps with marked positions of all EVC charging stations. All green markers indicate that EVC charging station is currently available. Red markers indicate that EVC charging station in unavailable.

When user clicks on green marker (available) they get pop-up box with option “Charge here”. When clicked on “Charge here” user gets redirected to charging page where they have option to start charging as shown on Figure 5.

Charging page allows you to customize charging of

your car. First option is to select charging point from dropdown menu. Second you can choose type of charging, three default values are slow, semi fast and fast. And optional you can define time period for charging. Start charging button initiates charging, sends customer SMS message with start time and all following activities as shown on 6, 7 and 8.

The administrator can see average power consumption during the day for selected charging station. Every column represents a 24 hour day. Under each column administrator can see daily use percentage of charging station. Red fields represent the time of day when the charging station was inactive (not charging) and green fields represent the time of day when the charging station was active (charging) (Figure 9).

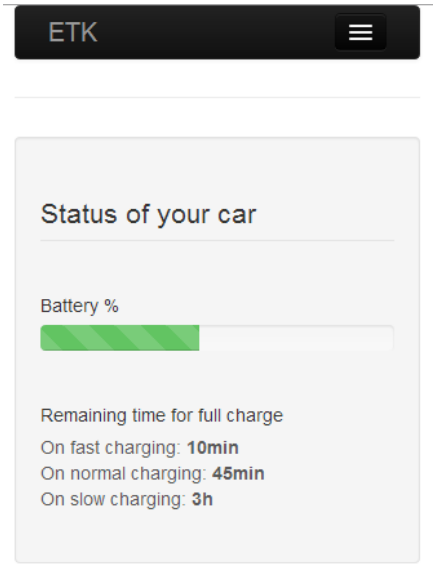


Figure 5 Car battery status

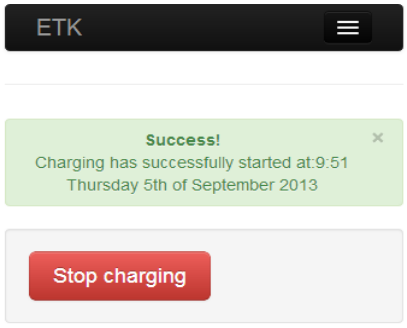


Figure 6 Stop charging process functionality

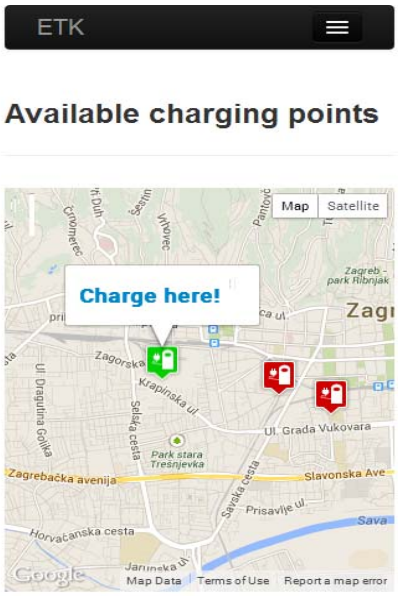


Figure 7 Available charging station points

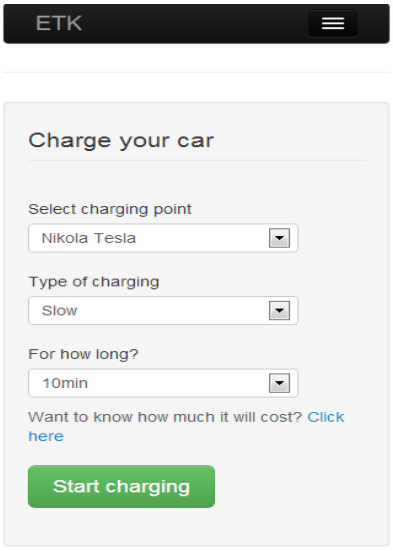


Figure 8 Start charging process functionality

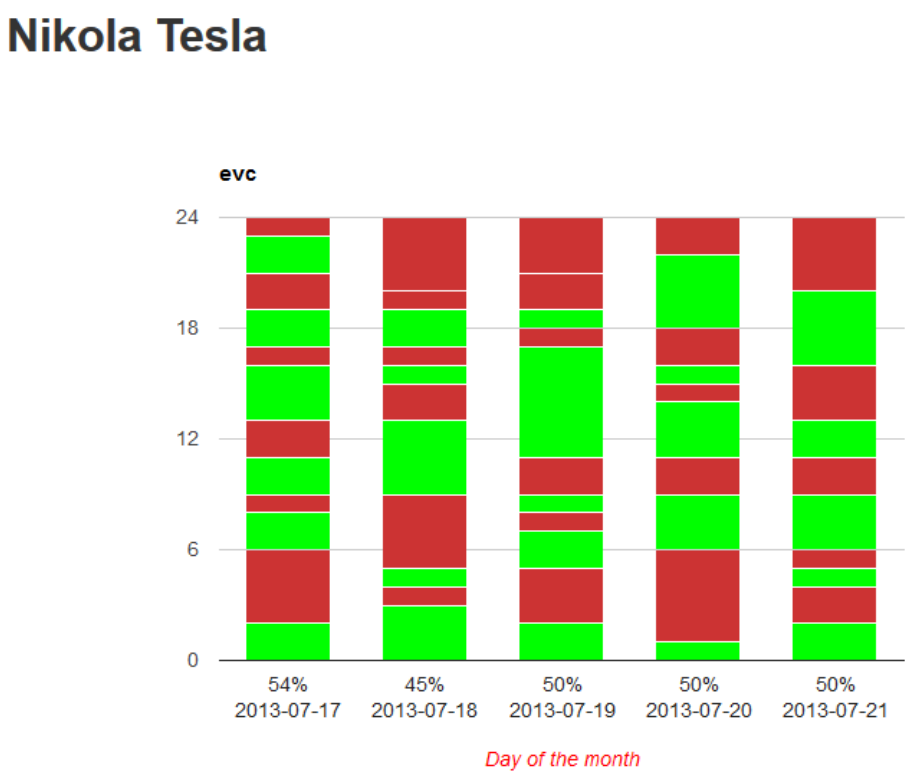


Figure 9 Charging point utilization overview

VI. FUTURE WORK AND DEVELOPMENT

Even though the whole system has been developed, a few points exist in which the system could be upgraded in the near future.

- One major upgrade is to integrate the system with Ericsson Service Portal in order to get user profile provisioning
- Integration with services from smart grid area.
- Other possibility for further development is to implement more features to our EVC application:
- Develop location services so that, based on your location, application can determine nearest charging station and give you directions to it
- Read current battery status of plugged in car

That upgrade has to be enabled on car as well on the charging station side. With information provided from battery we can determine estimated time of remaining charging time.

VI. CONCLUSION

Complete system for remote control of electric vehicle charging stations, for both customers and owners was developed as a result of this project. System consists of frontend web application and backend scheduler application on M2M electric vehicle charging station.

The system is able to collect availability data from M2M electric vehicle charging stations using scheduler application which communicates with M2M DM. Web application can pull that data from M2M DM and control the charging process by sending on/off signals to charging stations.
Light bulb was connected to charging station representing electric vehicle. When the bulb lights up, it represents charging process is on. When the bulb is turned off, it represents charging process is off.

Web application works on any computer or smartphone capable of internet access. It utilizes

administrator and user interface. M2M service enablement data and charging statistics are presented in web application interface. Administrator has additional features for managing and checking charging stations availability presented on both Google Maps and separate status tab in web application.

Goal of the project was to show that variety of devices can communicate using M2M DM..

ACKNOWLEDGEMENT

The authors thank M2M SE team members for help provided in regards to M2M SE Node and especially to Marko Mrkus for help provided in integrating described prototype system with M2M SE.

VII. REFERENCES

[1] Machine to Machine communications (M2M) ; M2M service requirements, European Telecommunications Standards Institute, 2010. ETSI TS 102 689 v1.1.1

[2] Machine-to-Machine communications (M2M); Functional architecture. European Telecommunications Standards Institute. 2011. ETSI TS 102 690 V1.1.1 (2011-10

[3] Internal M2M SE project documentation

[4] Ericsson Car Charging project materials

[5] International Energy Agency, Clean Energy Ministerial, and Electric Vehicles Initiative (April 2013). "Global EV Outlook 2013 - Understanding the Electric Vehicle Landscape to 2020". International Energy Agency. Retrieved 2013-04-20

Applications methods of the objects, and suitability for universal applications in computer graphics context

Nderim Zeqiri*

* Faculty of Applied Sciences– State University of Tetova, 1200 Tetovo, R. Macedonia, office, Mechatronics
nderimzeqiri@gmail.com

Abstract – The use of the primitive objects, can be modified and adapted for universal applications in computer graphics context. The present paper addresses some access and research in the field of computer graphics. The program code present the basics step to increase the study and to explore differently level by describing each of them by math formulas and graphical representations. Also, here is explained the methods how we can use all these recourses to implement to any system, where can be find the better solution in the context of universal applications and solutions.

So, to solve all from above questions, are given some characteristics objects and program code for respective representations. Also is given a direct approach for ADC (analog-digital conversion), by use the fundamental steps followed by this topics: sampling rate, quantization. Also is given statistical results by creating the respective images and respective view in the context of the statistical analyses.

Keywords: Computer graphics, motion, adaptations, program code, generation, ADC converter, samples

I. INTRODUCTION

To make a good system that must be a functional and well structured system is required basic knowledge of a relevant programming languages and communications systems. Depending on the application, one or another measure might be more relevant [1].

The present paper addresses some access, research and implementation for computer graphics and image processing.

The two major components in a high-speed digitizer's are the analog input path and the analog-to-digital converter (ADC) as presented in Figure 1.

The analog input path such as: attenuates, amplifies, filters, and/or couples give us an opportunity to optimize the digitization by the ADC converter or transformation [1] [12] [13].

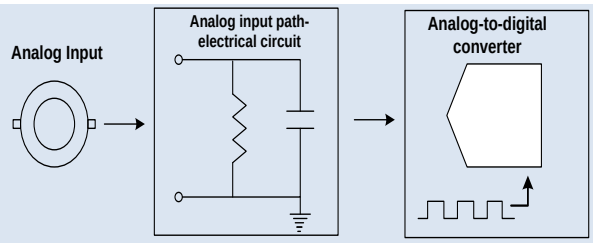


Figure.1. Analog input path and ADC converter.

The Analog to digital (ADC) converter, converts the analog input waveform to digital data. Here is used the sampling rate as a first step to obtain a discrete values on the basis of the Nyquist theorem [2][3][15]. The Nyquist Theorem explains the relationship between the sample rate and the frequency of the measured signal.

In the paper has been decribed how we can do the better solution in the way of implementations, and how we can make a system flexible to use in differently applications.

Also here is described and is given the adequate math formula for values generations, which are very important for fast generations. Using respective sampling rate can be obtained statistical result. The quantization method is a second aim of the paper. Also, in the paper is given how can do and realize the quantization process and statistics for signal to noise ratio.

II. PROGRAM CODE AND MATRIX TRANSFORM DETERMINATION

In the below step is given the code, where is described the manner and the way, how we can generate a primitive object in the windows form applications,

```
Graphics g = e.Graphics;
GraphicsState State = g.Save();
g.PageUnit = GraphicsUnit.Pixel;
Matrix matrixTransformes = new Matrix (1, 0, 0, -1, 0,
0);
matrixTransformes.Translate(this.ClientRectangle.Widt
h / 2, -this.ClientRectangle.Height / 2);
g.Transform = matrixTransformes;
g.DrawLine(new Pen(Color.FromArgb(255, 0, 0), 3), 0,
0, 100, 100);
g.Restore(StateBefore);
g.DrawLine(new Pen(Color.FromArgb(0, 255, 0), 3), 0,
0, 100, 100);
```

The given code can practically serve as a good basis of simulation for graphics context by uses the primitive objects, which are very important to create a good basis and solution for matrix transformer and translation [5][6]. A large number of simulations and projects, in terms of computer graphics can be used in educational institutions, and companies. The main code (given above), creates the possibility of expanding it at the request of users. Application software products have to be developed on the basis of the algorithm, from where we can create a software product, which is relevant and appropriate that finds wide application in the context of computer graphics [7][8].

III. IMAGE PROCESSING SYSTEMS ANALYSES AND VISUAL INFORMATION

Visual information is analyzed according to some research, which represents nearly 80% of all information we receive.
This is not just for the audience that human senses are efficient, but also from the fact that electromagnetic radiation, as well as their part and light are the main carriers of information about objects and processes found in nature [9][10].
Movement of visual information can be presented with the following scheme:

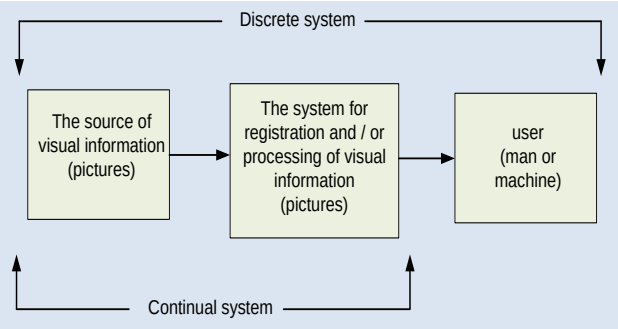


Figure.2. Control flow of visual information

In mostly of the differently system can be incorporated sensors systems which can be an intelligent sensor, for directly implementations to the computer control system, for respective visualizations scene [8] [10].
Received information the user may use in two ways.

- Directly: to control any process (for reaching a decision and take appropriate action).
- Indirectly: for observation, i.e. accumulation of visual information, in order that they will later be processed along with some other information.

Scene, meant limited two-dimensional fields:
Where each point (x, y) at any point of broadcasts (as primary or secondary source) electromagnetic radiation, power spectral density [12][13].

$$P(x, y, \lambda, t) \quad (1)$$

$$0 \leq P(x, y, \lambda, t) \leq A \quad (2)$$

Distribution of load accumulation in the field of cumulative interval T is:

$$\frac{dQ(x,y,t)}{dt} = k \int_{\lambda 1}^{\lambda 2} P(x, y, \lambda, t)\varepsilon(\lambda) \quad (3)$$

This is an analogue aspect, and the goal of researchers in recent years has been the transition to digital system because analog systems needs report definitions S/N (signal/noise), the band and especially the creation of issuing high stability of these characteristics. Therefore processing of digital image processing goes on, especially with the development of advanced computer system.
Transfer of analog monochrome figure – in digital image, called digitalization and includes three procedures [9]:

- Discretization or 2D (sampling) of analog figure.
- Discretization of luminance samples (samples) or quantization.
- Coding

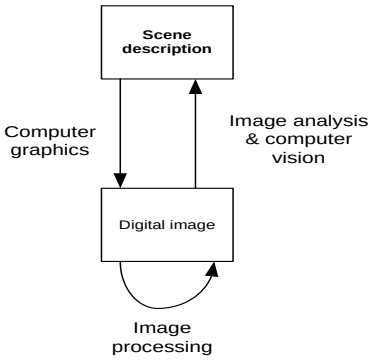


Figure.3. Computer graphics and Image Processing

IV. THE SAMPLING RATE AND RELEVANT VALUES

Sample rate is not directly related to the bandwidth specifications of a high-speed digitizer. Sample rate is the speed at which the digitizer's ADC converts the input signal, after the signal has passed through the analog input path, to digital values that represent the voltage level. This means that the digitizer will sample the signal after any attenuation, gain, and/or filtering has been applied by the analog input path, and converts the resulting waveform to digital representation [13]. The sample rate of a high-speed digitizer is based on the sample clock that tells the ADC when to convert the instantaneous analog voltage to the digital values, Figure 4. National Instruments high-speed digitizers support a variable effective sample rate derived from the maximum sample rate of the device. For example, the NI 5112 has a maximum sample rate of 100 Megasamples/second (MS/s) and can be set to rates of (100MS/s)/n, where n = 1,2,3,4,.... [2][14]

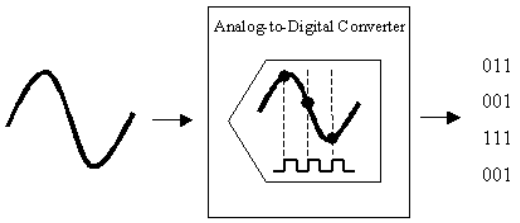


Figure.4. Nyquist theorem, sample rate and ADC

It is recommended that the bandwidth of our digitizer be 3 to 5 times the highest frequency component of interest in the measured signal to capture the signal with minimal amplitude error (bandwidth required = (3 to 5)*frequency of interest). The theoretical amplitude error of a measured signal can be calculated from the ratio of the digitizer's bandwidth in relation to the input signal frequency (R). A complete simulation system must model all components of a computer system. However, only a few components play a determining factor in the speed of the simulation etc [5][8][9].

V. THE QUANTIZATIONS AND STATISTICS FOR SNR (SIGNAL TO NOISE RATIO)

For digital signal processing, continuous-amplitude samples from the sample-and-hold are quantized and mapped into n-bit binary digits. For quantization to n bits, the amplitude range of the signal is divided into $M = 2^n$ discrete levels, and each sample is quantized to the nearest quantization level, and then mapped to the binary code assigned to that level [2].

The quantization of the signal into 4 discrete levels [4][15]. Quantization is a many-to-one mapping, in that all the values that falls within the continuum of a quantization band are mapped to the centre of the band. The mapping between an analog sample xa(m) and its quantized value x(m) can be expressed as,

$$x(m) = Q[x_a(m)] \quad (4)$$

Where $Q[x_a(m)]$ - is quantizing function. The performance of a quantized is measured by signal to quantization noise ratio SQNR per bit. The quantization noise is defined as,

$$e(m) = x(m) - x_a(m) \quad (5)$$

Now consider an n-bit quantized with an amplitude range of ±V volts. The quantization step size is, $\Delta = 2V/2^n$. Assuming that the quantization noise is a zero-mean uniform process with an amplitude range of ±Δ/2 we can express the noise power as,

$$E[e^2(m)] = \int_{-\Delta/2}^{\Delta/2} f_{E(e(m))}\, de(m) = \frac{1}{\Delta}\int_{-\Delta/2}^{\Delta/2} e^2(m)de(m) = \frac{1}{\Delta}\cdot\frac{V^2 2^{-2n}}{3} \quad (6)$$

Sampling and Analog–to–Digital Conversion where $f_{E(e(m))}$=1/ Δ is the uniform probability density function of the noise. Using Equation (6) signal–to–quantization noise ratio is given by,

$$SQNR(n) = 10Log_{10}\left(\frac{E[x^2(m)]}{E[e^2(m)]}\right) = 10Log_{10}\left(\frac{P_{Signal}}{V^2 2^{-2n}/3}\right) = 10log_{10}\ \ 3\text{-}10\ \ log_{10}\left(\frac{V^2}{P_{Signal}}\right) + 10log_{10}2^{2n} = 4.77 - \alpha + 6n \quad (7)$$

Where Psignal is the mean signal power, and α is the ratio in decibels of the peak signal power V^2 to the mean signal power Psignal.

After the sampling we have a sequence of numbers which can theoretically still take any value on a continuous range of values. Because this range in continuous, there are infinitely many possible values for each number, in fact even uncountable infinitely many. In order to be able to represent each number from such a continuous range, we would need an infinite number of digits - something we don't have. Instead, we must represent our numbers with a finite number of digits, that is: after discrediting the time-variable, we now have to discredit the amplitude-variable as well. This discretization of the amplitude values is called quantization. Assume, our sequence takes on values in the range between −1... + 1. Now assume that we must represent each number from this range with just two decimal digits: one before and one after the point. Our possible amplitude values are therefore: −1.0, −0.9, . . . , −0.1, 0.0, 0.1, . . . , 0.9, 1.0. These are exactly 21 distinct levels for the amplitude and we will denote this number of quantization levels with each level is a step of 0.1 higher than its predecessor and we will denote this quantization step size as q. Now we assign to each number from our continuous range that quantization level which is closest to our actual amplitude: the range −0.05... + 0.05 maps to quantization level 0.0, the range 0.05...0.15 maps to 0.1 and so on. That mapping can be viewed as a piecewise constant function acting on our continuous amplitude variable x. In the Table 1 is given the statistics for signal to quantization noise ratio - where Psignal is the mean signal power, and α is the ratio in decibels of the peak signal power V^2 to the mean signal power Psignal. Therefore, from Equation (7) every additional bit in an analog to digital converter results in 6 dB improvement in signal–to–quantization noise ratio.

Number-Examples	$\boldsymbol{\alpha = \left(\frac{V^2}{P_{Signal}}\right)}$ **is the ratio in decibels**	**n**	**The ratio in decibels of the peak signal power to the mean signal power Psignal** $\boldsymbol{10log_{10}\left(\frac{V^2}{P_{Signal}}\right)}$	$\boldsymbol{10log_{10}2^{2n}}$	**Signal to quantization noise ratio** $\boldsymbol{SQNR(n) = 4.77 - \alpha + 6n}$
1	1	2	0.00	12.04	16.81
2	2	3	3.01	18.06	19.82
3	3	4	4.77	24.08	24.08
4	4	5	6.02	30.10	28.85
5	5	6	6.99	36.12	33.90
6	6	7	7.78	42.14	39.13
7	7	8	8.45	48.16	44.48
8	8	9	9.03	54.19	49.92
9	9	10	9.54	60.21	55.43
10	10	11	10.00	66.23	61.00
...	...	…	…	…	…

Table.1. The quantization and statistics for signal to noise ratio.

Note, that this mapping also includes a clipping operation at ±1: values larger than 0.95 are mapped to quantization level 1, no matter how large, and analogous for negative values.

Each intensity value is a number, for digital storage the intensity values must be quantized [14][16]. The quantization can limit the number of different intensities that can be stored, can limit the brightest intensity that can be stored, how many intensity levels are needed for human consumption [2] [3]:

- 8 bits usually sufficient
- some applications use 10 or 12 bits

VI. DIGITAL IMAGE DEFINITIONS

A digital image a[m,n] described in a 2D discrete space is derived from an analog image a(x,y) in a 2D continuous space through a sampling process that is frequently referred to as digitization. The 2D continuous image a(x,y) is divided into N rows and M columns. The intersection of a row and a column is termed a pixel. The value assigned to the integer coordinates [m,n] with {m=0,1,2,…,M–1} and {n=0,1,2,…,N–1} is a[m,n] [13]. In fact, in most cases a(x,y) – which we might consider to be the physical signal that impinges on the face of a 2D sensor – is actually a function of many variables including depth (z), color (λ), and time (t) [15][16].

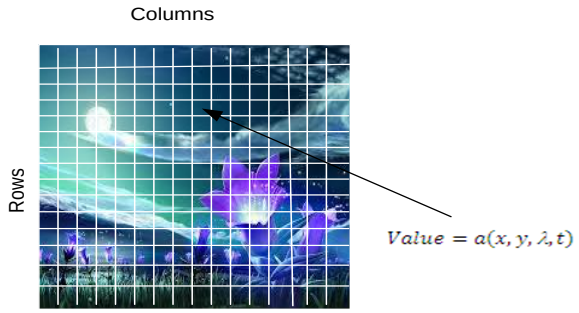


Figure.5. Digitalization of a continuous image. The pixel at coordinates [m=10, n=3] has the integer brightness value 110

The image shown in Figure 5 has been divided into N = 16 rows and M = 16 columns. The value assigned to every pixel is the average brightness in the pixel rounded to the nearest integer value. The process of representing the amplitude of the 2D signal at a given coordinate as an integer value with L different gray levels is usually referred to as amplitude quantization or simply quantization.

There are standard values for the various parameters encountered in digital image processing. These values can be caused by video standards, by algorithmic requirements, or by the desire to keep digital circuitry simple. Table 2 gives some commonly encountered values.

Parameter	Symbol	Typical values
Rows	N	256,512,525,625,1024,1080
Columns	M	256,512,768,1024,1920
Gray Levels	L	2,64,256,1024,4096,16384

Table.2. The standard values for the various parameters encountered in digital image processing.

On the basis of the Table 1 - as Author's own generations, we see quite frequently cases of $M = N = 2^K$ where {K = 8,9,10,11,12}. This can be motivated by digital circuitry or by the use of certain algorithms such as the (fast) Fourier transform [16] [17].

The number of distinct gray levels is usually a power of 2, that is, $l = 2^B$, where B is the number of bits in the binary representation of the brightness levels. When B>1 we speak of a gray-level image; when B=1 we speak of a binary image. In a binary image there are just two gray levels which can be referred to, for example, as "black" and "white" or "0" and "1".

VII. COLOUR VISION ANALYZE

Exist three types of cone, each responds to a different spectrum, roughly can be defined as red, green, and blue; each has a response function r(λ), g(λ), b(λ); different sensitivities to the different colors, roughly 40:20:1, so cannot see fine detail in blue, overall intensity response of the eye can be calculated:

$$y(\lambda) = r(\lambda) + g(\lambda) + b(\lambda) \qquad (8)$$

$y = k \int P(\lambda)\, y(\lambda) d\lambda$ - is the perceived luminance many different spectra will induce the same response in our cones the values of the three perceived values can be calculated as:

$$r = k \int P(\lambda) r(\lambda) d\lambda \,, g = k \int P(\lambda)\, g(\lambda) d\lambda$$
$$b = k \int P(\lambda)\, b(\lambda) d\lambda \qquad (9)$$

k is constant, P(λ) is the spectrum of the light incident on the retina; two different spectra (e.g. P1(λ) and P2(λ)) can give the same values of r, g, b; can thus fool the eye into seeing (almost) any color by mixing correct proportions of some small number of lights.

Implications of vision on quantization humans can distinguish, at best, about a 2% change in intensity not so good at distinguishing color differences

For TV - 10 bits of intensity information, 8 bits is usually sufficient. Why use only 8 bits? Why is it usually acceptable? This is adequately for movie film – which is 14 bits of intensity information [14][15].

VIII. STORING IMAGES IN MEMORY

To store image in memory is the main aims, because, after sampling rate, and diskretization, all the imgaes converted into binary number, respecitvely can be stored in memory. So, for example, on the basis of the Figure 6, 8 bits has become as a standard for greyscale images, 8 bits = 1 byte, an image of size *WxH* can therefore be

stored in a block of *WxH* bytes, one way to do this is to store pixel[x][y] at memory location *base + x +W x y* base; memory is 1D, images are 2D. [11][12][15]

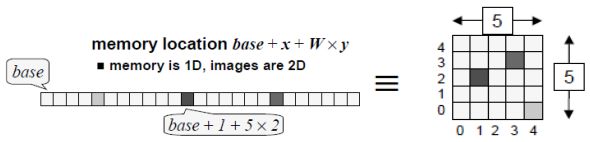


Figure.6. Storing images in memory

IX. CONLUSION

The present paper addresses some access, research and management, implementation of computer graphics. Also in the paper is treated the image processing, sampling rate, quantization and coding process. By creating programming codes as a starting point and visual access methods for generating a certain values, which can be used to advantage parts and segments of any object, image, and to a certain body dimensions. Mobility of primitive and others complex objects forms requires the idea and development of the interpolation method. After the sampling we have a sequence of numbers which theoretically can take any value on a continuous range of values. The conversion from analog to digital values, give us the opportunity to consider this transformations as better solution in the context of performance and as fast solution, because by use formula (7) is optimal range in the aspects to obtain discrete value. Quantization is a many-to-one mapping, in that all the values that falls within the continuum of a quantization band are mapped to the centre of the band.

In the context of color vision analyze- is emphases that exist three types of cone, each responds to a different spectrum, roughly can be defined as red, green, and blue; each has a response function r(λ), g(λ), b(λ); different sensitivities to the different colors, roughly 40:20:1. Also, is described one way how we can memorize any image into memory; adequately is described memory locations and its bases.

These values are practical because they may soon be generated through the respective program and differently programs can be incorporated to establish respectively connection with software components.

Adaptation is about the ways of communication and connection of external equipment. By using this system can be managed the communications system, because this process can be a part of processing, all these values are converted into discreet value, and they are ready for computer processing systems [16][17]. In the section, Waveforms, quantization & and signal to noise ration and statistics are given the solution to consider an n-bit quantizer with an amplitude range of ±V volts. The quantization step size is $\Delta = 2V/2^n$. Assuming that the quantisation noise is a zero-mean uniform process with an amplitude range of ±Δ/2 we can express the noise power, the formula and calculated statistics are very flexible, and extremely convenient for different systems.

REFERENCES

[1] Trading Bit, Message, and Time Complexity of Distributed Algorithms Johannes Schneider and Roger Wattenhofer. 25th International Symposium on Distributed Computing (DISC), Rome, Italy, September 2011.

[2] R.J. Marks II, Handbook of Fourier Analysis and Its Applications, Oxford University Press, (2009), Chapters 5-8

[3] R.J. Marks II, Editor: Advanced Topics in Shannon Sampling and Interpolation Theory, Springer-Verlag, 1993.

[4] D. Burger and T. Austin, "The SimpleScalar Tool Set Version 2.0," Univ.of Wisconsin, Madison, Tech. Rep. CS-TR-97-1342, 1997.

[5] M. Rosenblum, E. Bugnion, S. Devine, and S. A. Herrod, "Using the SimOS machine simulator to study complex computer systems," *ACM Trans. Modeling Computer Simulation*, vol. 7, no. 1, pp. 78–103.

[6] R. N. Ibbett, "Hase DLX simulator," *IEEE Micro*, vol. 20, no. 3, pp. 57–65, May/Jun. 2000.

[7] J. L. Bechennec, "ASF: A teaching and research object-oriented simulation tool for computer architecture design and performance evolution," in *Proc.Workshop Computer Architecture Education (WCAE-98)*, Barcelona, Spain, Jun. 27, 1998, pp. 93–96.

[8] Potashnik, Michael, and Adkins, Douglas. Cost Analysis of Information Technology Projects in Education: Experiences from Developing Countries. Education and Technology Series, vol. 1, no. 3. Washington:World Bank, 1996.

[9] Bounds On Contention Management Algorithms Johannes Schneider and Roger Wattenhofer. 20th International Symposium on Algorithms and Computation (ISAAC), Honolulu, USA, December 2009

[10] Stephan Holzer, Yvonne Anne Pignolet, Jasmin Smula, and Roger Wattenhofer. 6th International Workshop on Algorithms for Sensor Systems (Algosensors), Wireless Ad Hoc Networks and Autonomous Mobile Entities, Bordeaux, France, July 2010.

[11] Fundamentals of Interactive Computer Graphics Foley & van Dam [1Y104], Addison-Wesley, 1982

[12] Digital Image Processing Gonzalez & Woods [U242] Addison-Wesley, 1992.

[13] Computer Graphics: Principles & Practice Foley, van Dam, Feiner & Hughes [1Y919] Addison-Wesley, 1990

[14] Classification of Services in the Digital Economy Weber, Rolf H., Burri, Mira 2013, XIX, 135 p. 6 illus, ISBN 978-3-642-31635-7

[15] Jonathan M. Blackledge (2003). Digital Signal Processing: Mathematical and Computational Methods, Software Development and Applications. Horwood Publishing. ISBN 1-898563-48-9.

[16] Digital Signal Processing Alan V. Openheim (Author), Ronald W. Schafe, ISBN-10: 0132146355.

[17] Fast Fourier Transform and Its Applications, E. Brigham, Publication, 1988, ISBN-10: 0133075052.

On-line Rapid Diagnostics of the Human Immune Status

T.A. Hryhorova*, L.V. Vorobiov **
* Kremenchuk Mykhailo Ostrohradskyi National University/Department of Higher Mathematics and Computer Science, Kremenchuk, Ukraine
** Medical Center VIKOM/Immunologist, Kremenchuk, Ukraine
gasta1@yandex.ru

Modern information technology and the existing rapid (express) diagnostics method allow you to monitor the state of health. The method is based on the thermovision diagnostics and enables determining the human immune status. The designed software implements this method. At the time of the diagnosis the system generates recommendations for eliminating an immune deficiency. These recommendations are made based on the results of the calculations of the integral temperature indicator used in the expert module.

The system consists of the web and desktop applications, local and global databases, which use Sqlite and MSSql technologies. The access to the global database is realized by the web-service with the help of a ServiceAdapter mediator, which implements the data storage interface and encapsulates the web-service behavior. The designed program of the express diagnostics of the immune status can be easily scaled as a local application, on web, or cloud.

Results of statistical data analysis from the data warehouse can be presented in the form of diagrams, graphs, tables and allow a qualified health professional to evaluate the health status of the population of the city, region, etc.

The developed software system is accessible on-line to everyone.

I. INTRODUCTION

Immunity is responsible for protecting a human health, and immune deficiency is a signal that health is at risk, and the necessary measures should be taken immediately. Diagnostics of the immune status will allow an early detection of any abnormalities. It will allow to restore vital functions of the human body and prevent developing serious diseases. Prevention of the immune deficiency is important for reduction of the overall morbidity, reducing chances of developing chronic diseases, and decreasing the number of incidences of atypical courses of the diseases. It would also lead to a more of efficient immunization routine, and reduce a number of adverse reactions to vaccinations. Additionally, it would improve reparative function of the human body, lengthen the life of AIDS patients, improve a maternal and child health, and the overall population dynamics.

Currently, there are two ways to diagnose immune deficiency - a laboratory diagnostics and express diagnostics. The first method requires an immunology laboratory, time, and experienced personnel, which make it difficult to use it for prevention of the immune deficiency among the population. The second method is based on thermovision diagnostics, which was developed by Professor Michael Vogralik. According to his data, sensitivity and level of reliability of the thermovision diagnostics, were 84% - 90% [1].

The Vogralik method is as follows. There is a verified link between the immune deficiency and the heat production of the immunity organs (thymus gland and spleen). In order to determine the status of the metabolic and functional activity, the level of the heat production of the immunity organs needs to be measured. The IR thermometer – pyrometer with a measuring accuracy up to 0.1 degrees of Celsius - is needed for such measurement. For a thymus gland, a deep temperature is measured in the organ projection. For a spleen, a temperature in the remote locations that reflexively related to the spleen is measured. It is possible to determine the activity of immunity organs by taking into account both quantitative and qualitative immune system parameters. One of the easily controllable parameters of the immune response is a measure of lymphocytes, which may be obtained from the total blood analysis.

One of the methods that allows to make an immunocorrection without any side effects is the use of the physical therapy that uses reflex mechanisms of action on immunocompetent organs. The duration of the immune status normalization is up to 6 months, and this term satisfies the rules of the clinical supervision, the practice of anti-recurrence treatment, and disease prevention. Taking into consideration the indicators of lymphocytes and leukocytes, it is possible to rationalize the choice of the physical impact factor to eliminate immune deficiency.

In 1995 a special device for express diagnostics was created in Russia based on the Vogralik method. In 2004 a local version of a software was developed for the out-patient clinic in the Automobile Plant KrAZ and Medical Center VIKOM using the same method. During more than 9 years of practical experience of using express method (a metabolic control of the immune status) in the Medical Center VIKOM the number of clinical visits was reduced by 1.8 - 2.4 times [1]. After few years the method of the software operation was improved, namely, changes in the heat production level calculations were introduced. As a result, it became necessary to create an

information system of the express diagnostics of human immune status with the support of local, web and cloud applications [2]. This system provides an opportunity to use express diagnostics applying two methods - by a specialist during a clinical visit (local application) or self-diagnostics using the public access to the system through Internet resources.

II. Approaches and Program Development Phase

Information model is represented by the following information entities:

- Users working with the system are divided into two categories - a qualified health professional and the end user who wants to take the express diagnostics independently in the on-line mode.
- Input data for both types of users remain unchanged.
- The output results are formed based on the input data and predetermined system rules. The results may be presented in different document formats (MSWord or HTML).
- The statistical data stored in the data warehouse are represented in the form of charts, graphs, and tables, so a qualified health care provider can assess the health status of the population of the city, region, etc.

A generalized model of the express diagnostics information system has been developed (1). It consists of the off-line applications that work with the local databases, the on-line application that works with the support of a global database, and a system kernel. The model includes an expert system module, a statistics module, an output documents module, a data access module, and a kernel configuration module.

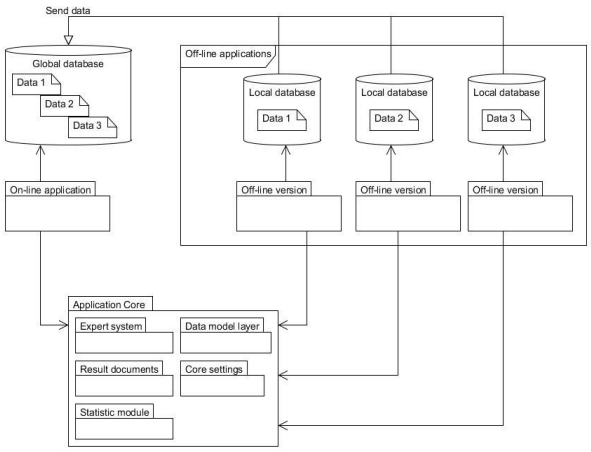


Figure 1 - Model of the information system

Information system architecture (2) combines a multi-level architecture of the distributed application (a service oriented application with respect to access to data warehouses for both local and distributed applications). It is built on a single system kernel, which supports the operation of the basic system functions.

A component - oriented approach was used to develop the program [3].

An important practical result of applying the component approach concept is an ability to improve the strategy of reusing the code and reduce the cost of designing and implementing the software. The code with a higher level of abstraction does not require a substantial modification in adapting to the changing conditions or new data types.

The main advantage of the chosen concept is a unification of the different categories of objects. Abstract classes and methods allow an uniform handling of the heterogeneous data, and it is not required to develop an additional code to adapt to new classes and data types. The development of the system can be performed using interfaces by adding new components and replacing some of the old components without rebuilding the entire system. To expand the system, you can create new components that provide additional services.

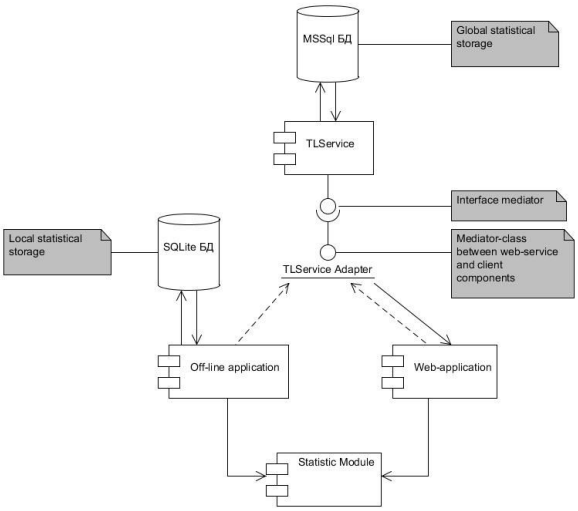


Figure 2 - Architecture of the information system of rapid diagnostics

The developed implementation model presented in the form of a component diagram (3) simulates the physical aspects of the information system of express diagnostics.

The implementation model of the information system of express diagnostics consists of the following components:

- The software kernel. It is responsible for the model levels and their interfaces, expert system, resulting documents, and statistical data processing system.
- Resource library. It includes additional choices of language (English, Ukrainian, etc.), images, and all other resources. Library allows to change graphics, string and file application data dynamically - without recompiling the program.
- Web-service. It plays a role of a global data repository and is used for the statistical data collection (for local and on-line versions).

- On-line version of the system. It provides the ability to pass the express-testing without installing a local version. It serves as an interface for the administration tasks including viewing statistics. On-line version has a research module, a statistics module, data validation, and program localization.
- Local version. It has the widest variation of settings, the ability to save a document in different formats, the licensing system, and other functions.
- Licensing system. It is responsible for the state of the product according to the status of the license. It serves to protect against unlicensed use of the product.
- Generator license. It uses a licensing system to generate license keys.
- Library of localization. It depends on the technology chosen. Typically it is used by external libraries for localization of local applications.

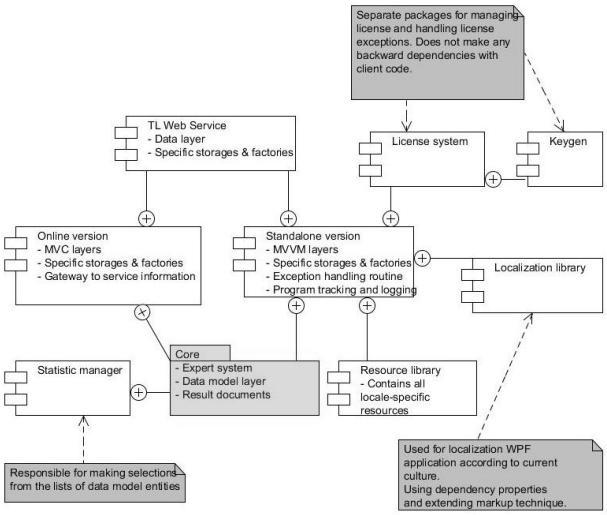


Figure 3 - Component Diagram

The project was implemented using Microsoft.Net technology [4].

The designed system works with two databases - local and global, and each of them uses a separate technology (Sqlite and MSSql). Web-service is the entry point to the global database [5]. It is accessed through a mediator ServiceAdapter that implements an express diagnostics results data storage interface and encapsulates the logic of working with the web-service. Both types of applications use the data storage to send and receive data from a global data repository.

Interfaces and general base classes are included in the kernel without reference to the platform and the component type. Setup classes are also in the kernel, and they can be used to configure the kernel to work with local or web-application.

For less connectedness of the system the kernel was distributed to the following atomic parts:

- data model and interfaces;
- system settings;
- system of document creation;
- expert system;
- statistics system;
- extend functions of Net Framework [6].

Given that the kernel supports both the off-line and on-line applications, it was not possible to use the built-in settings of the Net [6]. To maintain such a function, it was necessary to realize a settings mechanism that saves its state between software runs.

Expert system architecture directly depends on the data model, since it is necessary to operate data entities during the express diagnostics.

The rules are formed by immunology specialist and are loaded into the program model in the interactive mode. The expert system is initialized by storage of these rules (XML for both versions), and corresponding objects are created. Thereafter, the expert system is ready for the express diagnostics. This mechanism allows achieving the following goals:

- an ability to change the rules in the runtime;
- an ability to localize the rules and, as a result, the entire source document;
- a possibility of using a unit-testing to ensure the correctness of code writing.

It should be noted that some rules may lead to the emergency completion of the express diagnostics, and some - to results cleanup. Activity diagram (4) describes this behavior.

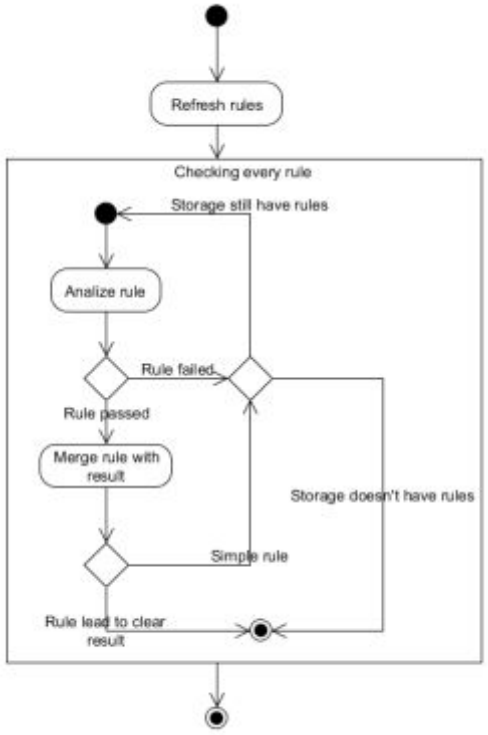


Figure 4 - Activity diagram of the expert system

The system uses a resource library, which includes a mechanism for localization of other languages, images, and all other resources. It allows dynamically change the graphics, string and file data used by the program.

When designing a document creation system, it was taken into account that it can display both the resulting express diagnostics documents and statistical documents. The option to save documents in various formats is provided in the documents interface.

Full specific business logic is applicable only to the off-line version and completely detached from the software kernel, or simply extends it. For example, the licensing system is used only in the off-line version. It has been implemented in order to protect the product from unauthorized use without the knowledge of the customer.

The express diagnostic results and the results of the statistical data analysis can be output in two formats - MS Word and HTML.

Some of the interfaces require a step by step data input. Given this, it is expedient to organize a wizard that encapsulates the logic that relates to the incremental data input. During the design and implementation phase the following have been taken into consideration: the need for the input data validation, editing of the Wizard's steps, and changing an algorithm of the express diagnostic completion. These needs were met by designing an architecture that allows transmitting a controller interface into Wizard object. Two controllers have been created - for the first launch Wizard and Wizard for the express diagnostics.

Geographic Information System web-service of Yandex is used (5) to visualize the distribution of the immune status parameters of the population by country, region, city, etc. Polygons were used to enable the selection of the individual regions on the map. They are located on a server as XML documents.



Figure 5 - Geodistribution indicator of immune status by regions of Ukraine (1 - hyperfunction, 2 - norm, 3 - hypofunction, 4 - limit of normal)

Research module is a part of the kernel that is common for both off-line and on-line applications. It includes expert system and system of the temperature and blood test analysis indicators. These systems are used for obtaining a qualitative and quantitative immunity results. In the system of the temperature indicators analysis the calculations of the integrated temperature indicator used by expert system were modified. To verify data entry, a mechanism of client and server validation is used. The application allows keeping a history of the express diagnostics that makes it possible to see the dynamics of the human immune system for any length of time. The system has a choice of various settings, ranging from the visual settings (for an average user) to the configuration of the expert system (for immunology specialist). A global database collects anonymous information from the on-line application and transmits it from local databases and off-line applications over the Internet.

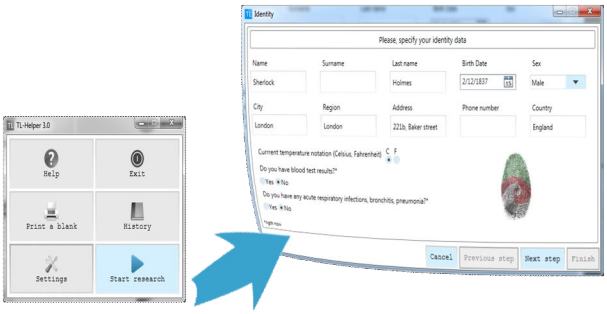


Figure 6 – Program menu and input data window

There are seven global commands (6) for the whole application - Show history, Show settings, Display help, Show local statistics, Start research, Print a blank, and Exit the application. They are available in any application window.

III. CONLUSION

The program of the express diagnostics of the immune status designed by authors can be easily scaled as a local application, on web, or on cloud. It has a flexible architecture and solves a wide spectrum of problems with minimal connectivity between components.

An approach to the design and implementation of an information system is based on the component-oriented programming technique. This approach makes it possible to implement a combining software architecture in a base of common kernel for different types of applications. The idea is to select the components in software kernel that do not depend on the type of application and deployment platform. This kernel can be initialized and configured by the client code according to the needs of the target platform. Therefore, the proposed approach can be used in projects with similar architecture.

The developed software system realized an express method of the self-diagnostics of the immune deficiency and its prevention. A medical health provider can work with the off-line version of the system where he or she can use the local database with the patient's diagnostics results to monitor the patient's immune status during the course of

the treatment. A global database is available on-line only to the health care authorities that registered in the system with administrator rights. A global database contains the anonymous diagnostics data for all people who used the on-line system for diagnostics, and these data can be used for monitoring the immune status of the population on the level of city, region, or country.

A limitation of the presented method of the express diagnostics is in that it cannot replace the full immunology laboratory results since it only provides an information on the immunity status of the patient - the presence or absence of the immune deficit.

The product is successfully operated in the establishment of the practical health care Medical Center VIKOM in the city of Kremenchug, Ukraine and is available on-line at www.tl-helper.com.

[1] L.V. Vorobiev, "Experience with the physiological reflex in the diagnosis and correction of immunodeficiency states", Alergologiya i Immunologiya, Vol. 4, no. 2, pp. 41–46., 2003.

[2] David S. Linthicum, Cloud Computing and SOA Convergence in Your Enterprise, Addison-Wesley, Boston,.USA., 2010

[3] Eric J. Braude, Software engering: An Object-Oriented Perspective, Piter, Saint Petersburg,.Russia, 2004.

[4] Matthew MacDonald, Pro WPF in C# 2008: Windows Presentation Foundation with .NET 3.5, Translated by Artemenko, U.A., OOO "I. D. Williams", .Moscow, Russia., 2008

[5] V.A. Gladycin, K.V. Krinikin, and V.V. Yankovskiy, Service-oriented architecture. Standards, algorithms, protocols, SPbGETU "LETI", Saint Petersburg,.Russia.R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press., 2004.

[6] Matthew MacDonald, Adam Freeman, and Mario Szpuszta, Pro ASP.NET 4 in C# 2010, Translated by Artemenko, U.A., OOO "I. D. Williams", .Moscow, Russia., 2008.

INFORMATIONS SYTEMS

Information system infrastructure planning - threshold setting based on risk analysis

I. Golub and B. Radojević*
* Croatian Academic and Research Network (CARNet), Zagreb, Croatia
Ivana.Golub@CARNet.hr, Branko.Radojevic@CARNet.hr

Abstract - Information systems infrastructure planning performed based on the data collected from network monitoring and asset management systems can be developed further to include risk assessment and analysis as an input to the threshold setting methods in the two-dimensional load-age balance graph. The two-dimensional load-age balance graph is used to present generic value for network devices load gathered as measured values through the network managements system, and age - gathered from the asset management system. Threshold lines are defined in such graph to present areas of different interest from the perspective of low-high age, low-high load, or some other selection criteria. This paper explores how the threshold values can be set to reflect risk assessment methodology of the owner/maintainer/user.

The position of each networking device in the graph is unique, disregarding the threshold values. However, the total score for the system, determined based on the position of a device in the graph area, depends on the defined threshold. The paper analyzes the impact of the threshold change based on the risk assessment criteria to the total score for a system. The total score can then be used to compare different systems, or to compare the same system in time.

I. INTRODUCTION

Information systems infrastructure planning has to be based on the information about the current state of the system received from the system itself (e.g. via monitoring system), and information about the plans for further development of services and users needs. It is usually not in the domain of the system maintainers to have information about neither the future services development nor the users growth. The only data that can be gathered and examined are those from the system itself.

Risk in a general sense represent "a probability or threat of damage" or "loss or any other negative occurrence" "caused by external or internal vulnerabilities, that may be avoided through preemptive action" [1] or the possibility that something bad or unpleasant will happen [2]. In the sense of an information system, a risk can be seen in the degraded level of service, or as the total unavailability of the service offered to users (customers).

Risk assessment has its two dimensions - the probability for an event to happen and the level of danger that the event can cause to the system. In addition, as explained in [3], [4], the risk has its objective and subjective side, and both determine the assessment of a risk in a system. While the "risk as analysis" (as defined in [4] starts from the information gained through measurement systems of the current state of an information system, the "risk as feelings" [4] has its final decision in the determination of the risk levels and categories.

The model for the determination of the current stage for an information system is presented in [5] and is re-printed in the Figure 1.

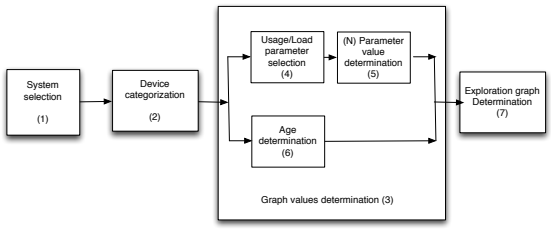


Figure 1. Exploration Determination Model (Source [5])

Based on that model, using the data from a monitoring system, and as explained in the [5] it is easy to design two-dimensional load-age graph with the current status of each observed device represented with one dot in the graph. X axis will contain generic value for the load of the system-relevant components (e.g. CPU and memory load and the number of used ports for networking equipment, average CPU and memory load for servers) expressed in percent of the maximum load and represented as the sum of values for each measured parameters. Y-axis contains the age calculated as the current year minus year of the equipment acquisition. Thus created graph represents current state of the observed system. However, in order to be able to interpret the data and use it for the purpose of system development and planning, it is necessary to put this data into the relevant context for the system.

II. RISK ASSESSMENT AND THRESHOLD DEFINITION

Each information system can be seen differently from different user groups - the system owner's view does not necessarily have to be the same as the view of a user or a view of a service owner or the information system

maintainer. Such different views on the requirements and the levels of service will also reflect the definition of risk. A general interpretation, in accordance with the definitions found in on-line dictionaries can be summarized to the probability that the service will be degraded or unavailable. Transferring that expression into the view of a system maintainer, the risk can be seen through the system performance degradation, or system unavailability. The first from the two parameters for risk determination - the probability for an event to occur - can be learned through the system observation, understanding, as well as the previous knowledge of the system, based on the system behavior analysis. The second parameter - the potential loss produced with an event has to be determined for an individual system. For example, the same system can be established in one scenario for a small number of users (e.g. as an redundant system) with a small number of non-vital services, and in another scenario for several times more users with vital business-critical services. In those two cases, even if the same risk indicator is identified for both systems, the potential loss in the case of system degradation will not be the same.

For an information system and the presented exploration determination model, the risk assessment is determined based on the reached level of individual parameters at which there is a higher probability that a system degradation of malfunctioning can happen.

For example, in an information system, the failure may occur when there is a high CPU or memory load. In addition, in the networking part of the system, the possibility for future development and growth is reduced with the decreased number of available interfaces.

Which level of CPU/memory load is critical for which system, and especially potential loss that can thus be caused to the system owner or user - cannot be seen directly from the graphs or current status of the system, but is based on the knowledge of the system and attached services.

The risk can be expressed using Boolean (bipolar) representation or a range representation. In the Boolean representation, values are expressed for each field either is there is NO risk (the state is "good"), or there is a risk (the state is "bad"). In the later case, existence of risk is a message to the maintainer to perform an action to determine the cause of the potential risk and undertake actions to reduce the recognized risk to the minimum.

Range representation of a risk would allow an analyst to define more then two risk levels. For example, levels can be defined as in the Table 1, or there could be even more levels defined for risk assessment used for the range representation.

TABLE I. RISK LEVELS IN THE RANGE REPRESENTATIONS

Risk level	Risk Level Description
0	No risk
1	Small
2	Medium
3	High

For the analysis of the two provided approaches, the same example data will be used as presented in [5]. In this paper, a part of a data center was analyzed from the perspective of load and age for each relevant device. The analysis was separated for servers and networking equipment since those two device categories have different set of parameters relevant for the determination of the current state. Average CPU and memory load was observed versus equipment age for server farm and results represented in the two-dimensional graph where maximum load range reaches the level of 200%. In the case of networking equipment, together with the load of CPU and memory, the percentage of used network interfaces was also examined as a third parameter versus the equipment age, and data presented at the graph with max x axis value of 300% (total sum of load percentage for all three observed parameters).

Starting graphs as presented in [5] are provided in Figures 2 and 3.

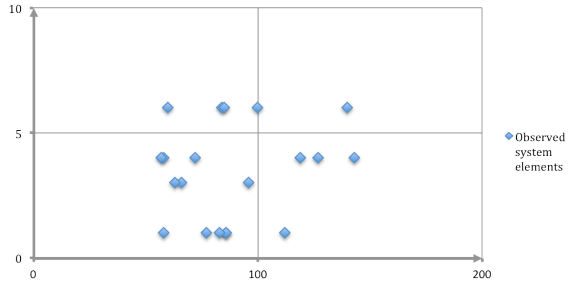


Figure 2. Exploration graph for the servers

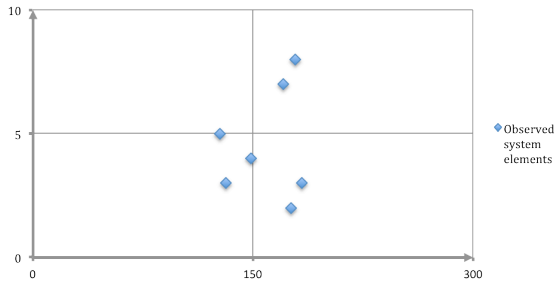


Figure 3. Exploration graph for the networking equipment

Both figures are taken from the same test case and presents also thresholds as initially defined at the 50% of the load scale and at the 5 years of custom amortization value. In order to focus on the risk assessment, the definition of thresholds and the interpretation of data in thus yielded graph, it is sufficient to follow just one of the two presented examples - either for server or for the networking equipment. Therefore, all further analysis will be demonstrated using data from Figure 3.

When the thresholds are used to represent the risk assessment, then the risk level is defined for each axis separately. The threshold levels are drawn at the graph to present the risk level, dividing the graph into several fields.

III. RISK ASSESSMENT FROM THE EXPLORATION GRAPH

In the case of Boolean representation, there will be two values for each axis. Such case is already represented at Figures 2 and 3. At X axis, field in the range [0-150) is field with NO risk, and the field [150-300] is the field that contains the risk. At Y axis, filed [0-5) is NO risk field, and [5-10] brings the risk. Looking at the whole graph, only the lower left field near the coordination system origin is a NO RISK field. At Figure 3, there is only one device that completely belongs to that field. For any other of the remaining three fields there is at least some level of risk assigned to it. The further from the coordination system origin a point is - the bigger is the risk for the maintainer and system owner. In praxis, that means that for each of the points (devices) that are outside of the NO RISK area additional examination should be started in order to determine the risk cause, and to define the techniques that could be applied in order for the risk to be minimized or removed. The final goal should be to "move" the state for each individual device as close to the coordinating system origin as possible. The Figure 4 presents the NO_RISK/RISK definition for the devices in the graph from Figure 3, based on the Boolean representation of the risk.

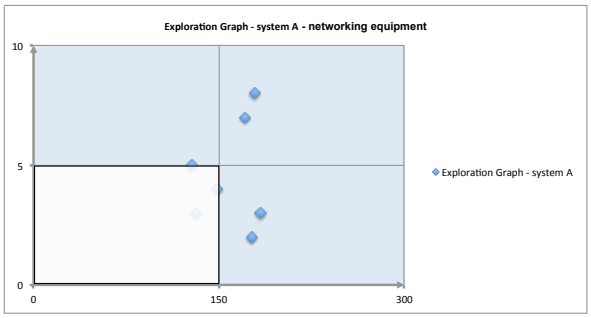


Figure 4. Exploration graph for the networking equipment with white NO-RISK area and blue RISK areas

In the four-fields representation of young/old and low/high loaded devices it was easy to recognize for each device, which risk its position in the graph it might carry. For the old equipment, there is a risk of absence of maintenance, no spare parts, lack of support for new features or new technologies. In the field of high load, there is a risk that a small increase of load might cause system instability or unavailability. With this risk representation in the same graph, it is even easier to recognize all the devices that can potentially be a cause of problem for the system.

In the case with range risk representation, as exampled in Table 1, such granulation should again be applied to each axis separately. In addition, the risk assessment has to be performed separately for each observed parameter that determines the overall load. The values can differ for each of those parameters, to reflect the correct status and development of the system.

As an easier example, consider that the ranges for X-axis parameter will be the same for each parameter at the each risk level, as presented at the table 2.

TABLE II. RISK LEVELS IN THE RANGE REPRESENTATIONS

Risk level	Risk Level Description	Age	Load level for 1 parameter	Load for 3 parameters
0	No risk	0-3	0-30 %	0-90
1	Small	3-5	30-50 %	90 - 150
2	Medium	5-7	50-70 %	150 - 210
3	Large	>=7	>=70 %	>=210

It is important to emphasize that the numbers in the Table 2 are examples of values that could be selected for each risk level for each parameter. For a selected system, the values have to be defined in accordance with the system use policy and taking into considerations development plans from the perspective of users as well as the system owners.

Values for thresholds for Y axis are given in column 3 and for X axis in column 5, and the graph with corresponding multiple thresholds is provided in Figure 5..

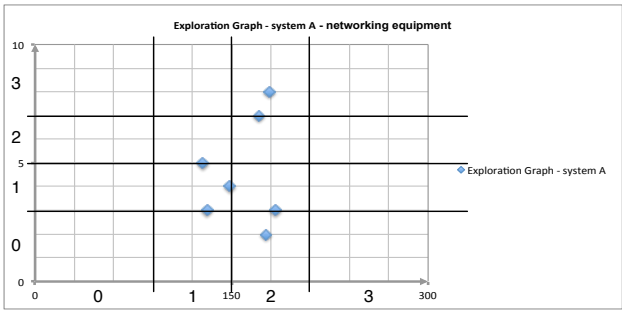


Figure 5. Exploration graph for the networking equipment with several thresholds corresponding to range risk representation

At X and Y-axis each field is presented with the corresponding risk level. Again, it can be seen that only lower left field closest to the coordinating system origin is risk free area, while all the other fields carry some risk level associated with it. Only the devices that will be in the NO RISK field will not require any action at the moment, based on the time of the analysis. (It is possible though that at the very next analysis period a device will change its place based on the load/age values and that might require additional analysis and actions.) All the other devices in all other areas of the graph will require additional analysis to determine the cause for the potential risk and possible actions that could be taken in order to bring each device status closer to the coordinate system origin. For the upper example NO RISK and risky areas are presented at Figure 6.

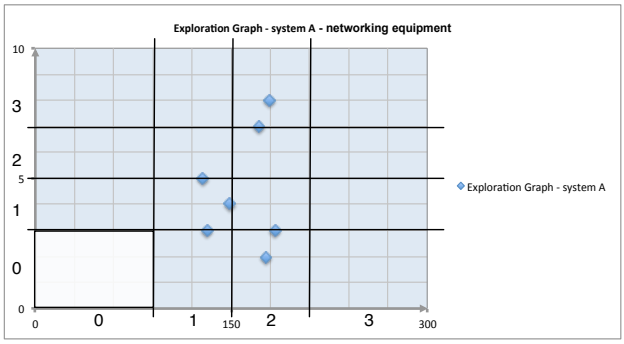


Figure 6. Exploration graph with white NO RISK area and blue risk areas for the networking equipment with several thresholds corresponding to range risk representation

In the case when the risk range values are different for each parameter, for the same risk range (0, 1, 2 or 3) the range where this risk applies for an observed parameter has to be defined, and the upper level for the threshold range is determined by summing the values for all observed parameters. With the Y axis the situation is clear since age is the only parameter observed.

Additionally, for each field in such multiple thresholds graph, the risk value can be calculated as the sum of the X-axis risk and the Y-axis risk. On this way, for the graph presented in Figure 6 there will be up to six different risk categories, as presented at Figure 7. Despite the different values, the conclusion can again be simply expressed that the risk is higher for the points in the graph that are more distant from the coordinate system origin.

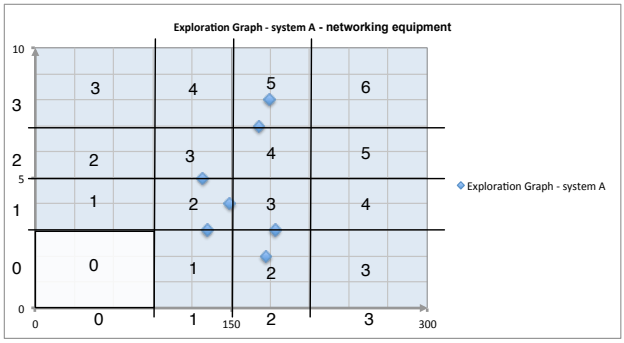


Figure 7. Exploration graph with white NO RISK area and six categories of blue risk areas based on the risk calculation

Although there were four different categories of risk (no risk, low, medium and high) calculated for each parameter, and different thresholds defined in each of those risk groups, risk recognition is at the end categorized only at two levels - is there a risk for a devices or not. Lack of risk means that at the time when the analysis is performed there is no need for individual devices in the NO RISK area to be analyzed further based on its load and current age. Devices with its points in the risk area - no matter in which one of the individual multiple threshold fields - will require additional analysis.

So, instead of multiple thresholds per axis and multiple risk levels, it is sufficient to define only one level per axis that will represent the point where there might be a possibility for an event occurrence that might jeopardize current status of the system so that the system will no longer be available to provide a service, either at the acceptable level (so that the service will be degraded) or at all (so that the service will not be available).

IV. Related work

The research presented in this paper is continued work from [5], based at the same model for the information system infrastructure planning. This model relies on the data from the network/data infrastructure monitoring system, present in each network/service provider system, as those presented in [6]. The representation at two levels - Boolean or range representation, used for the categorization of risk representation - is used in [7] for the representation of link quality. Analog tomography used for individual links is similarly used for the representation of individual parameters that combined influence the total load for a device. Presented model can as well be used as a part of a decision support system, like the one presented in [8].

V. Conclusion

Presented model for the information system infrastructure planning can very easily and successfully be used to graphically indicate elements in the system whose current status represents a potential risk for the system owner or maintainer. Data are gathered from the monitoring and asset management system(s), and presented at a time when the analysis is performed. Since data are gathered from a live system that can be a subject of a constant change, the results are is valid for the time when the analysis is performed, with a possible extension depending on the dynamics of the system. Risk levels are calculated and determined based on the knowledge of the system and cannot be determined statically and uniquely for any generic system, since it depends on the number and activity of users, number and importance of services, as well as the development plans. Therefore, a risk levels for each system and for each parameter has to be carefully determined.

The results presented in this paper has shown that it is more important to define correct risk levels for a two-level (Boolean) risk representation, then to introduce multiple levels of risk, because the main decision is the system at risk or not will again be just a two-choices answer that will split the graph into just two fields - one containing the devices without recognized risk status and the other with the devices whose status brings certain risk into the current system.

For all devices whose risk level is above the NO RISK status (above zero risk) will require additional analysis in order to define the risk cause and possible actions that could be taken in order to reduce the potential risk and move the device-representing point closed to the origin in the coordinating system of the exploration graph.

All the analysis was presented at the graph for the networking equipment (from the sample case from [5]). However, the logic with two or multiple risk levels per axis can equally be applied for the servers example with two load parameters.

This analysis has again proven the simplicity of the two-dimensional graph representation of the system status, and the information about the potential risk can easily be seen by any viewer, for different levels of the technical knowledge of the system, and management hierarchy.

REFERENCES

[1] http://www.businessdictionary.com

[2] http://www.merriam-webster.com

[3] Slovic P., Finucane, M. L., Peters, E., MacGregor, D. G., "Risk as Analysis and Risk as Feelings: SOme Thourghts about Affect, Reason, Risk, and Rationality", 2004, Risk Analysis, 24(2), 311-322. available (02/2014) at: http://faculty.psy.ohio-state.edu/peters/lab/pubs/publications/2004_Slovic_Finucane_etal_risk_as_feelings_502.pdf

[4] Slovic, P., Peters, E., Finucane, M. L., & MacGregor, D. G. (2005). Affect, risk, and decision making. Health Psychology, 24, S35-S40., available (02/2014) at: http://faculty.psy.ohio-state.edu/peters/lab/pubs/publications/2005_SlovicPetersFinucaneMacGregor_HealthPsych_special_issue_538.pdf

[5] Golub, I. Radojević B., "Information systems infrastructure planning based on collected data from monitoring tools", Information & Communication Technology Electronics & Microelectronics (MIPRO), 2013 36th International Convention, pp. 456-460, May 2013, available (02/2014) at: https://www.researchgate.net/publication/257368138_Information_systems_infrastructure_planning_based_on_collected_data_from_monitoring_tools/file/9c96052525d3fdf972.pdf?ev=pub_int_doc_dl&origin=publication_detail&inViewer=true

[6] R. Boutaba and J. Xiao, Network Management: State of the Art, in Proceedings of IFIP World Computer Congress (WCC'02) TC6 Stream on Communication Systems: The State of the Art, pp. 127-146, 2002, available (02/2014) at: http://rboutaba.cs.uwaterloo.ca/Papers/Conferences/2002/Boutaba02.pdf

[7] Zarifzadeh S., Gowdagere M., Dovrolis C., "Range Tomography: Combining the Practicality of Boolean Tomography with the Resolution of Analog Tomography", IMC'12, November 14–16, 2012, Boston, Massachusetts, USA, available (02/2014) at: http://www.cc.gatech.edu/~dovrolis/Papers/final-sajjad-imc12.pdf

[8] Y. Beres, D. Pym, S. Shin, "Decision Support for Systems Security Investment", Manuscript, HP Labs , 2010

[9]

Interactive application in service of popularizing architectural heritage

M. Jozić, T. Oštarčević, M. Šalković, D.Šimunić
Fakultet elektrotehnike i računarstva/Zavod za radiokomunikacije, Zagreb, Hrvatska
matej.jozić@fer.hr

The main aim of this paper is making connections between modern technologies, architecture and art, i.e. how to increase the interest for cultural sights and facilitate the retrieval of desirable information with use of modern technologies. Smart phones dominate worldwide mobile phone market and they are mainly used for display and information retrieval. Furthermore, not only that smartphones offer information accessibility at any moment, they also demand interaction with user, so the aspiration for information's is increasing. As a solution, an android application the Story of Zagreb was chosen, which allows display of information, photographs and locations of cultural sights in Zagreb. Desirable data can be reached with choosing certain cultural sight from the list or by sensing QR code, two dimensional code, which describes that cultural sight, with the camera of smartphone.

I. INTRODUCTION

Story of Zagreb is a project, which has purpose of promoting cultural sights of Zagreb.

Its database contains number of sights for various purposes the catedral, many churches, stadiums, museums, libraries, theaters and other attractive buildings in Zagreb.

Application allows user to get familiar with a particular building or to improve existing knowledge of it, so it's useful for tourists, but also for the residents of Zagreb aswell. It includes descripitons of development of construction, its function both historically and today, special features, a variety of anecdotes, legends, etc. There are also some other interesting features that make the building attractive.

In addition, it is possible to go through the photographs of selected buildings which include representations from various perspectives, view inside and out, view the details (various images, altars, monuments, etc.), important events (inaugurations, anniversaries, etc.) Each image contains a brief description of display.

Application also offers user an option of locating buildings on a map of Zagreb, which allows easier reach to the desired destination.

To search for available sites, it is not necessary to know their names, it is sufficent to choose a picture of current interest. Each sight is available to search and contain generated unique QR code that it describes. That allows user who is already on the point of his interest to simply read the QR code within the application and the application will automatically connect the user with the chosen sight.

A. *Implementation*

The application is developed for portable devices (mobile phones and tablets) which use the Android operating system.

Since this is a very popular operating system, the application is therefore available to a large number of users.

The system is available in several versions, and application is based on version 2.2 (called Froyo) which is compatible with the majority (96% to be more accurate) of Android portable devices.

Within the application, ZbarScanner is also available. It is a library which is integrated in application, whose role is reading QR codes. With that, user doesn't need to install a special application on his device to read QR codes, and it becomes automatically available when you install the application 'StoryOfZagreb'.

All sights, which are available by searching the application, are located in the database that can be further expanded, if it is necessary, regardless of whether it is a complete adding of new sights or partial modification of existing information specific sites.

II. DEMONSTRATION

A. *Home Screen*

During the launch of appliaction, the application opens the home screen, shown in Figure 1, which enables the function described below.



Figure 1 - The layout of the home screen when launching application StoryOfZagreb

- Language selection – currently there are two available languages: Croatian and English. Depending on the choice, all the options, as well as the content and available descriptions will be displayed in the chosen language.
- Reading QR code – an option that allows you to search available sights that are contained in the application, by reading (scanning) QR codes and attractions. While chososing this method of search, the camera is turning on, on the portable device of the user, and after focusing QR code in the center of the camera, the application automatically detects sights and shows the user, informations associated with it.
- Search – another way of searching the available sights that are contained in the application. While choosing this method of search, the images and names of all available sights are opening. After choosing the sight, the application show informations related to that chosen sight.

B. *Additional features*

In Figure 2, it is seen the display of the selected sights containing text description. For example, Croatian National Theater is chosen, and on the home screen, the language will be Croatian.

Application allows user returning to the previous, home screen or to continue searching, or on the other hand, selecting additional options related to the current selection.

Additional features include:

- Picture Gallery – display of images of sights in whole or detailed view of some of its parts, as well as images that are in some way connected with it. Each image contains a further description of what it shows (in the language selected on the home screen). Part of the picture galleries for the Croatian National Theatre building, is shown in Figure 3. In the example, the chosen langauge is Croatian.
- Show on map – this option opens map of Zagreb in the browser you use, in which location of the chosen sights is marked. This option is intended for users who search on the basis of the available sights, rather than through QR code, and users who are not currently located in front of thesights. Display on the map for the building of Croatian National Theathre, is shown in Figure 4.

C. *Display of design of the application*



Figure 2 - Layout of textual description of Croatian National Theatre and offered additional opportunities

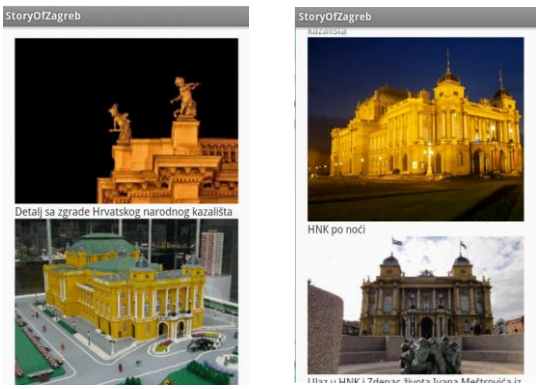
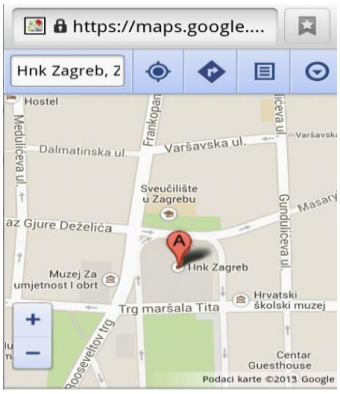


Figure 3 - Part of picture gallery for Croatian National Theatre



III. CONCLUSION

This paper has shown how art and technology, despite popular opinion, may very well cohabitate. The interweaving of art and modern technology lately, is experiencing bigger attention and delivers interesting and useful gain.

With use of new tehcnologies in terms of this project, it enables the fast and efficent retrieval of information and popularization of destinations.

This application, except the fact that promotes tourism industry, also plays an important role in preserving the heritage of the urn and its opening up to the rest of the world.

LITERATURE

[1] Zlatko Karač, Alen Žunić, "Antologijski arhitektonski vodič Zagreba: 100 izabranih zgrada", Zagreb : Arhitektonski fakultet Sveučilišta : UPI-2M plus, 2012

[2] Jeff Brown, ZBar Code reader, http://zbar.sourceforge.net/, listopad 2013.

[3] Zigurd Mednieks, Laird Dornin, G. Blake Meike, Masumi Nakamura, "Programming Android, 2nd Edition Java Programming for the New Generation of Mobile Devices", O'Reilly Media, rujan 2012.

Adresiranje i usmjeravanje e-dokumenata u sustavu informacijskog posrednika za uslugu e-Računa

Danijel Matić*, Alan Čičak *, Boris Vrdoljak**
* InfoDom d.o.o., Zagreb, Hrvatska
** Sveučilište u Zagrebu Fakultet elektrotehnike i računarstva, Zagreb, Hrvatska
danijel.matic@infodom.hr, alan.cicak@infodom.hr, boris.vrdoljak@fer.hr

Sažetak – Elektroničko poslovanje je izvršavanje svih elemenata poslovnih aktivnosti ekonomskog procesa elektroničkim putem, odnosno stvaranje dodatnih vrijednosti koristeći se ICT tehnologijama. Informacijski posrednik je fizička ili pravna osoba koja u ime drugih obavlja otpremu, prijem i prijenos elektroničkih dokumenata.

U radu se predlaže generička arhitektura sustava informacijskog posrednika pod utjecajem normi i smjernica za adresiranje i usmjeravanje u kontekstu poslovnih funkcija upravljanja odnosima s kupcem i upravljanja uslugama i operacijama. Opisan je i prijedlog strukture interne poruke prilikom obrade u sustavu informacijskog posrednika. Isporuka dokumenata zahtijeva uspostavu dodatne strukture, tzv. omotnice oko dokumenta. U omotnici se pohranjuju informacije iz poslovnog konteksta pošiljatelja potrebne za adresiranje i usmjeravanje dokumenta prema krajnjem odredištu. Adresiranje i usmjeravanje u sustavu informacijskog posrednika koristi prilagođene podatkovne strukture temeljene na normama OASIS ebXML Messaging (ebMS), OASIS UBL 2.1, te CEN CWA smjernicama što omogućuje informacijskom posredniku uspostavu različitih scenarija razmjene poruka.

I. UVOD

Elektroničko poslovanje je oblik poslovanja koji prilikom razmjena informacija između poslovnih partnera koristi informacijske i komunikacijske tehnologije. Cilj je elektroničkog poslovanja što više automatizirati poslovne procese radi povećavanja efikasnosti i efektivnosti, te smanjenja troškova poslovanja.

Elektronički računi i ostali poslovni dokumenti u elektroničkom poslovanju razmjenjuju se putem poruka (eng. messages). Poruka je informacijski paket koji se koristi prilikom razmjene, a sastoji se od zaglavlja (eng. header) u kojem se nalaze informacije potrebne za dostavu poruke na krajnje odredište (eng. Addressing and Routing information) i sadržaja (eng. Payload) koji uključuje jedan ili više elektroničkih dokumenata [9].

Adresiranje u elektroničkom poslovanju je upravljanje logičkim identifikatorima sudionika u razmjeni poruka te identifikacija informacija potrebnih za dostavu poruke na krajnje odredište.

Da bi poruka bila isporučena na krajnje odredište potrebno je identificirati put, tzv. rutu kojim mora putovati prema krajnjem odredištu. Usmjeravanje je identifikacija puta poruke. Prilikom usmjeravanja koriste se procesi adresiranja s ciljem određivanja svih potrebnih jedinstvenih identifikatora primatelja i pošiljatelja, njihovih adresa za prijem i dostavu, ali i identifikatora međuodredišta u situacijama kada se na ruti nalazi više od jednog informacijskih posrednika.

U elektroničkom poslovanju koje se operativno temelji na razmjeni elektroničkih dokumenata, procesi adresiranja i usmjeravanja su neophodni. Adresiranje će jedinstveno identificirati primatelja i pošiljatelja elektroničkog dokumenta, a usmjeravanje poruka omogućuje razrješenje gdje i kako je potrebno isporučiti poslovni dokument. Adresiranje i usmjeravanje poruka je naročito važno u visokoautomatiziranim procesima razmjene.

Za pouzdanu i efikasnu elektroničku komunikaciju poslovni partneri moraju definirati i razmijeniti skupove identifikatora potrebnih za uspostavu elektroničke komunikacije te iste implementirati u svoje informacijske sustave. Prije implementacije identifikatora poslovni i pravni zahtjev je da isti postaju dio poslovnog odnosa, a s time i dio formalnog ugovora između dvije poslovne strane. U tom kontekstu ugovoreni identifikatori između dvije strane se teško mogu propagirati u ostatak poslovnog okruženja.

Bilateralno ugovaranje prema svim poslovnim partnerima je dugotrajan proces. Radi olakšavanja ugovaranja elektroničkog poslovanja samo ugovaranje kao poslovni proces se prepušta trećima tj. informacijskim posrednicima. Uloga informacijskog posrednika postaje olakšavanje i skraćivanje procesa ugovaranja [10] elektroničke razmjene dokumenata.

Prilikom adresiranja informacijski posrednik mora imati mogućnost upotrebe različitih identifikacijskih shema i identifikatora. Proces adresiranja informacijskog posrednika nije ograničen na upotrebu jedne identifikacijske sheme.

Upotreba različitih identifikacijskih shema dovodi do izazova u implementaciji općenitog sustava adresiranja koji će efikasno i pouzdano identificirati sudionike razmjene i omogućiti određivanje rute puta elektroničkog dokumenta kroz mrežu.

Izazovi na koje je potrebno odgovoriti su:

- koje identifikatore koristiti za adresiranje i kako specificirati njihove identifikacijske sheme,
- kako mapirati identifikatore na odredišne adrese na koje je potrebno isporučiti elektronički dokument,
- kako implementirati procese adresiranja i usmjeravanja poruka u sustav informacijskog posrednika.

Kao odgovor na ove izazove prilikom implementacije sustava informacijskog posrednika potrebno je definirati zajedničko zaglavlje koje pruža konzistentno sučelje i neophodno je za operativnu uspostavu pravne, organizacijske, semantičke i tehničke interoperabilnosti među sudionicima razmjene.

Na stručnom skupu o registru korisnika e-računa održanom 16. studenog 2012. godine na Fakultetu elektrotehnike i računarstva u Zagrebu, tim koji radio na pilotskom projektu „FER e-račun“ [15] je u svrhu jednostavnijeg preslikavanja identifikatora na odredišne adrese predložio uvođenje središnjeg nacionalnog registra korisnika e-računa. Kako nije došlo do značajnijeg napretka u implementaciji središnjeg registra, u ovom radu neće se razmatrati scenariji adresiranja i usmjeravanja koji uključuju takav registar.

U ovom radu je iznesena upotreba i utjecaj normi za adresiranje i usmjeravanje elektroničkih dokumenata na arhitekturu sustavu informacijskog posrednika. Prikazan je proces razmjene elektroničkih dokumenata prema modelu 4-strane. U poglavlju III. opisana je generička arhitektura sustava informacijskog posrednika te funkcija adresiranja i usmjeravanja. U poglavlju IV. predložena je struktura poruke, opisano zaglavlje poruke koje omogućava adresiranje i usmjeravanje unutar sustava informacijskog posrednika. Nakon toga je opisana uloga kanonskog oblika e-dokumenta unutar sustava informacijskog posrednika.

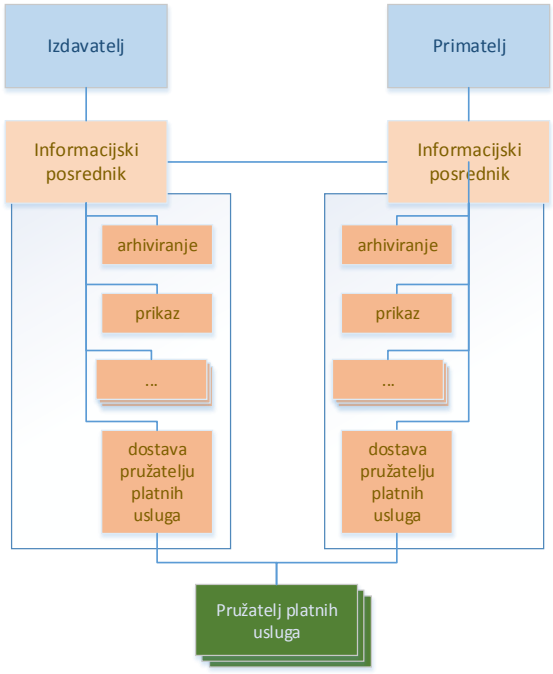
II. ZAHTJEVI NA ARHITEKTURU INFORMACIJSKOG POSREDNIKA

Informacijski posrednik će preuzeti aktivnosti ugovaranja razmjene elektroničkih dokumenata. Primatelj i pošiljatelj sklapaju bilateralne ugovore [10] [11] za uslugu razmjene elektroničkih dokumenata pojedinačno s informacijskom posrednikom. Usluga razmjene mora biti dostupna različitim profilima korisnika, npr. pravne ili fizičke osobe, velike ili male organizacije, organizacije visokog, srednjeg ili niskog stupnja informatizacije i sl. Zavisno od profila korisnika korisnici osim usluge razmjene elektroničkih dokumenata imaju potrebu za dodatnim uslugama čije funkcionalnosti nemaju realizirane u vlastitim IT sustavima, npr. arhiviranje elektroničkih dokumenata, dostava naloga za plaćanje pružatelju platnih usluga, mjesto za zaprimanje i slanje elektroničkih dokumenata (elektronički sandučić) i ostalo.

Kompleksnost se povećava ako korisnici ugovaraju razmjenu svaki sa svojim informacijskim posrednikom. Informacijski posrednici moraju ugovarati interkonekciju što će omogućiti dostavu elektroničkog dokumenata prema korisnicima koji se nalaze kod drugog informacijskog posrednika i obrnuto. Informacijski korisnik operativno radi po dva modela [4] [9] u odnosu na ishodište i krajnje odredište:

- Model 3-strane – u ovome modelu pošiljatelj i primatelj imaju ugovor s istim informacijskim posrednikom. Informacijski posrednik funkcionira u isto vrijeme s jedne strane u ime pošiljatelja, a s druge strane u ime primatelja.
- Model 4-strane – je model u kojem i pošiljatelj i primatelj imaju različite informacijske posrednike koji u njihovo ime razmjenjuju elektroničke dokumente.

Prijem i dostava elektroničkih dokumenata prema korisnicima različitih profila odvijat će se upotrebom različitih komunikacijskih kanala i protokola [3]. Svaki korisnik ugovara s informacijskim posrednikom komunikacijski kanal, protokol i tipove elektroničkih dokumenata, formate poruka, logičke identifikatore i ostalo.



Slika 1. Model 4-strane za razmjenu elektroničkih dokumenata.

III. GENERIČKA ARHITEKTURA SUSTAVA INFORMACIJSKOG POSREDNIKA

Sustav informacijskog posrednika potrebno je minimalno promatrati kroz nekoliko pogleda:

- Procesni pogled – koje je poslovne procese potrebno uspostaviti za isporuku usluge korisniku.
- Funkcionalni pogled – koje funkcije je potrebno osigurati za podršku uspostavljenim poslovnim procesima.

Ovdje identificirajte konkretne sponzore. Ako nema sponzora, izbrišite ovaj tekstualni okvir. (sponzori)

- Podatkovni pogled – kako i u kojim strukturama omogućiti pohranu podataka i informacija potrebnih za realizaciju neophodnih funkcija.

Svaki pogled će rezultirati uspostavom jedne od arhitektura i to procesne arhitekture, funkcionalne arhitekture i podatkovne arhitekture.

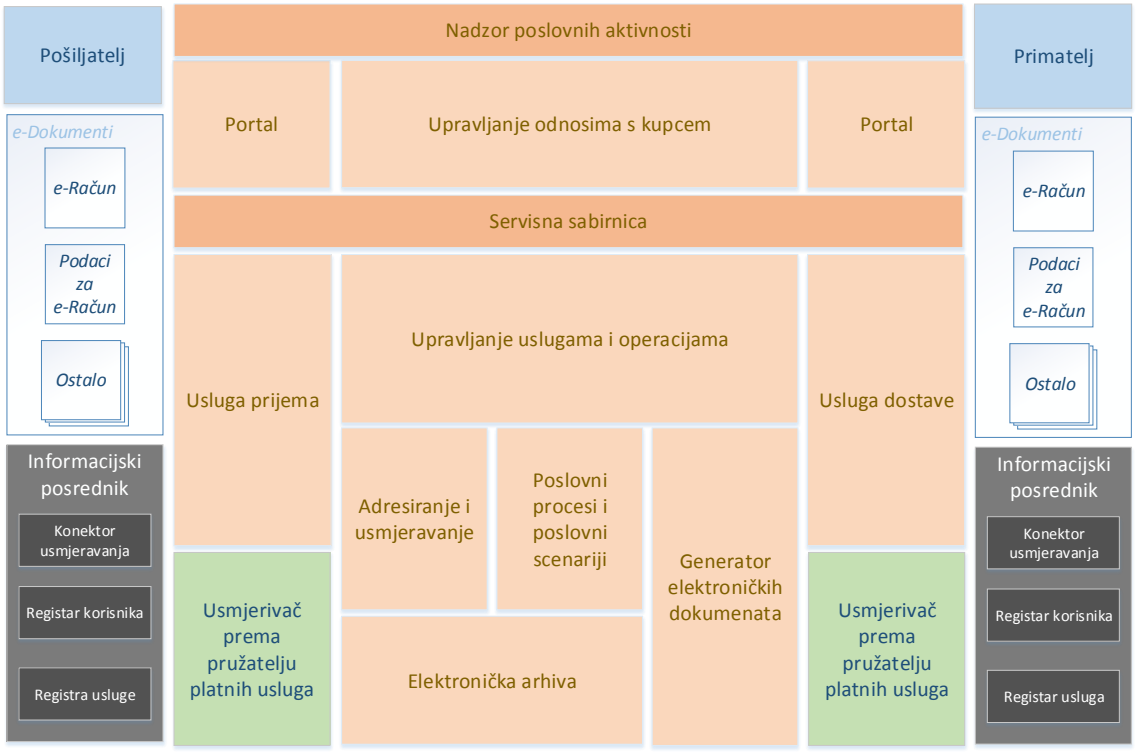
Procesna arhitektura informacijskog posrednika mora minimalno sadržavati ključne poslovne procese za uspostavu poslovanja usmjerenog na isporuku usluga. Uz poslovnu arhitekturu kao potporu izvršavanju poslovnih procesa potrebno je osigurati adekvatnu funkcionalnu i podatkovnu arhitekturu. Funkcionalna arhitektura definirat će potrebne funkcije za realizaciju poslovnih procesa. a podatkovna arhitektura smanjuje složenosti prilikom uspostave sustava, integracije s drugim informacijskim sustava, samog razvoja sustava i naknadnih prilagodbi.

Poslovanje informacijskih posrednika zasniva se na isporuci elektroničkih usluga korisnicima. Slično kao u telekomunikacijskoj industriji informacijski posrednik grupira usluge u proizvode, proizvode u kataloge proizvoda te uspostavlja tarifne modele nad katalogom proizvoda. Procesna arhitektura sustava informacijskog posrednika mora podržavati poslovne procese iz domene odnosa s kupcima (eng. Customer relationship management) i upravljanja uslugama i operacijama (eng. Service management and operations).

Kao ključna komponenta za uspostavu procesne arhitekture može se koristiti Business Process Framework (eTOM) [12], koji kao podatkovnu arhitekturu koristi Information Framework (SID) [13], što je dobro polazište za izgradnju podatkovne arhitekture informacijskog posrednika. SID osigurava referentni model i zajednički rječnik podataka.

Generička funkcionalna arhitektura (Slika 1.) implementira minimalni skup potrebnih funkcija za realizaciju poslovnih procesa iz domene odnosa s kupcem i upravljanja uslugama i operacijama. Glavne funkcionalne komponente su:

- Portal – dio sustava čija je uloga pošiljateljima i primateljima omogućiti pristup informacijama vezanim uz proces isporuke, statusu predanih elektroničkih dokumenata, pregledu ugovorenih usluga i servisa. Funkcionalnost portala mora implementirati različite funkcije zavisno o profilu korisnika. Npr. privatni i poslovni korisnici, izdavatelji i primatelji i sl.
- Upravljanje odnosima s kupcem (eng. Customer Relationship Management) – implementira funkcionalnosti potrebne za upravljanje odnosima s korisnicima, pokriva ugovaranje, osiguravanje nivoa kvalitete isporuke usluga (eng. Customer QoS i SLA management), kataloga produkata, registra korisnika, te naplate korištenja usluge.
- Upravljanje uslugama i operacijama (eng. Service Management & Operations) – funkcija koja implementira katalog servisa, konfigurira i aktivira servise, sakuplja informacije o upotrebi servisa (proces medijacije). Kao potporu operativnom radu minimalno mora uspostaviti funkcije potpore procesima rješavanja problema, te osiguravanja zadovoljavajućeg nivoa isporuke usluge.
- Usluga prijema, usluga dostave (eng. Receiver Services, Delivery Serivces) – uspostavlja različite vrste komunikacijskih kanala i protokola. Svaki kanal i protokol inkapsulira se u tzv.



Slika 2. Generička arhitektura informacijskog posrednika

konektor, npr. ERP konektor, Peppol konektor, konektor za informacijske posrednike.

- Adresiranje i usmjeravanje (eng. Addressing & Routing) – funkcija adresiranja će jedinstveno identificirati primatelja i pošiljatelja elektroničkog dokumenta, a usmjeravanja poruka će identificirati gdje i kako je potrebno elektronički dokument isporučiti. Usmjeravanje će razriješiti rutu kroz koju je potrebno poslati poruka na konačno odredište.
- Poslovni procesi i poslovni scenariji (eng. Business Processes &, Scenarios) – zavisno od ugovorenih proizvoda od strane korisnika sustav informacijskog posrednika omogućava izvršavanje različitih poslovnih scenarija korištenjem dostupnih servisa. Uspostavlja repozitorij i konfigurator poslovnih procesa.
- Ostale funkcionalne komponente: Elektronička arhiva (eng. e-archive) ,Servisna sabirnica (eng. Enterprise Service bus), Nadzor poslovnih aktivnosti (eng. Business activity monitoring), Generator elektroničkih dokumenata (eng. Eletronic Document Generator), Usmjerivač prema pružatelju platnih usluga (eng. Payment Service Provider gateway).

A. Funkcija Adresiranje i usmjeravanje

Prilikom isporuke do konačnog odredišta moguć je scenarij u kojem se na putu poruke nalazi jedan ili više drugih informacijskih posrednika

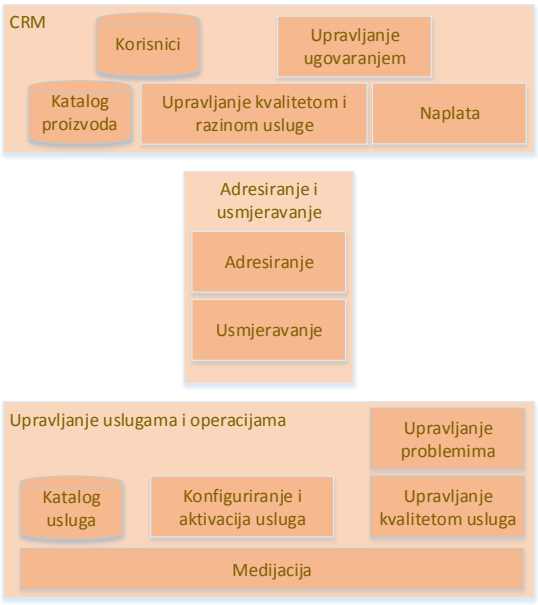
Adresiranje i usmjeravanje je poveznica između dva funkcionalna bloka sustava. Jedan je Upravljanje odnosima s korisnikom a drugi je Upravljanje uslugama i operacijama.

Upravljanje odnosima s korisnicima implementira Registar korisnika u kojem su pohranjeni identifikatori korisnika po različitim identifikacijskim shemama. Upravljanje uslugama i operacijama implementira Katalog Usluga, te konfiguraciju svih usluga koje pojedinog korisnik, kao i procese aktivacije pojedinih usluga.

Adresiranje koristi podatke iz Registra korisnika. Registar korisnika omogućava identifikaciju korisnika identifikatorom po jednoj shemi, dohvat svih podataka o korisniku, te preslikavanje identifikatora korisnika na drugu identifikacijsku shemu, npr. identifikacija korisnika po OIB-u, te dohvat njegovog GLN-a po GS1 identifikacijskoj shemi.

Usmjeravanje koristi podatke iz Konfiguratora usluga. U Konfiguratoru usluga su pohranjene instance pojedine definirane usluge iz Kataloga usluga. Instanca usluge kao objekt pohranjuje sve potrebne konfiguracijske vrijednost koje se odnose za pojedinog korisnika. Temeljem vrijednosti instanca Usluga pronalazi se ruta za isporuku elektroničkog dokumenta od pošiljatelja prema primatelju.

Npr. u jednostavnom scenariju kada i pošiljatelj, Korisnik A, i primatelj, Korisnik B, imaju ugovoren odnos s istim informacijskim posrednikom Korisnik A će imati ugovorenu uslugu slanje elektroničkog dokumenta. Instanca Usluge slanja će sadržavati podatke o



Slika 3. Pozicioniranje adresiranja i usmjeravanja u arhitekturi informacijskog posrednika

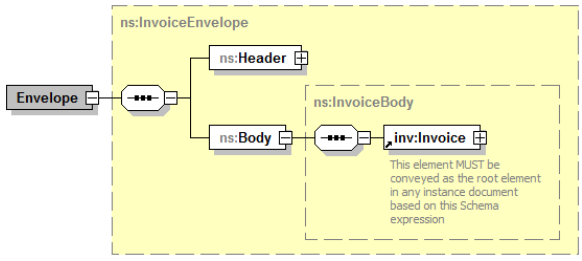
komunikacijskom kanalu i protokolu, prijemnoj lokaciji, formatu poruke i druge. Korisnik B će imati ugovorenu uslugu primanja. Instanca Usluge primanje će sadržavati informacije o komunikacijskom kanalu i protokolu, prijemnoj lokaciji, formatima poruke koju može zaprimiti i drugo. Adresiranje će temeljem identifikatora u zaprimljenoj poruci u Registru korisnika a temeljem identifikacijske sheme dohvatiti potrebne informacije o Korisniku B. Informacije će omogućiti ispravno usmjeravanje poruke prema prijemnoj lokaciji Korisnika B.

IV. ZAHTJEVI NA IMPLEMENTACIJU ADRESIRANJA I USMJERAVANJA

Jedan od kritičnih zahtjeva na adresiranje i usmjeravanje je podržavanje modela 4-strane [11]. U ovom modelu poruka od pošiljatelja putuje prema njegovom informacijskom posredniku koji je usmjerava prema informacijskom posredniku primatelja koji je prosljeđuje na krajnje odredište, tj. samom primatelju. Sve sudionike razmjene je potrebno jedinstveno identificirati. Problem je kompleksniji ako je potrebno preusmjeravati poruke zavisno o vrsti sadržaja koji se s njima prenosi.

Prilikom adresiranja različiti sustavi koriste različite identifikacijske sheme. Kao što je navedeno u [6] u različitim zemljama implementirani su različite identifikacijske sheme. Npr. Norveška, Nizozemska i Švedska koriste službeni porezni broj i upotrebu meta identifikatora sukladno ISO/IEC 6523. ePRIOR sustav podržava upotrebu internacionalnog identifikatora GLN od GS1 organizacije. Njemačka koristi identifikatore informacijskog posrednika čija je struktura sukladna ISO 3166 standardu [14]. Za Republiku Hrvatsku je u [6] specificirana identifikacijska shema prema poreznom broju (OIB-u). Prema identifikacijskoj shemi porezni broj za Republiku Hrvatsku ima kod 9934.

Implementacija adresiranja osim identifikacije po nekoj od identifikacijskih shema omogućuje i



Slika 4. Struktura poruke [7]

preslikavanje vrijednosti iz skupa vrijednosti jedne identifikacijske sheme na vrijednosti iz skupa vrijednosti druge identifikacijske sheme. Preslikavanje će dohvatiti jedan ili više dodatnih identifikatora po različitim identifikacijskim shemama. Norma [5] specificira formalni mehanizam za identifikaciju poslovnih subjekata u poslovnim XML dokumentima. Mehanizam adresiranja minimalno bi morao implementirati ISO 6523 [14], ISO 9735 [14] i ISO 20022 [14] kataloge shema za imenovanje identifikacijskih domena. Specifikacija [5] definira formalni mehanizam za referenciranje identifikacijskih shema korištenjem URN notacije.

U slučaju da je identifikacijska shema registrirana u ISO 6523, ISO 9735 ili ISO 20022 koristi URN format:

*urn:oasis:tc:*ebcore*:partyid-type:catalog-identifier:scheme-in-catalog:scheme-specific-identifier*

U ostalim slučajevima koristi URN format:

*urn:oasis:names:tc:*ebcore*:partyid-type:unregistered:mycatalog:myscheme*

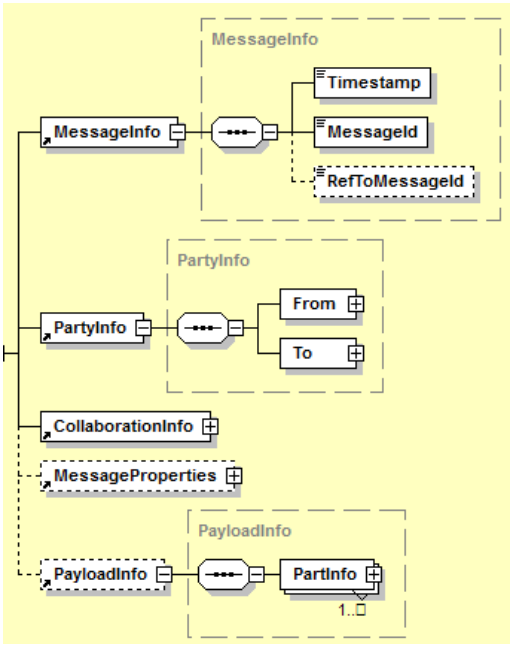
Usmjeravanje, na temelju vrijednosti identifikatora iz zaprimljene poruke, određuje odredište ili odredišta na koja zaprimljena poruka biti treba dostavljena. Prilikom određivanja odredišta usmjeravanje koristi funkcionalnosti adresiranja u svrhu preslikavanja identifikatora iz zaprimljene poruke na identifikatore odredišta. U zaglavlje poruke dodaju se novi identifikatori i poruka se prosljeđuje prema odredištu. Unutar sustava informacijskog posrednika za elektronički račun e-dokumenti imaju ovojnicu. Slika 4 prikazuje strukturu ovojnice. Ovojnica sadrži zaglavlje (Header) i tijelo poruke (Body). Zaglavlje poruke sadrži podatke o poruci, podatke usmjeravanja, opis tijela poruke itd. Slika 5 prikazuje strukturu zaglavlja poruke. Tijelo poruke sadrži kanonski oblik zaprimljene poruke ili e-dokument.

Zaglavlje poruke opisano u [7] ima sljedeće elemente:

- MessageInfo sadrži podatke o poruci: vremensku oznaku kreiranja zaglavlja (Timestamp), jedinstveni identifikator poruke (MessageId) i identifikator povezane poruke (RefToMessageId).
- PartyInfo sadrži podatke kreatora i primatelja poruke: identifikator organizacije (PartyId) i ulogu organizacije (Role).
- CollaborationInfo sadrži podatke o suradnji između stranaka: poveznica na ugovor (AgreementRef), usluga koja djeluje na poruku (Service), operacija ili aktivnost unutar usluge (Action) i identifikator konverzacije (ConversationId).
- MessageProperties sadrži svojstva poruke. Kao dijelovi zaglavlja svojstva sadrže informacije koje kvalificiraju podatke poruke ili omogućavaju povezivanje poruke s nekim poslovnim procesom. Time se može omogućiti učinkovitije praćenje, korelacija, usmjeravanje ili validacija podataka poruke bez pristupa sadržaju poruke.
- PayloadInfo identificira nosive podatke povezane s porukom. PayloadInfo može identificirati podatke unutar poruke ili podatke dostupne preko URL-a.

Primjer korištenja različitih informacijskih shema u elementima From i To iz zaglavlja poruke:

```
<eb3:PartyInfo >
  <eb3:From>
    <eb3:PartyId
type="urn:oasis:names:tc:ebcore:partyid-
type:unregistered:x2xplatformid:0001">3DF5848-85D5-
48FB-831F-A4DF30C7A128
    </eb3:PartyId>
    <eb3:Role>Supplier</eb3:Role>
  </eb3:From>
  <eb3:To>
    <eb3:PartyId
type="urn:oasis:names:tc:ebcore:partyid-
type:iso6523:9934">47475070790</eb3:PartyId>
    <eb3:PartyId
type="urn:oasis:names:tc:ebcore:partyid-
type:unregistered:x2xplatformid:0003">
          5006346999
              </eb3:PartyId>
    <eb3:Role>Buyer</eb3:Role>
  </eb3:To>
</eb3:PartyInfo>
```



Slika 5. Struktura zaglavlja poruke [7]

V. KANONSKI OBLIK ELEKTRONIČKOG DOKUMENTA

Sadržaj koji se razmjenjuje može biti u različitim formatima. U upotrebi su različite norme po kojima se generiraju dokumenti, npr. UN/CEFACT ASTMA, OAGi, UBL i druge. Potrebno je omogućiti standardni način razmjene sadržaja. Zbog optimizacije različiti tipovi dokumenata, npr. račun, narudžbenica, pisma, podnesci, akti i drugo, unutar sustava informacijskog posrednika bi trebali imati vlastiti standardni format, tj. kanonsku formu. Npr. preporuka za standardni format za e-Račun može biti zasnovana na normi UBL 2.1 [1]. Ovakav pristup rezultira optimizacijom pri transformacijama dokumenata iz jednog u drugi format te pojednostavljuje ukupnu implementaciju ostalih funkcionalnih komponenti sustava informacijskog posrednika. Informacije koje se pohranjuju u zaglavlje poruke prilikom zaprimanja dolaze iz poslovnog konteksta pošiljatelja. Poslovni kontekst pošiljatelja može biti zapisan u sadržaj poslanog dokumenta ili kao skup dodatnih podataka izvan sadržaja. Identifikatori iz sadržaja e-dokumenta ne trebaju biti osnova za usmjeravanja zbog nekoliko razloga. Ako je e-dokument kriptiran informacijskom posredniku identifikatori mogu biti nerazumljivi. Ukoliko informacijski posrednik ne čita sadržaj poruke neće moći dostaviti poruku prema identifikatorima iz same poruke. Za slučaj kada se podaci nalaze izvan sadržaja dokumenta prilikom razmjene potrebno je implementirati Business Document Header [2] strukturu. Iz Business Document Headera podaci poslovnog konteksta se preslikavaju u zaglavlje poruke, a u sadržaj poruke postavlja se sam dokument ali i Busines document Header. Zaglavlje poruke se koristi kao objekt koji nosi sve potrebne informacije za daljnje procesiranje.

VI. ZAKLJUČAK

Prilikom razmjene elektroničkih dokumenata potrebno je implementirati različite identifikacijske sheme u procesima adresiranja i usmjeravanja poruka. U internim procesima potrebna je čak i implementacija dodatnih internih identifikatora. Uspostava standardiziranog oblika poruke za internu obradu postaje obavezan zahtjev u sustavima informacijskih posrednika. Struktura interne poruke mora omogućiti jednostavna preslikavanja u ili iz eksternih struktura u kojima je pohranjen poslovni kontekst i elektronički dokument. Na taj način se osigurava prilikom razmjene jedinstven kontejner za različite elektroničke dokumente bez obzira na njihov format. S druge strane standardiziranje eksterne strukture omogućuje izvoru i konačnom odredištu automatizaciju obrade u njihovim poslovnim aplikacijskim sustavima.

U radu je predložena generička arhitektura sustava informacijskog posrednika pod utjecajem normi i smjernica za adresiranje i usmjeravanje u kontekstu poslovnih funkcija upravljanja odnosima s kupcem i upravljanja uslugama i operacijama. Predložena je struktura interne poruke prilikom obrade u sustavu informacijskog posrednika i zahtjevi na adresiranja i usmjeravanje, te preslikavanje različitih identifikatora.

LITERATURA

[1] OASIS Universal Business Language Technical Committee, "Universal Business Language Version 2.1", OASIS: OASIS Standard, http://docs.oasis-open.org/ubl/os-UBL-2.1/UBL-2.1.html, 04 January 2014

[2] UN/CEFACT Applied Technology Group (ATG), "UN/CEFACT Business Document Header Technical Specification 3.0 ODP step 3 Working Draft Revision [A7]", UN/CEFACT (United Nations Centre for Trade Facilitation and Electronical Business), 16. November 2010, http://www1.unece.org/cefact/platform/display/ATG/SBDH+-+ODP3+-+1st+working+draft , 04 January 2014

[3] OASIS ebXML Messaging Service Technical Committee, "AS4 Profile of ebMS 3.0 Version 1.0", OASIS: OASIS Standard, 23 January 2013, http://docs.oasis-open.org/ebxml-msg/ebms/v3.0/profiles/AS4-profile/v1.0/os/AS4-profile-v1.0-os.pdf , 12 January 2014

[4] CEN Workshops on eInvoicing, "CEN CWA 16050: Framework for the emerging network infrastructure of eInvoicing service providers throughout Europe", CEN (European committee for standardization): CEN CWA 16050, ftp://ftp.cen.eu/PUBLIC/CWAs/eInV2/CWA%2016050.pdf , 12 January 2014

[5] OASIS ebXML Core (ebCore) Technical Committee, "OASIS ebCore Party Id Type Technical Specification Version 1.0", OASIS: OASIS Committee Specification 01, 28. Sptember 2010, http://docs.oasis-open.org/ebcore/PartyIdType/v1.0/PartyIdType-1.0.html , 12 January 2014

[6] PEPPOL, "PEPPOL Transport Infrastructure ICT – Models Policy for use of Identifiers", 20. March 2012, https://joinup.ec.europa.eu/svn/peppol/PEPPOL_EIA/1-ICT_Architecture/1-ICT-Transport_Infrastructure/13-ICT-Models/ICT-Transport-Policy_for_using_Identifiers-220.pdf , 12 January 2014

[7] OASIS ebXML Messaging Service Technical Committee, "OASIS ebXML Messaging Services Version 3.0: Part 1, Core Features", OASIS: OASIS Standard, 01 October 2007,http://docs.oasis-open.org/ebxml-msg/ebms/v3.0/core/ebms_core-3.0-spec.pdf , 12 January 2014

[8] OASIS ebXML Messaging Service Technical Committee, "OASIS ebXML Messaging Services Version 3.0: Part 2, Advanced Features", OASIS: Committee Specification 01, , 19 May 2011, http://docs.oasis-open.org/ebxml-msg/ebms/v3.0/part2/201004/ebms-v3-part2.pdf , 12 January 2014

[9] CEN Workshops on eInvoicing, "CEN CWA 16464-1: Electronic invoicing - Part 1: Addressing and Routing", CEN (European committee for standardization): CEN CWA 16464-1, May 2012, ftp://ftp.cen.eu/CEN/Sectors/List/ICT/CWAs/CWA16464-1.pdf, 12 January 2014

[10] CEN Workshops on eInvoicing, "CEN CWA 16464-2: Electronic invoicing - Part 2: Model Interoperability Agreement for Transmission and Processing of Electronic Invoices and other Business Documents", CEN (European committee for standardization): CEN CWA 16464-2, May 2012, ftp://ftp.cen.eu/CEN/Sectors/List/ICT/CWAs/CWA16464-2.pdf , 12 January 2014

[11] CEN Workshops on eInvoicing, "CEN CWA 16464-3: Electronic invoicing - Part 3: Conformance Criteria for Interoperability between Electronic Invoicing Services", CEN (European committee for standardization): CEN CWA 16464-2, May 2012, ftp://ftp.cen.eu/CEN/Sectors/List/ICT/CWAs/CWA16464-3.pdf 12 January 2014

[12] Telemanagemt fortum (TMForum), "The Business Process Framework (eTOM)", http://www.tmforum.org/DownloadRelease135/15584/home.html , 12 January 2014

[13] Telemanagemt fortum (TMForum), " Information Framework (SID)",http://www.tmforum.org/InformationFramework/1684/Home.html , 12 January 2014

[14] ISO (International Organization for Standardization), ISO Standards, http://www.iso.org/iso/home.html , 12 January 2014

[15] L. Humski, B. Vrdoljak, Z. Skočir,"Concept, Development and Implementation of FER e- Invoice System", *SoftCOM 2012 - 20. International Conference od Software, Telecomunications & Computer Networks*, Split, FESB, 2012.

INFORMACIJSKE PRETPOSTAVKE ZA IZGRADNJU INFRASTRUKTURE ZA ŠIROKOPOJASNI PRISTUP

Dalibor Mesarić[1], Mario Štimac[2], Franjo Ambroš[3]

[1]Elektrotehnički fakultet Osijek, Hrvatska
d2mesaric@etfos.hr
[2]GEOprem d.o.o. Osijek, Hrvatska
mario.stimac@geoprem.hr
[3]GEOprem d.o.o. Osijek, Hrvatska
franjo.ambros@geoprem.hr

Sažetak - Digitalnom agendom Europa 2020 definiran je način provedbe širokopojasnog pristupa. Hrvatska vlada je 2011. godine usvojila strategiju i provedbeni program iz tog područja. Novina u pristupu pri realizaciji ovih dokumenata je uvođenje lokalne samouprava kao mogućeg aktivnog subjekta pri planiranju i izgradnji širokopojasne infrastrukture za koju su predviđeni fondovi EU iz područja infrastrukture koji u velikom postotku investicije sudjeluju u realizaciji takovih projekata. Kako bi se osigurale sve pretpostavke za racionalno ulaganje, potrebno je analizirati potrebe korisnika, opisati postojeću telekomunikacijsku infrastrukturu kroz katastar vodova te njen potencijal za realizaciju širokopojasnog pristupa. Osim toga, prostorne planove treba adekvatno pripremiti za rezervaciju infrastrukturnih koridora, a postupkom imovinsko pravnog uređenja osigurati vlasničke pretpostavke za realizaciju ove zadaće. Bez adekvatne informatičke podrške, koja će dati sveobuhvatni pregled problema, nije moguće potpuno definirati i izraditi projektni zadatak. U radu je izložen koncept ulaznih parametara potrebnih za bazu podataka temeljem koje se donose odluke o prioritetima ulaganja u širokopojasni pristup, te prikaz web pristupa podacima.

I. UVOD

Europska komisija je u okviru Strategije Europa 2020 sredinom 2010. inicirala Digitalni plan za Europu (engl. *Digitalna Agenda for Europe*) kao jednu od sedam ključnih inicijativa predviđenih tim strateškim programom. Hrvatska Vlada krajem 2011. godine donosi dokument: „Strategija razvoja širokopojasnog pristupa u Republici Hrvatskoj u razdoblju od 2012. do 2015. godine“ i „Provedbeni program Strategije razvoja širokopojasnog pristupa u Republici Hrvatskoj za razdoblje od 2012. do 2013.“[1].

U Osijeku je početkom 2012. godine na Elektrotehničkom fakultetu Osijek održan stručni skup uz sudjelovanje Ministra graditeljstva i prostornog uređenja i Hrvatske agencije za poštu i elektroničke komunikacije (HAKOM), s ciljem pripreme lokalne zajednice za provođenje ovog strateškog dokumenta koji, prema ocjenama svih, daje tehnološki iskorak usporediv s elektrifikacijom koja je provedena u prošlom stoljeću.

Na skupu je navedeno kako ovaj tehnološki iskorak ima i neke specifičnosti karakteristične za Hrvatsku. Naime, pravna regulativa iz područja elektroničke komunikacijske infrastrukture (EKI) nije predviđala da, investitor u ovu infrastrukturu, može biti lokalna samouprava što proizlazi iz duha ove Agende. U međuvremenu je HAKOM predložio, a Vlada donijela: „Uredbu o mjerilima razvoja elektroničke komunikacijske infrastrukture i druge povezane opreme“ [2] te time otklonila ovu dilemu. Drugi uočen problem je neadekvatna prostorno planska dokumentacija temeljem koje se planira izgradnja lokalne infrastrukture. Treći mogući usporivač procesa je neuređeno imovinsko pravno stanje na nekretninama na kojima se planira gradnja EKI. Uočeno je kako ne postoje aktualni strateški dokumenti razvoja lokalne sredine temeljem kojih planiramo buduće potrebe.

Tim koji je tada formiran pri Elektrotehničkom fakultetu Osijek, odmah je započeo sa pripremom dokumenata temeljenih na strategiji širokopojasnog pristupa. Napravljen je idejni projekt „Slavonska mreža“ koji će se, u vidu pilot projekta, realizirati na području 5 županija istočne Hrvatske (Osječko-baranjska, Vukovarsko-srijemska, Brodsko-posavska, Požeško-slavonska i Virovitičko-podravska). One predstavljaju 22,1% teritorija Hrvatske na kojem živi preko 800 tisuća stanovnika u 998 naselja, teritorijalno organiziranih u 22 grada i 105 općina. Veličina početne investicije je procijenjena na 21.5 milijuna €.

Projektna ideja je kandidirana na natječaj raspisan u okviru Programa pripreme i provedbe razvojnih projekata prihvatljivih za financiranje iz fondova Europske unije u 2013 godini – potprogrami Priprema regionalnih razvojnih projekata. Ministarstvo regionalnog razvoja i

fondova Europske unije je krajem 2013. godine donijelo odluku kojom se projektu odobrava tehnička pomoć u pripremi.

Kako bi se projekt racionalno vodio potrebno je izraditi informacijski sustav podataka sa svim parametrima koji utječu na poslovne odluke. Bitne karakteristike takovog sustava su modularnost i 24-satni pristup informacijama. Modularnost uključuje podatke razvrstane u bazama davatelja informacija koje se koriste pri izradi detaljnih projekata. Ovakvu logiku imaju geografsko informacijski sustavi (GIS) dok 24-satni pristup osigurava web servis. Autori ovog rada, sa suradnicima, razvijaju informacijsku potporu temeljenu na web pristupu kojom će se omogućiti racionalnije planiranje i planiranje izgradnje elektroničke komunikacijske infrastrukture, nastojeći iskoristiti dostupne javne baze podataka direktnim uvidom, ili uz prilagodbu.

II. INFORMACIJE NA KOJIMA SE TEMELJI PROJEKT

Realizacija Strategije širokopojasnog pristupa temelji se na prikupljanju raznih podataka o prostoru, imovinsko pravnim odnosima na trasama budućih koridora infrastrukture, stanju prostorno planske dokumentacije, lokacijama potencijalnih korisnika usluga, podataka o postojećoj infrastrukturi, razini dostupnih brzina širokopojasnog pristupa kako za fiksnu tako i za mobilnu mrežu i podataka o potencijalnoj izgradnji ostale infrastrukture. Važno je napomenuti da sve podatke prate i tzv. Metapodatci koji opisuju kvalitetu i strukturu podataka bez kojih često dolazi do krivog tumačenja ulaznih parametara, što se u konačnici odražava na kvaliteti realizane ideje.

A. Koordinatni sustavi

Prije nego detaljnije analiziramo izvore podatke potrebno je ukazati na potrebu vezanja svih podataka na određenu lokaciju/adresu. Pretpostavke za to su definirani koordinatni sustavi koji imaju matematičku osnovu preslikavanja geometrijskih oblika sa zemljine površine na ravninu. Napominjemo da koordinatni sustavi imaju svoju praktičnu svrhu te da je poznata matematička osnova velikog broja koordinatnih sustava. Samo podaci prikupljeni u jednom koordinatnom sustavu mogu se jednoznačno povezati s lokacijom. Ukoliko želimo koristiti podatke u jednom a prikupljeni su u drugom koordinatnom sustavu, neophodna je transformacija podataka. U Hrvatskoj postoje podaci o prostoru prikupljeni u više koordinatnih sustava (geodetski datumi). To je naša stvarnost o kojoj moramo i te kako voditi računa. Specifičnosti pojedinog koordinatnog sustava nije predmet ovog rada nego upozorenje da je od 1. 1. 2010. godine službeni koordinatni sustav (geodetski datum) Hrvatski terestički referentni sustav/trans Mercator, čija skraćenica glasi: HTRS96/TM [3]. Obveza upotrebe ove projekcije je pri izradi topografskih karata i katastarskih planova. Kako su od 1929. do 2010. godine na našim prostorima bila u primjeni 2 koordinatna sustava Gauss-Krügerove projekcije (5. i 6.), velik broj podataka treba transformirati u novi koordinatni sustav. Pri toj transformaciji, uz veliko povećanje karte, na ekranu računala se uočavaju značajna vizualna odstupanja do kojih dolazi zbog karaktera same transformacije. Pri ocjeni tih odstupanja mora se voditi računa o ekonomičnosti. Ponekad su odstupanja od nekoliko metara u granicama tolerancije, dok u nekim slučajevima tražimo centimetarsku točnost.

B. Karte, izvori podataka o prostoru

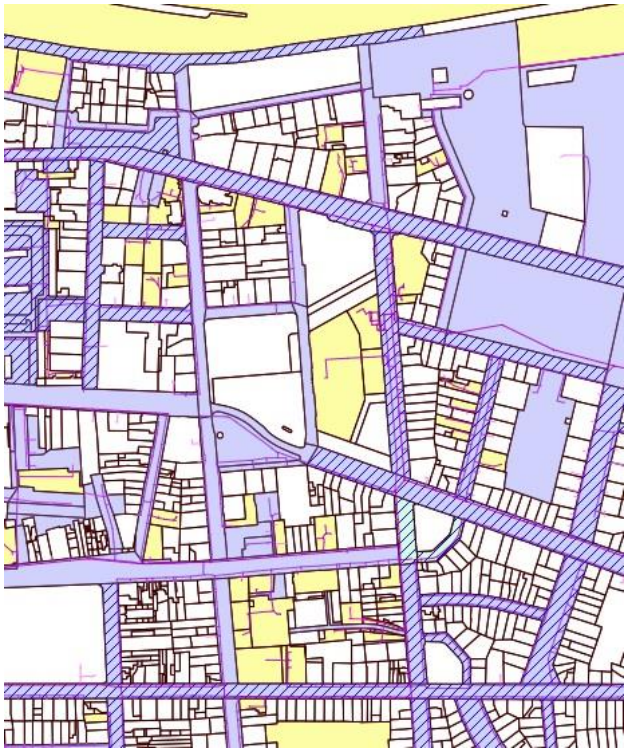
Izvori podataka za planiranje zahvata u prostoru su kartografski prikazi u raznim mjerilima (sl. 1). Izrada karte je složen i dugotrajan postupak te je potrebno ocijeniti aktualnost svake dostupne karte. Informacije za procjenu uporabne vrijednosti karte, ukoliko se radi o tiskanim primjercima, obično se nalaze na obodu karte. Potrebno je obrati pozornost na mjerilo jer ono određuje stupanj uopćavanja, vrijeme prikupljanja podataka, jer je ono indikator zastarijevanja, podatke koji ukazuju na vezu s kartama sitnijeg mjerila i dr. Kako je list karte prikaz (isječak) određenog područja koje se u odabranom mjerilu može prikazati u nekom standardnom formatu, potrebno je obratiti pozornost i na međusobnu vezu listova. Radi lakše dostupnosti i korištenja, karte se prevode u digitalni oblik. Najčešći oblici konverzije su skeniranje, geokodiranje i vektorizacija. Geokodiranje je postupak kojim se slikovni oblik karte smješta u prostor te time nastaje hibridni model koji je pogodan kao podloga za prikaz tematskih sadržaja. Vektorizacija je rastavljanje karte u elementarne dijelove, a potom spajanje u digitalni oblik. Smatramo da se o karti kao proizvodu mora posvetiti više pozornosti. Korištenjem karte u digitalnom obliku, ona se vraća u realan prostor. Međutim, podaci o „realnom prostoru“, konverzijom analogne u digitalnu kartu, nisu prikupljeni točnošću koja se očekuje (koju pruža računalo). Za razliku od analogne karte, gdje su dostupni podaci za svaki list karte temeljem kojih se procjenjuje upotrebljivost karte, korištenjem digitalnih prikaza prethodi ocjena upotrebljivosti temeljem Metapodataka.



Slika 1. Detalj list HOK 1:5000

C. *Katastarski planovi - izvori o podacima pravnih režima na nekretninama*

Katastarski planovi, za razliku od drugih kartografskih prikaza, težište usmjeravaju na granice raznih pravnih režima. Najčešće govorimo o granicama vlasništva. Katastarski planovi prikazuju katastarske čestice kao dijelove prostora čiji je vlasnik jedna (samovlasništvo) ili više osoba (suvlasništvo). Pri planiranju EKI, moraju se urediti imovinsko pravni odnosi s vlasnikom nekretnine, pri čemu su to najčešće fizičke osobe, tvrtke, lokalna ili regionalna samouprava odnosno država. Karakteristika uređivanja imovinsko pravnih odnosa za potrebe infrastrukture je kompleksnost odnosa zbog velikog broja katastarskih čestica na kojima se linijska infrastruktura gradi, neusklađeni podaci o posjednicima i vlasnicima koji se vode u javnim evidencijama (Katastar/ Zemljišna knjiga) sa stanjem na terenu. Hrvatska ima preko 14 milijuna katastarskih čestica, a dosadašnja elektronička komunikacijska infrastruktura je izgrađena na 8.72% katastarskih čestica, od kojih je 66% u privatnom vlasništvu. Na česticama u privatnom vlasništvu je oko 25% duljina svih trasa EKI. [4]. Koridori za infrastrukturu zaslužuju daleko veću pozornost nego što im se do sada poklanjala. Ukoliko se želi racionalnije i brže graditi mora se pojednostaviti pravna regulativu po pitanju prostornog planiranja i gradnje infrastrukture. [5].
Zbog potrebe uvida u imovinskopravno stanje razrađen je model provjere usklađenosti javnih evidencija i budućih planova izgradnje (sl. 2). Ulazni podaci, za ovu provjeru, su javni podaci katastra dostupni kao e-posjedovni list, digitalni katastarski planovi, čija je dostupnost uz plaćanje moguća preko Državne geodetske uprave, baza

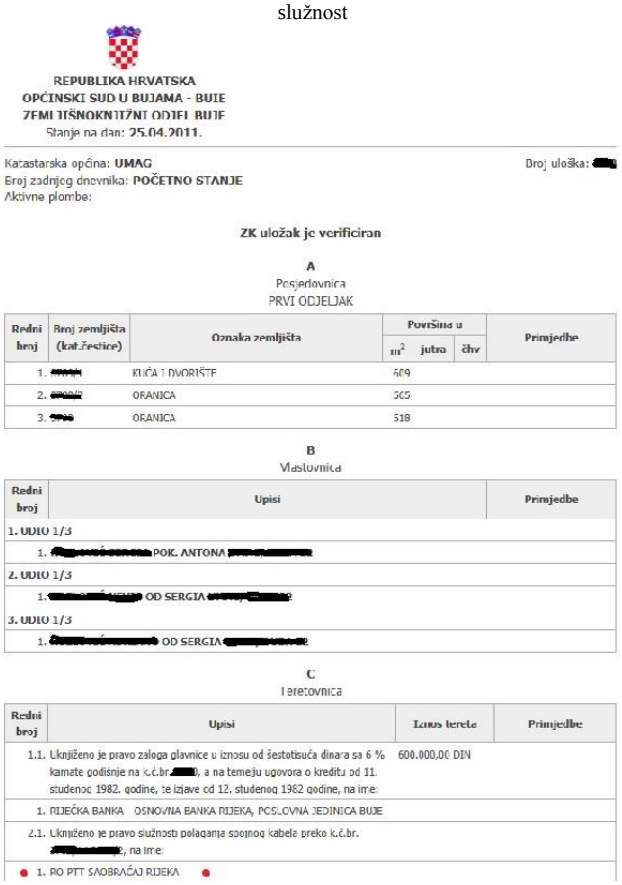


Slika 2. Sklopljeni ugovori za upis služnosti

Ministarstva pravosuđa: e-zemljišno-knjižni izvadak te podaci iz katastra vodova upravitelja voda. Za ovu je analizu razvijen program koji vizualizira neusklađenosti ili potencijalno problematične upise koji bi mogli biti prepreka pri kandidiranju projekta za financiranje iz EU fondova. Ovom se analizom otkrivaju i neka upisana prava (pravo služnosti) koje je prednik vlasnika infrastrukture stekao u prošlosti (sl. 3 i 4.).
Detaljnije o aktivnostima oko uređivanja imovinsko pravnih odnosa na infrastrukturi u radu [6]



Slika 3. Naznačene katastarske čestice (ulice) na kojima je uknjižena služnost



REPUBLIKA HRVATSKA
OPĆINSKI SUD U BUJAMA - BUIE
ZEMLJIŠNOKNJIŽNI ODJEL BUJE
Stanje na dan: 25.04.2011.

Katastarska općina: **UMAG** Broj uloška: [redacted]
Broj zadnjeg dnevnika: **POČETNO STANJE**
Aktivne plombe:

ZK uložak je verificiran

A
Posjedovnica
PRVI ODJELJAK

Redni broj	Broj zemljišta (kat.čestice)	Oznaka zemljišta	Površina u m²	jutra	čhv	Primjedbe
1.	[redacted]	KUĆA I DVORIŠTE	609			
2.	[redacted]	ORANICA	365			
3.	[redacted]	ORANICA	518			

B
Vlastovnica

Redni broj	Upisi	Primjedbe
1. UDIO 1/3		
1.	[redacted] POK. ANTONA [redacted]	
2. UDIO 1/3		
1.	[redacted] OD SERGIA [redacted]	
3. UDIO 1/3		
1.	[redacted] OD SERGIA [redacted]	

C
Teretovnica

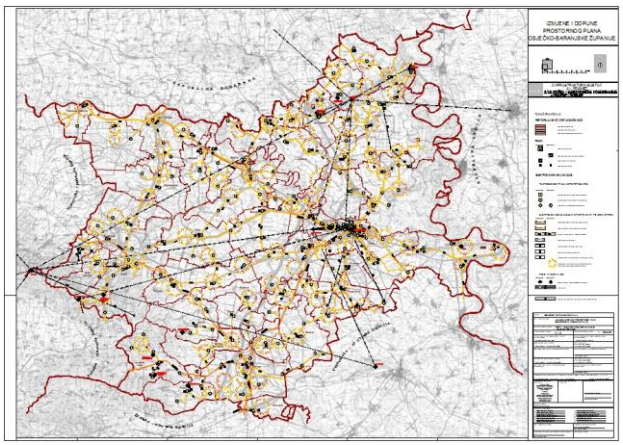
Redni broj	Upisi	Iznos tereta	Primjedbe
1.1.	Uknjiženo je pravo zaloga glavnice u iznosu od šestotisuća dinara sa 6 % kamate godišnje na k.č.br. [redacted], a na temelju ugovora o kreditu od 11. studenoc 1982. godine, te izjave od 12. studenog 1982 godine, na ime:	600.000,00 DIN	
1.	RIJEČKA BANKA OSNOVNA BANKA RIJEKA, POSLOVNA JEDINICA BUJE		
2.1.	Uknjiženo je pravo služnosti polaganja spojnog kabela preko k.č.br. [redacted], na ime:		
1.	PO PTT SAOBRAĆAJ RIJEKA		

Slika 4. Zemljišnoknjižni uložak s upisom služnosti

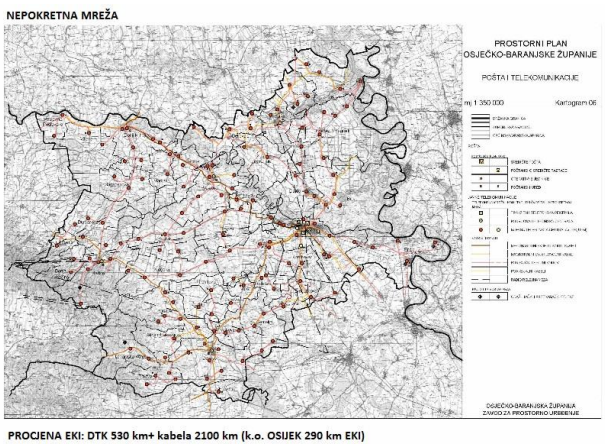
D. Prostorno planska dokumentacija

Planiranje infrastrukture na nekom području podrazumijeva osiguranje koridora ili lokacije za izgradnju na kome će se graditi infrastruktura ili objekt u funkciji infrastrukture. Za potrebe mobilne mreže, lokacije baznih stanica su dogovorno planirane u sklopu županijskih prostornih planova (sl. 5), koridori infrastrukture su obično rezervirani u planovima nižeg reda, izuzetak su magistralni pravci koji prelaze granice lokalne ili regionalne samouprave (sl. 6).

Istraživanja o troškovima izgradnje podzemne infrastrukture [7] ukazuju da je udio troškova kopanja oko 70% kod gotovo svakog voda. Ovo upućuje na moguću uštedu ukoliko se sinkronizira izgradnja više tipova infrastrukture istovremeno. O toj činjenici bi trebalo voditi računa već pri izradi prostorno planske dokumentacije. Informacijski sustav bi trebao signalizirati da su počele pripreme za projektiranje infrastrukture na određenom području. Zakoni Republike Hrvatske predviđaju različite investitore za svaku infrastrukturu. Procjenjujemo da postojeći kaos: kopanje – zatrpavanje – asfaltiranje pa ponovo kopanje treba prekinuti jer on osim neracionalnosti nepotrebno iritira građane. Transparentno planiranje (uvid preko weba), doprinijeti će većoj koordinaciji pri planiranju i izgradnji infrastrukture.



Slika 5. Lokacije baznih stanica u Osječko-baranjskoj županiji



Slika 6. Prostorni plan pošte i telekomunikacija Osječko-baranjske županije

E. Podaci o postojećoj infrastrukturi

Za realizaciju strategije širokopojasnog pristupa prostor kao resurs može ograničavati izgradnju potpuno nove infrastrukture, zbog toga je neophodno da se analiziraju postojeći kapaciteti koji su izgrađeni za potrebe telekomunikacija. Prije svega tu je telefonska kanalizacija koja se u gradovima počela graditi prije više od 70 godina. Zatim je tu distributivna telefonska kanalizacija (DTK) koja se gradi više od 20 godina (sl. 7). Osim toga tu su i optički kabeli raznih infrastrukturnih operatera ili vlasnika.

Kroz cijelo razdoblje izgradnje i eksploatacije telekomunikacijske mreže dokumentacija o položaju i kapacitetima se odvojeno vodila, što zbog praktičnosti a što zbog zaštite tajnosti. Tako je ustrojavanjem katastra vodova samo prostorni položaj infrastrukture bio predmet katastarske evidencije dok je tehnička dokumentacija koja je prikazivala kapacitete bila samo kod upravitelja voda. Za osmišljavanje projekta širokopojasnog pristupa interesantan je kapacitet kanalizacije i DTK te optički kabeli koji su u toj kanalizaciji.

Od 1969. godine postoji obveza osnivanja i vođenja katastra vodova [8] temelj kojega je geodetska izmjera vodova. Od 1973. ovo područje je regulirano Zakonom o katastru vodova [9]. Sjeverni dio Hrvatske ovu je zakonsku obvezu ozbiljnije shvatio od južnog dijela Hrvatske. Ova bi činjenica mogla sjevernom dijelu Hrvatske pomoći u realnijem osmišljavanju realizacije strategije širokopojasnog pristupa.



Slika 7. Distributivna telefonska kanalizacija

Promjenom tehnologije s „bakra na optiku“ promijenili su se tehnološki zahtjevi prema DTK kao infrastrukturi. Uz tehnologiju mikro-cijevi kapaciteti se mogu višestruko povećati tako da taj resurs treba staviti na raspolaganje svim ostalim operaterima pa i davateljima digitalnih usluga koje nemaju javni karakter, ali mogu značajno unaprijediti kvalitetu života (umrežavanje raznih senzora onečišćenja zraka, buke, usluga učenja, e-doktor i dr.). To u budućnosti implicira na mijenjanje odnosa koji proizlaze iz prava puta, u odnos partnerstva lokalne zajednice, infrastrukturnih operatera, davatelja i korisnika usluga.

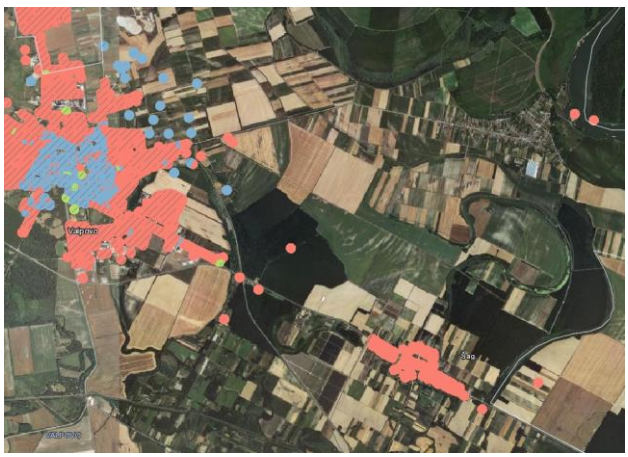
Posebno je interesantna činjenica da je najveći infrastrukturni operater HT Hrvatski telekom od 2009. godine investirao preko milijardu kuna u optičke kabele (FTTH) koji su danas „mrtvi kapitali“.

Još jedan segment koji utječe na uobličavanje svih elemenata digitalne usluge je kućna instalacija zgrade. Ukoliko se u proces ulaska u digitalno doba ne uključe vlasnici zgrada, upitna je realizacija strategije širokopojasnog pristupa. Tu lokalna zajednica mora preuzeti aktivnu ulogu izgradnje kućne instalacije po svim zgradama prema važećim propisima i standardima. Podaci o zgradama i vlasničkoj strukturi zgrade bitni su za osmišljavanje koncepta obnove kućne instalacije. Važno je napomenuti da dopunama Zakona o vlasništvu i drugim stvarnim pravima [10] od 2008. godine montaža i održavanje elektroničke komunikacijske infrastrukture i povezane opreme, spada u redovnu uporabu zgrade, što omogućava da 51% vlasnika donosi odluku o radovima na zgradi.

F. Dostupnost korisnika maksimalnim brzina

Kako je podatak o dostupnosti određenoj brzini širokopojasne usluge podatak koji mijenja prioritete kod investiranja u ovu infrastrukturu iz fondova EU potrebno to prokomentirati HAKOM ovu kartu.

Karta Područja dostupnosti širokopojasnog pristupa (sl.8) dostupna je na stranicama HAKOM-a [11] uz ogradu od prezentiranih podataka. Kako na brzinu preko postojeće bakrene infrastrukture utiče više elemenata: udaljenost od digitalnog čvora (centrale), kvaliteta spojeva u kabelu,



Slika 8. Karta Područja dostupnosti širokopojasnog pristupa

broj istovremenih korisnika to ukazuje da bez mjerenja na određenom području ne postoji realna slika o dostupnim brzinama. Ugovori kojima operateri ugovaraju pakete usluga s korisnicima teško bi prošli detaljniju provjeru udruge potrošača. Što znači brzina koja u periodu mjesec dana korištenja nije prešla 10% maksimalne ugovorene brzine.

Na karti su podaci dostupnosti i za fiksnu i za mobilnu mrežu. Nije potpuno jasno da li se u ocjeni područja koje se priprema za izgradnju infrastrukture za širokopojasni pristup smatra da postoje dva operatera fiksni i mobilni. Ukoliko se radi u dva komercijalna operatera područje spada u područje koje nema pravo na potporu iz fondova Europske unije.

III. MOBILNA MREŽA

Uloga mobilne mreže zaslužuje analizu njene dostupnosti i potencijala za digitalnim uslugama u ruralnom području. Nažalost potencijal ove usluge nije u Hrvatskoj korišten za razvoj privrede nego se koristi za zabavu i društvene mreže. Pokrivanje signalom je usmjereno prema koncentraciji korisnika a ne na pokrivanje teritorija. Rezultat čega je da su neki servisi ostali bez adekvatne telekomunikacijske logistike. Tako je u radu [12] iznijeta analiza rada mobilnog interneta preko kojeg servis CROPOS prima korekcijske parametre za određivanje pozicije u realnom vremenu s centimetarskom točnošću. Servis CROPOS je u službenu upotrebu pustila Državna geodetska uprava 9.12.2008. Ovaj servis je osim geodetima namijenjen i drugim korisnicima. Njegov potencijal u „preciznoj poljoprivredi“ nije prepoznat jer je mobilni Internet u ruralnom području često nedostupan. Usluga je usmjerena na govornu komunikaciju, tako da nedostaje kapaciteta za podatkovnim uslugama. Doradu ove mreže treba razmotriti za kandidiranje iz fondova EU u sklopu širokopojasnog pristupa.

Lokacije baznih stanica svih mobilnih operatera s podacima o usmjerenjima antena trebaju činiti bazu podataka temeljem koje će se praviti procjena realizacije strategije širokopojasnog pristupa.

IV. INFORMACIJSKA PODRŠKA ZA REALIZACIJU STRATEGIJE ŠIROKOPOJASNOG PRISTUPA

U posljednje vrijeme je mrežni pristup GIS-u sve popularniji. Prikaz dinamičkih mapa preko Interneta postaje rastući izazov. Vlada RH i lokalne samouprave sve češće postavljaju digitalne mape kao javne servise na svoje Web stranice. Trgovačka društva žele zemljopisnim podacima informirati svoje potencijalne kupce o mjestima trgovina.

Pri stvaranju GIS mapa teži se, nedovoljno obučenim korisnicima, pojednostaviti usluga dostupa podataka preko GIS okruženja. Jednostavnim označavanjem željenih procesa, korisnici aktivno sudjeluju u prikazivanju i istraživanju podataka korištenjem interaktivnih mapa preko pogodnosti korištenih web preglednika.

U GIS programe se ulažu velika sredstva za njihovo istraživanje i razvoj kako bi se tržištu ponudile nove, napredne mogućnosti. Ovdje se prvenstveno misli na obradu 3-D zemljopisnih podataka te na web prikaz temeljen na georeferenciranim podacima. Ovo se posebno odnosi na razvoj distribuiranih tehnoloških sustava obrade velikih količina podataka kao i obrade zemljopisnih podataka u realnom vremenu.

A. Primjena web prikaza podataka

Opredjeljujući se za 24 satni pristup podacima i istovremenom potrebom da se uključuje što veći broj korisnika, nameće se kao rješenje web pristup tipa: „geoportala otvorenog koda“ koji podržava MapGuide Open Source. Aktualnost karata i katastarskih planova moguće je provjeriti uvidom u digitalni ortofoto plan (DOF5). Isto koristi i HAKOM za karte područja dostupnosti širokopojasnog pristupa. Moguće je uključiti Google Maps i Street View. Preveliki zahtjevi prema informatičkim tehnologijama mogu dovesti do nerealnih očekivanja i usmjeravanja prema interpretaciji, a ne prema kvaliteti prikazanih podataka. Vizualizacija podataka može ukazati na moguće nelogičnosti podataka unutar baze, tako da o tome trebamo posebno voditi računa. Dosadašnja praksa svjedoči o neusklađenosti slike i zbilje. Ovo je uobičajena zabluda kod GIS tehnologija. Korisnik često odustaje prije moguće prilagodbe prikazanih podataka jer želi savršeno prezentirani traženi podatak. Zato je informacijska podrška u ovoj fazi namijenjena krugu stručnjaka koji će osmišljavati strategiju i na osnovu prezentiranih podataka određivati prioritete.

V. ZAKLJUČAK

Provedba Strategija razvoja širokopojasnog pristupa u RH od 2012. do 2015. godine je na području istočne Hrvatske započela projektom „Slavonska mreža“ – kojeg je 2013. g. pokrenuo Elektrotehnički fakultet Osijek u suradnji s pet županija regije Slavonija i Baranja. U pripremnim etapama ovog projekta uočen je niz problema i organizacijskih prepreka koje treba riješiti na nacionalnom planu. To su:

- Izrada informacijskog sustava kojim će se utvrditi standardi korištenja postojećih podataka i meta podataka kako bi se dobio realan uvid u stanje prostora i koridora.
- Utvrđivanje modela i organizacijske razine katastra vodova; ne samo širokopojasnog pristupa već svih infrastrukturnih instalacija u prostoru)
- Izgradnja informacijske podrške koja će biti kompatibilna s postojećim sustavima u jedinicama lokalne samouprave
- Izgradnja baze podataka o prostoru i elektroničkoj komunikacijskoj infrastrukturi kojom će se koristiti svi ovlašteni subjekti

Zahvati u prostoru koji iziskuju toliko ulaznih parametara kao što je izgradnja infrastrukture za širokopojasni pristup - ne mogu biti osmišljeni i realizirani bez informatičke podrške. Ovim se pristupom mogu paralelno prikupljati i analizirati podaci te modelirati postupci provedbe uz manje troškove i primjerene rokove izgradnje.

LITERATURA

[1] Strategija razvoja širokopojasnog pristupa u Republici Hrvatskoj u razdoblju od 2012. do 2015. godine i Provedbeni program Strategije razvoja širokopojasnog pristupa u Republici Hrvatskoj za razdoblje od 2012. do 2013. Narodne novine: 144/2011.

[2] Uredba o mjerilima razvoja elektroničke komunikacijske infrastrukture i druge povezane opreme, Narodne novine: 131/2012.

[3] Odluka o utvrđivanju službenih geodetskih datuma i ravninskih kartografskih projekcija Republike Hrvatske, Narodne novine: 110/2004

[4] F. Ambroš, O.Antoni, P. Božičković: Javne evidencije katastra i gruntovnice kao temelj geodetskih elaborata izvlaštenja za elektroničku komunikacijsku infrastrukturu, VI simpozij ovlaštenih inženjera geodezije, Neiskorišteni potencijali geodezije – napuštena i nova područja djelovanja, rad u zborniku (175-183),Opatija, 2013.

[5] J. Bajt: Uspostava katastra vodova kao jedinstvenog javnog registra linijskih infrastrukturnih građevina, VI simpozij ovlaštenih inženjera geodezije, Neiskorišteni potencijali geodezije – napuštena i nova područja djelovanja, rad u zborniku (184-191),Opatija, 2013.

[6] D. Salopek, F.Ambroš: Iskustvo Hrvatskog telekoma d.d. u imovinsko pravnom uređenju svoje elektroničke komunikacijske infrastrukture, Simpozij o inženjerskoj geodeziji, Hrvatsko geodetsko društvo, Slavonski Brod, 2012., uvodno predavanje

[7] F. Ambroš, M. Ivanović D. Mesarić: Izgradnja komunalne infrastrukture i razvoj optičke mreže na području Slavonije i Baranje, 11. skup o prirodnom plinu, toplini i vodi, 4. međunarodni skup o prirodnom plinu, toplini i vodi, rad u zborniku(134-142) Osijek, 2013.

[8] Pravilnik o metodama i načinu rada pri premjeru podzemnih instalacija o objekata, Službeni list SFRJ 49/1969.

[9] Zakon o katastru vodova, narodne novine SRH 44/1973.

[10] Zakon o vlasništvu i drugim stvarnim pravima, Narodne novine 91/1996 te dopuna 146/2008.

[11] url 1: www.hakom.hr

[12] G. Toplek, D. Bilajbegović, M. Štimac, F. Ambroš: Mobilni Internet u funkciji CROPOS-a i povećanju geodetske produktivnosti, 3. CROPOS konferencija, zbornik radova (97-102), Opatija, 2013.

Upotreba naprednih informacijskih sustava u kontekstu ubrzanja izrade studija izvodljivosti za potrebe izgradnje širokopojasne infrastrukture u RH

Damir Medved*, Saša Vojvodić*, Saša Drezgić**
* Ericsson NT d.d., Zagreb, Hrvatska
,** Ekonomski fakultet u Rijeci, Rijeka, Hrvatska
damir.medved@ericsson.com; sasa.vojvodic@ericsson.com; sdrezgic@efri.hr

Uz tradicionalne isporučitelje telekomunikacijskih sustava: elektroničke komunikacijske infrastrukture (EKI) i usluga pojavljuju se u posljednje vrijeme i posve novi tip ponuditelja telekomunikacijske infrastrukture. To su jedinice lokalne samouprave (JLS): gradovi i općine koji investiraju u gradnju nove komunalne infrastrukture EKI.

Liberalizacijom telekomunikacijskog tržišta došlo je do promjene vlasništva nad javnom telekomunikacijskom infrastrukturom – privatnopravni gospodarski subjekti koji su postali vlasnici, prvenstveno vode brigu o ekonomskim interesima svojih tvrtki. No razvoj i gradnja elektroničke komunikacijske infrastrukture javni je interes koji država i JLS moraju poticati, promicati i zaštititi osobito radi osiguranja istih uvjeta i ravnomjernog razvoja na svim nastanjenim područjima vodeći se pri tome načelima održivog razvoja, te zaštiti drugih javnih i privatnih interesa.

Gradovi i općine koji namjeravaju pokrenuti projekte nove EKI imaju mnogobrojne izazove – traže se novi izvori financiranja potrebni za daljnji razvoj zajednice, ali i bolju kontrolu i naplatu stavki koje im po zakonu pripadaju. Potencijalno velika sredstva iz budžeta Europske Unije (EU) su na raspolaganju, no pitanje je imaju li JLS dovoljno kapaciteta i znanja za pripremu i realizaciju takvih novih projekata? Malo je vremena preostalo za projektiranje infrastrukture ukoliko se želi smanjiti zaostatak za sumjerljivim zemljama iz Europske Unije. Manji gradovi nemaju vlastitu IT infrastrukturu i dovoljno stručnih kadrova za efikasno upravljanje životnim ciklusom izgrađene infrastrukture.

Koristeći svoja značajna iskustva u procesu izgradnje komunikacijskih sustava odnosno mrežne infrastrukture, te dobro definirane poslovne procese, Ericsson Nikola Tesla (ENT) predlaže „industrijalizaciju" procesa kreiranja projekata i aplikacija za EU fondove.

I. UVOD

Liberalizacijama telekomunikacijskih tržišta i postupcima privatizacije došlo je do promjene vlasništva nad javnom telekomunikacijskim infrastrukturama – privatnopravni gospodarski subjekti koji su postali vlasnici prvenstveno vode brigu o svojim korporativnim privatnim ekonomskim interesima, te objektivno zapostavljaju razvoj širokopojasne infrastrukture u ruralnim ili ekonomski neatraktivnim područjima gdje je povrat uloženog kapitala predug ili ne zadovoljava financijske kriterije privatne korporacije. No razvoj i gradnja javne elektroničke infrastrukture od javnog je interesa koji država i JLS moraju poticati, promicati i zaštititi osobito radi osiguranja istih uvjeta i ravnomjernog razvoja na svim nastanjenim područjima vodeći se pri tome načelima održivog razvoja, te zaštiti drugih javnih i privatnih interesa.

Telekom operateri odnosno davatelji usluga trebaju pouzdanu infrastrukturu da bi bili konkurentni, profitabilni, da bi odgovarali potrebama klijenata – ne samo danas - već i pri ispunjavanju budućih zahtjeva tržišta i poslovnih ciljeva. Sposobnost davatelja usluga da integrira vlastitu telekomunikacijsku infrastrukturu s aplikacijskim alatima sustavima operativne podrške (Operation Support Systems – OSS) u širem smislu, određuje koliko su dobro dostupne informacije usvojene, korisne, upravljane i sačuvane. Stoga za učinkovito i pravovremeno donošenje odluka informacije iz brojnih sustava moraju biti dostupne svim nadležnim osobama.

No, uz već tradicionalne korisnike OSS sustava pojavljuje se u posljednje vrijeme i posve novi tip korisnika – JLS. One su prepoznale važnost osiguranja nove elektroničke komunikacijske infrastrukture koja će svim stanovnicima zajednice omogućiti nove razvojne potencijale korištenjem inovativnih informacijsko komunikacijskih tehnologija (ICT). Nova širokopojasna informacijska infrastruktura postala je u zadnjim revizijama prostornih planova jednakovrijedna drugim infrastrukturama (energetskim, vodoopskrbnim, prometnim, itd.) čime se konačno zahtijeva da uz vodu, električnu energiju i prometnice, građanima treba osigurati i širokopojasni pristup na Internet.

Evolucija ponašanja privatnih i javnih korisnika komunikacijskih usluga podržana je eksponencijalnim razvojem sofisticiranih korisničkih uređaja i pripadajućih multimedijskih aplikacija. Isporuka ili razmjena audio-video sadržaja visokih i super visokih rezolucija putem EKI, zahtjeva osiguranje vrlo brzih **simetričnih**, minimalno 100 Megabita u sekundi (Mbps) komunikacijskih infrastruktura koje će osigurati odgovarajuću kvalitetu i korisničko iskustvo. Nažalost,

postojeća bakrena kabelska infrastruktura ima svoja tehnološka ograničenja i ne može osigurati tražene brzine i kvalitetu.

Trenutne digitalne (Digital Subcsriber Line, DSL) i kabelske (Data Over Cable Service Interface Specification, DOCSIS) linijske tehnologije, ovisno o kvaliteti bakrenih i kabelskih mreža, dosežu maksimalne brzine u prijemu od 50 Mbps, što nije dovoljno već sada, a kamoli za buduće potrebe razvitka društva. Potrebne simetrične brzine omogućava svjetlovodno vlakno, pa se ono treba planirati, vezati i ugrađivati kao integralni dio cjelokupne komunalne infrastrukture (struja, voda, kanalizacija, plin, prometnice).

II. POSLOVNI MODELI

Postoje različiti poslovni modeli za tržište komunalnih širokopojasnih mreža [1], tj. ne postoji jedinstveni poslovni model koji bi bio pogodan za sve JLS.

U pravilu, u poslovnom modelu postoje tri osnovne razine:

- prva razina određuje tko eksploatira pasivnu infrastrukturu (distribucijska telekomunikacijska kanalizacija, zdenci, šahtovi, svjetlovodni kabeli i dr.)
- druga razina određuje tko osigurava i eksploatira aktivnu mrežnu infrastrukturu (usmjernici, komutatori, prijenosni svjetlovodni komutatori i sl.)
- treća razina određuje tko nudi pristup mreži, usluge i sadržaj.

Mrežni operator osigurava pasivnu infrastrukturu, najčešće svjetlovodnu kabelsku mrežu. Tipičan mrežni operator je komunalna tvrtka u vlasništvu JLS.

Komunikacijski operator osigurava aktivnu infrastrukturu (komutacije, usmjernike, svjetlovodne mreže, i dr.). Komunikacijski operator iznajmljuje fizičke kapacitete od mrežnog operatora.

Davatelj usluga, osigurava usluge na mreži. Davatelj usluga plaća komunikacijskom operatoru pristup krajnjem korisniku i ima prihod od krajnjeg korisnika.

Dodjela različitih odgovornosti jednoj od razina u poslovnom modelu proizvodi različite poslovne scenarije koji pokazuju kako javne organizacije, gradska uprava, komunikacijski operatori i davatelji usluga međusobno surađuju na dobrobit krajnjih korisnika, građana i zajednice.

Posve je jasno da bez obzira na odabrani model JLS rade iskorak u, za njih, posve novo i nepoznato područje i da će trebati efikasno i ekonomski održivo rješenje za upravljanje životnim ciklusom infrastrukture. Bez obzira na nedostatak sredstava u JLS, otvaranje mogućnosti za financiranje širokopojasne infrastrukture iz sredstava EU strukturnih fondova daje nade da će se očekivane investicije u narednom razdoblju ipak realizirati.

Važno je napomenuti da usmjeravanje sredstava EU strukturnih fondova za financiranje telekomunikacijske infrastrukture omogućava kompenziranje značajnog financijskog jaza kada je riječ o prvoj razini poslovnog modela, odnosno pasivnoj infrastrukturi. Dio troškova pasivne infrastrukture koje treba nadoknaditi JLS može se financirati na temelju prihoda iz druge i treće razine poslovnog modela, odnosno iz prihoda od komunikacijskog operatora te davatelja usluga.

Zahvaljujući činjenici da telekomunikacijska infrastruktura može ostvariti određenu razinu vlastitih (komercijalnih) prihoda ovakvi sustavi idealni su za uspostavljanje odgovarajućeg modela javno-privatnog partnerstva (JPP) [2], [3]. Naime, zbog neutralizacije financijskog jaza sredstvima EU značajno se smanjuju rokovi povrata na vlastita ulaganja JLS u ovim projektima. U situaciji ekonomsko-financijske krize, u kojoj nema raspoloživih proračunskih sredstava, JLS mogu putem JPP modela financiranja prenijeti trošak nepokrivenog dijela financiranja pasivne infrastrukture na privatnog partnera. Na taj način zaokružuje se poslovni model u kojem se financira javna telekomunikacijska infrastruktura (u vlasništvu JLS) bez korištenja ograničenih proračunskih sredstava. Naravno, stupanj samofinanciranja ovisit će o visini donacije EU te mogućnostima komercijalnog iskorištenja druge i treće razine infrastrukture.

III. SITUACIJA

A. Izazovi lokalne zajednice

Kako smo vidjeli, potencijalno velika EU sredstva su na raspolaganju, no pitanje je su li JLS dovoljno stručno i organizacijski sposobne za pripremanje, apliciranje i realizaciju novih projekata vezanih uz širokopojasne distribucijske mreže? Malo je vremena preostalo za projektiranje infrastrukture ukoliko se želi smanjiti zaostatak za sumjerljivim zemljama iz Europske Unije. Manji gradovi nemaju vlastitu IT infrastrukturu i dovoljno stručnih kadrova za efikasnu pripremu i upravljanje životnim ciklusom novo izgrađene infrastrukture.

Mnoge JLS (županije, Grad Zagreb, gradovi i općine) kao nositelji planiranja vlastitih komunalnih infrastruktura trebaju projektante obvezati na kvalitetno razvojno planiranje EKI odnosno same preuzeti odgovornost za unaprjeđivanje širokopojasne infrastrukture na svojim područjima.

Dodatni argument za veći angažman JLS jest nepostojanje jedinstvene baze postojećeg stanja EKI i povezane opreme što je temelj svakog kvalitetnog planiranja. Podaci neophodno potrebni za planiranje razvoja sporadično se prikupljaju od telekom operatora, postojećih raznolikih podatkovnih izvora u JLS (npr. katastar, komunalna poduzeća, itd.). Tako prikupljeni podaci dobiveni u različitim formama i formatima često nisu konzistentni, ne odražavaju stvarno stanje na terenu i bitno usporavaju bilo kakve zahvate u prostoru te poskupljuje sve aktivnosti.

B. Strategija i ciljevi Europske Unije

Europa 2020 strategija je rasta EU-a za sljedeće desetljeće. U svijetu koji se stalno mijenja želimo da EU ima pametno, održivo i inkluzivno gospodarstvo. Navedena tri prioriteta, koja se međusobno jačaju, trebala bi pomoći Europskoj uniji i državama članicama ostvariti

visoke razine u području zapošljavanja, produktivnosti i kohezije.

Unija je, konkretno, postavila pet ambicioznih ciljeva: u području zapošljavanja, inovacija, obrazovanja, socijalnog uključivanja i klime/energije koje je potrebno ostvariti do 2020. godine. Svaka je država članica prihvatila vlastite nacionalne ciljeve u svakom od navedenih područja. Strategiju određuju konkretne aktivnosti na razini EU-a i pojedinih država članica.

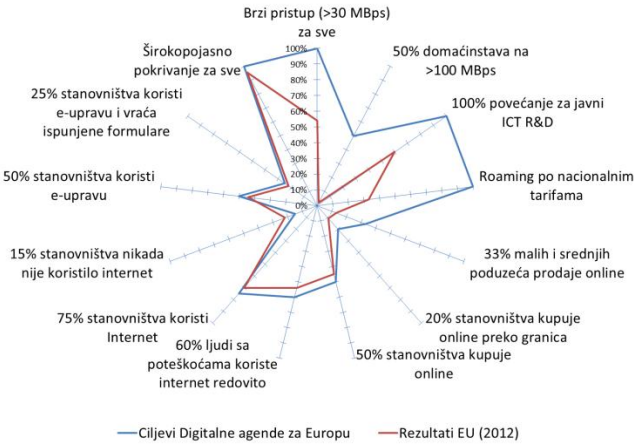
Na temeljima strategije razvoja Europe [4], a kroz program Digitalne agende za Europu (DAE) unaprjeđuje se razvoj digitalnih tehnologija koje će pomoći ostvarenju ciljeva strategije Europa 2020 [4][5]. Digitalna ekonomija raste sedam puta brže od ostale ekonomije, te je od strateške važnosti razvoja pojedine članice EU ispuniti ciljeve definirane DAE programom, prikazane Slika 1.

Hrvatska gotovo u svim parametrima mjerenja uspješnosti razvoja digitalne industrijalizacije zaostaje za prosjekom Europske Unije. Za dosezanje razina razvoja naših susjeda potrebno je uložiti značajno veći napor i angažirati značajnija financijska sredstva. Ipak, najvažnija karika brzih i kvalitetnih pomaka razvoja je efikasna organizacija cjelokupnog procesa od priprema projekata, natječaja, izgradnje i konačnog upravljanja novim infrastrukturama temeljem državne i regionalne strategije razvoja [6][7]. Uz novu infrastrukturu potrebno je osigurati i razvoj novih digitalnih e-usluga paralelnim programima.

Kako bi se spomenute ambicije pretvorile u stvarnost Europska komisija aktivno radi na uklanjanju barijera. Spomenut ćemo samo neke, našem radu, interesantne akcije.

Najznačajnija aktivnost je sufinanciranje izgradnje širokopojasnih mreža velikih brzina kroz instrumente EU i privlačenje kapitala za ulaganja u širokopojasnu infrastrukturu. Razlog fokusa na područje sufinanciranja leži u činjeničnom stanju slabog razvoja EKI (Slika 1. uzrokovana percipiranim visokim rizicima javno-privatnih modela, dugim povratima investicija, nedovoljnim iskustvom javnog partnera, te još uvijek nedostatnim dokazima koji opravdavaju novi poslovni model kojim širokopojasna infrastruktura postaje nova državna imovina.

Nadalje, države članice trebaju poduzeti mjere, uključujući i zakonske odredbe, kako bi se olakšalo ulaganje u širokopojasnu infrastrukturu. Sustavno uključivanje potencijalnih investitora prilikom izgradnja komunalnih infrastruktura (voda, plin, kanalizacija i sl.), jasna regulativa prava puta, te dostupnost podataka o postojećoj infrastrukturi značajno će smanjiti trošak gradnje.



Slika 1. Ciljevi i rezultati (2012 g.) Digitalne agende za Europu

C. EU fondovi i koristi za lokalnu zajednicu

U svrhu potpore ostvarivanja ciljeva Digitalne agende Europska komisija je:

- usvojila revidirane smjernice za primjenu EU pravila o državnim potporama u sektoru širokopojasnog pristupa [4], posebno dijela koji sadrži jačanje obaveza otvorenog pristupa i unaprjeđenje pravila transparentnosti.
- usvojila prijedlog Uredbe o smanjenju troškova za građevinske radove, koji čine do 80% troškova izgradnje širokopojasne mreže. Izbjegavanje nepotrebnih kopanja prema prijedlogu Uredbe Komisije, može se uštedjeti između 40 i 60 milijardi eura ili do 30% ukupnih investicijskih troškova.

Važna nova komponenta koja se pojavljuje kao prilika za jedinice lokalne samouprave da osiguraju sredstva subvencioniranja izgradnje putem državnih potpora, ali i putem Europskih strukturnih fondova: Europskog poljoprivrednog fonda za ruralni razvoj (*Agricultural Fund for Rural Development*, EAFRD) i Europskog fonda za regionalni razvoj (*European Regional Development Fund*, ERDF). ERDF je namijenjen financiranju projekata za smanjivanje razlika u razvoju pojedinih zemljopisnih područja koja uključuju postavljanje nove elektroničke komunikacijske infrastrukture.

Budžet fonda je oko 50 milijardi eura godišnje temeljen na sufinanciranju (50-85% - ovisi najviše o socioekonomskim parametrima države) i namijenjen regijama čiji je BDP po stanovniku manji od 75% prosjeka EU. Sam fond ima 6-godišnji ciklus sa sljedećim bitnim miljokazima:

- Europska Komisija (European Commission, EC) definira prioritete fondova na razini EU kroz Community Strategic Guidelines dokument
- Države članice na temelju tok dokumenta transformiraju prioritete u nacionalne prioritete kroz National Strategic Reference Framework (NSRF) dokument
- Države članice na temelju NSRF izrađuju Operativne Planove (OP) za svaku svoju regiju

- EC mora odobriti NSRF i OP prije bilo kakve implementacije

Aplikacija za sredstva ERDF slijedi sličan proces kao i za IPA: nakon što su odobreni OP, operativno tijelo Republike Hrvatske (npr. ministarstvo) otvara natječaj za prijavu projekata. Budući korisnici predaju projektni sažetak koji se nakon prve selekcije pozitivno ocjenjuje ili odbacuje. Ako je projekt pozitivno ocijenjen, onda se od predlagatelja traži potpuni prijedlog projekta sa svim financijskim i operativnim detaljima. Operativno tijelo daje konačnu ocjenu, te se pristupa potpisivanju ugovora. Za svaku godinu države članice potpisuju ugovor s EC o sufinanciranju odobrenih projekta u narednoj proračunskoj godini.

D. Ključne koristi za JLS

Investiranje u širokopojasne mreže donosi brojne koristi za JLS primarno u domeni razvoja gospodarstva, otvaranje novih poslovnih prilika i novih radnih mjesta. Utjecaj izgradnje širokopojasne infrastrukture na porast bruto društvenog proizvoda (BDP) u JLS je nedvojben pri čemu upravo sredstva EU pomažu da se infrastruktura uopće izgradi. Glavne direktne, indirektne i inducirane koristi koje donosi nova infrastruktura za zajednicu su:

- novi izvor prihoda lokalnim zajednicama od prodaje kapaciteta i najma infrastrukture
- e-građanin u uređenoj e-zajednici: e-uprava, e-zdravstvo, socijalna skrb, sigurnost, e-obrazovanje, transport, turizam: dostupnost svih tijela gradske i ostalih razina vlasti potrebnih za realizaciju usluga potrebnih građanima i poslovnim korisnicima
- nove mogućnosti poslovanja i trgovanja putem novih e-poslovnih mogućnosti
- otvaranje novih radnih mjesta, omogućavanje i povećanje inovativnosti i konkurentnosti.
- povećanje kvalitete i standarda života
- zaustavljanje i okretanje negativnih depopulacijskih trendova

E. Izazovi projektiranja i izgradnje nove širokopojasne infrastrukture

Za uspješan završetak svakog projekta, pa tako i izgradnje širokopojasne infrastrukture, potrebno je odlučiti o metodama, procesima i alatima koji će se koristiti. Neophodno je u samom početku izabrati tehnološko i topološko rješenje nove infrastrukture ne zaboravljajući na činjenicu da projekt ne smije biti tretiran izolirano, te da je on samo djelić veće geografske i komunalne cjeline. Kompatibilnost metoda, procesa i alata u tom širem kontekstu i van okvira JLS mogla bi donijeti korist cijeloj zajednici.

Dizajniranje mreža može biti usko grlo i usporiti ostale povezane procese. Bez centraliziranog pristupa dizajnu mreža, koji bi trebao biti stalno dostupan i korišten tijekom cjelokupnog procesa implementacije od projektiranja preko izgradnje do prodaja i upravljanja instaliranih kapaciteta, bitno će biti otežano dovršavanje pojedinih radnih segmenata ili zadataka. Slobodni decentralizirani pristupi nužno sa sobom donose problematiku različitosti formata, dokumentacija, teškog pristupa informacijama partnera na projektima te eventualnu potrebu usklađivanja ili integracija između JLS i županija na državnoj razini. Problematika takvog pristupa zahtijevat će puno iteracija i usklađivanja kako bi se pojedini zadaci mogli izvršiti.

Pogledajmo izazove koje donosi projektiranje širokopojasnih mreža za JLS. Tradicionalni procesi su prespori, zahtijevaju veliku količinu manualnog rada, izvori podataka su nesinkronizirani odnosno neažurni te su moguće brojne pogreške:

- Novodizajnirana topologija mreža nije usklađena sa postojećom infrastrukturom (voda, struja, kanalizacija, ceste, plin, postojeća EKI) ili socioekonomskom situacijom u JLS
- Dizajn i stvarno izgrađena mreža i pripadajuća dokumentacija nisu usklađeni najboljom praksom te sa standardima i smjernicama koje propisuju EU i nadležna tijela države
- Informacije iz projekta mreže teško su dostupne ili čitljive svim sudionicima projekta, te je vrlo visoka mogućnost pogrešaka tijekom izgradnje
- Ne postoji konzistentan proces upravljanja projektom što rezultira probijanjima rokova izgradnje i troškovnika
- Loši ili nedostupni podaci zahtijevaju češće nadzore i više ponekad potpuno nepotrebnih izlazaka na teren
- Različiti alati i pristupi dizajniranju otežavaju međuprojektnu suradnju (posebice na razini regija ili klastera), a dugoročno su neodrživi
- Ne postoji nesmetani protok informacija između projektanata, izgradnje, upravljanja i održavanja mreže što zahtjeva velike količine manualnog rada, a time i česte pogreške u interpretaciji informacija.

Navedeni izazovi mogu se riješiti unificiranim pristupom. Primjenom zajedničkih metoda, procesa i alata tijekom svih segmenata projekta ali i na svim sličnim projektima moguće je smanjiti ili potpuno ukloniti navedene izazove, te ubrzati pripremu i izgradnju novih mreža.

GIS baziranim alatom za projektiranje i upravljanje mrežama pozitivno će se utjecati na kvalitetu i troškove projekta: boljim načinom uvida u stvarnu situaciju na terenu, ujednačavanjem dizajna, boljim načinom dostave planova izgradnje tehničarima na terenu, mogućnost uvida u blizinu postojećih infrastruktura, te brze procjene troškova izgradnje.

Geografski informacijski sustav (GIS) je sustav za upravljanje prostornim podacima i osobinama pridruženih njima. U najstrožem smislu to je računalni sustav sposoban za integriranje, spremanje, uređivanje, analiziranje i prikazivanje geografskih informacija.

Ključ industrijalizacije leži u upravljanju podacima koje treba prikupiti, konfigurirati i koristiti. Zapitajmo se

koje podatke treba spremiti, kako i u kojem sustavu, te kako ćemo osigurati njihovu dostupnost partnerima u projektima i drugim korištenim sustavima? U zajedničkom GIS sustavu informacije o tome gdje je planirana ili se nalazi oprema, kakvih kapaciteta i zauzeća, kako je povezana, koji je trošak gradnje, te kada će biti spremna dostupne su u svakom trenutku.

Specijaliziranim GIS alatom namijenjenim za upravljanje životnim ciklusom telekomunikacijske infrastrukture automatizirat će se pojedini specifični zahtjevi procesa dizajna, izgradnje i upravljanja. Definiranje vlastitih pravila izgradnje automatizirati će planiranje i dizajn novih mreža, shematski prikazi osigurat će kvalitetnije snalaženje u planiranju i rješavanju problema, a alat za terenske ekipe osigurat će lakši i sigurniji rad, te stalnu konzistenciju podataka. Nakon izgradnje proširenje funkcionalnosti na upravljanje instaliranom imovinom smanjiti će i ubrzati donošenje odluka smanjujući trošak održavanja.

IV. NACIONALNI PROGRAM

Ministarstvo pomorstva, prometa i infrastrukture (MPPI) pokrenulo je niz aktivnosti usmjerenih prema razvoju širokopojasne infrastrukture u Hrvatskoj. Krajem 2013. godine objavljena je javna rasprava na temu o okvirnom nacionalnom programu za razvoj infrastrukture širokopojasnog pristupa u područjima u kojima ne postoji dostatan komercijalni interes za ulaganja. Okvirnim programom predviđeno je imenovanje jednog tijela javne vlasti nositeljem Okvirnog programa (NOP), koji će koordinirati provedbu Okvirnog programa na nacionalnoj razini te nadzirati sukladnost pojedinačnih projekata s Okvirnim programom, kao odobrenim programom državnih potpora. Pojedinačne projekte koji će slijediti iz Okvirnog programa u većini slučajeva će voditi tijela javne vlasti na lokalnoj ili područnoj (regionalnoj) razini (općine, gradovi i županije) kao nositelji pojedinačnih projekata (NP-ovi). Osim toga, uz koordinativnu ulogu NOP-a, u ključnim provedbenim aktivnostima projekata predviđena je i podrška HAKOM-a, kao nacionalnog regulatornog tijela za sektor elektroničkih komunikacija.

V. METODOLOGIJA UNIFICIRANOG PRISTUPA

Koristeći svoja značajna iskustva u procesu izgradnje komunikacijskih sustava odnosno mrežne infrastrukture, te dobro definirane poslovne procese Ericsson predlaže „industrijalizaciju“ procesa kreiranja projekata i aplikacija za EU fondove.

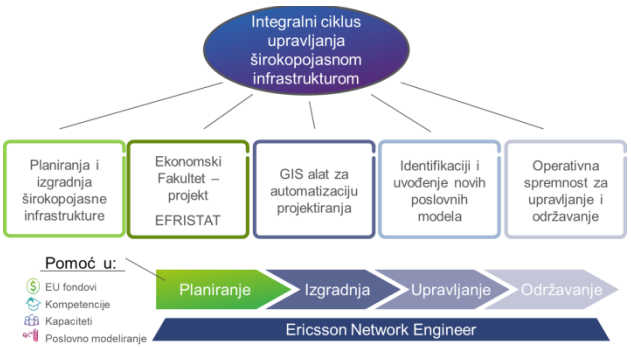
Predloženo rješenje ne prejudicira konkretan alat ili kompaniju nego praktično vizualizira prednosti unificiranog pristupa na projektiranju i upravljanju mrežama, te pripremi natječajne dokumentacije. Posve je jasno da odabir konkretnog tehničkog rješenja za predloženi sustav (softveri i implementacija) u cijelosti mora biti odrađen u skladu sa EU pravilima i pravilima javne nabave.

Analizom problema i prepreka koje smo ranije adresirali nužno je JLS pomoći u sljedećim segmentima:

- Kompetencija – za planiranje i dizajn mreža
- Kapacitet – za izgradnju i održavanje
- Poslovno modeliranje – za odabir optimalnog modela korištenja izgrađene infrastrukture
- Prikupljanje i održavanje potrebnih podataka

Osim toga nužno je predložiti rješenja koja će biti repetitivnog tipa odnosno replikabilna za sve JLS bez obzira na veličinu.

Konceptualno rješenje uz primjenu komercijalnih proizvoda je prikazano na Slika 2. Dvije su ključne pretpostavke – unifikacija ulaznih podataka uz pomoć sustava EFRISTAT Ekonomskog fakulteta u Rijeci [9] te korištenje Ericsson Network Engineera kao sustava za projektiranje i optimizaciju svjetlovodne infrastrukture, čime će se realizirati unifikacija procesa na razini svih jedinica lokalne samouprave i dobiti konzistentni rezultati.



Slika 2. Uvođenje „industrijskog“ principa u koncept upravljanja mrežnom infrastrukturom kao funkcija primijenjenog polja.

Projekt EFRISTAT je baza podataka gradova i općina u RH koja objedinjuje znanstvene te stručne potencijale domaćeg gospodarstva te unaprjeđuje upravljanje lokalnim javnim sektorom. Osnovnu podlogu predstavlja formiranje baze podataka, i izračuni pokazatelja o gradovima i općinama, koja sadržava sljedeće:

- ekonomski podaci
- fiskalni podaci (kategorije prihoda, kategorije rashoda, zaduženost, pokazatelji efikasnosti i djelotvornosti)
- socijalni podaci (siromaštvo, demografski podaci, kriminalitet)
- organizacijsko-administrativni podaci (kvaliteta administracije, podrška lokalnom poduzetništvu i sl.)
- svi ostali podaci i pokazatelji relevantni za lokalne jedinice.

Svrha ovakvog tipa baze podataka je potencijalnim ulagačima osigurati detaljan uvid u mogućnosti ulaganja unutar pojedinih gradova i općina, a jedinicama lokalne uprave i samouprave podršku za učinkovitije upravljanje vlastitim ekonomskim razvojem.

Korištenjem EFRISTAT-a uz tehnološka rješenja, također je moguće ubrzati proces izrada predstudija o izvedivosti, koja se izrađuje tijekom faze identificiranja, osigurava utvrđivanje svih problema te ocjenu svih mogućih rješenja, kao i toga da je odabrano najbolje rješenje na temelju faktora kvalitete. Studijom i Europska

komisija i zemlja korisnica dobivaju dovoljno informacija na kojima će se temeljiti njihovo prihvaćanje, izmjena ili odbijanje predloženog projekta.

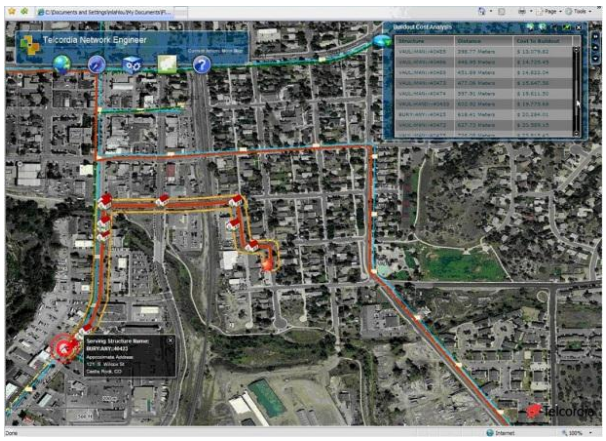
Iz EFRISTAT-a se pripremaju ulazni podaci za Ericsson Network Engineer odnosno za proces projektiranja i optimizacija. Podaci uključuju generalne informacije o korištenju prostora u JLS (radne zone, turistička područja) te detaljne informacije o pozicijama tzv. Demand Pointova odnosno točaka koje je potrebno povezati na svjetlovodnu infrastrukturu. Osim rezidencijalnih korisnika u fokusu su sljedeće lokacije od posebnog interesa:

- Objekti samih JLS
- Obrazovnih i odgojnih ustanova (fakulteti, škole, vrtići, domovi)
- Knjižnice svih razina
- Bolnice, ambulante, privatne ordinacije
- Vatrogasci, policijske postaje
- Komunalna društva
- Lučke uprave, kapetanije
- Nadzor prometa

Nakon odabira tehnologije pristupa se izradi idejnog rješenja u Ericsson Network Engineeru i inicijalnog troškovnika potrebnog za budžetiranje i izradu studija isplativosti.

VI. JAVNA DOSTUPNOST I TRANSPARENTNOST IDEJNIH RJEŠENJA NAPRAVLJNEIH ZA POTREBE JLS

Kao logičan nastavak prethodnih aktivnosti i primjer praktične implementacije koncepta je i računarstvo u oblaku u kontekstu aplikacija e-uprave putem koje će sve do sada navedene aktivnosti biti transparentno prezentirane široj društvenoj zajednici i na raspolaganju širem krugu potencijalnih korisnika.



Slika 3. WEB prikaz podataka JLS

Kako su korištene GIS tehnologije već odavno sposobne prikazivati rezultate na WEB-u (primjer Slika 3.) a danas je koncept pružanja informacija kroz oblak postao prevladavajuća forma i poželjan način konzumacije ponuđenih informacija. Prednost rješenja u oblaku je što se sustav može ponuditi većem broju raznih korisnika u vrlo kratkom vremenu i na konzistentan način čime se optimiziraju troškovi i na strani davatelja usluge i na strani korisnika.

JLS prije svega ne moraju imati vlastitu opremu i stručne kadrove za održavanje aplikacija – oni se pojavljuju samo kao korisnici sustava. Ukoliko sve JLS koriste isti sustav omogućeno je vrlo jednostavno i koherentno praćenje svih projekata, a posebice važno – uvođenje ključnih indikatora efikasnosti korištenja sredstava iz EU te praćenje napretka i izvođenja projekata.

VII. ZAKLJUČAK

Iskorak državne uprave, županija i JLS, a pogotovo u smislu uvođenja vrhunskih OSS tehnologija, iz tradicionalnog telekom segmenta u domenu podrške JLS predstavlja značajnu pomoć lokalnoj upravi i samoupravi na putu ka efikasnijem korištenju sredstava iz strukturnih fondova EU te ne predstavlja „nemoguću misiju“. U kriznim vremenima kakva su danas posebno je važno za cijelu društvenu zajednicu korištenje ekonomski optimalnih tehničkih rješenja i alata, pri čemu se pokazuje da najmodernije OSS tehnologije, ako su pravilno koncepcijski iskorištene i implementirane, ne moraju biti nedostupne čak i vrlo malim korisnicima koji nemaju posebno domensko znanje.

Rješenje je u korištenju jedinstvenog procesa i zajedničke baze projekata u geografskom informacijskom sustavu dostupne svim dionicima i u svim fazama razvoja, od idejnih do realiziranih rješenja.

LITERATURA

[1] Željko Popović, Izgradnja digitalnih gradova, ENT Revija 2008 http://www.ericsson.com/hr/etk/revija/Br_1_2008/digitalni_grad.pdf

[2] Delivering next generation access through PPP, European PPP expertise centre www.eib.org/epec/resources/epec_broadband_en.pdf

[3] Guide to broadband investment, EU Regional Policy, 2011 http://ec.europa.eu/regional_policy/sources/docgener/presenta/broadband2011/broadband2011_en.pdf

[4] Europa 2020, http://ec.europa.eu/europe2020/index_hr.htm

[5] Digital Agende for Europe, https://ec.europa.eu/digital-agenda/en

[6] Strategija razvoja širokopojasnog pristupa u Republici Hrvatskoj u razdoblju od 2012. do 2015. Godine, http://www.mppi.hr/default.aspx?id=8977

[7] Okvirni nacionali program za razvoj infrastrukture širokopojasnog pristupa u područjima u kojima ne postoji dostatan komercijalni interes za ulaganja, http://www.mppi.hr/default.aspx?id=10457

[8] EU Guidelines for the application of State aid rules in relation to the rapid deployment of broadband networks, Official Journal of the European Union, 26.1.2013

[9] EFRISTAT, Ekonomski fakultet Rijeka i CLER, http://www.cler.hr/

[10] Jed Kolko, Does broadband Boost Local Economic Development, PPIC, January 2010 www.ppic.org/content/pubs/report/r_110jkr.pdf

[11] Erik Bohlin, The economic impact of broadband speed, Columbia University, October 14, 2011

[12] Indrek Petersoo, Life in E-stonia, EBR #2 2012, www.e-stonia.com

[13] Josip Brumec, Javna informacijska infrastruktura, Fakultet organizacije i informatike Varaždin http://www.mingo.hr/userdocsimages/trgovina/Brumec_Javna_informacijska_infrastruktura.pdf

QUALITY MANAGEMENT

Real Time Assuring QoE in the Lowest OSI/ISO Layers During Delivering of IPTV Services

N. Goran*, M. Hadžialić ** and A. Begović*
* BH Telecom JSC, Sarajevo, Bosnia and Herzegovina
** University of Sarajevo, Faculty of Electrical Engineering, Bosnia and Herzegovina
nermin.goran@bhtelecom.ba; mesud.hadzialic@etf.unsa.ba; alen.begovic@bhtelecom.ba

Abstract - this paper compares the values of data link layer QoS (Quality of Service) indicators of several access lines, which can serve as a measure of the physical link quality or QoPh (Quality of Physical Layer), in real environment without reference video signal (No Reference), during the delivering of IPTV (Internet Protocol Television) multicast services to multiple users. It also focuses on the fact that an increase in channel capacity reduces the number of signal degradation indicators in the data link layer, because it increases the resistance of useful information on the interference. The paper proves the fact that at the same time, improving the physical layer conditions, the values of network QoS indicators, such as packet loss or delay, is reduced which affects the final grade at the end-user QoE (Quality of Experience) of IPTV (Internet Protocol Television) services.

***Keywords* — Access network; Quality of Service; Quality of Experience; Quality of Physical Layer; Channel estimation;**

I. INTRODUCTION AND RELATED WORK

Functional dependency between QoS (Quality of Service) parameters and QoE (Quality of Experience) is a challenge while defining the simplest system of prediction of a service QoE level evaluation with no reference original signal. The authors consider analytical dependence in the [1] and [2], and emphasize that MOS is a typical user oriented evaluation of any service which can be determined with a subjective evaluation given by individual users or with an objective measure of service features delivered along with the service. They also claim that the appearance of basic QoS indicators, such as packet loss, delays, delay variations, wrong packet order and traffic limits, implies the appearance of problems in application layer, e.g. the appearance of video artifacts, different video signal degradations or in some cases, waiting for a service, which is reflected to the awaited QoE level. They have discovered that the relation between QoE and QoS parameters can be presented as certain function dependence. The relating function is called IQX (exponential interdependency of quality of experience and quality of service) hypothesis and has been successfully tested on QoS network parameters. The authors in [3] have developed a slightly different correlation model of QoS and QoE which numerically estimates IPTV (Internet Protocol Television) QoE with usage of survey results. They have classified the traffic towards IP QoS classes [4] and determined how big is the impact of IP QoS parameters on IPTV service for every class. They have used weighting factors to estimate total QoS, and then, based on the modified IQX hypothesis, the QoE values have been obtained. The authors in [5] also claim that the main question to be solved at quality QoS/QoE correlation is deep packet inspection. The authors in [6] have examined the layer QoE through defining integral relation between QoE and observed network and application layer parameters, not as independent entity. This relation is defined as the function dependent on a joint factor AQoS (Application QoS*)* and NQoS (Network QoS) in a new QoE evaluation frame. AQoS parameters, especially during the delivery of video content, have also been considered by the authors in [7-9]. Testing of quality of experience of IPTV service in relation to NQoS parameters as well as different types of testing scenarios are presented in [10]. The authors in [11] see the solution to this problem in modified design of QoS, while on the other hand [12] and [13] present cross-layer design solutions which, from physical layer of the media and QoS demands of certain applications, take into account rate calculations, strength, physical media coding and their adjustments to application demands at a given channel and given network conditions. The authors in [14] see the solution to this gap towards optimization of the relation between QoS and QoE in correlating the physical layer with QoS parameters of the application layer. They used VoIP over LTE technology as an experimental service, and in their work they correlated parameters of the first and second layer of LTE radio configuration with IP performance parameters. The connection with QoE has been realized by MOS results with the use of PESQ algorithm. However, the authors haven't given a unique analytical solution for the mentioned relation and they also haven't put the correlation between physical layer parameters and the QoE level evaluation. In the report [15], certain demands have been formed for IPTV service delivered over DSL in a form of technological improvements based on DSM, vectoring, improved protection from impulse noise, SRA, AL-FEC and measurements. The authors in [16] have worked on the influence of impulse noise on QoS parameters as well as appearance of error seconds caused by impulse noise. Although end-to-end QoE in ITU recommendation implies monitoring these values to the user, very few authors, e.g. in [16] and [17], deal with home and access networks. In other words, the authors emphasized that in order to fulfil the above mentioned demands, it is necessary to observe each technological segment that contributes proper QoE evaluation. Therefore models that correlate these segments have great importance in end-to-

end QoS insurance. Considering a particular technology and some type of protective coding, the physical link quality or QoPh can be discussed. The basis of this paper is the novel method and analyzing the physical layer influence on a service quality performance and quality of experience of a service delivered to the user, at the same time. Technology, as well as usage of certain type of media for information transfer, is being developed, but in order to consider service quality at access network domain, it is highly important to consider the physical channel which connects the user with the first point in telecommunication system. The evaluation of total QoS is emphasized in this paper, but the focus is on the correlation between parameters of physical layer and appearance of packet loss at network layer as a QoS indicator and as a basic parameter of QoS/QoE correlation.

II. MATHEMATICAL OBSERVATIONS

In order to determine quantitative influence of physical channel on quality of service delivery in access network domain i.e. from the customer to the first point of a telecommunication system it is necessary to define and discover physical layer parameters as well as data link layer parameters that represent reflection of occurrences in the channel. This is a very important determination of total customer QoE estimation, because after this it is possible to estimate delivered QoE, indirectly with measurable QoS indicators, now available from the calculated electrical channel characteristics. The mentioned indicators relating to jitter, delay and packet loss are changeable depending on the channel conditions, therefore there is the function:

$$\mathrm{QoE} = f_1\left(\mathrm{AQoS}, \mathrm{NQoS}, \mathrm{DQoS}, \mathrm{QoPh}\right) \qquad (1)$$

Not one system has been defined which can completely control the QoS parameters for service delivery. Therefore it is necessary to provide an overview of deep packet inspection of bottom up OSI/ISO layers. In other words, the solution must involve estimation of occurrences in the last mile which is, in its biggest part, subject to disturbances of different sources in form of stationary and non-stationary noise, so it is necessary to identify the parameters which have to be modelled with higher OSI/ISO layers, in order to provide more precise QoE estimation. Quality of IPTV service delivery to end users depends on timing of transport stream that passes through numerous routers, compression technics in higher OSI/ISO layers and psychological and psychophysical possibilities of users. However, we want to discover and consider influence of the last mile on QoE. That can be done only by considering optimal line rate on physical layer that can be accomplished only with good channel conditions. Also considering appearance of bit errors, caused by stationary and non-stationary noise especially in OFDM channel, and error control system can help in quality evaluation of physical layer [18]. Undoubtedly, while evaluation the quality of physical layer, it is necessary to consider the relation between signal and noise (SNR), needed energy or bit power, bandwidth and transfer function of observed channel. It is obvious from all the things mentioned that, considering the above parameters, it is necessary to accomplish the maximization of needed line bit rate of a service. Besides that, the choice of error-correcting code defines the quality of packet loss correction. During the transmission of DMT signal in DSL systems, channel capacity can be presented by the following equation:

$$Rc = \frac{1}{2N}\log_2\left(\prod_{n=1}^{N}\left[1+\frac{SNR_n}{\Gamma \bullet mar_n}\right]\right) \qquad (2)$$

Main significance of this formula is that it provides a high-level way of thinking about how SNR as the performance of a communication system depends on the basic resources available in the channel, without going into the details of specific modulation and coding schemes used. Maximum bit rate can be accomplished by minimization of average SNR, because low values of that relation generate appearance of symbol errors. At the same time low SNR values lead to reduction of SNR margin that is necessary to provide symbol error probability less than 10e-6. Error probability rises if SNR margin falls below 6 dB. If there is bigger depth of SNR margin variation, increase of symbol errors will occur. Error control system will not succeed in recovering all wrong bits and that will lead to indicators appearance of that phenomenon in form of CRC (Cyclic Redundancy Check) bits. Every code violation (CV) or CRC in DSL super frames will cause packet loss in higher OSI/ISO layers. At the same time, system will try to solve the originated problem with increasing of interleaving, but that will raise delay. From all above mentioned, we can conclude that there is functional dependency between indicator of QoS and SNR or SNR margin, i.e. we can write:

$$\mathrm{QoS}_{pack_loss} = f_2\left(mar, \mathrm{CV}\right) \qquad (3)$$

In case of extremely low SNR margin, the coded system cannot improve error probability values. In case of Gaussian noise, the coded system can achieve maximum value of channel capacity, but it cannot if signal fades with changing of coupled transfer function, especially in long lines. Therefore, if huge variations of SNR margin appear, it can cause typical bit error when we use coded systems as well as uncoded systems. One major challenge for researchers is observing IPTV delivery in real time and timeframe for consuming that service from customer side. Considering [15], existing of an error second in one hour of IPTV service consummation can be allowed for ordinary users. An error second will appear only if one or more code violations appear. The number of error seconds increases with the number of code violations. This phenomenon depends on depth of negative changes of SNR or SNR margin as well as actual Ra and maximum possible rate in channel Rc. So, there is a function:

$$ES = f_3\left(mar, \mathrm{CV}, \mathrm{R}_a / \mathrm{R}_c\right) \quad (4)$$

These violations, in DSL case, are correlated to packet loss, because they appear in superframes. Every 68 DSL frame has one CRC indicator. That superframe carries on IP packets that contain MPEG-2 TS packets. Depending on actual rate and transferred service, that super frame can carry on 10 to 20 IP packets and at the same time 70 to 140 video packets. This can be a linear indicator of quantitative measure of disturbances in physical channel. It also introduces a reflex of data link layer when SNR margin changes dramatically. As we mentioned earlier, packet loss, especially in UDP cases, can indicate one important QoS parameter that can help in QoE evaluation. The most known correlation that defines dependency of QoE from QoS is IQX hypothesis. It is exponential interdependency of quality of experience and quality of service and it can be depicted as:

$$\mathrm{QoE} = \alpha \cdot e^{-\beta \cdot \mathrm{QoS}} + \gamma \quad (5)$$

Authors have got parameters α, β and γ, by nonlinear regression analysis and they also have been arguing good correlation with real values. This correlation was being checked, as we say in the introduction, for different network cases as testing of audio, Web, video and data packets. Considering packet loss, IQX hypothesis in network layer can be shown as follows:

$$\mathrm{QoE} = 3,01 \cdot e^{-4,473 \cdot \text{packet loss}} + 1,065 \quad (6)$$

As there is linear interdependency between appearance of CRC and packet loss, IQX hypothesis can be identified in data link layer. In that case next equation is given:

$$\mathrm{QoE} = 3,01 \cdot e^{-\frac{\mathrm{CV}}{\mathrm{CV}_{tot}} \cdot t} + 1,065 \quad (7)$$

It can be noticed, considering the quantity of CRC, that QoE level can be evaluated in real time, but only in DSL domain, i.e. in access network domain. Certainly, this is not total QoE value because packets travel through a huge number of routers as well as different head-end and super head-end systems.

III. Test Environments

In order to validate the above equations, we have chosen an experimental method in which the IPTV service delivery is observed in real time on three independent DSL lines. SDTV IP service has been delivered via MPEG-2 transport stream, UDP transport protocol, IP network protocol, ATM and ADSL2+ data link and physical layers. Three slightly longer lines have been disturbed by stationary and non-stationary noise and their values of SNR and maximum attainable rate have been compared. Stationary noise was caused by neighboring DSL lines in L0 state, while non-stationary noise was generated by neighboring DSL lines which changed their states from L0 to L2. We measured SNR margin, actual rate and attainable rate in physical layer. In data link layer, we also observed the appearance of CRC code violation of DSL super frames. In network layer, packet loss was observed and correlated with all the mentioned data.

IV. Results

Overview of the results of the mentioned measurements, from test environment is given in the Table 1. SNR_margin determines average variability of SNR margin of all subcarriers that caused the code violation appearance; CV determines the number of code violations and CV_{tot} total number of potential CVs. At the same time, IQX hypothesis determines the value obtained in the equation result (6) and IQX_DLL the value of equation (7). In order to compare the obtained results, it is necessary to construct a graph presented in Figure 1

TABLE I. Measured Values

SNR_margin	3,26	2,9	2,1
CV	249	449	4711
CV_{tot}	211764,7	211764,7	211764,7
$\mathrm{CV}/\mathrm{CV}_{tot}$	0,0012	0,0021	0,0222
Packet loss	0,12%	0,21%	2,24%
IQX hypothesis	3,73	3,50	1,39
IQX_DLL	3,56	2,08	1,065

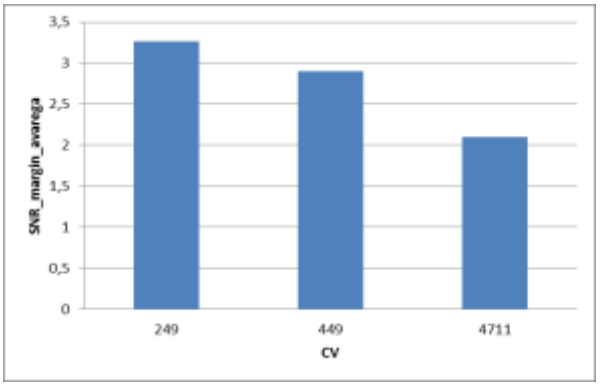


Figure 1. Relation between SNR_margin and CV

which determines the relation between the appearance of code violation and average noise margin in observed period. Each appearance of the error caused a packet loss with an average value of given in Table 1. Figure 2 presents the relation between the appearance of average packet loss and QoE calculated by IQX hypothesis. The relation between measured value CRC (i.e. CV) and total number of potential errors (i.e. in every superframe there are code violations) is presented with $\mathrm{CV}/\mathrm{CV}_{tot}$. Figure 3 presents the same relation and anticipated QoE calculated by modified IQX hypothesis in data link layer. If Figure 2 and 3 are compared, it can be confirmed that, with the help of modeling of physical and data link layer

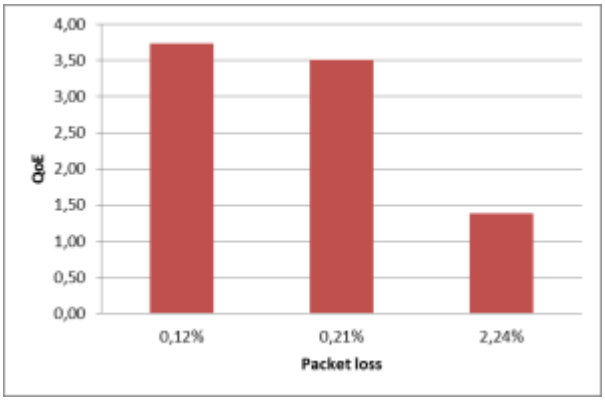


Figure 2. QoE dependency on packet loss

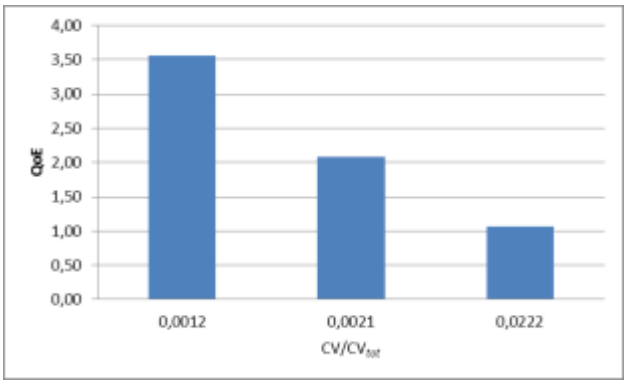


Figure 3. QoE dependency on CV/CV$_{tot}$

parameters, it is possible to evaluate QoE in domain of access network more precisely than in the case of IQX hypothesis in network layer.

V. CONCLUSION

Without considering physical layer, we cannot get complete and accurate model that can serve for QoE evaluation process, especially, if total QoE has to be evaluated in E2E (end-to-end) solutions. In this case, access network domain, from user terminal equipment to operator's CO, cannot be avoided. Channel capacity, SNR, bit power, modulation and coding have huge importance in defining possible rate in access network that is needed for a service. Variations of SNR and changing of actual and maximum possible rate relation can be used for evaluation of stability of conditions that is necessary for IPTV service delivery. In this paper we have described the way for evaluation that can show the amount of influence of physical layer parameters. These results can be used by a monitoring system for decreasing necessary time for maintenance of particular lines. For future work we plan to get precise function between all parameters of physical and data link layer needed for determination of bit error propagation through OSI/ISO layers.

REFERENCES

[1] Hoßfeld, M. Fiedler, Th. Zinner, "The QoE Provisioning-Delivery-Hysteresis and Its Importance for Service Provisioning in the Future Internet", in Proceedings of the 7th Conference on Next Generation Internet Networks (NGI), Kaiserslautern, Germany, Jun. 2011, pp. 1-8

[2] Markus Fiedler, Tobias Hoßfeld and Phuoc Tran-Gia, "A Generic Quantitative Relationship between Quality of Experience and Quality of Service", in IEEE Network, vol. 24, issue 2, pp 36-41, April 2010

[3] H. J. Kim and S. G. Choi, "A study on a QoS/QoE correlation model for QoE evaluation on IPTV service," in Proceedings of the 12th International Conference on Advanced Communication Technology, Phoenix Park, Korea, Feb. 2010, pp. 1377–1382

[4] Y Maeda, "QoS standards for ip-based networks ", in IEEE Communications Magazine, vol.41, no. 6, pp. 80, Jun. 2003

[5] R. Schatz, T. Hoßfeld, L. Janowski, S. Egger, "From Packets to People: Quality of Experience as a New Measurement Challenge", in Data Traffic Monitoring and Analysis, Lecture Notes in Computer Science, vol. 7754, pp. 219-263, 2013

[6] M. Siller, J.C. Woods, "Improving quality of experience for multimedia services by QoS arbitration on a QoE framework", in Proceeding of International Conference on Visual Information Engineering (VIE 2003), Jul 2003, pp. 238-241

[7] Mu Mu, A. Mauthe, F. Garcia, "A Discrete perceptual impact evaluation quality assessment framework for IPTV services" in Proceedings of the 2010 IEEE International Conference on Multimedia and Expo (ICME), Suntec City, Singapore, Jul. 2010, pp. 1505-1510

[8] D.Collange, J.L. Costeux, "Passive Estimation of Quality of Experience", in Journal of Universal Computer Science, vol. 14, no. 5, pp. 625-641, 2008

[9] C.E. Vegiris, K.A. Avdelidis, C.A. Dimoulas, G.V. Papanikolau, "Live Broadcasting of High Definition Audiovisual Content Using HDTV over Broadband IP Networks", in International Journal of Digital Multimedia Broadcasting, vol. 2008, Article ID 250654, 18 pages

[10] P. Calyam, P. Chandrasekaran, G. Trueb, N. Howes, R. Ramnath, D. Yu, Y. Liu, L. Xiong, D. Yang, "Multi-Resolution Multimedia QoE Models for IPTV Applications", in International Journal of Digital Multimedia Broadcasting, vol. 2012, Article ID 904072, 13 pages, Apr. 2011

[11] Mu Mu, A. Mauthe, F. Garcia, "A Utility-Based QoS Model for Emerging Multimedia Applications" in Proceedings of the Second International Conference on the Next Generation Mobile Applications, Services and Technologies (NGMAST '08), Cardiff, UK, Sept. 2008, pp. 521-528

[12] G. Gardikis, L. Boula, G. Xilouris, A. Kourtis, E. Pallis, M. Sidibe, D. Negru, "Cross-layer monitoring in IPTV networks", in IEEE Communications Magazine, vol. 50, no. 7, pp. 76-84, Jul. 2012

[13] J. Sen, S. Bhattacharya, "A Survey on Cross-Layer Design Frameworks for Multimedia Applications over Wireless Networks", in International Journal of Computer Science and Information Technology, vol. 1, no. 1, pp. 29-42, Jun. 2008,

[14] F. Javier Rivas, A. Díaz, P. Merino, "Obtaining More Realistic Cross-Layer QoS Measurements: A VoIP over LTE Use Case", in Journal of Computer Networks and Communications, vol. 2013, Article ID 405858, 10 pages, 2013

[15] Broadband forum, Technical Report MR-180, "Achieving Quality IPTV over DSL", 2012

[16] J. A. Andersson, M. Kihl and S. Höst, "Impact of DSL link impairments on higher layer QoS parameters", in Proceedings of 8th Swedish National Computer Networking Workshop, Stochholm, Sweden, June 2012, pp. 95-98,

[17] C. Develder, P. Lambert, W. Van Lancker, S. Moens, R. Van de Walle, J. Nelis, D. Verslype, S. Latré, N. Staelens, N. Vercammen, B. Vermeulen, B. Masschelein, T. Van Leeuwen, J.-F. Macq, K. Struyve, F. De Turck, B. Dhoedt, "Delivering scalable video with QoS to the home", in Telecommunication Systems, vol. 49, no. 1, pp 129-148, Jan. 2012,

[18] N. Goran, M. Hadzialic, A. Begovic, "An impact of access network physical impairments on IPTV QoE in DSL environment", in Proceedings of 36th MIPRO, Opatija, Croatia, May 2013, pp. 380-383

Measuring XML schema standard quality change through versioning

Maja Pušnik*, Marjan Heričko* and Boštjan Šumak*
* University of Maribor, Faculty of electrical engineering and computer science, Institute of informatics, Smetanova ulica 17, SI-2000 Maribor, Slovenia
maja.pusnik@um.si, marjan.hericko@um.si, bostjan.sumak@um.si

Abstract - Quality measurements of XML schemas are difficult to perform objectively, providing all aspects of quality demands. However, an approximate evaluation of XML schema's quality and validation, whether a higher version of an XML schema standard is in fact an improvement, is useful information for XML schema users. The paper addresses who are the users of XML schemas, what aspects of XML schemas are vital to its quality and how to measure those aspects. We present 3 quality aspects in more detail: (1) structure type, (2) re-use possibilities and (3) balance of building blocks. A table of all included variables, that were measured, is provided. Based on results of over 200 XML schemas, quality assessment are given, including data standardization methods. Measurement results of XML schema's versions are compared, providing insight into quality growth (or decline) through versioning.

I. INTRODUCTION

Quality measurements of XML schemas are difficult to perform objectively, providing all aspects of quality demands. Since the XML schemas constantly change, the quality has to be consistent through newer versions, preferably higher than the quality of the previous XML schema.

The primary role of XML schemas is the definition of XML data and supporting rules regarding the use of constantly changing XML data, an important part of information technologies. XML schemas and related technologies present an important part of IT solutions [2]. Using XML has spread from the field of e-business and data exchange to data presentation into various levels of contemporary information solution architectures. The quality of XML schema is visible within its structure and also in the contents of elements and attributes.

By definition, XML schemas are extensible, flexible and reusable in all kinds of environments, representing a foundation to the complex infrastructure of e-business, which is concealed to the end-user. If done properly, their quality reflects on time and cost reduction, creating lower-cost services, improving user experience, creating more flexibility, shorter time to market and many other features, which play an important role in enabling high performance of the end product, operating with the XML document and consecutively the XML schema [8][9].

Due to the widespread use of XML schemas and its constant change, the question of its quality is often open, particularly from the aspect of structure (and content) of XML schemas, which indirectly influence the quality of data that XML schema describes. Therefore measuring XML schemas quality as it is growing is one of the basic research challenges in our paper.

TABLE I. VARIABLES WITHIN METRICS

N	*Abbreviation*	*Explanation*
1	NE	number of elements
2	Nat	number of all attributes
3	Nt	number of all data types
4	Nr	number of restrictions
5	Ng	number of all groups
6	Nan	XML schema document size
7	Nt_s	number of simple data types
8	Nt_c	number of complex data types
9	Nt_i	number of derived data types
10	NE_U	number of unbounded elements
11	Nre_all	number of references on elements
12	Nra_all	number of attribute references
13	Nrt_all	number of all used data types
14	Nrg_all	number of group references
15	Nri_all	number of external XML schemas
16	LOC	lines of code

This paper is organized in 5 chapters. The first two introduce the research area and related work, exposing the lack of research in the field of measuring quality of XML schemas and its versions. The third chapter presents the research model and explains the definition of XML schema quality aspects, as well as a set of three hypotheses. The fourth chapter explains the quality change through versioning. Conclusions and future work are given in chapter five.

II. LITERATURE REVIEW

There have been several attempts to evaluate and measure XML schemas. Few of them are summed in [1]; however none of them includes measurements of the same XML schema through versions. Significantly related work was also done in [4] and [6], where attempts to measure XML schemas and software in general were made. The

subject was addressed in other papers, not included in this overview, however the background are mainly software metrics, which do not necessary always apply to needs of XML schema quality (and complexity) measurements.

Dynamic creation and adaptation of XML schemas schedules and presents an additional research challenge that requires the use of new approaches and solutions, universal and specific according to a domain. The aim of this paper is definition of an approach for evaluating the quality of XML Schema, based on a new set of metrics and comparing its results through versions. The formulas are derived from research work made in [11].

III. RESEARCH MODEL

An evaluation of XML schema's quality and validation, whether a higher version of an XML schema standard is in fact an improvement, is very useful information for XML schema users. The paper addresses what aspects of XML schemas are vital to its quality and how to measure those aspects.

We present 3 quality aspects in more detail: (1) structure type, (2) re-use possibilities and (3) balance of building blocks. A table of all included variables, that were measured, is provided in TABLE I. Based on results of over 200 XML schemas, quality assessment are given, including data standardization methods.

A. *Quality aspects*

The results of a systematic review of literature in the field of measuring XML schemas showed that several metrics were applied to XML schema evaluation, extracted mainly from the methods of software engineering measurements, focusing mostly on the complexity of XML Schemas. To include a variety of parameters addressing complexity and quality, we searched different fields on quality measurement and defined the following three, shown in Figure 1.

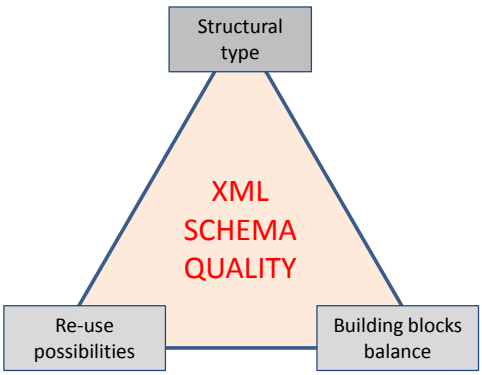


Figure 1. Quality aspects of XML schemas

Structure type: Other authors have researched measuring the structure of XML schemes for calculating the complexity and quality in [4] and [5]. The authors present a number of metrics, taken mainly from "quality model" ISO standard and link them into a single formula. Each variable is further multiplied, however the factors are not justified, values are not on the same scale, so the formula cannot be applied, but we have analyzed and partly used it in our calculation formula of quality. Within the complexity calculations a conclusion can be made that the higher the value of the individual variable, the greater the complexity. According to XML schema needs we redefined metrics into the following composite metric (1) with the following parameters:

$$Q1 = N_{ri_all} + \frac{N_E}{N_{at}} + \frac{N_r}{N_{t_s}} + \frac{N_{t_s}}{N_{t_c}} + \frac{N_{An} + N_{t_i} + N_{E_U}}{N_E} \quad (1)$$

Re-use possibilities: The equation was inspired by [7]; summing up existing results and defining a set of metrics for measuring re-use of software. The metric includes parameters that allow the re-use and are inherently global (2).

$$Q2 = \frac{N_{re_{all}} + N_{ra_all} + N_{rg_all} + N_{ri_all} + N_{rt_all} + N_{t_i}}{N_E + N_{At} + N_g + N_t} \quad (2)$$

Balance of building blocks: In this metric we combine the parameters indicating the minimum XML schemas building blocks, where the concept of minimalism is defined as the level, where one can anticipate that there is no other set of less building blocks, however still descriptive full (3).

$$Q_3 = \frac{N_{An} + N_E + N_{At}}{LOC} + \frac{N_{rt_all}}{N_t} \quad (3)$$

The XML schema **quality index** combined all three aspects equally (4). Each included XML schema was evaluated individually by each metric and a combined quality index was calculated. Versions of the same XML schemas were compared.

$$Q_i = \frac{1}{3}(Q_1 + Q_2 + Q_3) \quad (4)$$

The proposed parameters and metrics are relatively non-complicated to obtain and calculate, however no such tool was found available within our support tools survey. Therefore a personalized web application "XML schema metrics" was implemented (under supervision of Institute of informatics Maribor), measuring and reporting all vital XML schema variables, enabling a much faster gathering of data, hindering hand-made mistakes.

B. *Hypothesis*

In this study, we addressed the following research question: The newer version of XML schemas is of higher quality as the previous one. We addressed the research question with the following hypothesis:

H1 –The quality of structural view is higher in case of a higher version.

H2 – The quality of re-use possibilities is higher in case of a higher version.

H3 - The quality of building blocks balance is higher in case of a higher version.

C. *Standardization methods*

The calculated parameter values included several value spans, from zero to over 1000. So that individual

metrics could be compared, standardization of variables was conducted. All the variable results that were used within metrics were transformed to a scale of 0 to 1, where 0 represented the least desirable value for each parameter and 1 the most desirable value. The transformations were based on linear programming, assuming that the growth relationship is without exception linear. Each of the parameters underwent a resize process with the help of one of the three scaling techniques (Figure 2, Figure 3, Figure 4).

If the most desirable outcome of the parameter is as maximal as possible, the maximal value holds the estimation 1 and it linearly drops to 0 (Figure 2).

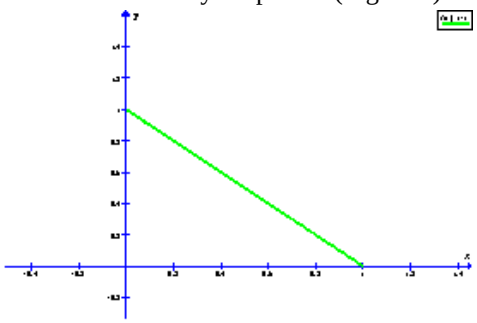


Figure 2. Maximum value best option

If the most desirable outcome of the parameter is as minimal as possible, the minimal value holds the estimation 1 and it linear goes to 0 (Figure 3).

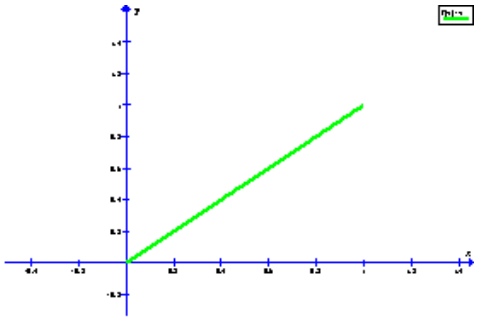


Figure 3. Minimum value best option

If the most desirable outcome is of average value then the average value holds 1; the extreme minimal and maximal values hold value 0 (Figure 4).

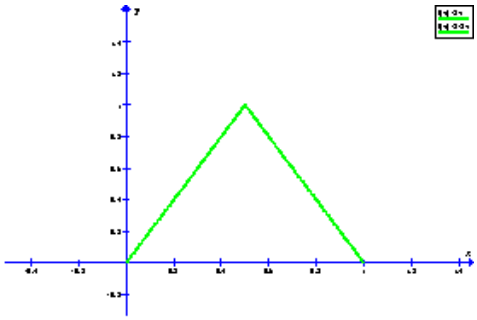


Figure 4. Average value best option

IV. QUALITY CHANGE THROUGH VERSIONING

Measurement results of XML schema's versions were compared, providing insight into quality growth (or decline) through versioning. We overviewed several types of versioning notation in order to compare the right XML schemas.

A. *Definition of versions*

In [10] the author explains the possible XML schema versioning techniques: (1) changing the internal schema version attribute, (2) changing the schema's target namespace, (3) changing the name of the schema or (4) changing the location of the schema. Each technique has its advantages and disadvantages:

- **Changing the schema version attribute**; although an easy task the validator often ignores the attribute field and can be used only as a cue to the human user (Technique 1).
- **Changing the schema's target namespace** demands a suitable change in instance documents, which will not validate until they are changed to designate new target namespace, attribute, even if the change to the schema would not impact the instance. The problems occur with included XML schemas (Technique 2).
- **Changing the (file) name of xml schema;** is the most often case of versioning, however additionally demanding a change in instance documents, facing with identical problems as in technique 2 (Technique 3).
- **Changing the location of the schema** is used in conjunction with case 3 (Technique 4).

XML schemas, included in our research, mostly used technique 3, in combination with technique 1 or technique 2.

B. *Comparission of XML schema versions*

The formulas were tested on 250 XML schemas; regression analysis was conducted in order to verify significance of included variables [11]. However, comparing versions of included XML schemas was reduced to 10% of all XML schemas, due to the lack of availability of more than one version of XML schemas

All three quality aspects as well as joint quality index were compared among the found XML schema pairs. In average, the newer version improved the structural aspect for 2%. The aspect of re-use was reduced for 1% and optimality of building blocks was improved by 2%. In average a newer version of XML schema was improved by no more than 1%.

C. *Hypothesis answers*

Although the average quality of XML schemas with newer version was higher (Figure 5), all hypotheses could not be confirmed.

H1 – The quality of structural view was in fact averagely higher by 2% in some cases even as high as 21%. Hypothesis ACCEPTED.

H2 - The quality of re-use possibilities is higher in case of a higher version was averagely not correct. Averagely the newer version had a lower quality evaluation by 1%, in some cases even as low as 33% lower. Hypothesis REJECTED.

H3 - The quality of balance of building blocks is higher in case of a higher version was true averagely by 2%; however several newer versions had a significantly lower quality level as low as to 70%. Therefore the hypothesis can be only PARTIALLY ACCEPTED.

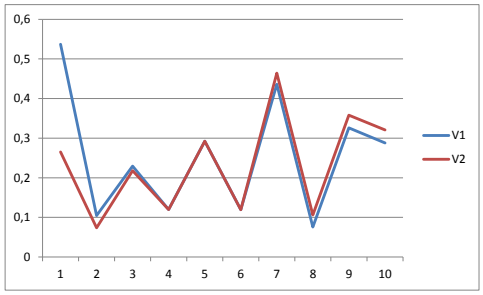


Figure 5. Quality of two successive XML schema versions

In average, the newer version had a 1% improvement on its quality evaluation (TABLE II). Several newer versions had a poorer quality evaluation by 27%, maximum evaluation of quality index was better by 3%, pointing out that newer XML schemas are at best only slightly higher than the older versions; most of them in fact have a poorer quality evaluation.

TABLE II. RESULT IMPROVEMENT IN NEWER VERSION

Aspect	***Min***	***Average***	***Max***
Structure	-2%	2%	21%
Re-use	-33%	-1%	8%
Balance	-70%	2%	12%
Average	-27%	1%	3%

V. CONLUSION

A comparison of two version of the same XML schema was made for several schema pairs. The newer XML schema was averagely slightly improved compared to the previous XML schema, however only by 1 %. The largest improvement was in the structure quality aspect and the balance of the building blocks aspect. The lowest improvement was in re-use of XML schema elements. The results show, the newer versions focus mainly on adding new content or changing older content; improving their structure or enhancing re-use of its building block is not a priority while re-shaping XML schemas.

Since XML schemas develop and change through their life cycle, quality of each newer addition to XML schema is an important aspect, influencing not only XML documents but also higher levels of IT development. The versioning process should be created with greater care, providing improvement not only in its up-to-date contents, but more so in a consistent structure. Each newer version should be tested for its quality before released to achieve at least the same level of quality as its predecessor. As results showed, this is not always the case, opening a concern in the future of XML schema versioning.

The study has also limitations. The first limitation is a small set of XML schema version pairs. Once a new version is released, the older is often no more available. The study should be repeated on a larger set of latest XML schemas, in order to enhance its relevance. Additionally, we compared only two versions. A research should be extended to comparing all versions of one XML schema and monitoring its growth or decline in quality index.

In our future work we will address the deficiencies described as limitation of this study. A larger set of XML schemas will be collected, including all available versions. Quality comparison will be made among all XML schema versions, enabling a more precise evaluation of quality change within XML schema life cycle. Metrics will also be expanded with other approaches such as entropy as a measure of quality and abstract data type structuring.

REFERENCES

[1] Zhang, Y. (2008). Literature Review and Survey: XML Schema Metrics.

[2] Wes Rishel. (2011). Does XML Schema Earn its Keep? The Gartner Blog Network. http://blogs.gartner.com/wes_rishel/2011/12/31/ok-xml-schema-does-earn-its-keep-in-hl7/

[3] Standard ISO/IEC 9126 Software engineering

[4] McDowell, A., Schmidt, C., Yue, K. (2004). Analysis and Metrics of XML Schema. Proceedings of the International Conference on Software Engineering Research and Practice, SERP'04, v 2, p 538-544, 2004.

[5] Burris, E. (2012), Hierarchical Nature of Software Quality, Programming in the Large, The Practice of Software Engineering, http://programminglarge.com/hierarchical-nature-of-software-quality/.

[6] Narasimhan, V.L., Hendradjaya, B. (2007). Some theoretical considerations for a suite of metrics for the integration of software components. Information Sciences, Volume 177, Issue 3, 1 February 2007, Pages 844-864. http://dx.doi.org/10.1016/j.ins.2006.07.010

[7] Washizaki, H., Fukazawab, Y. (2005). A technique for automatic component extraction from object-oriented programs by refactoring. Volume 56, Issues 1–2, April 2005, Pages 99–116. http://dx.doi.org/10.1016/j.scico.2004.11.007

[8] Richi Nayak, Wina Iryadi: XML schema clustering with semantic and hierarchical similarity measures, Knowledge-Based Systems 20 (2007) 336-349.

[9] Mustafa Atay, Artem Chebotko, Dapeng Liu, Shiyong Lu, Farshad Fotouhi: Efficient schema-based XML-to-Relational data mapping, Information Systems, Volume 32, Issue 3, (2007), 458-476

[10] Pulvermacher, M (2001)., ML Schemas: Best Practices Versioning, Stylus Studio® Developer Network, http://www.stylusstudio.com/xmldev/200109/post10080.html (24.2.2014).

[11] Pušnik, M., Šumak, B., Heričko, M., Budimac, Z., Redefining Software Quality Metrics to XML Schema Needs, SQAMIA 2013, Novi Sad, Serbia, September 15-17, 2013, p. 87-9

N-BEB: New Backoff Algorithm for IEEE 802.11 MAC Protocol

Mohammad Shurman[1], Bilal Al-Shua'b[2], Mohammad Alsaedeen[2], Mamoun F. Al-Mistarihi[3], Khalid A. Darabkh[4]
[(1)]Jordan University of Science and Technology/Network Engineering and Security Department, Irbid, Jordan
[(2)]Jordan University of Science and Technology/Computer Engineering Department, Irbid, Jordan
[(3)]Jordan University of Science and Technology/Electrical Engineering Department, Irbid, Jordan
[(4)]The University of Jordan/Computer Engineering Department, Amman, Jordan
alshurman@just.edu.jo, alomaribelal@yahoo.com, mkalsaedeen12@cit.just.edu.jo,mistarihi@just.edu.jo, k.darabkh@ju.edu.jo

***Abstract*- IEEE 802.11 standard primarily utilizes Distributed Coordination Function (DCF) which is based on carrier sense multiple access with collision avoidance (CSMA/CA). IN DCF, when a collision occurs, then a backoff algorithm will take place to randomize medium access for the nodes. The default backoff mechanism used in 802.11 is the binary exponential backoff (BEB) algorithm. In this paper, we analyze the behavior of various backoff algorithms used in IEEE802.11 standard: BEB, I-BEB, and E-BEB, then we propose the New Binary Exponential Backoff (N-BEB) algorithm to improve channel access fairness while preserving the channel throughput. Our N-BEB algorithm improves the selection process of the contention window CW, based on the number of successful and unsuccessful transmissions. Simulation results confirm fairness improvement, power consumption enhancement, reduction of dropped packets and high delivery ratio for proposed proptocol over BEB, I-BEB, and E-BEB protocols.**

***Keywords:* IEEE802.11 back off, BEB, QoS, chancel fairness, wireless MAC.**

I. INTRODUCTION

Much research literature and several algorithms address the problem of fairness among senders in wireless networks. Some of these algorithms, despite solving the fairness problem, suffer a significant reduction of the channel throughput, increase delay or reduce packet delivery ratio. In this paper, we proposed a novel approach to solve the fairness problem in wireless networks without degrading other metrics. Throughput and fairness in wireless channels are inversely related, for instance, some research maximizes the throughput [1] but keeps the CW constant, while other research modifies the CW with respect to the size of the region and the state of the channel. Yet, others modify the CW with respect to the conditions of the network load.

The quality of service (QoS) of the IEEE 802.11 protocol is a critical issue for some applications such as multimedia. Therefore, a key element is that we preserve the bandwidth, packet loss rate, and delay for such applications to achieve an acceptable performance in wirless networks.

II. RELATED WORK

Wireless medium access control (MAC) protocol suffers from several problems such as hidden terminal, exposed terminal, as well as fairness problem. The sharing of the channel causes hidden and exposed terminal problems while the fairness problem is a result of using ill-suited backoff algorithm. In binary exponential backoff (BEB) algorithm [2], the CW is used to determine the value of the backoff time (BT) for each node, whereas CW value is changed in BEB as follows:

- When a node sends a packet successfully, the value of the CW will return to CW_{min}:

$$CW = CW_{min} \quad (1)$$

- When a node sends a packet and a collision occurs, CW will be doubled:

$$CW = 2 * CW \quad (2)$$

- When a node sends a packet unsuccessfully for seven times, the packet will be dropped.

BEB algorithm is the main cause of the fairness problem in IEEE802.11, when a node in BEB sends its packet successfully, it will choose a random number in the range of (0, 31), knowing that unsuccessful transmission will double its CW to reach a large value such as 64, 128, 256….and so on. These values are considered very large with respect to CW_{min}. Hence, the chosen random number will probably be larger than the one chosen earlier, in other words, each node with a small CW will have a higher probability to access the channel and vice versa [2], that is to say, the node that accesses the channel for the first time has a higher probability in accessing the channel in the next trial. Conversely, the node,with a large value for the CW, has a lower probability to access the channel and send its packets.

Improved binary exponential backoff (I-BEB) algorithm [3] is an adaptive algorithm which improves the QoS of IEEE 802.11 protocol. This adaptive algorithm is dependent on a dynamic operation to modify the value of the CW based on the number of successful sending times. In I-BEB, when the node sends a packet successfully, it will execute the function *rst_cw_I_BEB()*. In this function, if the amount of successful transmissions is less than a predefind threshold quantity *c*, the CW value of those nodes which send data packet successfully is divided by the parameter *d*. When the CW is less than another predefined threshold quantity *G* (CW<*G*), their protocol sets CW value equals to *G* (i.e $G = 0.125 \times CW_min$)[3].

If the amount of successful transmissions is greater than *c*, this algorithm will increment CW to a large value (i.e., CW+e×CW_min), giving other nodes the opportunity to use the channel. Both processes are repeated in a continual manner, giving a more evenly distributed opportunity to each node for channel access. They proposed the best values for the parameters in this algorithm as: *e*=8, *d*=4, *c*=12 [3].

To compaire with our approach, we implemented and tested I-BEB algorithm against multiple scenarios using NS2 simulator. After analyzing the results, we found that I-BEB achieves high fairness but suffers the highest number of dropped packets among all studied algorithms (BEB, E-BEB, N-BEB), and this leads to channel overhead increase and network utilization reduction, since a large number of packets will be lost.

Enhanced binary exponential backoff (E-BEB) algorithm[4] uses a distinct tactic when successful transmission occurs. In this algorithm, a new variable is added which differentiates it from BEB and I-BEB. This variable is called "*BT_TH*" and used to calculate the backoff time. Unlike the traditional technique that calculates the backoff time by *BT=CW*Slot time*, the *BT_TH* in E-BEB algorithm, is calculated as:

$$BT_TH = \frac{CW_{min}}{\sqrt{CW_{max}}} \times CW \times slot\ time \qquad (3)$$

We also implemented and tested E-BEB under NS2 against varios scenarios, and we found that E-BEB achieves low fairness with high processing, in addition to power dissipation increase, which is a crucial factor in the wireless network lifetime.

III. PROPOSED ALGORITHM

In our proposed algorithm, which is based on the original BEB algorithm and its generations, a new procedure is followed. Our algorithm enhances the opportunity for any given node intending to send packets to access the channel. In the instance, when the node cannot access the channel, our proposed algorithm allows the CW to increase its value two times consecutively. However, these increments depend on specified parameters were calculated by intensive simulationexperiments to achieve the best results. After two unsuccessful transmissions, the node starts decreasing the value of the CW.

When a transmission is successful as shown in Fig. 1, the CW is calculated by CW=CW / ($\sqrt{2} \cong 1.4$).This operation is repeated during the first nine successful transmissions. Afterwards, we reset the number of successful transmission to 1 and CW is set to CW×1.5. The value of 1.5 is selected based on extensive simulation results, which gives opportunity to other nodes to use the channel specially those nodes that collide more than two times. This step introduces an efficient fairness between the nodes.

In case of an unsuccessful transmission as explained in Fig. 2, the CW is calculated by CW=CW×($\sqrt{2}$). This operation will be repeated during the first two unsuccessful transmissions. Thereafter, the number of unsuccessful transmission is set to 1 and CW to CW/1.5. This reduction in CW value allows this node to retransmit its data successfully.

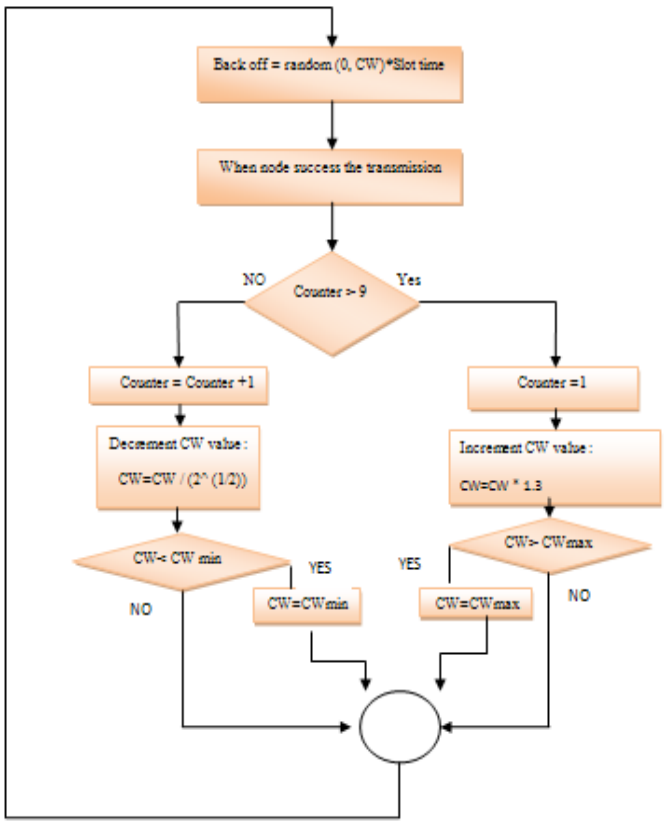


Fig.1 Proposed algorithm (Successful case)

IV. SIMULATION AND RESULTS

We used NS2 [5] simulator to implement and test our proposed algorithm using parameters in Table 1, and the same scenarios mentioned in I-BEB and E-BEB papers:

Scenario (1) : there are 10 nodes distributed on the 500×500 planar domains at random; the maximum connections are 4, the transport layer adopts the UDP protocol. In the application layer, CBR (Constant Bit Rate) is adopted as flows, the top node's speed is 10m/s, residence time is 10 Sec, the simulation time is 100 Sec [3].

Table 1 Simulation parameters

Simulation parameters	
Transport Protocol	UDP
Max-speed	10m/s
Slot time	20 µs
ACK	20 bytes
RTS	25 bytes
CTS	20 bytes
Packet size	1000 bytes
DIFS	40 µs
SIFS	10 µs
CW_{min}	31
CW_{max}	1023
Bandwidth	2 Mbits/s

We implement BEB, I-BEB, and E-BEB as a baseline comparison with our proposed algorithm (N-BEB), using scenario(1) as mentioned in I-BEB paper [3]. As demonstrated in the simulation results figures, we notice that our proposed algorithm achieved the lowest number of dropped packets during the simulation, reduced the channel overhead and increased the packet delivery ratio (PDR). Additionally, N-BEB achieved the least end-to-end delay.

Our algorithm is a bit more complex than I-BEB and BEB, since the original BEB has two cases: successful, where CW is reset to $CW_{min,}$ so the node can access the channel as soon as possible, and unsuccessful, whereby the BEB algorithm simply duplicates the CW, while the I-BEB algorithm focuses primarily on success, and the algorithm assesses the number of successful transmissions to determine whether to increase or decrease the CW; however, in the cases where transmission did not attain success consecutive manner, the CW is reset to minimum, making both of these algorithms simpler than our algorithm.

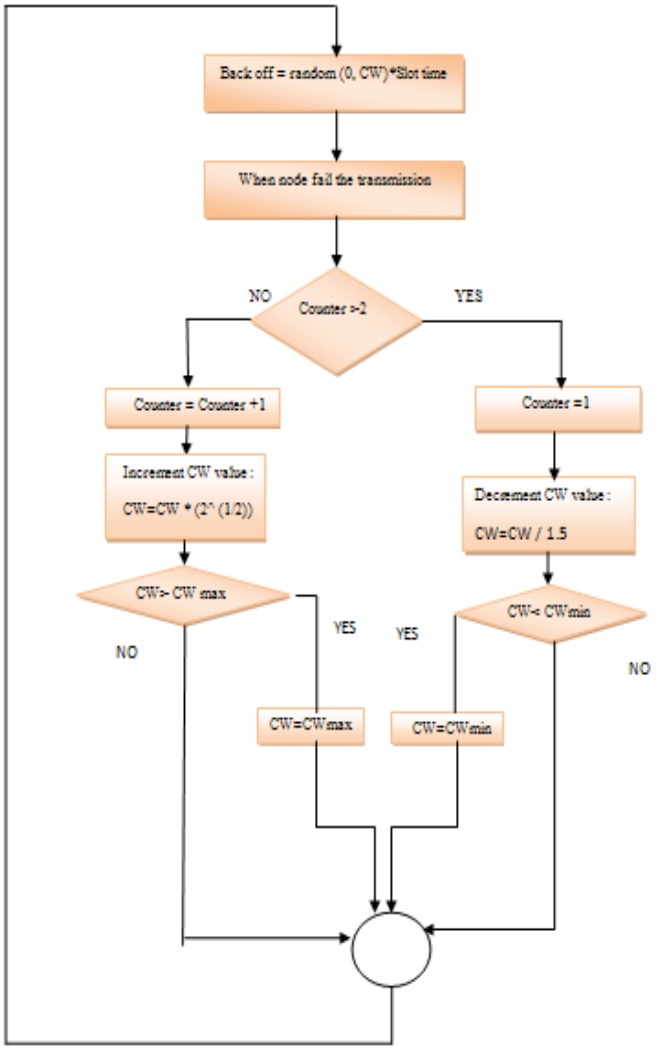


Fig. 2Proposed algorithm (Unsuccessful case)

In our algorithm, we considered both cases of successful and unsuccessful transmissions.We take into consideration the number of unsuccessful and successful transmissions and adjust the value of CW accordingly, thus, N-BEB is considered a little more complex than BEB and I-BEB.

In E-BEB algorithm, we note that its higher complexity stems from taking into account the value of the backoff time, which is then used to recalculate the value of the CW. By doing so, this algorithm adds a new computational complexity, this computational complexity increases power consumption which is critical issue in ad hoc network. Our algorithm considers power consumptionof nodes and tries to save power by condensing the calculation needed to perform the task but it still requires more computation than BEB and I-BEB.

Fairness index can be quantified by fairness index (FI) using the following formula:

$$fairness\ index\ (FI) = \frac{(\sum_{i=1}^{n} x_i)^2}{n \times \sum_{i=1}^{n} (x_i)^2} \qquad (4)$$

where *n* is the number of the same priority flows and *Xi* is the throughput of flow *i*.

Scenario (2): 8 nodes share same channel. Distance between neighbor nodes is 50m; the nodes are static, at 1.0 Sec simulation time, nodes (0, 2, 4 and 6) start simultaneously sending data to its neighboring nodes (1, 3, 5 and 7) respectively[4].

The result we obtained was 98% fairness, the reason behind this high result is that the node is static and each node can connect to only its neighbor node, for this reason we make a new scenario based on mobile nodes as shown in Fig (3), the result is shown in Fig (9) and this scenario is explained next.

Scenario (3): in this scenario 8 nodes share one channel at the same time, all nodes are single-hop, we assume that all nodes are mobile and fall within the range of each other. At time 1.0 Sec, nodes (1, 4, 6, 8) start simultaneously sending data to their neighboring nodes (2, 5, 7, 9) respectively.

We test this scenario on the four algorithms: BEB, I-BEB, E-BEB, and our N-BEB. In the two cases of static and mobile nodes, we can see from the result that our proposed algorithm has higher value of fairness than other algorithms as shown in Fig (9).

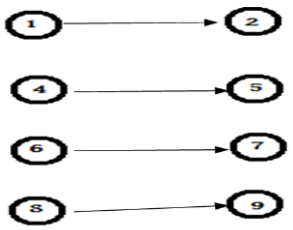
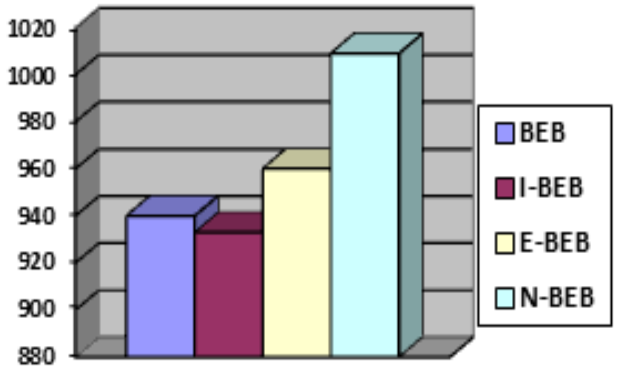
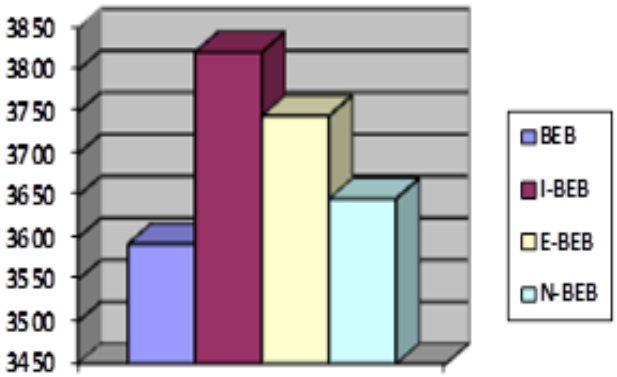


Fig.3 scenario(2)



(A)



(B)

Fig. 4: (A) Number of successful sent packets
(B) Number of overall sent packets

Fig. 4(A) shows the number of successfulsent packets for N-BEB algorithm along with the other mentioned algorithms. It is clear that the number of successfulsent packets in our prposed algorithm is the largest among all mentioned algorithms.

Fig. 4(B) shows the overall sent packets for N-BEB along with the other mentioned algorithms. It is noticeable that the number of sent packets in our proposed algorithm is less than I-BEB and E-BEB, because the number of dropped packets (shown in Fig. 5) is reduced and retransmissions are not needed, which also implies a reduction in power consumption. The number of sent packet in the proposed algorithm is greater than the original BEB, knowing that our algorithm reduces the end-to-end delay, allowing more packets to be sent and received during the same network period of time.

Fig. 5 compares the dropped packets for N-BEB algorithm with the other described algorithms. The proposed algorithm N-BEB, has the least number of dropped packets when compared to the other algorithms and as a result, the channel overhead is reduced significantly, thus, a higher network utilization is accomplished, since a lower number of packets will be lost and consequently retransmitted.

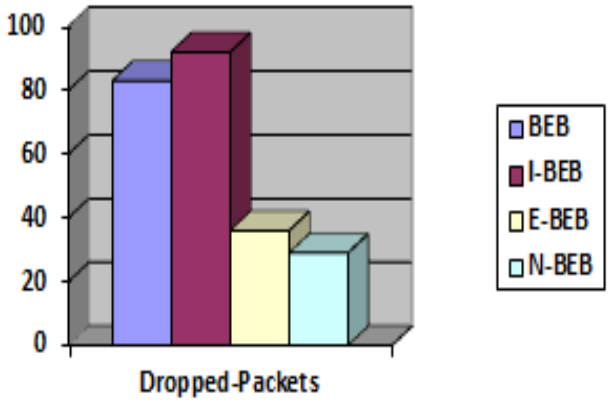


Fig. 5 Dropped packets

To make sure that N-BEB is scalable, we applied N-BEB on a network with different number of nodes (10 to 60 nodes). Here, we see that our algorithm is scalable and can be applied to networks with various numbers of nodes as shown in Fig. 6.

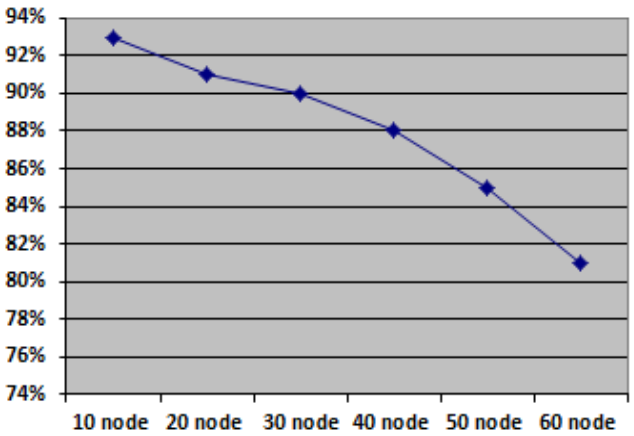


Fig.6 N-BEB scalability with respect to PDR

Fig. 7 shows the delivery ratio for the proposed algorithm, N-BEB, when being compared to the other algorithms. It is obevious that the proposed algorithm has the highest delivery ratio due to the higher channel availability for the sender than the other algorithm as a result of retransmissions reduction.

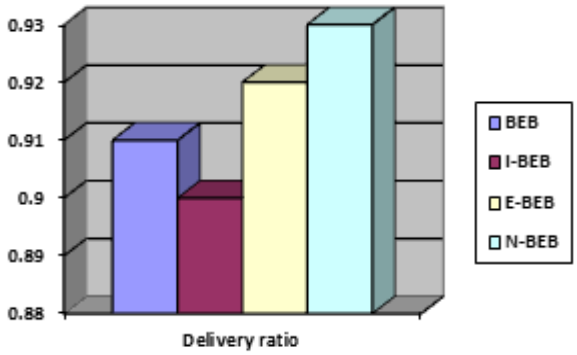


Fig. 7 Packet delivery ratio

End-to-end delay for N-BEB and the other mentioned algorithm is shown in Fig. 8. Among all algorithms, N-BEB has the least delay because the opportunity to access the channel is significantly increased.

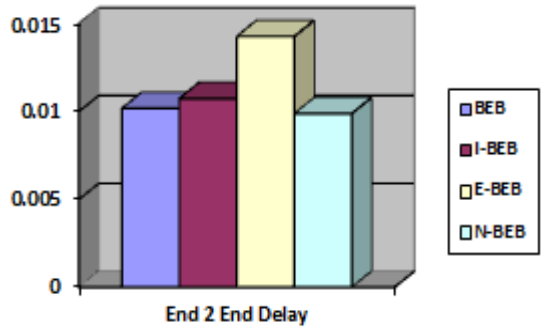


Fig. 8 End-to-end delay

Fig. 9 shows the value Fairness index of N-BEB along with the other mentioned algorithms according to scenario(3). It is clear that the value fairness index for our proposed algorithm is the highest among other algorithms..

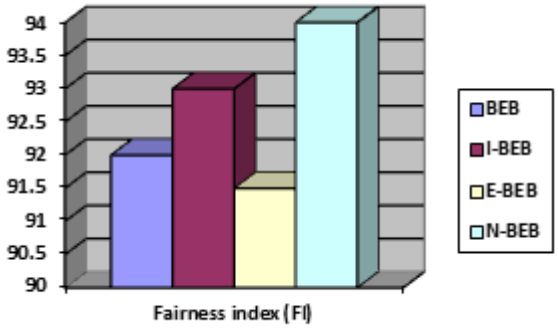


Fig. 9 Fairness index

V. CONCLUSION

In this paper, we proposed a new binary exponential backoff (N-BEB) algorithm that mainly aims to solve the problem of fairness without neglecting the improvement of the throughput. In addition to attaining the best fairness among other tested and relevant algorithms, the N-BEB achieved the highest packet delivery ratio and least number of dropped packets. It deserves pointing out that our proposed algorithm manages a massive improvement in the end-to-end delay, which found to be the least among comparative algorithms. When making a comprehensive evaluation and as far as the complexity is concerned, we found that N-BEB complexity is greater than I-BEB and BEB, but lower than E-BEB.

REFERENCES

[1] SachinGarg, Martin Kappes, and A. S. Krishnakumar, "On the Effect of Contention- Window sizes in IEEE 802.11b Networks,"*Technical report*, Avaya Labs Research, June 2002.

[2] Chunyu Hu, Hwangnam Kim, and Jennifer C. Hou, "An Analysis of the Binary Exponential Backoff Algorithm in Distributed MAC Protocols," *In Tech. Rep. No. UIUCDCS-R-2005-2599 UILU-ENG-2005-1794, July 2005.*

[3] Kun Wang and Changming Sun, "An Improved Backoff Algorithm of Ad Hoc Networks," *In proceedings of International Conference on Computational Intelligence and Software Engineering (CiSE 2009)*, pp. 1-4, Wuhan, December 2009.

[4] Mohammed Al-Hubaishi, Tariq Abdullah, RaedAlsaqour, and Amine Berqia, "E-BEB Algorithm to Improve Quality of Service on Wireless Ad-Hoc Networks," *Journal of Applied Sciences, Engineering and Technology,* vol. 4, no. 9, April 2012.

[5] S. McCanne and S. Floyd, "ns Network Simulator," ISI University Web reference, http://www.isi.edu/nsnam/ns/.

Multi-objective parameter optimization of S-MAC protocol

Enio Kaljić* & Samin Konjicija**
Department of Telecommunication*
Department of Automatic Control and Electronics**
Faculty of Electrical Engineering at the University of Sarajevo
Zmaja od Bosne, 71000 Sarajevo, Bosnia and Herzegovina
Email: ekaljic@etf.unsa.ba, skonjicija@etf.unsa.ba

***Abstract*—In the process of design of protocols for medium access control in wireless sensor networks, it is necessary to find a compromise between energy saving, time sensitivity, scalability, adaptability and complexity. In this research, as an object of optimization S-MAC protocol is chosen. The main advantage of the S-MAC protocol is a state of sleep, which significantly reduces power consumption. Although the S-MAC protocol is described with a large number of parameters, in this paper we have selected three parameters that directly affect the time frame protocol: the duty cycle, the duration of the synchronization interval, and duration of the data interval. To assess the quality of the selected parameters, three measures were selected: total electricity consumption in the network, delay in the delivery of packets to the application layer, and the total achieved bandwidth. Using ns-2 network simulator, a model for testing the influence of parameters of S-MAC protocol on network performance has been developed. The parameters values are selected using the NSGA-II algorithm implemented in MATLAB software package, and the whole process of finding of non-dominated solution set is fully automated. In addition to multi-objective optimization of parameters, an experiment with single-objective optimization using aggregated utility function and the standard GA is performed. Finally, the analysis of the obtained set of non-dominated solutions is performed and a comparison with the results of the single-objective optimization is done.**

Keywords — S-MAC, WSN, multi-objective, single-objective, NSGA, optimization, ns-2, protocol.

I. Introduction

The performance of communications networks depends on various factors at all layers of the OSI/ISO reference model (physical layer, data link layer, network layer, transport layer, session layer, presentation layer and application layer). Although the OSI/ISO reference model provides separate views on different network layers, overall behaviour of the network can not be inferred looking at single layer, because as there is interaction between peer layers and between adjacent layers in OSI/ISO hierarchy. When optimizing the performance of communication networks, the problem is usually solved in two ways (or in two directions, i.e. cross-layer) from the lowest to the highest layer, or vice versa. Direction of optimization is selected depending on which layer has the most impact on the overall network performance.

Cross-layer design approaches presented in [1] can be classified depending on layer interactions (e.g. interaction between Medium Access Control (MAC) layer and Network layer, Physical layer and MAC, etc.). One example of these approaches in design, Energy Aware Routing Protocol, proposed in [2] has the disadvantage in that it only regards one parameter as the performance measure.

When it comes to optimize the performance of wireless sensor networks (WSN), where they perform relatively simple applications, lower layers of the OSI/ISO reference model have the greatest impact on the overall network performance. Since the physical layer is largely standardized (conventional modulation technique, coding, pre-defined radio channels, etc.), the logical choice to start the optimization is the data link layer. The task of the data link layer is to control medium access and establishing a logical connection between two nodes in the network. The behaviour of this layer is defined by the design of MAC protocol. Important factors in the design of MAC protocols in wireless sensor networks are:

- Energy saving - the main goal of the design of the MAC of protocols for wireless sensor networks is energy saving in the wireless sensor nodes which increases the lifetime of the network [3]. An important factor is also the QoS in the network, which is manifested through measures such as throughput, delay, fairness, bandwidth utilization etc. To fulfil the mentioned objectives it is necessary to design an efficient MAC protocol.
- Time sensitivity - another important requirement in the design of the MAC WSN is to ensure the real-time transmission for applications that require it [4].
- Scalability and flexibility - the MAC protocol of wireless sensor networks must be highly scalable due to the large number of sensor nodes and the different functionalities that the network can support. Also, the MAC protocol must be adaptable to changes in topology that occur due to failure or node mobility. The network should be self-organizing and capable of functioning normally regardless of unreliability of individual nodes. To make the network more reliable, decentralized network architecture is desirable [4].
- Complexity - refers to the power consumption in the execution of complex algorithms and protocols. One of the most important requirements in the design of WSN MAC is the simplicity [5].

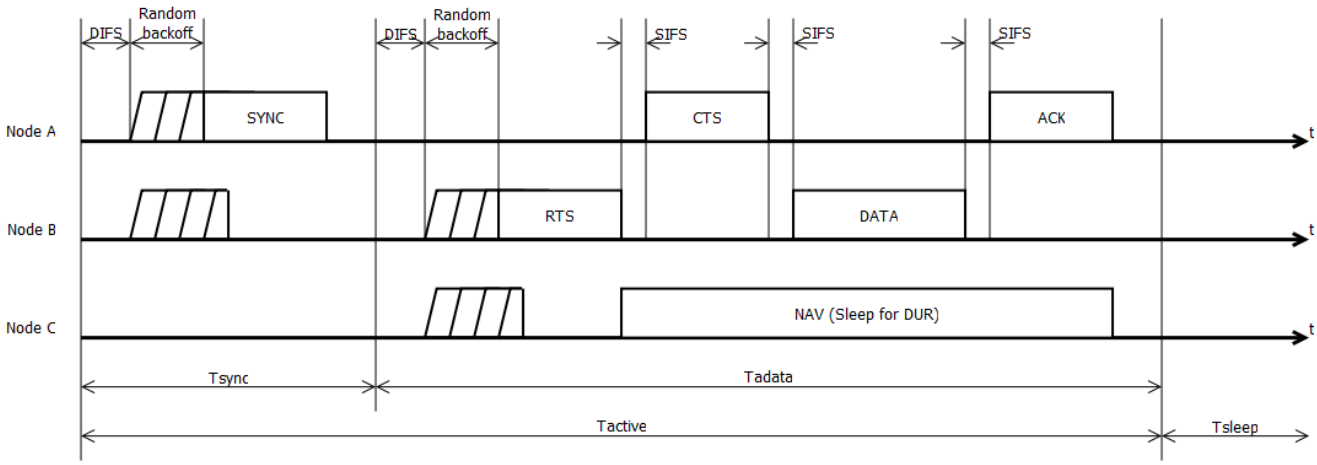


Fig. 1. Timeframe of S-MAC protocol

In the process of design of the MAC protocols it is necessary to find a compromise between the above set of requirements, since some of them are in opposition to each other.

The protocols used within the data link layer of a WSN in the general case are classified into two categories: protocol based on contention for medium and Time Division Multiple Access (TDMA) protocols. Although the IEEE 802.11-based protocols are mainly used for simplicity and resistance to the problem of hidden nodes they are energy inefficient because they do not solve the problems of constant listening to the media and listening to media in the IDLE state. Details about the disadvantages that eliminates certain protocols from the above categories in compared to S-MAC protocol are described in [6]. An overview of all existing MAC protocols for WSN along with their pros and cons is also given in [7]. The main advantage of the S-MAC protocol is the SLEEP state which significantly reduces power consumption. S-MAC also easily adapts to changes in topology. In addition, there is no need for a central node or strictly precise synchronization [8].

In the S-MAC protocol (Fig. 1), one cycle of network operation is divided into two intervals: interval of activity (T_{active}) and sleep interval (T_{sleep}). These periods alternate and have the same time of occurrence for all nodes in a virtual cluster (nodes are fully synchronized). The network can consist of multiple virtual clusters. The active period is divided into two sub-periods consisting of time slots in which the first is used to exchange messages to synchronize the cluster (T_{async}), and the other is used for data transfer (T_{adata}). In both sub-periods nodes access the medium using CSMA/CA mechanism with random counters that each node internally runs if the medium is free. To avoid the problem of hidden nodes in the network are used RTS/CTS control packets by which a particular node informs the other nodes about the duration of its transmission. The nodes that receive the RTS packet, update their Network Allocation Vector (NAV) counters on the duration of transmission of the node that transmits and go to sleep. In this way the energy saving is achieved.

Although the S-MAC protocol is described with a large number of parameters, in this paper three parameters were chosen which have a direct impact on the time frame of the protocol: duty cycle (i.e. the ratio of the interval of activity T_{active} and the total period $T_{active}+T_{sleep}$), duration of T_{sync} and the duration of the interval T_{data}. To assess the quality of the selected parameters three measures were chosen: total energy consumption in the network, average delay in delivery of packages at the application layer, and the total achieved bandwidth at the application layer.

In [9], the effect of duty cycle on network performance is investigated. However, impacts on delay and energy consumption are evaluated seperately. Furthermore, impact of contention window sizes is neglected.

Therefore, this paper will give a much broader look at problem that has been studied earlier, but with multi-objective approach.

II. A FORMAL MODEL OF THE PROBLEM SITUATION

In this paper, the problem situation is the optimization of the parameters of the S-MAC protocol. Problem situation can be formally presented with a set of alternatives in three-dimensional problem space:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \in E^3 \tag{1}$$

where x_i are alternative elements (with the given constraints based on empirical assumptions) which correspond to the following parameters of the S-MAC protocol:

- $x_1 \in [1, 99]$ - Duty cycle.
- $x_2 \in [7, 127]$ - Window size for contention during media access in SYNC phase ($SYNC_CW$). This parameter linearly affects the duration of the interval T_{sync}.
- $x_3 \in [7, 127]$ - Window size for contention when accessing the media in the DATA phase ($DATA_CW$). This parameter linearly affects the duration of the interval T_{adata}.

The objectives of the problem situation can be expressed using three objectives:

- $q_1 \simeq E$ - The total energy consumption in the network.
- $q_2 \simeq T$ - Mean end-to-end delay in the delivery of packets of the application layer.
- $q_3 \simeq B$ - Total achieved bandwidth at application layer.

Total energy consumption in the network is calculated as:

$$E = \sum_{i=1}^{N} E_i \tag{2}$$

where E_i is the power consumption of the i-th node in the network and N is the number of nodes in the network, and is calculated as:

$$\begin{aligned} E_i =& P_{IDLE} \cdot T_i^{IDLE} + P_{TX} \cdot T_i^{TX} \\ &+ P_{RX} \cdot T_i^{RX} + P_{SLEEP} \cdot T_i^{SLEEP} \end{aligned} \tag{3}$$

Constants P_{IDLE}, P_{TX}, P_{RX} and P_{SLEEP} represent power dissipation in idle mode (listening to the media), when sending packets, when receiving packets and in a state of sleep, and variables T_{IDLE}, T_{TX}, T_{RX} and T_{SLEEP} time the i-th node spent in the corresponding mode.

Mean end-to-end delay is calculated as the average of delay of all packets:

$$T = \frac{1}{M} \sum_{i=1}^{M} (t_i^R - t_i^S) \tag{4}$$

where M is the number of sent and received packets, t_i^R the time of receipt of the i-th packet, and t_i^S is the sending time of the i-th packet.

Achieved bandwidth is calculated as:

$$B = \frac{\sum_{i=1}^{M} P_i}{T_{sim}} \tag{5}$$

where M is the number of received packet, P_i is the size of the i-th packet and T_{sim} duration of the simulation, i.e. time of the activity of traffic generator.

Aspiration level can be defined for each objective:

- Achieve the minimum of energy consumption ($\max_{x \in \Omega}\{q_1\}$).
- Achieve the minimum of end-to-end delay ($\max_{\mathbf{x} \in \mathbf{\Omega}}\{q_2\}$).
- Achieve the maximum of bandwidth ($\max_{\mathbf{x} \in \mathbf{\Omega}}\{q_3\}$), or achieve the minimum of unused bandwidth ($\max_{\mathbf{x} \in \mathbf{\Omega}}\{\frac{q_3^{max} - q_3}{q_3^{max}}\}$).

For multi-objective problem can be defined vector function $\mathbf{Q} : E^3 \rightarrow V^3$ that point of problem space $\mathbf{x} \in E^3$ maps to point of value space $\mathbf{v} \in V^3$, where:

$$\mathbf{Q} = \begin{bmatrix} q_1 \\ q_2 \\ q_3 \end{bmatrix}, \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \in \mathbf{\Omega} \subset E^3, \mathbf{v} = \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} \in V^3. \tag{6}$$

If we want to solve the multi-objective problem as a single-objective, then one of the methods is using the aggregate utility function. Aggregate utility function (AUF) performs the mapping three objective functions into a single function as follows:

$$\begin{aligned} U(x_1, x_2, x_3) &= w_1 q_1(x_1, x_2, x_3) \\ &+ w_2 q_2(x_1, x_2, x_3) + w_3 q_3(x_1, x_2, x_3) \end{aligned} \tag{7}$$

where w_1, w_2 and w_3 are weights which determine the preference of each of the objectives. Now the optimization problem is reduced to search for the minimum/maximum of aggregate utility function.

III. Parameters optimization

Fig. 2 shows the software architecture of the system for optimization of parameters. For the realization of the simulation model of the S-MAC protocol, the network simulator NS-2, which is written in C++ programming language with OTcl API was used. To create a simulation model scripting language Tcl was used which represents the interface between the user and OTcl API. In this paper, for the purpose of the S-MAC protocol optimization, a simple simulation model was realized, in which eight of wireless sensor nodes generate traffic at a constant rate and send it to a wireless access point (i.e. a node which performs the function of the receiver).

For the analysis of trace file that is produced by NS-2, a tool designed to analyse large text files AWK was used. The tool is controlled with AWK scripting language with syntax similar to C language. Based on the proposed formal model of the problem situation, three AWK scripts for calculation of energy consumption, end-to-end delay and achieved bandwidth were written. By executing these scripts, three files containing the evaluated measures of performance are generated.

In MATLAB, algorithms for a multi-objective and single-objective optimization of parameters based on genetic algorithm (GA) are implemented. In phase of fitness evaluation for selected individual from the population, MATLAB runs the NS-2 simulator with the parameters encoded within the individual and the result, obtained after the execution of the simulation and AWK scripts, is returned as a value of fitness. Upon completion of the NSGA-II, individuals from the last generation represent the estimated Pareto front.

A. Multi-objective approach using NSGA-II

MATLAB supports GA-based multi-objective optimization using the tool *gamultiobj* which implements a variant of the algorithm NSGA-II with controlled elitism. NSGA-II is described in detail in [10]. Elitist GA favours individuals with better fitness function value (i.e. better rank). GA with controlled elitism favours individuals that enhance the spread of the population, even if these individuals have a lower value of the fitness function. It is very important to maintain a spread of population to converge to Pareto front. The spread is maintained by controlling the elite members of the population during the execution of the algorithm.

B. Single-objective approach using AUF

To demonstrate the main advantage of multi-objective approach, the problem is also solved as a single-objective using

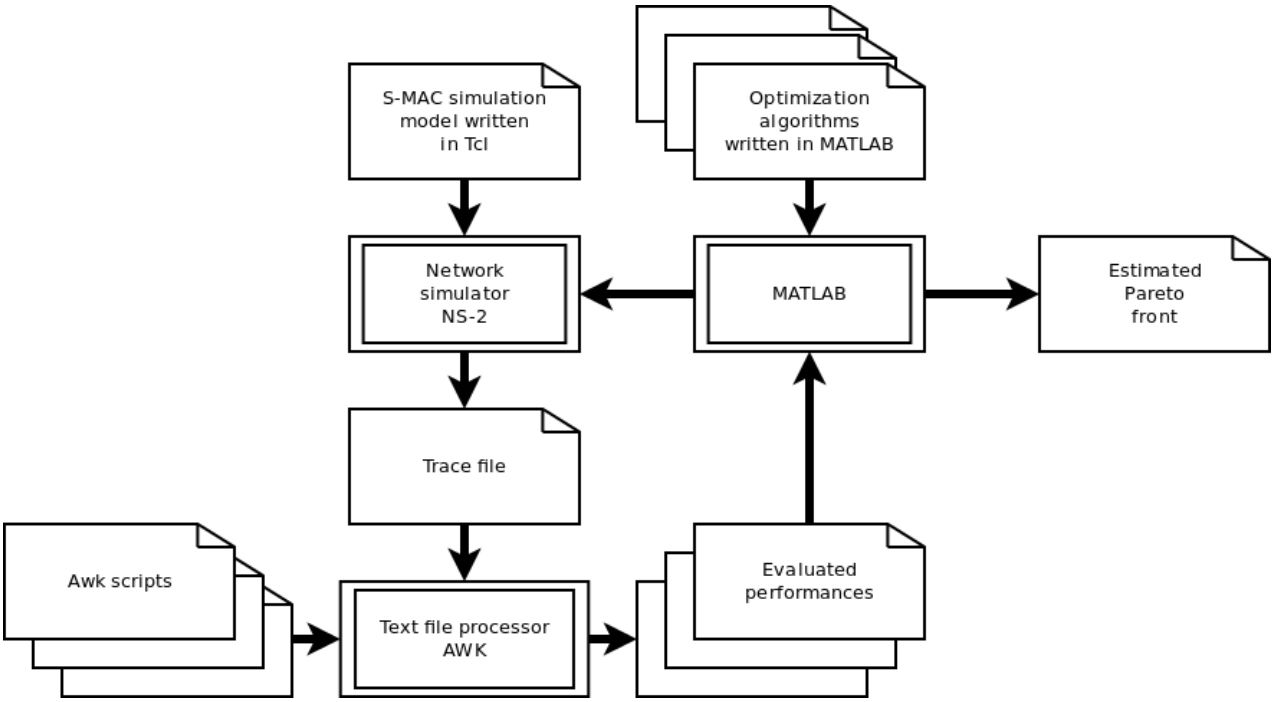


Fig. 2. Software architecture of the optimization system

aggregated utility function. Taking into account the values obtained by all three objectives in multi-objective approach, weight ratios, which also normalize objective functions, are determined, and the following function is obtained:

$$U(x_1, x_2, x_3) = q_1(x_1, x_2, x_3) + 100 \cdot q_2(x_1, x_2, x_3) + q_3(x_1, x_2, x_3) \quad (8)$$

With this mapping one thing is clear - part of the information is lost, because the three-dimensional problem is reduced to one-dimensional. The impact on results will be discussed in next section. As already mentioned before, now the optimization problem is reduced to search for the minimum/maximum of aggregate utility function. With regard to the objective set so that the aspiration level represents a minimum for each of the objectives, in single-objective optimization it is necessary to find a minimum of aggregate utility function.

IV. Results

In the following results obtained using both mentioned approaches will be presented and discussed.

A. Multi-objective approach using NSGA-II

During the execution of NSGA-II algorithm it is possible to monitor parameters such as the average spread of individuals on the Pareto front (Fig. 3) and the average distance between individuals. These two parameters provide information about the distribution and proximity of solutions to the Pareto front. The average spread can be used as a stopping objective of the algorithm. When this value converges to a small value (ideally zero), the solutions are uniformly distributed on the Pareto front.

Taking into account the distribution of non-dominated solutions, layout of Pareto fronts can be guessed, and the using of linear interpolation even shown (Fig. 4). Fig. 5 shows the non-dominated solutions in the problem space. Elements of alternatives in the problem space do not take the full range of possible values.

Table I shows several chosen solutions based on different objectives.

The first solution has a minimum value at the first objective (energy consumption). In the problem space, this solution has a pretty large duration of duty cycle (approximately 84%), while the ration of sizes of the windows: to exchange SYNC, and to exchange DATA packet is approximately 1:2. At first glance, it can be concluded that the high value of the duty cycle could ensures low energy consumption. This is reasonable if we take into account that nodes that do not transmit information go to sleep, ie power save mode, even if the active phase is in progress (due to the mechanism of the NAV of the counter).

The second solution has a minimum value at the second objective (delay). In the problem space, this solution has a larger window sizes for the exchange of SYNC and DATA packets. By increasing the contention window size, the more nodes have a greater chance to send a packet, so the delay is reduced. However, too large windows lead to the frequent occurrence of collisions, and it reflects on energy consumption.

The third solution has the minimum value of the third objective (unused bandwidth). In the problem space, this solution has a moderate duty cycle (approximately 43%), and the ratio of sizes of the windows similar to second solution.

Fourth chosen solution has a moderate value by all three objectives (neither objective does dominate). In the problem

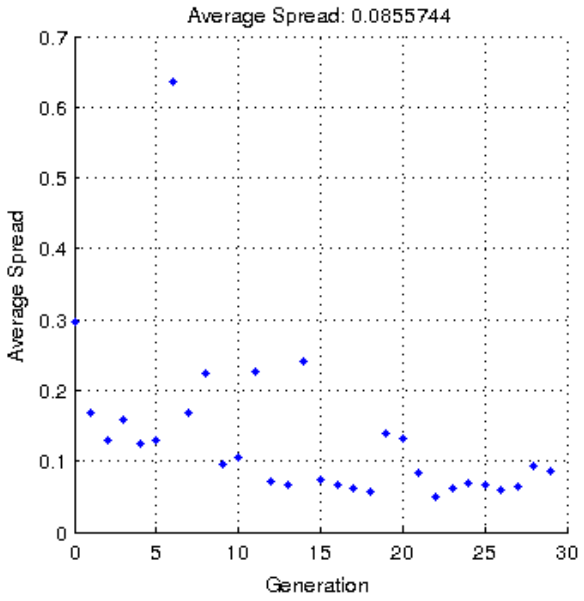


Fig. 3. Average spread

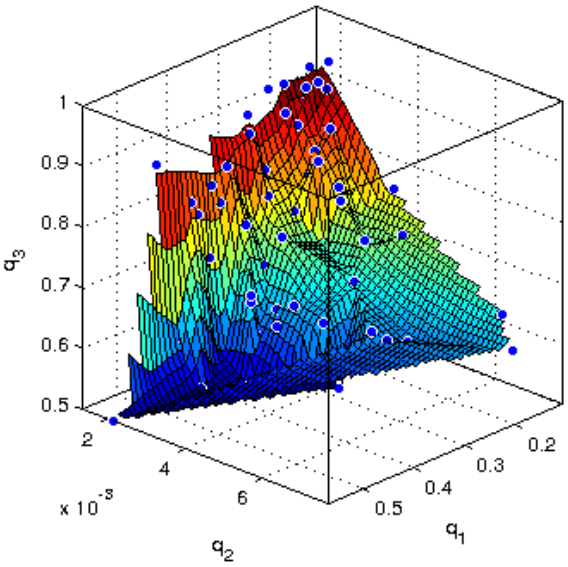


Fig. 4. Linear interpolated Pareto Front (NSGA-II)

space, this solution also gives moderate values for all three parameters. If we take into account that wireless sensor networks do not transfer a large amount of information, when choosing a concrete solution third objective can be neglected. Then selection of solution is reduced to a compromise between energy consumption and delay.

B. Single-objective approach using AUF

After optimization by a single objective, the GA has generated population of solutions that are close to minimum of aggregate utility function. The minimum value of the aggregate utility function is 1.2139, the average value of 1.7035, and the

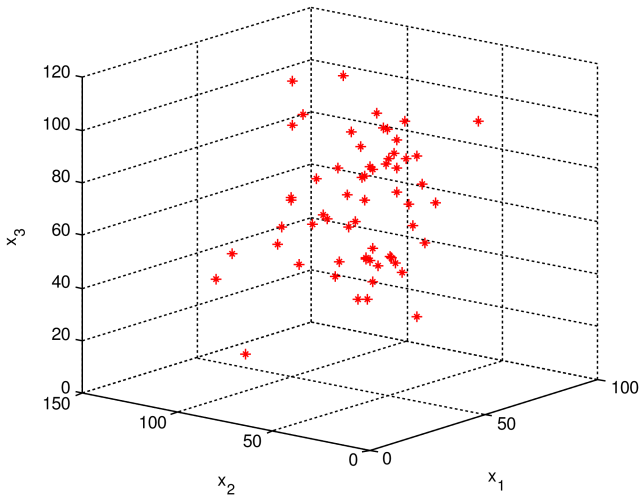


Fig. 5. Non-dominated solutions in problem space (NSGA)

TABLE I
CHOSEN SOLUTIONS (NSGA-II)

Selection objective	x_1	x_2	x_3	q_1	q_2	q_3
$min_{\mathbf{x}\in\mathbf{\Omega}}\{q_1\}$	84.3495	44	96	0.1076	0.0017	0.9185
$min_{\mathbf{x}\in\mathbf{\Omega}}\{q_2\}$	69.2336	123	104	0.2033	0.0014	0.8994
$min_{\mathbf{x}\in\mathbf{\Omega}}\{q_3\}$	43.3275	82	62	0.5648	0.0021	0.5003
Mean by all	55.4211	68	48	0.2627	0.0025	0.7041

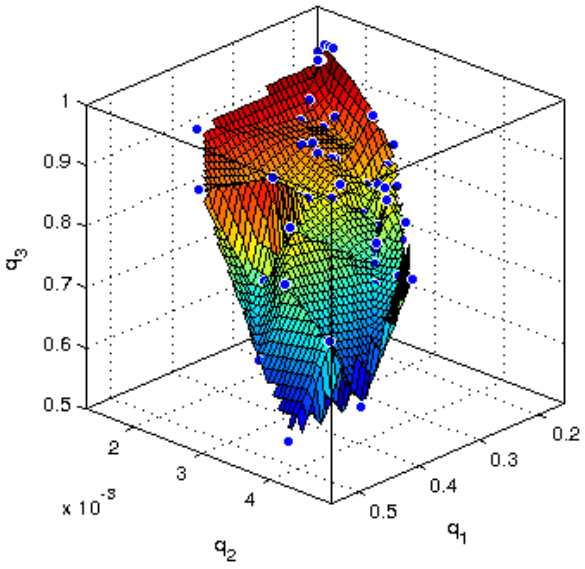


Fig. 6. Linear interpolated Pareto Front (AUF)

standard deviation is 1.3965. From this we can conclude that a large number of particular solutions lie close to minimum of aggregate utility function.

For easier comparison of the solutions obtained by solving the multiple-objective problem, for the each individual, value of each objective is determined and shown in the space of values (Fig. 6). By linear interpolation of points in the space

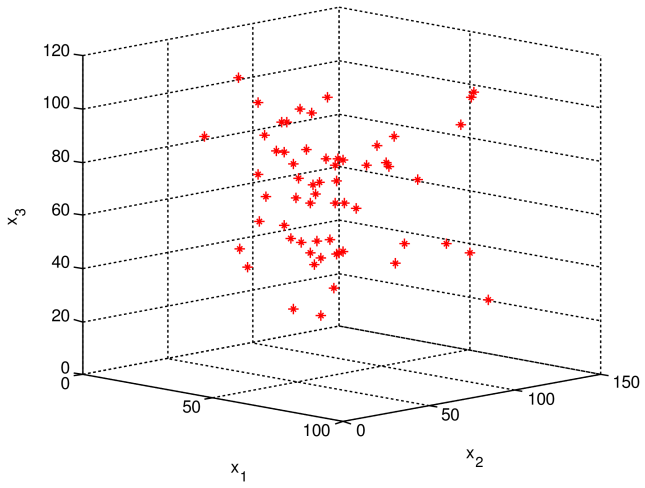


Fig. 7. Non-dominated solutions in problem space (AUF)

TABLE II
CHOSEN SOLUTIONS (AUF)

Objective	x_1	x_2	x_3	q_1	q_2	q_3	U
$\min_{x\in\Omega}\{U\}$	42.46	35	52	0.1591	0.0013	0.9238	1.21
$\min_{x\in\Omega}\{q_1\}$	42.69	35	52	0.1590	0.0014	0.9412	1.24
$\min_{x\in\Omega}\{q_2\}$	42.46	35	52	0.1519	0.0013	0.9238	1.21
$\min_{x\in\Omega}\{q_3\}$	42.44	35	67	0.5426	0.0043	0.5732	1.54
Mean	98.53	35	15	0.2818	0.0032	0.7312	1.46

of values, surface on which points lie is generated.

Fig. 7 shows the non-dominated solutions in the problem space. Elements of alternatives in the problem space do not take the full range of possible values.

Table II shows several chosen solutions. When selecting different objectives were used. The first solution achieves a minimum of aggregate utility function. Second, third and fourth solution, a minimum by the first, second and third objective, respectively. The fifth solution has moderate values for each objective. Low spread of obtained solutions proves the assumption about the loss of information when mapping multi-objective problem in single-objective.

The solutions of single-objective problem are more concentrated than the solutions of multi-objective problem. In the solution set of single-objective problem, there are points in the problem space that do not exist in the set of solutions of multi-objective problem. Due to the small elitism of the single-objective GA it can be assumed that these solutions do not lie near the Pareto front (dominated solutions).

V. CONCLUSION

In this paper we dealt with the problem situation relating to the choice the parameters S-MAC protocol to meet various objectives. The duty cycle and the size of contention window are selected as parameters. Energy consumption, delay, total achieved / unused bandwidth were chosen as the objectives. Aspiration level of multi-objective problem is the minimum for each objective. Using the NS-2 network simulator, model was developed to test the influence of the S-MAC protocol parameters on the network performance. Estimation of network performance is obtained by using the three selected measures. Optimization is performed in two ways: solution of multi-objective problem using NSGA-II, and solution of single-objective problem using AUF.

The entire process of finding a set of non-dominated solution is fully automated. Finally, the analysis of the obtained set of non-dominated solutions is performed and a comparison with the results of single-objective optimization is done. The analysis of results has led to some interesting conclusions:

- In both approaches, the obtained solutions in the problem area are covering 60-70% of the scope of allowed values of the elements of the problem situation.
- It is shown that the resulting solutions are a compromise between the chosen objective, which legitimate multi-objective approach to solving problems.
- It is shown that the mapping of multi-objective problem in single-objective leads to the loss of part of the information, which is reflected in the diversity of the obtained solutions (less coverage of Pareto front).
- By solving the single-objective problem, part of the obtained solutions does not lies in near the Pareto front.

REFERENCES

[1] Tommaso Melodia, Mehmet C. Vuran, and Dario Pompili. The State of the Art in Cross-Layer Design for Wireless Sensor Networks. EuroNGI Workshop: 78-92, 2005.

[2] Rahul C. Shah and Jan M. Rabaey. Energy Aware Routing for Low Energy Ad Hoc Sensor Networks. Proc. IEEE Wireless commun. and Network Conf., vol. 1, pp. 350—55, .2002.

[3] Robert Abo, Kamel Barkoui and Hafedh Zayani, A probabilistic analysis of energy efficiency of wireless sensor network duty-cycled MAC protocols: application to S-MAC and Eco-MAC, In Proceedings of the 6th ACM workshop on QoS and security for wireless and mobile networks (Q2SWinet '10), ACM, New York, NY, USA, 2010.

[4] Koubaa, A., Alves, M., Tovar , E., Lower Protocol Layers for Wireless Sensor Networks: A Survey, IPPHURRAY Technical Report, HURRAY-TR-051101, 2005.

[5] Simarpreet Kaur and Leena Mahajan, Power Saving MAC Protocols for WSNs and Optimization of S-MAC Protocol, International Journal of Engineering Business Management, W.H. Ip (Ed.), ISBN: 1847-9790, InTech, 2012.

[6] Rajgopal Kannan, Ram Kalidindi, S. S. Iyengar, and Vijay Kumar, Energy and rate based MAC protocol for wireless sensor networks, SIGMOD Rec. 32, 4, 2003.

[7] M. Riduan Ahmad, Eryk Dutkiewicz and Xiaojing Huang, A Survey of Low Duty Cycle MAC Protocols in Wireless Sensor Networks, Emerging Communications for Wireless Sensor Networks, Anna Foerster and Alexander Foerster (Ed.), ISBN: 978-953-307-082-7, InTech, 2011.

[8] Joseph Kabara and Maria Calle, MAC Protocols Used by Wireless Sensor Networks and a General Method of Performance Evaluation, International Journal of Distributed Sensor Networks, vol. 2012, Article ID 834784, 11 pages, 2012.

[9] Ramakrishnan, S.; Hong Huang; Balakrishnan, M.; Mullen, J., "Impact of sleep in a wireless sensor MAC protocol," Vehicular Technology Conference, 2004. VTC2004-Fall. 2004 IEEE 60th, vol.7, no., pp.4621,4624 Vol. 7, 26-29 Sept. 2004

[10] Deb, K.; Pratap, A.; Agarwal, S.; Meyarivan, T., "A fast and elitist multiobjective genetic algorithm: NSGA-II," Evolutionary Computation, IEEE Transactions on, vol.6, no.2, pp.182,197, Apr 2002

Primjena FMEA metode u upravljanju kvalitetom telekomunikacijskih usluga

Mr.sc. Nermin Karahmetović
BH Telecom d.d. Sarajevo - RD Tuzla, Tuzla, Bosna i Hercegovina
nermin.karahmetovic@bih.net.ba

Sažetak - U implementaciji sistema za upravljanje kvalitetom (QMS, TQM i sl.) neophodna je primjena adekvatnih kombinacija metoda i alata. Kao jedna od metoda koja kompanijama stoji na raspolaganju je i FMEA metoda. Iako je prvobitno bila namjenjena za upravljanje kvalitetom u proizvodnim djelatnostima, u zadnje vrijeme se nerijetko primjenjuje i u uslužnim kompanijama. U ovom radu prezentirat će se mogućnost primjene ove metode u unapređenju telekomunikacijskih usluga i procesa koji prethode pružanju usluga. Svrha FMEA metode jeste poduzimanje akcija za eliminisanje ili smanjivanje broja grešaka ili nedostataka, počevši od onih koji imaju najveći značaj i prioritet. Dobiveni rezultati pružaju smjernice pri rangiranju grešaka na način da ukazuju koji problem treba prvo rješavati. Prema dobivenim rezultatima pristupa se prvo rješavanju grešaka s najvećim rizikom. U istraživanju FMEA metoda je primjenjena na procese koji učestvuju u isporuci paket usluga, na način da je izvršena analiza ključnih procesa i aktivnosti. Cilj analize procesa i aktivnosti uključenih u stvaranje paket usluge je povećanje njihove efikasnosti i efektivnosti, te maksimiziranje vrijednosti (kvaliteta) koja se paket usluzi dodaje u svakom procesu i njegovim aktivnostima. Rezultati dobiveni primjenom metode ukazuju na konkretne mjere i mjerljive pokazatelje uočenih nedostataka te u konačnici pokazuju da je moguće značajno povećati efikasnost posmatranih procesa.

I. UVOD

FMEA metoda (*eng. Failure Mode and Effect Analysis*) omogućava identifikaciju i sprječavanje problema na proizvodu/usluzi ili u procesu prije nego što i nastanu, a fokusirana je na prevenciju grešaka, odnosno smanjivanje mogućnosti da se greška dogodi, a u konačnici ima pozitivan uticaj na povećanje zadovoljstva korisnika.
Postoje četiri osnovne vrste analize utjecaja i posljedica grešaka: FMEA sistema, FMEA dizajna, FMEA procesa i FMEA usluga.

Neki autori spominju samo dvije vrste, i to: FMEA dizajna (DFMEA- Design Failure Mode and effect Analysis) i FMEA procesa (PFMEA-Process Failure Mode and effect Analysis) [1].

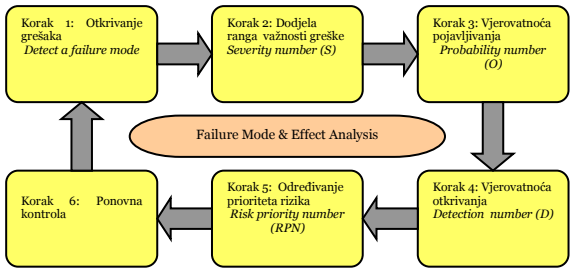
Metoda je prepoznatljiva po jednostavnosti i terminologiji kojom se koristi prilikom primjene, kao što su: greška, posljedica, rizik, uzrok greške, važnost, vjerovatnoća ponavljanja, vjerovatnoća otkrivanja i pojam faktora rizika.

U fazi dizajna usluge, procesa i sl. FMEA se koristi da bi se spriječile moguće greške, a u kasnijim fazama se koristi za sprovođenje kontrole.

FMEA je specifična metoda koja omogućava ocjenjivanje sistema, dizajna, procesa ili usluge svim mogućim oblicima u kojima se potencijalne greške mogu dogoditi. Za svaku uočenu ili prepoznatu grešku (bez obzira je li ona već uočena ili je potencijalna), radi se procjena vjerovatnoće pojavljivanja, važnosti i vjerovatnoće otkrivanja [2].

Analiza mogućih grešaka i njihovih posljedica predstavlja sistemski način osiguranja da sve moguće greške u dizajnu, procesu ili one koje prethode pružanju usluge budu uzete u obzir.

U suštini, FMEA je metoda kojom se maksimizira zadovoljstvo korisnika kroz potpuno eliminiranje ili djelimično smanjivanje uzroka problema koji se mogu pojaviti. Sa aspekta primjene (FMEA metode) izrazito je važno definisati da korisnik nije samo krajnji potrošač/kupac, već to može biti i naredni proces ili naredna operacija u istom procesu. Za optimalne rezultate u primjeni FMEA metode potrebno je da se započinje u najranijim fazama dizajna i sprovodi se kroz cijeli životni vijek usluge.



Slika 1. FMEA ciklus

Analiza mogućih grešaka i njihovih posljedica koristi se [3]:

- U toku dizajna ili redizajna procesa, proizvoda ili usluge, nakon provođenja QFD (Quality Function Development) metode, a prije nego li je dizajn finaliziran.
- Kada se postojeći proces, proizvod ili usluga primjenjuje na novi način.
- Prije razvoja planova kontrole novih ili modificiranih procesa.
- Kada se planiraju ciljevi unaprjeđivanja za postojeće procese, proizvode ili usluge.
- Kada se analiziraju greške u postojećim procesima, proizvodima ili uslugama.
- Povremeno kroz životni vijek procesa, proizvoda ili usluga.

FMEA usluga primjenjuje se za analizu usluge prije nego što usluga bude pružena korisniku. Budući da su osnovne

karakteristike usluge neopipljivost, nedjeljivost, nemogućnost uskladištenja, heterogenost i sl., kod primjene FMEA usluga treba posebnu pažnju posvetiti upravo ovim karakteristikama.
Primjenom ove metode za upravljanje kvalitetom usluga, najviše dolazi do izražaja raznolikost shvatanja pojma kvaliteta od strane korisnika kao krajnjeg potrošača, izražena kroz subjektivan osjećaj zadovoljstva ili nezadovoljstva pruženom uslugom.

II. OSNOVNI ELEMENTI PRIMJENE METODE

Shodno rečenome, cilj primjene FMEA metode je minimiziranje vjerovatnoće pojavljivanja grešaka. Slika 1. prikazuje korake koji su prisutni u primjeni metode. Kako bi se navedeni cilj adekvatno i sveobuhvatno realizirao, kroz različite faze internog i eksternog istraživanja prikupljaju se neophodni podaci, te se njihovom detaljnom analizom i obradom definišu određeni zaključci, te se predlažu mjere za unapređenje zadovoljstva korisnika paket usluge.
U teoriji i praksi implementacija FMEA metode neznatno varira zavisno od specifičnosti svake organizacije/kompanije. Ona u svom prvom koraku podrazumjeva formiranje jednog šireg tima, sastavljenog od stručnih ljudi različitih funkcija i znanja o procesima, uslugama, potrebama kupaca. Uključene funkcije su: dizajn, proizvodnja, testiranje, održavanje, nabava, prodaja, marketing i dr. Naredni korak je određivanje područja za primjenu FMEA. Primjenjuje li se na cjelokupni poslovni sistem, dizajn, proces ili samo na uslugu. Vrlo je važno da svi uključeni uposlenici koji provode primjenu metode poznaju područje svog djelovanja i da primjenjuju iste mjerne instrumente. Nakon toga se pristupa popunjavanju FMEA obrasca (Tabela 1.), kojeg svaka kompanija može prilagoditi svojim potrebama i specifičnostima. Naredni koraci se sastoje od prikupljanja informacija koje se unose u odgovarajuće kolone navedenog obrasca.
Za svaki od navednih uzroka se identifikuju, opisuju i unose postojeće kotrolne tačke ili procesi kontrole (kolona 7.). To mogu biti testovi, procedure ili mehanizmi koji se trenutno koriste da bi se izbjegle ili na vrijeme uočile pogreške i kako one nebi stigle do korisnika.

Za svaku od identifikovanih kontrola određuje se rang otkrivanja „D“ (*eng. detecting rating – D*).

Pomenutim rangom (kolona 8.) se procjenjuje koliko dobra može biti kontrola, odnosno koliko efikasna može biti u otkrivanju uzroka grešaka ili samih grešaka nakon što se pojavila a prije nego je stigla do korisnika. Vjerovatnoća otkrivanja se isto tako rangira na skali od 1 do 10, gdje 1 označava u potpunosti pouzdanu kontrolu, a 10 u potpunosti nepouzdanu kontrolu ili da kontrola uopšte ne postoji. Rang vjerovatnoće otkrivanja unosi se za svaki uzrok greške pojedinačno.

Nakon utvrđivanja ranga važnosti (S), ranga vjerovatnoće pojavljivanja (O) i ranga vjerovatnoće otkrivanja (D) uzroka greške, slijedi izračunavanje RPN faktora (kolona 9.) za svaku potencijalnu grešku pojedinačno. RPN (*eng. Risk Priority Number*) predstavlja faktor prioriteta rizika, i izračunava se po sljedećoj formuli:

$$RPN=S \times O \times D \quad (1)$$

Nakon izračunavanja RPN faktora, pristupa se izračunavanju CRIT faktora (*eng. Critical Factor* – tabela 1. kolona 10.) koji predstavlja faktor kritičnosti, a prema sljedećoj formuli:

$$CRIT=S \times O \quad (2)$$

Dobiveni rezultati pružaju smjernice pri rangiranju grešaka na način da ukazuju koji problem treba prvo rješavati. Prema dobivenim rezultatima pristupa se prvo rješavanju grešaka s najvećim rizikom. Nakon izračunavanja RPN faktora za svaku potencijalnu grešku, dobivene vrijednosti mogu se sabirati, te na taj način dobiti ukupni RPN faktor za cijeli posmatrani proces, uslugu ili sistem.

Nakon izračunavanja RPN faktora i faktora kritičnosti, rangiranja mogućih pogrešaka i njihovih posljedica, identifikuju se preporučene mjere (kolona 11.) tj. koraci za njihovo otklanjanje ili smanjenje ranga RPN i CRIT faktora. Ove mjere mogu biti promjene u procesu ili dizajnu usluge, uvođenje kontrolnih tačaka tamo gdje ih nije bilo i sl. Poduzete mjere imaju za cilj da u budućnosti izbjegnu pojavljivanje grešaka, smanjenje troškova, uštede u vremenu, povećanje zadovoljstva korisnika, te povećanje konkurentnosti.

TABLICA I. OBLIK/FORMA FMEA OBRASCA

B2:R11																
Ocjena: Korisnik:	Telekom operator		Dio naziv / broj		101		Uzorak dio:	A							
Verzija god.:	2011.		Opis procesa:		Prijem i skladištenje terminalne opreme						Ključni Datum:				
Program: Program:	Abc123		Tim za realizaciju:		A.A., N.N., M.M., B.A.						Izvorni datum:	Najnoviji datum revizije:			
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
												Ciljni rezultati planiranih mjera				
Opis procesa/ usluga	**Potencijalni neuspjeh**	**Potencijalni učinci neuspjeha**	**S**	**Mogući uzroci**	**O**	**Trenutne kontrole procesa**	**D**	**RPN**	**CRIT**	**Preporučene akcije**	**Odgovorne osobe i ciljni datum završetka akcije**	**Poduzeta/planirana mjera**	**S**	**O**	**D**	**RPN**
Prijem i kontrola prijema terminalne opreme (mobilinih telefona) od dobavljača	Mogućnost prijema neispravnih ili uređaja sa skrivenim nedostacima te puštanje takvog uređaja u prodaju	Nepotrebno velik broj reklamacija na prodate uređaje i kroz to nezadovoljstvo korisnika.	8	Tvornička greška na uređaju, neadekvatno skladištenje i rukovanje u transportu ili skladištu, neadekvatna kontrola kod dobavljača i sl.	7	Uzorkovanje i kontrola po ugovoru sa 3 ili 1% zavisno od dobavljača,kontrol a nakon otkrivanja grešaka.	7	392	56	Kontrola prijema i uzorkovanje po sistemu ANSI/ASQC Z1.4.(jednostruko uzorkovanje). Vođenje evidencije i zapisa o kontroli, vođenje evidencije i zapisa o izmjerenim vrijednostima vlage na ambalaži i samom uređaju.	Rukovodioc i u prodaji, komisija za prijem, magacinsko osoblje i dr.	Kontrola prijema i uzorkovanje po sistemu ANSI/ASQC Z1.4.(jednostruko uzorkovanje). Vođenje evidencije i zapisa o kontroli, vođenje evidencije i zapisa o izmjerenim vrijednostima vlage na ambalaži i samom uređaju.	4	5	6	120

Također je vrlo važno da se evidentira tko je odgovoran za provođenje tih mjera kao i krajnji datum realizacije predviđenih mjera (kolona 12.).
Prilikom određivanja rizika i prioriteta pojedinih mjera odnosno akcija, nije dovoljno samo oslanjati se na RPN i faktor kritičnosti, već je potrebno uključiti i ličnu procjenu i sud svih članova tima. Nakon što se završi sa provođenjem predloženih mjera, rezultati se evidentiraju u FMEA obrazac (kolona 13.). Ukoliko su preduzete mjere uspješno sprovedene, one rezultiraju nižim rangom važnosti (S), nižim rangom vjerovatnoće pojavljivanja (O) te nižim rangom vjerovatnoće otkrivanja (D), što u konačnici rezultira nižim RPN faktorom.
Na kraju se u Obrazac unose i te nove vrijednosti S, O, D, RPN (kolona 14, 15, 16, 17).
U svakom procesu definiraju se ulazni i izlazni zahtjevi, kontrolne mjere i potrebni resursi za uspješnu realizaciju svih operacija u procesu. U primjeni procesne FMEA metode definicija korisnika/potrošača nije jedina pretpostavka, već se kao potencijalni „korisnik“ može posmatrati i naredni procesni korak, odnosno operacija.
Bez jasne definicije procesa i procesnih operacija nije moguće ostvariti maksimalne koristi od primjene FMEA metode, s obzirom da postoji mogućnost zanemarivanja bitnih dijelova procesa i fokusiranja na područja koja ne zahtjevaju veliku pažnju. Iz tog razloga, poželjno bi bilo da uz opis procesa bude i grafički prikaz radi boljeg razumjevanja.
Analiza uticaja i posljedica grešaka može se posmatrati i kao analiza rizika sa osnovnim ciljem dobijanja odgovora na dva osnovna pitanja:

- koja su sva moguća odstupanja (greške) od propisanog zahtjeva, odnosno šta sve može poći pogrešnim pravcem i tako prouzrokovati potencijalne gubitke i nepotrebno trošenje resursa, i
- ukoliko postoji mogućnost nastanka grešaka, koja je vjerovatnoća nihovog pojavljivanja i koje posljedice donose u odnosu na uslugu, proces ili cjelokupan sistem.

U tom smislu vrši se detaljna analiza pritužbi korisnika u svim slučajevima koji su zabilježeni i radi se kvalitativna procjena važnosti za korisnika i kompaniju da se te greške izbjegnu u budućem radu. Za ovaj segment moguće je koristiti sve interne resurse, baze podataka i ankete uposlenika i korisnika. Dodatno, ukoliko je detektovano da se neka greška počela pojavljivati pristupa se procjeni vjerovatnoće njenog pojavljivanja, te se u taj podatak kvantificira sa posljedicama koje ta greška sa sobom donosi. Za procjenu vjerovatnoće pojavljivanja koriste se izvještaji i svi dosadašnji zabilježeni slučajevi otkrivanja, matematičke procjene i sl.

U tom smislu kao osnovna prednost ove metode je njena preventivna komponenta izražena kroz utjecaj na probleme prije nego se oni i pojave ili nakon već prvih pojavljivanja, jer se na taj način može smanjiti vjerovatnoća njihova pojavljivanja u budućnosti, a ako uzmemo u obzir činjenicu da je problem već poznat i sigurno je da će nastati, moguće je i minimizirati njegove posljedice. Na taj način, funkcija praćenja troškova nadopunjava se mogućnostima njihovog smanjivanja, a princip rješavanja problema, ustupa mjesto pronalaženju načina da se oni nikada i ne dogode.

III. METODOLOGIJA ISTRAŽIVANJA

Upravljanje kvalitetom telekomunikacijskih usluga primjenom FMEA metodologije u ovom istraživanju ima za cilj da se testiraju mogućnosti uključivanja uposlenika u procesima prodaje usluga pri unapređenju tih procesa, te da se u budućnosti ova metodologija uključi kao formalni alat u cjelokupno implementiranom QMS (Quality Management Sistem) sistemu BH Telecoma.

U istraživanju su obuhvaćeni procesi u prodaji paket usluge kod čijih je outputa postojalo najviše pritužbi krajnjih korisnika ili narednog korisnika tog procesa. Naime, radi se o usluzi kojom se GSM korisnicima pružaju usluge govora, prenosa podataka, i terminalne opreme.

Za potrebe istraživanja u primjeni FMEA metodologije angažovano je i obučeno 16 zaposlenika iz procesa prodaje i postprodajne podrške. Također, u procesu davanja prijedloga i procjene uspješnosti provođenja datih prijedloga učestvovali su i rukovodioci organizacionih cjelina u kojim se odvijaju posmatrani procesi.

Usaglašena je metodologija vrijednovanja kriterija i faktora u FMEA obrascima te ostavljena sloboda svim uljučenim za definisanje prijedloga mjera za poboljšanje.

Nakon unesenih podataka u FMEA obrasce i dostavljanja istih, izvršeno je kvantitativno vrednovanje unesenih podataka (S, O, D, RPN i CRIT) te je potom izvršeno usaglašavanje i konačno definisanje onih prijedloga i mjera koji su imali učestalost veću od 75%. Za kvantitativno vrednovanje korišteni su interni izvori podataka, statistički izvještaji te komparacija različitih internih baza podataka i mjesečnih izvještaja.

Za potrebe utvrđivanja ranga važnosti, vjerovatnoće pojavljivanja i vjerovatnoće otkrivanja mnoge kompanije i organizacije koriste razvijene kriterije (Tabela 2.) [4].

Analiza aktivnosti koje se realiziraju u procesima prodaje i podrške paket usluge treba da odgovore na osnovna pitanja efikasnosti tih procesa i njihovim efektima kod krajnjih korisnika. Analiza će obuhvatiti sve aktivnosti koje se direktno odnose na procese u koje su uključeni organizacioni dijelovi prodaje i podrške paket usluge a za koje je analizom potvrđena mogućnost unapređenja. Iako se određene aktivnosti vezane za paket usluge obavljaju i u drugim organizacionim dijelovima, tj. van okvira djelovanja cjelina prodaje i podrške, njihovo uključivanje koristit će se samo u slučajevima gdje su uposlenici iz prodaje i podrške naredni korisnici u lancu realizacije paket usluge.

U radu će se primjenom FMEA metodologije pokušati ukazati na mogućnosti unaprjeđenja procesa i aktivnosti za koje je praćenjem i analizom uočena potreba za unaprijeđenje, a sve s ciljem povećanja sveukupnog kvaliteta paket usluga.

Proces osmišljavanja paket usluga, različitih profila i sadržaja unutar ove usluge jedna je od bitnih faza stvaranja i prodaje paket usluge. Uključenost svih resursa (znanje, tehnologija, organizacija i dr.) preduslov je za uspješnost ove faze. U tom kontekstu rezultati istraživanja treba da posluže kao ulazni parametri u procesu kreiranja novih i unaprjeđenja postojećih usluga i servisa, a sve kako bi usluga bila što atraktivnija i pouzdanija. Dodatno, rezultati istraživanja i provedenih posmatranja poslužiti će kod definisanja potrebnih podataka i informacija u internim bazama podataka. Ažuriranje baza podataka kroz poboljšanje ažurnosti, za vanjske i interne korisnike preduslov je normalnog odvijanja procesa prodaje paket usluge.

Prijem i kontrola terminalne opreme koja ulazi u sastav paket usluge, aktivnost je koja mora biti u stanju da, pored kvantitativnog, obezbjedi i kvalitativni aspekt kontrole isporučene opreme. Osiguranju potrebnog kvaliteta, pored kontrole terminalne opreme, doprinosi i njeno adekvatno skladištenje, te će i ova aktivnost biti predmetom analize.

Upravljanje distribucijom terminalne opreme i ostalih komponenti potrebnih za pružanje cjelovite paket usluge, također, su značajne aktivnosti u cjelokupnom toku realizacije paket usluge. Optimalna iskorištenost raspoloživih kapaciteta (ljudskih, materijalnih i dr.) u stanju je da efikasno upravlja zalihama u cilju njihove raspoloživosti u skladu sa potražnjom.

Analizom svih prikupljenih podataka o reklamiranim slučajevima, obavljenih razgovora sa zaposlenicima zaduženim za različite aktivnosti unutar više odvojenih procesa, obilaskom svih prostorija za skladištenje, prostorija za prijem i otpremu terminalne opreme u procesu distribucije, pregledom transportnih sredstava, analizom procesa i samog čina prodaje terminalne opreme u sklopu paketa i dr., moguće je izolovati slabe tačke u svakom od odvojenih procesa a koji su u funkciji konačne isporuke terminalne opreme korisniku zajedno sa pripadajućom uslugom.

Način i mjere za otklanjanje nedostataka koji mogu direktno ili indirektno umanjiti kvalitet i funkcionalnost terminalne opreme, a samim time i kvalitet i funkcionalnost cjelokupnog paketa prikazati će se kroz FMEA metodologiju upravljanja kvalitetom, zajedno sa svim ostalim aspektima upravljanja kvalitetom.
Smanjenje RPN i CRIT faktora ostvaruju se kroz korekcije na izolovana procesa i više pratećih aktivnosti unutar procesa, a koji su uključeni u plasman paket usluge. Ti procesi su:

1. Prijem i kontrola prijema terminalne opreme,
2. Distribucija terminalne opreme u sklopu paket usluge,
3. Prodaja paket usluge,
4. Reklamacije i žalbe na kvalitet paket usluge.

Nakon opisa procesa koji su uključeni zaposlenici analizirali, u FMEA obrazac evidentirane su sve varijacije potencijalnih neuspjeha, a potom i efekti i učinci svakog neuspjeha. Nakon ovoga, pristupilo se izdvajanju i nabrajanju mogućih uzroka neuspjeha procesa, te mjere kontrole koje se trenutno obavljaju nad procesom ili aktivnostima unutar procesa. U skladu sa opisanom metodologijom određenim kategorijama dodjeljene su vrijednosti na osnovu kojih su izračunati RPN i CRIT faktor. U procjeni i kvantitativnom vrednovanju kriterija S, O i D uposlnici su se koristili izvještajima o reklamacijama, interim evidencijama o prijavama nedostataka i grešaka, izvještajima od strane komisije za kontrolu prijema terminalne opreme, anketama učesnika posmatranih procesa i njhovim kvantitativno/kvalitativnim procjenama i projekcijama. Izbor i preporuka korektivnih akcija i odgovornih zaposlenika treba da rezultira poboljšanjem kvaliteta kako samog procesa/aktivnosti, tako i njegovog *outputa.* Nakon što su

TABLICA II. KRITERIJI ZA ODREĐIVANJE RANGA VAŽNOSTI, VJEROVATNOĆE POJAVLJIVANJA I OTKRIVANJA

Rang	Kriterij		
	Važnost (S)	Vjerovatnoća pojavljivanja (O)	Vjerovatnoća otkrivanja (D)
1	Korisnici ne uočavaju grešku	Vrlo nevjerovatna. greška se pojavljuje manje od jednom u milion i pol slučajeva.	Gotovo sigurno otkrivanje greške.
2	Neki korisnici primjećuju grešku. Vrlo mali uticaj na uslugu ili proces.	Izuzetno rijetka. Jednom u 150.000 slučajeva.	Odlične prilike za otkrivanje greške: 99,99%
3	Većina kupaca primječuje grešku. Mali uticaj na uslugu ili proces.	Rijetko. Jednom u 15.000 slučajeva.	Vrlo viska vjerovatnoća otkrivanja greške: 99,9%
4	Korisnici blago uznemireni. Usluga ili proces pokazuje male nedostatke.	Nekoliko. Jednom u 2.000 slučajeva.	Visoka vjerovatnoća otkrivanja greške: 95%
5	Korisnici uznemireni. Nekritične komponente usluge ili procesa neadekvatne.	Povremeno. Jednom u 500 slučajeva.	Umjerena vjerovatnoća otkrivanja greške: 80%
6	Korisnici doživljavaju neugodnosti. Nekritične komponente usluge ili procesa neadekvatne.	Često. Jednom u 100 slučajeva.	Moguće je otkriti grešku: 50%
7	Korisnici prilično nezadovoljni. Djelimična nefunkcionalnost nekritičnih komponenti usluge ili procesa.	Vrlo često. Jednom u 20 slučajeva.	Mala vjerovatnoća otkrivanja greške: 20%
8	Korisnici vrlo nezadovoljni. Usluga ili proces nije u skladu sa očekivanjima.	Ponavljano. Jednom u 10 slučajeva.	Vrlo nevjerovatno otkrivanje greške: 10%
9	Sigurnost korisnika i usklađenost sa regulativom ugrožena, sa upozorenjem.	Uobičajeno. Jednom u 3 slučaja	Izrazito niska vjerovatnoća otkrivanja greške: 5%
10	Katastrofalni rezultati. Sigurnost korisnika ili usklađenost sa regulativom ugrožena, bez upozorenja.	Gotovo sigurno. Više od jednom u dva slučaja.	Gotovo sigurno da greška neće biti otkrivena ili ne postoje kontrolni mehanizmi koji mogu identifikovati grešku.

stepenu važnosti (S), vjerovatnoći pojavljivanja (O) i vjerovatnoći otkrivanja (D).

IV. REZULTATI ISTRAŽIVANJA

Na osnovu rezultata dobivenih tokom provedenog istraživanja i unesenih parametara u FMEA obrasce, odabrani su procesi kod kojih su uočene greške i nedostaci sa najvećim rizikom. Shodno rečenome, rezultati obavljenog istraživanja i primjene FMEA metodologije odredili su procese i aktivnosti koji trebaju biti predmetom unaprjeđenja i poboljšanja. Radi se o četiri evidentirane sve nabrojane mjere vrijednosti, izvršen je izračun ili procjena novih vrijednosti RPN i CRIT faktora.

Detaljnim posmatranjem i analizom toka realizacije prijema i transporta mobilnih uređaja, uočeni su brojni nedostaci čije otklanjanje ili minimiziranje može uticati na značajno smanjenje broja reklamacija na prodate mobilne uređaje i ostalu terminalnu opremu. Izmjerene vrijednosti su prikazane u Tabeli broj 3.

Nakon provedenih posmatranja i analiza, od strane zaposlenika koji su učestvovali u istraživanju, dati su prijedlozi mjera i urađena je procjena rezultata i očekivanja od primjene

predloženih mjera. Rezultati proračuna tih procjena i očekivanja dati su u Tabeli broj 4.

TABLICA III. IZMJERENE VRIJEDNOSTI POSMATRANIH PROCESA

Proces redni broj:	**1.**	**2.**	**3.**	**4.**
Važnost (S)	8	6	5	8
Vjerovatnoća pojavljivanja (O)	7	4	8	6
Vjerovatnoća otkrivanja (D)	7	7	6	7
RPN	392	168	240	336
CRIT	56	24	40	48

TABLICA IV. IZRAČUN VRIJEDNOSTI PREDLOŽENIH MJERA

Proces redni broj:	**1.**	**2.**	**3.**	**4.**
Važnost (S)	4	4	2	4
Vjerovatnoća pojavljivanja (O)	5	2	4	4
Vjerovatnoća otkrivanja (D)	6	4	4	5
RPN	120	32	32	80
CRIT	20	8	8	16

Analiza uočenih nedostataka u procesima i prijedlozi mjera za unaprjeđenje prikazanih numeričkim vrijednostima u tabelama 3. i 4. u svom opisnom dijelu rezimirana su kroz opisna opažanja, koja iznosimo u nastavku rada.

Prijem i kontrola prijema terminalne opreme - Kontrolom procesa i analizom prikupljenih podataka konstatovano je sljedeće:

- Potencijalni neuspjeh procesa ogleda se u mogućnosti prijema neispravnih ili uređaja sa skrivenim nedostacima te puštanje takvog uređaja u prodaju,
- Potencijalni učinci neuspjeha ogledaju se u nepotrebno velikom broju reklamacija na prodate uređaje i kroz to nezadovoljstvo korisnika,
- Mogući uzroci su tvornička greška na uređaju, neadekvatno skladištenje i rukovanje u transportu ili skladištu, neadekvatna kontrola kod dobavljača i sl.,
- Trenutne kontrole procesa se vrše na način da se uzorkovanje i kontrola po ugovoru sa 3 ili 1% zavisno od dobavljača, te kontrolom nakon otkrivanja grešaka,
- Preporučene akcije su kontrola prijema i uzorkovanje po sistemu ANSI/ASQC Z1.4.(jednostruko uzorkovanje). Vođenje evidencije i zapisa o kontroli, vođenje evidencije i zapisa o izmjerenim vrijednostima vlage na ambalaži i samom uređaju,
- Odgovorni uposlenici za provođenje preporučenih mjera su rukovodioci u prodaji, komisija za prijem, magacinsko osoblje i ostali uposlenici po potrebi.

Distribucija terminalne opreme u sklopu paket usluge - Kontrolom procesa i analizom prikupljenih podataka konstatovano je sljedeće:

- Potencijalni neuspjeh procesa ogleda se u mogućnosti nastanaka oštećenja na terminalnoj opremi u procesu distribucije, sporoj i neefikasnoj distribuciji i redistribuciji terminalne opreme, neefikasnom korištenju ljudskih i materijalno-tehničkih sredstava,
- Potencijalni učinci neuspjeha ogledaju se u oštećenjima i nastajanju kvarova na terminalnoj opremi, kašnjenju sa isporukom terminalne opreme, nepotrebno i neopravdano visokim troškovima distribucije i redistribucije,
- Kao mogući uzroci su otkriveni neadekvatan utovar i istovar, neprilagođena i neadekvatna transportna sredstva, postojeća organizacija poslova disrtibucije, angažman nepotrebnog broja uposlenika i odvajanje tih zaposlenika od redovnih radnih zadataka, troškovi putovanja (dnevnice i dr.), amortizacija vozila, troškovi goriva i sl.,
- Trenutne kontrole procesa se vrše na način da se vrši kvantitativna kontrola prijema/predaje distribuiranih asortimana, organizacija i upravljanje procesom distribucije vrši se na nivou Direkcije za prodaju i podršku i Regionalnih direkcija,
- Preporučene akcije realizovati kroz obezbjeđenje adekvatnih transportnih sredstava, da se lokacija utovara i istovara terminalne opreme obezbjedi u zatvorenim prostorijama bez direktnog atmosferskog uticaja (kiša, snijeg i sl.), organizaciju centralizovane distribucije iz Centralnog skladišta do priručnih sladišta regionalnih direkcija i istovremenu redistribuciju te obezbjediti aktivno upravljanje distribucijom,
- Odgovorni uposlenici za provođenje preporučenih mjera su rukovodioci u prodaji i distribuciji te magacinsko osoblje.

Prodaja paket usluge - Kontrolom procesa i analizom prikupljenih podataka konstatovano je sljedeće:

- Potencijalni neuspjeh procesa ogleda se u tome što se prodaja asemblirane usluge uglavnom odvija na način da korisnik ne dobije dovoljno informacija, kao i da korisniku nije prikladno i u potpunosti prezentirana terminalna oprema koja mu se prodaje,
- Potencijalni učinci neuspjeha ogledaju se u čestim prigovorima i žalbama korisnika da nisu dobili sve ili neke od informacija vezanih za terminalnu opremu,
- Mogući uzrok je da su šalterski radnici nedovoljno stručni i obučeni da daju sve informacije koje korisnici očekuju a odnose se na specifikaciju i funkcionalne karakteristike TO, da šalterska radna mjesta u tradicionalnim (poštanskim) šalter salama ne obezbjeđuju uslove za kvalitetnu prezentaciju terminalne opreme, da je starosna struktura zaposlenih u prodaji neadekvatna, te da je prisutan nesistematičan i nedovoljno definisan sistem obuke prodajnog osoblja po pitanju upoznavanja sa traženim specifikacijama terminalne opreme,
- Trenutne kontrole procesa se vrše kroz povremene i izolovane pojedinačne kontrole na prodajnim mjestima,
- Preporučene akcije podrazumjevaju intezivniju i obaveznu obuku zaposlenika sa ciljem upoznavanja osnova korištenja naprednih GSM servisa i aplikacija, inteziviranjem koncepta prodaje putem Telekom centara kao ambijenta koji omogućava i adekvatnu prodaju TO, obavezno propisanu konačnu kontrolu terminalne opreme i njenu kompatibilnost sa korisničkim paketom u momentu prodaje paketa i prije samog uručenja iste korisniku,
- Odgovorni uposlenici za provođenje preporučenih mjera su Top menadžment, menadžment prodaje i podrške, rukovodioci i zaposlenici u prodaji.

Reklamacije i žalbe na kvalitet paket usluge - Kontrolom procesa i analizom prikupljenih podataka konstatovano je sljedeće:

- Potencijalni neuspjeh procesa ogleda se u nepotrebnom i, za korisnika, dugom periodu rješavanja reklamacija, ne rješavanje reklamacije u skladu sa očekivanjima korisnika, višestruko pojavljivanje istih kvarova na TO, poništenje garancija na terminalnu opremu i sl.,
- Potencijalni učinci neuspjeha ogledaju se u nezadovoljstvu korisnika, neizvjesnoj ponovnoj kupovini, troškovima reklamacija, sudskim troškovima, gubitcima kroz izostalo korištenje GSM usluga za vrijeme trajanje reklamacije,

- Mogući uzroci su nekvalitetan prijem i obrada reklamacija, nepostojanje fleksibilnih i alternativnih rješenja i opcija, nekvalitetna servisna podrška, neadekvatan pritisak na servisne centre da kvalitetnije pristupaju rješavanju reklamacija, neupitna i bezuslovna konačna ocjena i nalaz servisnog centra a koja je uzrokovana neusklađenošću interesa telekom operatora i servisnih centara,
- Trenutne kontrole procesa se vrše po proceduri za rješavanje reklamacija TO, kontrolu kroz aplikaciju za reklamacije TO, te kontrolu obavljenu u servisnim centrima,
- Preporučene akcije podrazumjevaju obezbjeđenje mehanizama koji će omogućiti zamjenu ili popravak terminalne opreme o trošku telekom operatora tamo gdje ima interesa i gdje postoji opravdana dilema da korisnik nije kriv za nastali kvar, istrajavanje na konstantnom pritisku na servisne centre da otklanjanju kvarova na TO posvete više pažnje, integrisanje aplikacije za reklamacije u jedinstveni aplikativni sistem (korisnička baza - produkti),
- Odgovorni uposlenici za provođenje preporučenih mjera su menadžment za razvoj poslovanja, menadžment prodaje i podrške, menadžment EFP-a, i drugi po potrebi.

V. ZAKLJUČNA RAZMATRANJA

Upravljanje kvalitetom telekomunikacijskih usluga treba da obezbjedi brojne koristi, a njihov uticaj se, prije svega, treba posmatrati kroz uticaj kvaliteta na zadovoljstvo korisnika. Za razliku od upravljanja kvalitetom u procesima proizvodnje materijalnih dobara (proizvoda i roba), upravljanje kvalitetom usluga je mnogo suptilnije i zahtjeva intezivan monitoring svih procesa/aktivnosti integralno.

Praksa je pokazala da niti jedan implementirani QMS sistem nije održiv dugoročno i da zahtjeva stalnu nadogradnju kroz kritičko preispitivanje postojećeg sistema rada. Ukoliko se uzme u obzir jasno izražena dinamika promjena koja je prisutna u telekomunikacijskom sektoru, onda je jasno da je neminovna i intezivna primjena različitih metoda, modela i alata za upravljanje kvalitetom, sve šireg spektra, usluga.

Prezentirani primjer upravljanja kvalitetom telekomunikacijskih usluga primjenom FMEA metodologije pokazuje da je on itekako primjenjiv i u domenu usluga. Prisutni su brojni primjeri primjene ove metode u uslužnim i trgovinskim kompanijama [6]. Ono bitno što se izdvaja nakon primjene ove metodologije jeste to da se pokazalo da zaposlenici koji su aktivni učesnici i izvršioci procesa mogu biti izvrsni analitičari koji su sposobni da, ukoliko im se u ruke da adekvatan alat, prepoznaju manjkavosti procesa i značajno unaprijede procese u kojima učestvuju.

Kao konkretne zaključke izdvojit ćemo i sljedeća opažanja i mjere za unaprjeđenje, i to kako slijedi:

- Prijem i kontrolu prijema terminalne opreme potrebno je organizovati na način koji obezbjeđuje veću pouzdanost kod otkrivanja aparata sa greškom. Prijedlog za kontrolu je organizacija iste po sistemu ANSI/ASQC Z1.4. po principu jednostrukog uzorkovanja, normalnog nivoa kontrole kvaliteta. Izmjenom postojećih ugovora, predvidjeti kontrolu u skladištima primaoca opreme a ne u skladištima dobavljača.
- Disrtibuciju i redistribuciju terminalne opreme organizovati centralizovano, tj. na nivou cjelokupnog telekom operatora, gdje bi se sva distribucija i redistribucija obavljala iz jedne polazne tačke tj. od centralnog skladišta prema svim regionalnim i priručnim skladištima.
- Prodaju terminalne opreme u sklopu paket usluge organizovati na način da se aplikativna finalizacija prodaje obavlja tek nakon što je korisnik odabrao, pregledao i isprobao željeni aparat, te na otpremnici predvidjeti formu kojom korisnik i potvrđuje da je to tim redoslijedom izvršeno. Tek nakon ovih radnji i potpisivanja dokumentacije korisniku uručivati aparat uz ovjeru garancije. Na takav način u značajnoj mjeri bi se obezbjedilo: dodatna (prijeko potrebna finalna) kontrola terminalne opreme prije uručivanja iste korisniku, smanjenje broja slučajeva u kojima korisnik nakon završene prodaje (potpisivanja dokumenacije i ovjere garancije) traži povrat aparata sa obrazloženjem da nije želio taj model i sl.
- Putem nadležnih organizacionih dijelova telekom operatora potrebno je jasno definisati programe obuka i edukacija prodajnog i osoblja u podršci paket usluge, kako bi se i u tom segmentu smanjile mogućnosti za nastanak situacija u kojem navedeno osoblje nije u mogućnosti da na zahtjev korisnika ponudi potpunu i kvalitetnu informaciju ili još važnije, da ne ponudi netačnu informaciju korisniku što direktno proizvodi nezadovoljstvo kod korisnika paket usluge.
- Obezbjediti sistem podnošenja i evidentiranja svih vrsta pritužbi korisnika na način da sam proces prijave i evidencije pritužbe bude stimulirajući i lagodan za korisnika, a sve kako bi se obezbjedile dragocjene informacije o potencijalnim propustima i manjkavostima unutar procesa u pružanju paket usluga.
- Podršku za monitoring reklamacija na terminalnu opremu ostvariti integracijom postojećih aplikativnih rješenja za evidenciju i praćenje reklamacije u jedinstven aplikatini sistem, a koji će obezbjeđivati sve neophodne informacije za adekvatno upravljanje procesom reklamacija korisnika.
- Prije konačne implementacije predloženih mjera za poboljšanje kvaliteta paket usluge potrebno je formirati radni tim za izradu i testiranje svih funkcionalnosti procesa i njihovih *outputa*. Tek nakon tih aktivnosti pristupiti donošenju konačne odluke za njihovu primjenu.
- Detaljno razmotriti mogućnost implementacije garancije prodavca na isporučenu terminalnu opremu, tj. garancije koju bi telekom operatori nudili svojim korisnicima, a u skladu sa iskustvima kompanije *Asurion* iz SAD-a. Naravno, tačni uslovi pod kojima bi se vršilo osiguranje i reosiguranje trebaju biti predmetom opsežnih razgovora i pregovora sa osiguravajućim kompanijama.
- Ukrupnjavanjem pojedinačnih procesa, te jasnim definisanjem vlasnika svakog procesa (ime i prezime zaposlenika) omogućiti efikasniji monitoring, kontrolu i realizaciju svih aktivnosti i zadataka cjelokupnog niza procesa u prodaji paket usluge.

LITERATURA

[1] Lazibat T., Upravljanje kvalitetom, Znanstvena knjiga, Zagreb, 2009. god., str. 360.

[2] Stamatis D.H., Failure mode and effect analysis: FMEA from theory to execution, ASQ Quality Press, Milwaukee, WI, 2003.,str. 15.

[3] Tague N.R., The Qualty Toolbox, American Society for Quality, ASQ Quality Press, Milwaukee, WI, 2005.,str. 234.

[4] Tague, N.R., The Quality Toolbox, ASQ Quality Press, Milwaukee, WI, 2005., str. 240.

[5] Kilibarda M., Upravljanje kvalitetom u logistici, Saobraćajni fakultet Beograd, Beograd 2006.

[6] Doborović T., Primjena metode analize utjecaja i posljedica pogrešaka u vanjskotrgovnskom poduzeću, Magistarski rad, Ekonomski fakultet u Zagrebu, Zagreb 2004.

[7] http://www.asurion.com/web/asurion/handset-protection 01/2013

[8] http://www.efzg.hr/dok//TRG/tbakovic/Alati/metode_zupravljanje_kval-V2.ppt 04/2012

SOFTWARE ENGINEERING & TOOLS

Experiences from building a EUD business portal

Nikola Tanković*, Tihana Galinac Grbac**, and Mario Žagar***
* Superius d.o.o., Pula, Croatia
** Faculty of Engineering, University of Rijeka, Croatia
*** Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia
nikola.tankovic@superius.hr, tgalinac@riteh.hr, mario.zagar@fer.hr

Abstract - End user development (EUD) is the idea of providing end users, professionals outside computer science community, means to develop their own software. This idea has received special attention by domain experts who would like to easily modify software applications to their needs, without intervention of technical professionals and without learning conventional programming languages. The end user development is in fact a system property, closely related to system adaptability that should be built into a system to enable number of modifications at different system levels securing powerful expressions that would satisfy the end user needs. This is a challenging requirement for every software provider. There are a number of identified system perspectives, critical to EUD, and general guiding principles that should be addressed. However, in reality, building these properties into the concrete system, a number of complex and interrelated issues have to be solved. In this paper we report challenges that we addressed while implementing end user development capability into a software system within business domain.

I. INTRODUCTION

End User Development is focused around concept of providing means and tools for non-professional developers to participate in software development. Lieberman at el. give the following definition [1]: "End-User Development can be defined as a set of methods, techniques, and tools that allow user of software systems, who are acting as non-professional software developers, at some point to create, modify or extend a software artifact."

Spreadsheet is the most popular and well-known EUD tool, but there is an increase in the number of solutions such as scripting programming languages or graphical mashup tools that combine existing functionality into composite applications. The main reason for the popularity of EUD is the cost and time expense in traditional on-demand software development that many organizations face. The downside is the lack of quality in terms of reliability and security, especially in mission critical computing.

Our motivation for building a EUD business portal is to enable our customers with a tool for creating applications that implement simple business processes yet uncovered by their existing IS solutions. Such implementations are commonly non-critical with lower return of investment. This decision was made after a decade of software development experience in business domain designing and delivering various customer requests.

Building an end-user development system from technical perspective means combining and applying numerous existing software engineering concepts that are offered to improve the software lifecycle process. Usually the ideas behind these concepts are clear however their implementation is full of traps and obstacles.

Here we will elaborate on one particular case study how some of the most famous software engineering concepts are implemented and we elaborate the problems that have to be solved. The following concepts are addressed: automated customization of products to customers, automated production process, automatized delivery process and web oriented end user development process.

II. CONTEXT OF THE STUDY

We illustrate the concepts of this paper based on an example of software development in the supply chain management (SCM) domain. More precisely, the focus of this paper will be the software for supporting the continuous control of trading contracts. Trading companies are businesses usually selling different kinds of products to consumers. The products are ordered from different producers, offered in a retail stores and resold to buyers. One typical example is food shop. The relation between trading company and producer is defined by trading agreement. This contract includes metrics such as minimal number of products to be exposed on shelves, product positioning inside shop or minimum stock per product to be ensured. Contracts are subjects of control and future management that is performed by field agents dispatched by producers themselves. They ensure that for each of their products being sold, metrics agreed in contracts are being fulfilled across retail stores. In sequel, the software for supporting agents at managing trading contracts will be referred as Manage Trade software.

The company that is developing the Manage Trade software has long tradition and numerous customers. Our customers are before mentioned producers and distributors of goods. The core software system has been evolving for more than 10 years. A numerous improvements have been introduced in the software system. Here we will mention just few that have implications for our future elaborations.

1. When contract is signed between a producer and trading company, a number of agents are engaged to control and manage the contract execution. For that purpose, a number of software applications - one per agent - are developed for the purpose of one specific contract. Each software application has implemented one smaller derivation from the contract. For that purpose we introduced a feature of automated customization in our Manage Trade software system. That feature enables to automatically derive a number of customized software applications for every different contract.

2. Since we have many customers we have identified similarities in the agreements and contract management processes. A logical step was to automate software production with maximal reuse of existing software. Therefore, we developed our Domain Specific Language (DSL), a script language that describes the differences introduced with each contract by inheriting from common predefined constructs. That way all the specifics in applications for different customers are expressed easily.

3. We developed a Web based system coordinating and supplying data to applications from all agents by using a single server database instance for multiple contracts and producers.

4. Further step in improving our business was to automate the software delivery process. In that aim we implemented a Web based system that enabled Web based delivery of software applications to the agents of specific contract.

5. Our next step in improving our business is empowering our customers to develop their contract support system and thus reuse the concepts of the End user development (EUD) paradigm. A specific feature would be introduced in our Web based system to enable end user development. This feature we will refer as EUD portal.

III. EUD PORTAL – CONCEPT AND DEFINITIONS

EUD portal is a Web based system for definition, automatic construction and delivery of software applications in a certain domain of business. We observe and describe EUD portal use cases from different perspectives:

1) From *application development perspective* it is a web platform for modeling software applications using a provided DSL language. Resulting software applications

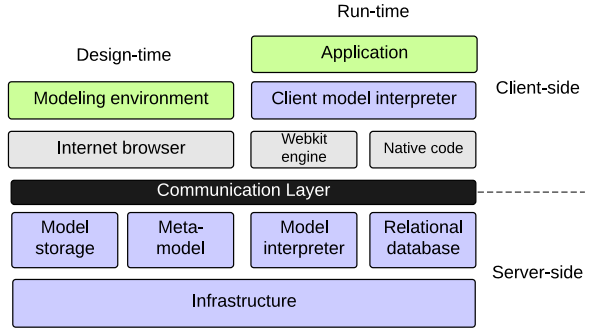


Figure 1. EUD portal by layers

are limited to elements and expressive level of DSL provided. In our case, we provide adequate DSL for Manage Trade domain where producers create applications to support their agents at work. The targeted end-user for this use case is a business expert who usually has some experience and expertise in simple programming such as creating Excel documents or scripting simple programs. Description of end-user programming interface used is out of the scope for this study and is still in development stage. We address only technical underlying challenges of building a portal with these properties.

2) From *runtime perspective*, the execution of Manage Trade applications, EUD portal leverages Software as a Service (SaaS) delivery model [2]. Each agent uses an application delivered and managed by EUD portal.

A. Application development perspective

Contracts between producers and retail shops vary in content. Agreements in these contracts evolve over time meaning that agent applications change. In addition, some producers also required additional functionality such as sales order creation or collecting additional data for their business intelligence systems. Given all that, our portal needs a lot of expressive power. To satisfy such a large variety of requirements, we pursued a model-driven approach, and defined a common AGM Meta-model typical for our tenants requirements presented at [3]. AGM meta-model possesses expressive power to model applications that can process a variety of structured data, business processes and validation rules to minimize data collection errors. AGM was defined from years of experience in SME business domains such as Manage Trade. The guidelines for defining a solid meta-model are provided at [4].

Figure 1 illustrates that design-time, or sometimes referred to as modeling-time, together with run-time environments reside on the same web infrastructure. Model storage is a specialized graph database Neo4J[1] containing end-user defined models. It is used by EUD portal as a central model repository of every modeled application, and for storing meta-model elements and definitions. Models are constructed using before mentioned DSL language for Manage Trade domain inside modeling environment provided by portal. This web-based environment is at a very basic and early stage, providing only a simple DSL editor to input or modify application definition. This area is a subject to future improvement with introduction of graphical modeling environment for a more user friendlier approach [5].

[1] Neo4J – open-source graph database, available at http://www.neo4j.org

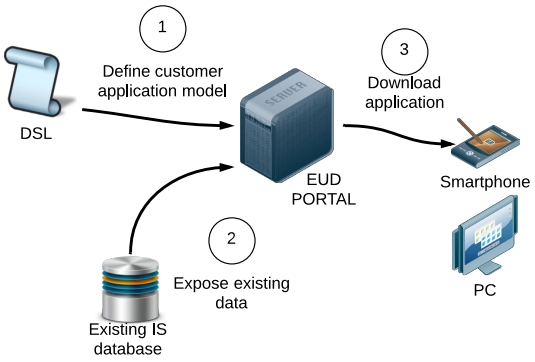


Figure 2. Basic steps involved in creating application

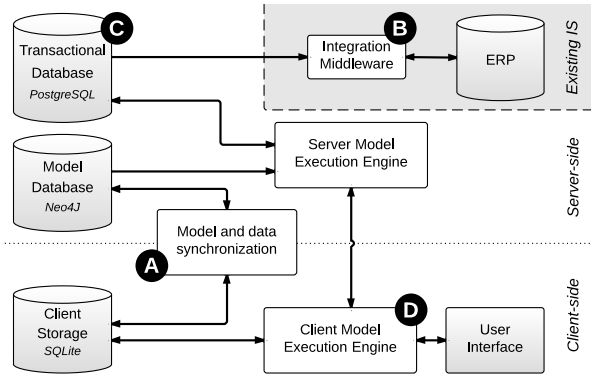


Figure 3. EUD portal architecture

Overview of complete application creation scenario is shown in Figure 2. Steps are simplified omitting ones like tenant registration or management of users. It gives the high-level picture of three-step application development:

1) In first step business expert in the role of end-user programmer models the application using the interface provided from EUD portal. This modeling step includes definition of data structure, data transactions that will be carried by agents, and a model of user interface. Transactions are modeled as a set of data collection forms. In Manage Trade domain agents use these forms to collect information that is analyzed to see if contracts are being met. Therefore, these forms are modeled according to contract contents.

2) Each organization exposes data from their existing IS through an ODBC [2] database connector or uses connectors to SAP ERP[3] or Microsoft Dynamics NAV[4] systems. The tools are provided by portal. This step is often completed with some help of professional developer or system integrator who has a deeper knowledge of SQL[5] language and database procedural languages. The structure of exported data in this step must conform to data structure defined in previous step.

3) Final application is automatically assembled and downloaded to each of agent devices. Manual download is required, but future updates are automatic and seamless.

EUD portal does not assemble each application as different executable specific to each agent, but rather provides a common executable for every agent, which is capable of interpreting provided model at runtime. We refer to this executable as a *client engine*. Model is synchronized to each client engine upon first usage and updated as necessary. Model interpretation in favor of generating end applications makes this an interpretative MDD approach. Differences between MDD approaches and benefits of each, together with rationale for choosing interpretative approach are described in [6].

B. *Runtime perspective*

1) From runtime perspective, EUD portal is a distributed software system composed of server and client components. Server-side components provide data storage services for assembled applications. Each agent continually collects information at retail stores that is stored in portal. This is achieved using a relational database. Single database instance is shared between agents and data is isolated accross companies.

To support data storage for multiple agents and organizations in a single database, we used a multi-tenancy design of the portal and underlying relational database. This means that each customer shares hardware and software resources leading to better resource utilization. A *shared-table design* [7], [8] is used to isolate data inside relational database instance. In our implementation, a single tenant correlates to single customer – producer or distribution organization.

EUD portal also possesses Web services and tools that provide integration with existing IS. In our business case, agents require immediate access to different information contained within existing Customer Relationship Management (CRM) and Material Management (MM) systems. Such information includes details about retail shops and products. Each agent also receives daily work schedules provided from these existing systems. They define a list of retail stores to visit and ensure contracts are being followed correctly. This work schedules are synchronized from outer systems to EUD portals central database and then further passed to each agent application.

Client side components are packaged as a hybrid mobile and desktop application [9]. Hybrid application composing means that applications are built using web technologies: JavaScript and HTML, combined with native modules developed separately for each platform we provide: PC, Android and iPhone. These native modules

[2] ODBC – stardadized middleware for accessing database management systems, originally developed by Microsoft

[3] SAP – SAP AG's Enterprise Resource Planning software, available at http://www.sap.com

[4] Dynamics NAV – ERP software product from Microsoft, available at http://www.microsoft.com/en-us/dynamics

[5] SQL – special domain specific language for managing data contoained in database management systems

are referred as *native wrappers*. They use Webkit[6] open-source engine to display HTML and JavaScript contents. That way we obtain functionality specific to native applications and still benefit from web technologies that facilitate software delivery and multiplatform execution [10]. Native functionality includes: access to file-system, photo camera, gyroscope, GPS module, device information and phone address book, depending on platform.

IV. CHALLENGES AND FUTURE DIRECTIONS

Many technical challenges arose during implementation phase. Those will be described in following subsections, each section covering different technical aspect of EUD portal. To position each problem according to EUD portal layer or component, we relate to Figure 1, which provides an overview on portal run-time and design-time layers, and Figure 3, which shows the main components used.

A. Communication between server and clients

As shown in Figure 3, labeled (A), data synchronization between portal and end applications is achieved with a distributed synchronization module residing both on server and client side. This empowers each client application to conduct computation over data independently, which unburdens server infrastructure resources. This is also very useful for supporting *offline* data processing, meaning that agent applications can process transactions on smartphones without data network, and postpone synchronization of completed transaction back to server after connection is reestablished. Modeling language provides expressions to filter data that is to be synchronized on each agent application.

It was quite difficult to achieve a stable synchronization mechanism given that resulting applications were designed by end-users so the synchronized data volume was often unpredictably large. This problem is even more present when data is sent over old generation mobile networks.

To resolve this issue, we are researching a mechanism that will predict the volume of data at run-time and warn the developers when such volume hits a certain threshold. That way, end-users can decide to lower the volume of data by filtering some less important data.

B. Integration

A key part of any business system is interoperability - efficient cooperation with existing systems. This is in our experience one of the greatest barriers for using EUD in business organizations. Many existing, on-premises IT systems are inaccessible or difficult to integrate with. A most common method of integration in our experience is by using intermediate tools to connect to their database systems. In Figure 3 this tools are noted (B). It can be observed that relational transactional database is connected with *integration middleware* for connecting to existing IS databases. These tools are configurable applications provided to be installed inside existing IS environments. The reason for on-premises installation is to evade strict firewall policies many of our customer posses and their inability to provide web services. Although SOA architectures are getting more common, our customers do not posses such solutions, so data is synchronized with their databases through ODBC drivers and a configurable set of SQL commands.

We already stated that for a stable integration, assistance from professional developer is required and this is quite a drawback for end-user development due to unpredictable costs and time factors when engaging professional help. After the transition to service-oriented architecture this problem will be easier to tackle, as there is many research on integrating web services with EUD principles. Dörner et al. [11] are researching end-user development integration with Service-Oriented Architectures (SOA) by making web services more understandable to end-users. Similar pattern can be applied here. However, our customers are from small and medium business (SMB) sector, and are still just considering migrating to SOA. Direct integration with their databases is still the most flexible and common solution.

C. Persistence Layer

Data persistence, marked (C) in Figure 3 is the most important part of any business transaction processing system and this was where we faced many challenges. When we did our research for choosing the right database to store EUD portal customers data, we had several main requirements:

- ACID properties,
- An open-source solution is preferred,
- Schema-less database is preferred; as it makes storing end-user designed data objects easier to implement.

We have chosen a relational open-source PostgreSQL[7] database. PostgreSQL, like any ordinary relational database, requires schema to be defined prior to usage. We gave up on this aspect in favor of ACID properties and reliability, as schema-less NoSQL[8] databases are still mainly not mature enough for business production usage, like storing important and sensitive data. Main deciding factors for PostgreSQL were durability, robustness and open-source code.

To achieve best hardware resources utilization, we implemented a row level multi-tenancy meaning that

[6] Webkit – open source layout component designed to render web pages, available at http://www.webkit.org/

[7] PostgreSQL – open-source object-relational database management system released under PostgreSQL license, available at http://www.postgresql.org

[8] NoSQL – common names for databases that store data in forms other than tabular relation used in relational databases, many of which do not require a schema to be defined

tenants share same tables with each row having tenant identification.

Since tenants had different schema structures, we had to share tables and columns across tenants. For that reason each column had the same universal variable length string type. This table and column abstraction on database side brought a small performance drawback because the benefits of relational schema were not fully utilized. To satisfy data isolation between tenants, we implemented a layer between the portal application code and database that does the mapping between user-defined schema and real database schema. This layer also brought a small degradation of data access performance. Similar approach has been implemented by the Force.com Salesforce1 solution [12] on top of Oracle relational database. We have not applied enough attention to properly address this issue and further research is required in this area. More research towards implementation of this feature natively inside databases is required, such as [13].

Another big issue was database scaling, ability operate effectively regarding to database size. We did not automatically track each tenant metrics on table sizes nor implement a mechanism for automatic index creation to rectify slow queries. Instead, we performed database optimization and administration manually. Taught by this production experience, we defined a set of characteristics that a fully automated EUD business system persistence layer should have:

1) Mechanism for keeping track of per-tenant meta-data and metrics: slow queries, overall data size, data fragmentation, and index efficiency. Basic solution is provided at Figure 4 where execution time is monitored together with requested query. Collected data is analyzed for decision such as automatic creation of additional indexes, data sharding and partitioning. Similar approach is applied at [14] on cost-driven basis.

2) Data sharding should take care not to divide data inside common tenant and to divide shards evenly. Such mechanism should track fastest growing tenants and reserve enough storage for further growth. If a single tenant gets larger than a single shard (database) it should be further split by relocating less accessed data.

Modern Platform-as-a-Service (PaaS) solutions offer APIs for managing hardware infrastructure. This could also be used for automatic new creation of additional database instances when total capacity is close to being exceeded with automatic data reallocation. An example of such solution is Amazon Auto Scaling[9] but it is limited in terms that it allows only predefined set of rules like defining peek times, or CPU usage thresholds when a scaling should occur. Scaling API should be combined with analysis of existing tenant meta-data and metrics for optimal resource allocation. Combined with a database that easily scales should provide a solid ground point towards building a database layer capable of handling large quantities of different tenant schemes. A good

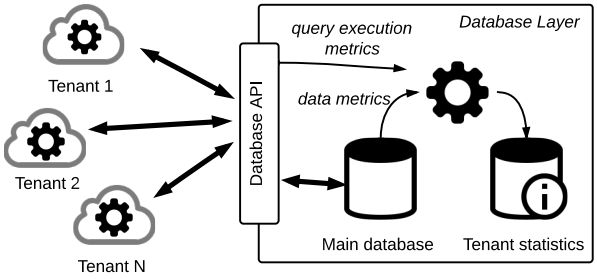


Figure 4. Mechanism for collection tenant data size and query performance metrics

database engines categorization with a special review to scaling is given in [15]. This work shows a promising future for NoSQL schema-less database solutions once they become more reliable.

D. Extending modeling capabilities

Certain transactions conducted by agents' required additional functionality that could not be expressed with our DSL language. Therefore, we provided means for writing simple JavaScript[10] functions extending provided modeling capabilities e.g. specific tax calculations, custom advanced validation controls or other types of more complex algorithms. They were executed on agent applications and can be modeled to execute on certain application events like startup of the application, beginning or end of transaction, periodically, and on different user interface actions.

Using these extension scripts in practice revealed a few additional problems:

1) *Security concerns* – client side JavaScript code is not safe from manipulation, so an additional JavaScript interpreter located at portal should revalidate those client-side calculations upon successful synchronization of data back into the portal database.

2) *Inefficiency* – certain scripts can perform much better on server side, e.g. manipulating a larger set of data. This requires more resources than client side computing can offer.

3) *Poor usability* – to write an extension, one needs some sort of assistance from professional developer who knows JavaScript. This was not suited for an end-user.

To resolve performance issues, an algorithm based on heuristics could predict functions to be executed on server side e.g. function with frequent access to large amount of data.

V. CONCLUSIONS

Implementing EUD portal in the web environment where technology is still not specialized enough for such dynamic use cases offers several technical challenges to surpass. Main concerns are infrastructure barriers, integration issues and scalability.

[9] AAS – Amazon Auto Scaling solution, available at http:// http://aws.amazon.com/autoscaling/

[10] JavaScript

Computational and network infrastructure is limited, but end-users do not possess understanding of those limitations which results in unstable and underperforming applications. Integrating EUD systems with existing solutions is another key barrier for end-users because it is often very complicated even for professional developers to understand and find a way to integrate with closed systems.

Regarding data persistence layer, greatest challenge is achieving multiple schema multi-tenant data storage on classical database management solutions. Having a user-defined data structure imposes a need for higher flexibility to change data structures in runtime. We proposed using a schema-less database or an abstraction layer over relational databases. A mechanism that would gather meta-data and usage statistics for each tenant is also needed to tackle self-optimization. Such mechanism should take decisions and actions over data partitioning, sharding and indexing.

To enable automatic scaling, platforms could be built on IaaS solutions that enable API's for dynamically creating new worker and storage nodes to take on extra traffic. Figure 5 shows an overview on key portal architecture characteristics: 1) multiplatform execution, 2) web-based WYSIWYG modeling environment, and 3) execution engine with ability to use infrastructure APIs for automatic scaling. Self-adaptivity in forms of carefully configured algorithms should take decision to scale by managing storage and computation resources from IaaS provider during runtime. This is crucial for EUD solutions where application resource footprints are highly dynamic and unpredictable.

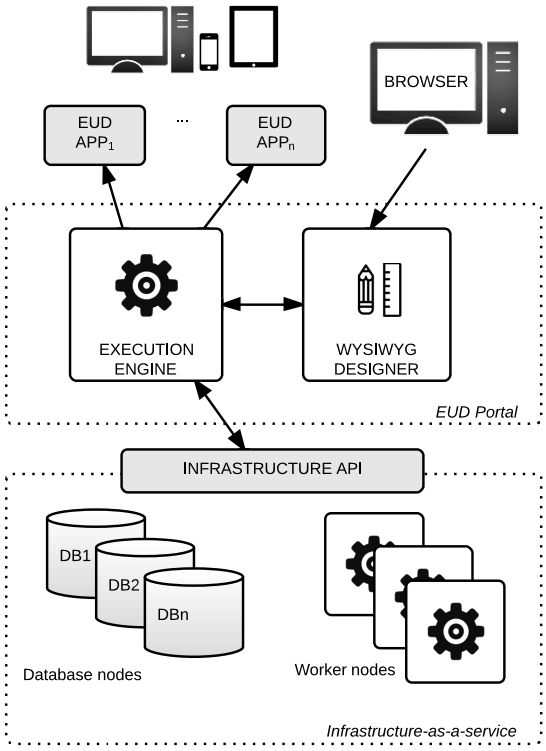


Figure 5. Proposed EUD portal architecture

VI. ACKNOWLEDGEMENTS

This work is supported in part by the Croatian Ministry of Science, Education and Sport, under the research project "Software Engineering in Ubiquitous Computing".

[1] H. Lieberman, F. Paternó, M. Klann, F. Paternò, and V. Wulf, "End-user development: An emerging paradigm," *End User Dev.*, vol. 9, pp. 1–8, 2006.

[2] M. Armbrust, I. Stoica, M. Zaharia, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, and A. Rabkin, "A view of cloud computing," *Commun. ACM*, vol. 53, no. 4, p. 50, Apr. 2010.

[3] N. Tankovic, D. Vukotic, and M. Zagar, *Executable Graph Model for building data-centric applications*. IEEE, 2011, pp. 577–582.

[4] R. Lagerström, J. Saat, and U. Franke, *Enterprise meta modeling methods–combining a stakeholder-oriented and a causality-based approach*, vol. 29. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 381–393.

[5] V. Deufemia, C. D'Souza, and A. Ginige, "Visually modelling data intensive web applications to assist end-user development," in *Proceedings of the 6th International Symposium on Visual Information Communication and Interaction - VINCI '13*, 2013, p. 17.

[6] N. Tankovic, D. Vukotic, and M. Zagar, "Rethinking Model Driven Development: analysis and opportunities." pp. 505–510, 2012.

[7] Z. H. Wang, C. J. Guo, B. Gao, W. Sun, Z. Zhang, and W. H. An, "A Study and Performance Evaluation of the Multi-Tenant Data Tier Design Patterns for Service Oriented Computing," in *2008 IEEE International Conference on e-Business Engineering*, 2008, pp. 94–101.

[8] J. Espadas, A. Molina, G. Jiménez, M. Molina, R. Ramírez, and D. Concha, "A tenant-based resource allocation model for scaling Software-as-a-Service applications over cloud computing infrastructures," *Futur. Gener. Comput. Syst.*, vol. 29, no. 1, pp. 273–286, Jan. 2013.

[9] A. Charland, B. Leroux, and B. A. Charland, "Mobile Application Development : Web vs . Native," *Commun. ACM*, vol. 54, no. 5, p. 49, May 2011.

[10] S. Xanthopoulos and S. Xinogalos, "A comparative analysis of cross-platform development approaches for mobile applications," in *Proceedings of the 6th Balkan Conference in Informatics on - BCI '13*, 2013, p. 213.

[11] C. Doerner, F. Yetim, V. Pipek, V. Wulf, and C. Dörner, "Supporting business process experts in tailoring business processes," *Interact. Comput.*, vol. 23, no. 3, pp. 226–238, 2011.

[12] C. D. Weissman and S. Bobrowski, "The design of the force. com multitenant internet application development platform.," 2009.

[13] O. Schiller, B. Schiller, A. Brodt, and B. Mitschang, "Native support of multi-tenancy in RDBMS for software as a service," in *Proceedings of the 14th International Conference on Extending Database Technology - EDBT/ICDT '11*, 2011, p. 117.

[14] S. Schulte, D. Schuller, P. Hoenisch, U. Lampe, R. Steinmetz, and S. Dustdar, "Cost-driven Optimization of Cloud Resource Allocation for Elastic Processes," *Int. J. Cloud Comput.*, vol. 1, no. 2, pp. 1–14, 2013.

[15] J. Pokorny, "NoSQL databases: a step to database scalability in web environment," *Int. J. Web Inf. Syst.*, vol. 9, no. 1, pp. 69–82, 2013.

Modeling of software failure cost in ERP systems

Frane Urem *, Krešimir Fertalj **, Ivan Livaja *
* University od applied science Šibenik/Department of managemet, Šibenik, Croatia
** Faculty of Electrical Engineering and Computing/Department of Applied Computing, Zagreb, Croatia
e-mail: frane.urem@compco.hr, kresimir.fertalj@fer.hr, ivan.livaja@vus.hr

Abstract - Enterprise resource planning (ERP) systems are the most important software for many types of businesses. Software defects inserted on working ERP system can cause significant cost to their users. Existing software reliability models usually don't consider any financial impact at all. A model proposed in this paper links the ERP application layer with critical business operations and proposes cost calculation. The described modeling also encourages users to easily quantify the consequences of software failures that will always appear in ERP system. The proposed methods are illustrated by a case study.

I. INTRODUCTION

In certain way ERP vendors and their customers have different perspective of software failure costs. It is relatively easy to determine failure costs to be borne by the ERP vendors during failure fixes. They will investigate reported problems, identify and remove software defects that have caused failures. That process defines the main costs for ERP vendor because it includes certain experts costs (developers, business domain experts, consultants etc.) for solution search and rework spent in defect removing from the current software release. Most of these costs ERP vendor can predict and charge users before (installation, configuration, training) or after (maintenance contract) an ERP implementation. Of course, the ERP vendor may suffer certain indirect costs connected with user dissatisfaction, legal contracts etc. but that is hard to capture and it is out of scope of software reliability modeling anyway.

User's perspective is different and more complex to define because every failure is linked with business and failure cost is dependent on affected business process. Existing business that is closely related to the installed ERP system can be seriously compromised in case of untimely failure fixes.

Therefore, in this paper proposed model links the ERP application layer with critical business operations and propose cost calculation from user's perspective. Described model is directly linked with ERP software failure cost for the customer but proposed procedures can be also useful for ERP vendors in order to focus on the customer's needs and know true risks of using their product.

II. DEFINITIONS

Standards as IEEE 1044 [1] are well-defined in classification of software anomalies such as problem in software, failure, defect and fault. These definitions are used in this paper and their interrelation is illustrated in Figure 1.

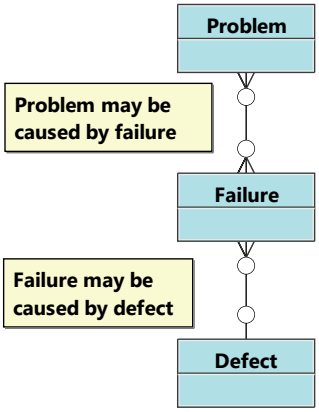


Figure 1 Relationship Problem – Model – Defect as an ER diagram

According to Figure 1, in every installed ERP system user can suffer various problems during the execution of different business operations. Problems may be caused in ERP system by software failure but it is not a rule because sometimes the cause may be something else such as insufficient training of users to work. According to definition from [1] failure is termination of the ability of an ERP system to perform required function or to perform it in previously specified limits. Defect is a source of software failures but sometimes user will not experience failures because defect is built in some functionality that is not actually used on existing ERP installation. In the past (as an example earlier version of IEEE 1044 standard published in 1993.) it was presumed that defect is a fault (error in software) but in complex software systems like ERP, defect term is more appropriate because defect can be something else like wrong configuration of ERP module. Software defect was initially defined by IBM [2], [3] but in IEEE 1044 standard defect attributes are more fully described.

III. MODEL

An adapted model from [4] is presented in Figure 2 and can be used to determine various ERP failure cost information. Initial model linked business layer to hardware/software system (called IT layer) deployed in a plant. Relevant business processes cost is linked with software function call with specific failure probability. Software reliability prediction is combined with financial calculations and software failure cost prediction is determined as a result.

In a Figure 2, original model is simplified because failure probability of ERP function call is not considered. Critical business operation is mapped with ERP function through software failure severity (caused by defect inserted in certain ERP function) which is later displayed in Table 1.

System is observed from user's perspective on two layers: business and ERP and some business process steps are assigned to corresponding functions implemented in ERP system. In case of success next optimal business process B will be executed. In case of failure another business process C will be executed with additional costs.

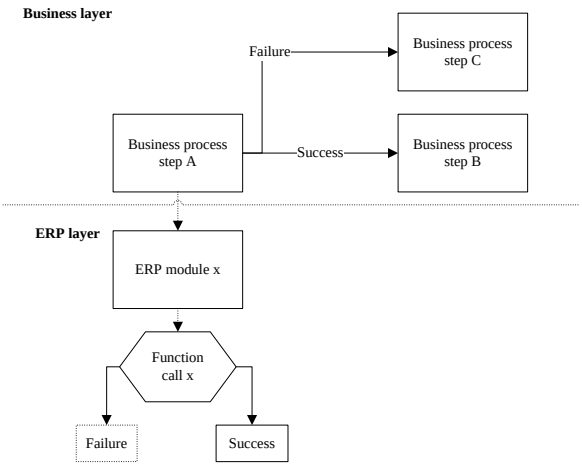


Figure 2 Connection between business layer and ERP system

Defect tracking system must be implemented as important prerequisite to apply presented model. Every failure is reported on defect tracking system with attributes detailed described at [1]. For failure cost analysis the following key attributes for reported failure must be available:

- Date/time when failure is observed
- Date/time when final disposition is assigned
- Defect reference

Referenced defects should be described with key attributes such as:

- Severity (the highest failure impact)
- Effect (the class of requirement that is impacted)
- Insertion activity (the activity during witch defect was inserted)

Severity can be simply correlated with expected frequency of failure. IEEE classification for defect severity [1] can be used as:

- Blocking (impossible to work on certain ERP module)
- Critical (certain critical business operation is disrupted)
- Major (certain critical business operation is affected but can proceed)
- Minor (nonessential business operations are disrupted)
- Inconsequential (no significant impact on business operations)

ERP module failure costs can be calculated according to (1) as:

$$CT = (ALC + DC) \times RT \times N \quad (1)$$

- CT - total failure cost
- ALC - additional labor costs caused by failure
- DC - downtime induced costs
- RT - resolution time
- N - the number of affected workplaces

Every failure is affecting a particular business operation and business domain experts define the financial costs according to referenced failure severity.

As a result, Table 1 can be made for every ERP module.

TABLE I. USER ASSESSMENT OF POSSIBLE COSTS FOR SPECIFIC ERP MODULE FAILURES WITH CERTAIN SEVERITY

COST DESCRIPTION:	*lost labor hours, downtime induced costs*			
MODULE:	*ERP module name*			
	CRITICAL BUSINESS OPERATIONS			
FAILURE SEVERITY	*Operation 1*	*Operation 2*	*...*	*Operation N*
Blocking				
Critical				
Major				
Minor				
Inconsequential				

Resolution time is calculating from failure report as the difference between dates when failure is observed and final disposition is assigned. Insertion activity can be used as a parameter for the case when user wants to assess the total cost in situation when failure is observed and corrective actions are in progress. For these cases Table 2 can be used. By contractual relations between the ERP vendor and customer it is usually possible to define and predict the response time when failures will occur during coding, configuration and documentation insertion activities. It is difficult to predict the time for corrective actions needed for removing of defects that were inserted

during requirements or design phase. Quite certainly the occurrence of such defects will seriously jeopardize business operations because their removal requires much more time.

TABLE II. USER ASSESSMENT OF POSSIBLE COSTS FOR SPECIFIC ERP MODULE FAILURES WITH CERTAIN SEVERITY

MODULE:	*ERP module name*	
DEFECT INSERTION ACTIVITY		
VALUE	**Definition (with examples)**	**Expected time for corrective action**
Requirements	Defect is inserted during requirements definition (incomplete use case specification, incorrect performance specification, incorrect security specification, incorrectly specified function etc.)	difficult to predict- business is at risk
Design	Defect is inserted during design activities (incorrect interface specification, incapable design etc.)	difficult to predict- business is at risk
Coding	Defect is inserted during coding or similar activities (incorrect data initialization on interface form, module interface does not work correctly, wrong reporting etc.)	contractually defined- controlled losses
Configuration	Defect is inserted during ERP module configuration (wrong interface to other ERP modules, incorrect interface to other software systems etc.)	contractually defined- controlled losses
Documentation	Defect is inserted in documentation for installation and operation	contractually defined- controlled losses

IV. CASE STUDY

The following case study illustrates proposed model for software failure cost on already implemented ERP system. One ERP module upgrade is analyzed on existing ERP system made by one Croatian ERP vendor. The user wanted to replace existing POS (Point Of Sale) module with the new one from the same vendor. Basically, it was an existing module upgrade with the following improvements:

- new user interface design optimized for use on the touch screen
- more detailed user roles
- better reporting integrated in a new interface

Nothing has changed on existing technology (.NET, Microsoft SQL Server) or equipment (only new touch screens were installed). The specified user was one of the first who decided to replace the old POS module but the new one has already been shipped to several smaller sites and there were no reported problems.

After installation the user has reported failures described in Table 3. All failures were categorized by ERP vendor and reported on vendor's helpdesk system.

TABLE III. REPORTED FAILURES

ID	**Title**	**Severity**	**Date observed**	**Date closed**	**Business operation**	**Referenced defect**	**Defect insertion activity**
1	Fiscalization service is working slower than in old module	Major	2.9.2013	4.9.2013	**Sales**	G2-TSPRM-006	Coding
2	Discount based on user role does not work correctly	Minor	2.9.2013	4.9.2013	**Sales**	G2-TSPRM-007	Coding
3	Cash register report does not print cash deposit amount correctly on POS printer (on screen report is working correctly)	Inconsequential	5.9.2013	9.9.2013	**Sales**	G2-TSPRM-008	Coding
4	After upgrade bar-code reader does not work correctly for Coca Cola products	Major	9.9.2013	10.9.2013	**Sales**	G2-TSPRM-009	Coding
5	After upgrade document replication from central location to POS does not work	Critical	10.9.2013	11.9.2013	**Replication**	G2-TSPRM-010	Coding

The user did not have serious problems with the observed module after the failures described in Table 3. According to Table 3 all defects were inserted during coding and every defect caused single failure. There were lot of problems caused by some reported failures but it was impossible to report all problems in a real time. As a result, it was impossible to objectively assess the costs incurred.

After an interview with the user, Table 4 and Table 5 were formed according to Table 1 and calibrated with available information about real costs that are incurred during failures from Table 3. Thus some costs were linked to lost labor hours according to the records of overtime hours caused by reported failures and some downtime costs were calculated as reduced earnings compared to regular working days. Some type of failures (like failures with blocking severity) was not reported and costs were estimated with worst case scenario. At the end, the resulting Tables 4 and 5 represent a projection of future costs that may be expected when certain types of failures will occur.

TABLE IV. USER ASSESSMENT OF LOST LABOR HOURS FOR SPECIFIC ERP MODULE FAILURES WITH CERTAIN SEVERITY

COST DESCRIPTION:	lost labor hours at one workplace, during one working day			
MODULE:	Point of sale			
CRITICAL BUSINESS OPERATIONS				
FAILURE SEVERITY	Sales/Returns	Replication (new products, prices, cash register ...)	Version upgrade	Loyalty programs
Blocking	16	2	1	2
Critical	16	2	1	2
Major	2	1	0	1
Minor	0,5	0,5	0	0,5
Inconsequential	0	0	0	0

TABLE V. USER ASSESSMENT OF DOWNTIME INDUCED COSTS FOR SPECIFIC ERP MODULE FAILURES WITH CERTAIN SEVERITY

COST DESCRIPTION:	downtime induced costs at one workplace, during one working day (EUR)			
MODULE:	Point of sale			
CRITICAL BUSINESS OPERATIONS				
FAILURE SEVERITY	Sales/Returns	Replication (new products, prices, cash register ...)	Version upgrade	Loyalty programs
Blocking	1.000,00	250,00	125,00	250,00
Critical	1.000,00	250,00	125,00	250,00
Major	250,00	125,00	0,00	125,00
Minor	50,00	50,00	0,00	50,00
Inconsequential	0,00	0,00	0,00	0,00

For an example, expected failure cost with major severity that affects sales operations for one day is calculating from (1) and Tables 4 and 5 (in this calculation salesperson work time corresponds to 7 EUR per hour) as:

CT = (2x7 EUR+250 EUR) x 1 x 10 = 2.640,00 EUR

The total costs of all reported failures from Table 3 are calculated in Table 6.

TABLE VI. TOTAL COST OF ALL REPORTED FAILURES FROM TABLE 3

Failure ID	**Cost**
1	2.640,00
2	505,00
3	0,00
4	2.640,00
5	2.640,00
TOTAL	**8.425,00**

Total estimated costs were amounted to more than 30% of planned upgrade investment. However, such costs are still within the limits expected. A newer POS version has introduced new features and added values so total investment was paid back soon after installation. Response time from ERP vendor in situations when failures are occurred was under contract limits so there was no reason to charge vendor for failure costs. In all ERP implementations or bigger upgrades failure cost can be expected but also prevented with stronger system testing. In described case study all defects were inserted during coding activities so failure costs were acceptable because repairs were made relatively fast.

It is also interesting to analyze worst case scenarios when blocking failures can be occurred during upgrade

according to Tables 4 and 5. For an example, the failure cost for one blocking failure in described ERP installation is more than 10.000,00 EUR per day. It seems impossible that issue will happen but failure number 4 from Table 3 occurred only because during system testing bar code function was not tested with Coca Cola products (QA tester used different products from other suppliers). In described case study some defects have simply resulted in failures although it has not happened to other users because they use the same ERP module in a different way. Therefore it is very important for user to provide a detailed system test on his location before any kind of ERP upgrade. This is particularly important when it comes to ERP module that can directly affect the business as in described case study.

V. CONCLUSION

Described modeling can be useful as a simple assessment of software failure costs that could occur in installed ERP system. In this model the critical business processes are linked to the severity of observed failures through the user assessment. In this way, ERP users can compare expected failure costs and investments required for their prevention. The initial failure cost estimation can be updated with the new failure records and the associated costs. The proposed model will be extended in future research with failure cost prediction according to previous failure occurrences.

REFERENCES

[1] IEEE Std 1044-2009, 2009, IEEE Standard Classification for Software Anomalies

[2] R. Chillarege, I. S. Bhandari, J. K. Chaar, M. J. Halliday, D. S.Moebus, B. K. Ray, and M.-Y. Wong, "Orthogonal Defect Classification: A Concept for In-Process Measurements," IEEE Transactions on Software Engineering 18, No. 11, 943–956, 1992.

[3] K. A. Bassin, T. Kratschmer and P. Santhanam, "An Objective Approach to Evaluate Software Development", IEEE Software Magazine, November/December, Vol.15 No.6,, 2002.

[4] F. Brosch, R. Gitzel, H. Koziolek Combining "Architecture-based Software Reliability Predictions with Financial Impact Calculations", Electronic Notes in Theoretical Computer Science 264 3–17, 2010.

Automatizirano projektiranje telekomunikacijske infrastrukture softverskim alatom baziranim na geografskom informacijskom sustavu (GIS)

Tomislav Pavičić*, Damir Medved*, Saša Vojvodić*, Neven Severinac* i Vlatko Roland**
* Ericsson NT d.d., Zagreb, Hrvatska
** Institut za GIS, Zagreb, Hrvatska
tomislav.pavicic@ericsson.com; damir.medved@ericsson.com; sasa.vojvodic@ericsson.com; vlatko@i-gis.hr; neven.severinac@ericsson.com

Sažetak - Jedinstvena platforma na kojoj se spremaju svi elementi projekta nove i/ili postojeće telekomunikacijske infrastrukture pojednostavljuje i ubrzava realizaciju novih projekata. Softverski alat baziran na GIS podacima, nedvosmisleno precizira trase i vizualizira administrativne pretpostavke za realizaciju projekta (npr. građevinske dozvole ili pravo puta).

Za kvalitetnu realizaciju projekata neophodni su ažurni GIS podaci. Prilikom izrade projekta svjetlovodne infrastrukture to su svi objekti od interesa kao što su namjena i korištenje prostora, kućanstva, ceste, zemljišne čestice te dostupnost postojeće telekomunikacijske i druge infrastrukture.

Alat za projektiranje mreže automatski izračunava potrebne količine materijala i trasa za realizaciju projekta te rezultate pohranjuje u GIS bazu. Iz baze se lako izračunava ukupna potrebna investicija, te se dodatnim funkcionalnostima lako upravlja i prati realizacija projekta na terenu.

Time je omogućena ključna komponenta uvođenja konzistentnosti u sve poslovne procese a posredno i u podatke, a to je implementacija upravljanja životnim ciklusom mrežne infrastrukture. Najveći izazov pri izradi ovakve vrsta projekata je osiguranje kvalitetnih ulaznih podataka.

Uređivanje i harmonizacija svih relevantnih podataka kroz izgradnju kvalitetno dizajnirane GIS baze omogućit će našem društvu efikasnije planiranje razvoja i izgradnje širokopojasnih mreža, lakše povlačenje sredstava iz EU fondova, pomoć pri donošenju odluka vezanih uz prostor, te transparentnu dostupnost društveno značajnih informacija.

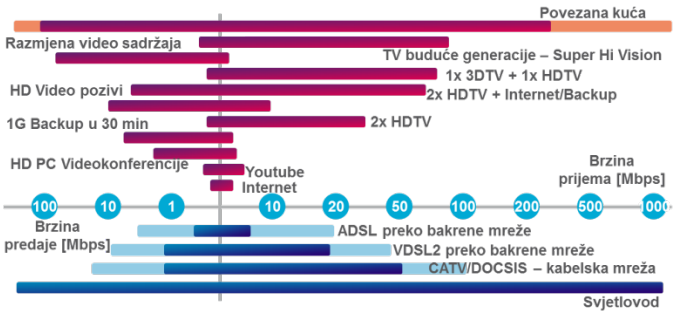
I. UVOD

A. *Potrebe*

Današnji način života okružen digitalnim sadržajima širi se u svim područjima društva povezujući svakoga i sve. Novo umreženo društvo omogućuje ljudima napredniju suradnju, otvara nove mogućnosti inovacija, potiče edukaciju, unaprijeđuje javni servis, a sudionici u umreženom društvu djeluju na načine za koje još prije nekoliko godina nismo mislili da će ikada biti mogući.

Da bi se ostvario potencijal umreženog društva, potrebna je nova, brža telekomunikacijska infrastruktura na osnovama jednostavnih, pametnih i skalabilnih rješenja fleksibilnih mreža i mreža spremnih za budućnost. To su mreže i mrežne tehnologije koje se razvijaju u svojim performansama od nekadašnjeg dial-up pristupa, ISDN-a (engl. Integrated Services Digital Network), preko danas raširenog ADSL-a (engl. Asymmetric Digital Subscriber Line) do sutrašnjeg sveprisutnog pristupa svjetlovodnim vlaknom.

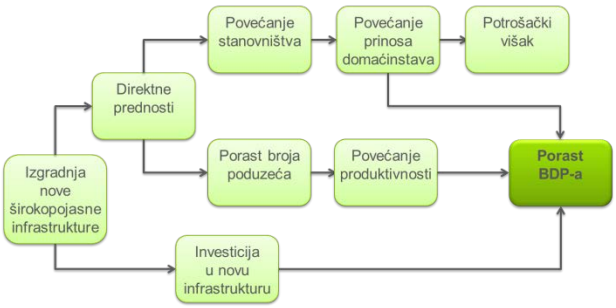
Promjene u mogućnostima, a time i ponašanju krajnjih korisnika digitalnih i sve obimijih video sadržaja zahtjevaju veće brzine i kvalitetu usluga koja je već sada dovela postojeću bakrenu telekomunikacijsku infrastrukturu do svojih limita. Postojeća digitalna DSL tehnologija doseže maksimalne brzine do 50 Megabita u sekundi (Mbps) u predaju, i jako je ovisna o kvaliteti bakrene veze. Ova brzina nije dovoljna za mnoge nove aplikacije, pa je potrebno planiranje i izgradnja novih svjetlovodnih mreža. Primjeri prijemnih i predajnih brzina ovsnih o tehnologijama i aplikacijama prikazani su slikom 1.



Slika 1. Brzine prijenosa u odnosu na tehnlogiju i uslugu

B. *Utjecaj nove infrastrukture na društvo*

Investicije u razvoj nove širokopojasne infrastrukture utječu na bruto domaći proizvod (BDP) direktnim angažmanom kapitala u izgradnji, te dugoročnim utjecajem promijenjenih ekonomskih i socijalnih aspekata (edukacija, zdravlje, uprava), te pozitivnog utjecaja na okoliš (promet, CO2), slika 2 [1].



Slika 2. Utjecaj nove infrastrukture na BDP

Mogućnosti koje pruža nova radna okolina pozitivno utječu na mogućnosti kreiranja novih radnih mjesta. Također ne zaboravimo na pozitivan efekt smanjenja emisije CO2, slika 3 [1].



Slika 3. Utjecaj nove infrastrukture na zapošljavanje

II. TEHNOLOGIJA I TOPLOGIJA SVJETLOVODNIH PRISTUPNIH MREŽA

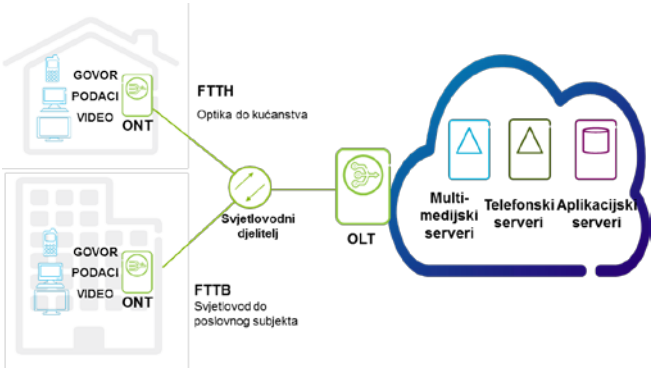
A. *Očekivanja od nove mreže – zašto projektiramo?*

Potrošačka potražnja za IPTV i drugim naprednim uslugama danas stvara potrebu za konvergentne, u potpunosti IP (engl. all-IP) bazirane svjetlovodne mreže visokog kapaciteta. Obećanje koje daje sve-svjetlovodna IP mreža uključuje i visoku učinkovitost i uštedu troškova te znatno veću propusnost. Tako ne treba čuditi da je većina telekomunikacijskih operatora već odabrala gigabitne pasivne svjetlovodne mreže (engl. Gigabit Passive Optical Network, GPON) [2] točka-do-više točaka (engl. Point-to-Multipoint, P2MP) sustava koji dijele jedno svjetlovodno vlakno za povezivanje više korisnika bez potrebe za aktivnom opremom u vanjskom postrojenju. GPON, Ethernet i ITU (engl. The International Telecommunication Union) standard, nudi isplativo rješenje koje osigurava visoku brzinu prijenosa podataka i napredne funkcionalnosti. Pasivne komponente instalirane izvan telekomunikacijskog središta ne zahtijevaju dodatno napajanje za rad. Kako više korisnika dijeli dio distribucijske mreže, potrebno je manje mjesta za aktivna sučelja i svjetlovodne distribucijske okvire u ormarima u telekomunikacijskom središtu. GPON pruža migracijski put prema all-IP mrežama i nudi potrebnu propusnost za isporuku nove napredne glasovne, video i podatkovne usluge.

B. *Pametna arhitektura mreže omogućuje trenutne usluge i usluge sljedeće generacije*

Umreženo društvo predviđa velike zahtjeve na kapacitet medija korištenjem velikog broja uređaja. Nova mreža treba biti svjesna uređaja, sadržaja, položaja korisnika i toka podataka, a takvu arhitekturu podržavaju GPON mrežna rješenja. Dedicirani prometni upravljači u PON arhitekturi omogućavaju nuđenje fleksibilnih paketa krajnjim korisnicima te na taj način koriste minimalni broj konekcija na najučinkovitiji način.

Slika 4. prikazuje osnovnu arhitekturu GPON tehnologije i način spajanja krajnih korisnika. Može se vidjeti da od GPON optičkog linijskog terminala (engl. Optical Line Terminala, OLT) koji se nalazi u telekomunikacijskom središtu ide samo jedno vlakno do pasivnog razdjelnika koji onda dijeli to isto vlakno na više krajnjih korisnika. Tako između telekomunikacijskog središta i krajnjih korisnika nema potreba za aktivnom opremom i napajanjem opreme električnom energijom, već je sva oprema pasivna.



Slika 4. Osnovna arhitektura GPON mreže

Neke od prednosti GPON tehnologije [3] i arhitekture mreže su:

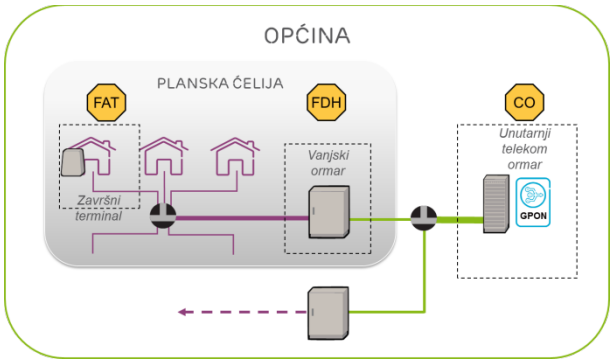
- Ušteda svjetlovodnih niti između telekomunikacijskog središta, tj. centralnog distribucijskog mjesta (engl. Central Office, CO) i djelitelja (engl. splitter)
- Samo se pasivna oprema (svjetlovodne niti i djelitelji) instalira izvan CO, dok je sva aktivna oprema smještena u CO ili kod krajnjeg korisnika
- Pouzdanje mreža s obzirom na potencijalne ispade elektroenergetskog sustava
- Kako se koristi struktura točka-do-više točaka iskorištava se manje portova na aktivnoj opremi u odnosu na točka-do-točka strukturu
- Podrška za scenarij plati kad je potrebno proširenje (engl. pay as you grow scenario)
- GPON tehnologija je tehnologija sa otvorenim standardima i spremna je za budućnost te nadogradnju tehnologije kao takve

C. *Dimenzioniranje*

Izabrana tehnologija i topologija nove mreže određuje arhitekturalno rješenje, te time utječe na izradu nominalnih planova distribucijskih i agregacijskih dijelova mreža. Razmotrimo dimenzioniranje distribucijske mreže na razini jedinice lokalne samouprave, to jest općine.

Područje općine ili aglomeracije dijeli se na planske ćelije čija se područja određuju gustoćom naseljenosti, sadašnjih i budućih potreba, te parametrima izabrane tehnologije. Unutar svake planske ćelije instalira se vanjski svjetlovodni distribucijski ormar (engl. Fiber Distribution Hub - FDH) kao središnje mjesto koncentracije svjetlovodne distribucijske mreže za pojedinu plansku ćeliju. Unutar FDH čvora instalira se svjetlovodni djelitelji tako da je svjetlovodna distribucijska mreža od FDH do krajnjeg korisnika izvedena u zvjezdastoj topologiji.

Svaki FDH spaja se na centralno distribucijsko mjesto pojedine općine na kojem će biti smješteni unutarnja telekomunikacijska oprema. Centralno distribucijsko mjesto (CO) se sa svakim FDH spaja direktnom vezom koja je izvedena sa svjetlovodnim kabelom odgovarajućeg kapaciteta.



Slika 5. Logička shema povezivanja na nivou općine

Distribucijski čvorovi (FDH) dimenzioniraju se za proračunati kapacitet uz minimalno +20% kapaciteta u rezervi za buduća proširenja i/ili nove korisnike. Svjetlovodna distribucijska mreža između FDH i krajnjeg korisnika u pravilu je maksimalne dužine do 1000 m po trasi sa ciljem što manje potrebe za varenjem svjetlovodnih vlakana. Svaki krajnji korisnik spaja se s jednom cjevčicom u koju se potom upuhuje svjetlovodno vlakno.

Općenito, korisnici se mogu podijeliti u dvije kategorije:

- Rezidencijalni/mali (engl. single dwelling unit, SDU) -> 1 svjetlovodno vlakno s 2 niti
- Poslovni/veliki (engl. multi dwelling unit, MDU) -> 1 svjetlovodno vlakno s 8 niti

Od toga se jedna nit terminira u korisničkom pristupnom terminalu i u distributivnom čvoru, dok ostale niti služe za buduće potrebe.

III. PROBLEM ULAZNIH PODATAKA

Dostupnost i kvaliteta ulaznih podataka značajno utječu na rezultate projektiranja mreže. Primjerice, manja cijena izgradnje i što jednostavnija pravna procedura može se osigurati imamo li dostupan podatak koje su parcele u vlasništvu jedinice lokalne samouprave, a dimenzioniranje rezervnog kapaciteta može se preciznije odabrati ukoliko je dostupan prostorni plan i strategija razvoja.

Poželjno je primijeniti GIS pristup gdje su ulazni podaci organizirani u prostornoj bazi podataka kako bi se moglo izvršiti prostornu analizu i dobiti što bolje rezultate. Većina podataka koji su bili na raspolaganju u dosadašnjim projektima a odnosili su se na kartografsku vizualizaciju (karte i planovi) nije bila dovoljne kvalitete za automatizaciju procesa projektiranja uz pomoć GIS sustava prikazano slikom 6.

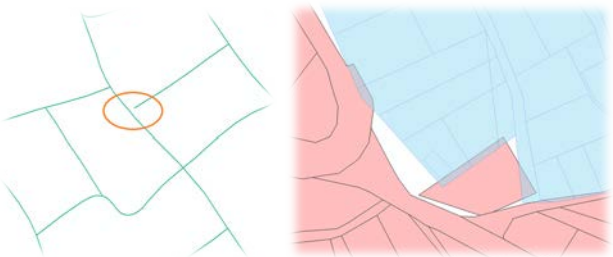


Slika 6. Proces dizajna svjetlovodne mreže

Ishodište problema leži u činjenici da je naglasak kod postojećeg procesa prikupljanja podataka bio na „crtanju“ a nije se vodilo računa o konzistentnosti nego samo o njihovoj vizualnoj reprezentaciji. Stoga je prvi korak u planiranju analiza, razvrstavanje, konverzija i uređivanje ulaznih prostornih i tabličnih podataka korištenjem GIS i ETL alata (Extract-Transform-Load).

Poligonski i linijski podaci moraju biti topološki ispravni, ne smije biti preklapanja poligona (overlap) niti „rupa“ između susjednih poligona (gaps). Linije moraju biti zaljepljene (snap) jedna na drugu, pojedina linija ne smije zatvarati petlje (self-intersect) sama sa sobom. Mora biti osigurana povezivost alfanumeričkih podataka (preko jedinstvenih identifikatora) s prostornim podacima.

Na slici 7. vidljive su topološke greške gdje linije cesta nisu zaljepljene jedna na drugu u čvorovima, te preklapanja i praznine na sloju katastarskih čestica



Slika 7. Topološke pogreške

U sljedećem primjeru, slika 8., adresa je upisana crnim tekstom a broj stanara na pojedinoj adresi crvenim tekstom. Situacija broj 1 (zaokruženo na slici) ukazuje na pogrešan unos naziva ulice (Splitska 1 umjesto Osječka 1). U situaciji broj 2 vidljiva je greška nastala prilikom unosa broja stanara (22 stanara u privatnoj kući tlocrtne površine 60m2). U planiranju bi se do te adrese dovelo 8 (MDU) umjesto 2 (SDU) niti ako greška ne bi bila

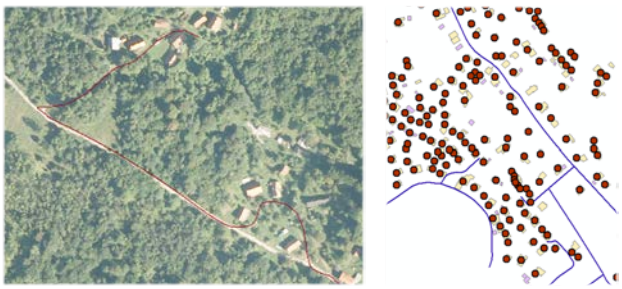
ispravljena. Situacija broj 3 pokazuje neupisanu adresu na stambenoj zgradi. Potrebno je izvršiti provjeru na terenu, upisati podatke u bazu i tek tada planirati mrežu.



Slika 8. Primjer netočnih adresa i broja stanara

Prilikom sređivanja podataka tipični problemi na koje se nailazi su starost podataka od deset i više godina, neažurne evidencije, nečitka slova hrvatske abecede (korištena stara kodna stranica 437). Vrlo često se prostorni podaci nalaze u „starom“ koordinatnom sustavu, 5. i 6. zoni Gauss Krugerove projekcije te ih je potrebno preprojicirati koristeći prikladne transformacijske parametre.

Problem uzrokuje i prostorna netočnost podataka s pogreškom koja iznosi i do desetak metara, neažurnost ili nepostojanje pojedinih podataka u bazi kao što je vidljivo na slici 9. Sloj cesta preklopljen sa snimkom iz zraka ukazuje na starost i netočnost podataka. U sljedećem primjeru vidljiv je velik broj adresa (crvene točke) ali ceste nisu ucrtane u bazi te ih treba dodatno unijeti.



Slika 9. Netočne i neucrtane ceste u bazi

IV. PROJEKTIRANJE UZ POMOĆ GIS ALATA

Dosadašnje metode projektiranja u kojima se nije koristio GIS bazirani alat, potrebno je zamijeniti sistemskim GIS rješenjima, transparentnom dostupnošću svih potrebnih ulaznih parametara koji će osigurati kvalitetu projekta i njegovu dugoročnu konzistenciju.

A. Neophodni ulazni podaci i koordinatni sustav

Za kvalitetan proces planiranja mreža potrebno je više vrsta informacija. Generalno se mogu podijeliti u poligonske, linijske, točkaste i alfanumeričke vrste podataka. Ulazni geoprostorni podaci trebaju biti pohranjeni u nekom od standardiziranih GIS formata (npr. kao *.shp datoteka). Ako se koristi ESRI ArcGIS platforma [4], dobro je da podaci budu konsolidirani u *.gdb prostornu bazu podataka (engl. File Geodatabase).

Odlukom Vlade RH definiran je koordinatni sustav kartografske projekcije HTRS96/TM kao referentni i od 2010. godine je ušao u službenu primjenu. HTRS96/TM sustav koristi se i za projektiranje svjetlovodne infrastrukture u GIS alatu.

Kao podloge za izradu projekata mogu se također iskoristiti podaci dostupni na javnim servisima DGU (URL: *http://geoportal.dgu.hr/wms*). Nakon uspješnog spajanja može se iz popisa slojeva odabrati željeni sloj čime se može provjeriti da su dostavljeni podaci i korištene podloge ažurni i u korektnoj projekciji.

Tematski slojevi koji se koriste pri planiranju su katastarske čestice i objekti u vlasništvu JLS, lokacije poduzetničkih zona, odnosno prostora buduće izgradnje - sadašnji i budući prostorni planovi (UPU, DPU i sl.) te lokacije potencijalnih prepreka (groblja, jezera, zaštićeni krajobraz).

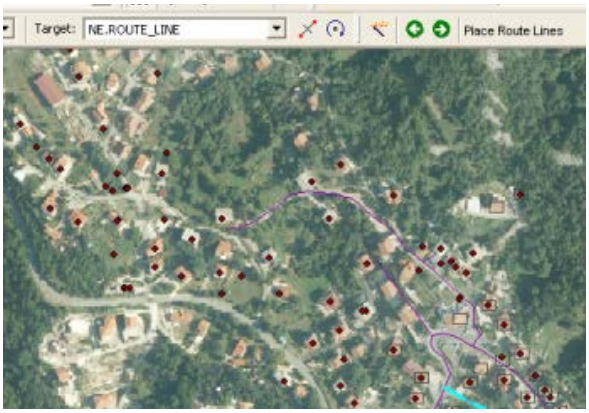
Neophodno je poznavati trase postojeće i planirane infrastrukture (prometnice, kanalizacije, vodovodi, plinovodi, javna rasvjeta) i lokacije objekata koji su u vlasništvu JLS, a mogu poslužiti kao preferirane ishodišne točke za planiranje optičke mreže.

Adresni model sa koordinatama kućnih brojeva koristi se za geokodiranje agregiranih statističkih podataka o potencijalnim korisnicima sustava, informacija o broju i tipu poduzeća u pojedinoj zgradi, broj zaposlenika i vrsta djelatnosti.

B. Softverski alat baziranim na geografskom informacijskom sustavu

Za bolje sagledavanje važnosti spomenutih prostornih i statističkih informacija u procesu automatiziranog projektiranja donosimo kao primjer kratak pregled procesa projektiranja FTTx mreža uz pomoć sustava Ericsson Network Engineer (ENE).

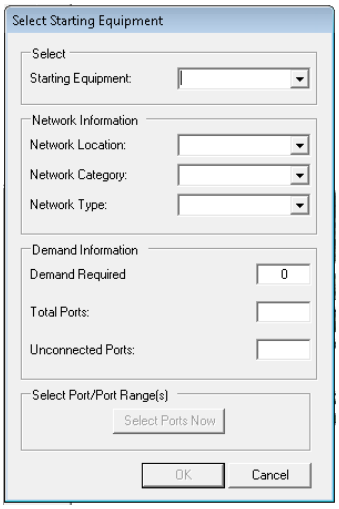
Nakon pripreme dostavljenih podataka (eventualne konverzije ili dorade, statističke obrade), pristupa se njihovom učitavanju u Ericsson Network Engineer te pokretanje procesa projektiranja mreže.



Slika 10. Prikaz radnog prozora Ericsson Network Engineer alata

Kako je prikazano na slici 10. . vidljive su točke od interesa za FTTx projekt te cesta koja u konkretnom slučaju služi kao ruta prema kojoj će sustav izvršiti optimizaciju elektroničke komunikacijske infrastrukture (EKI). Uobičajeno je da se kao ruta koristi postojeća komunalna infrastruktura čime se pojednostavljuje imovinsko pravna problematika i minimiziraju troškovi gradnje. Otvaranjem novog radnog naloga u aplikaciji, započinjemo proces projektiranja nove mreže na osnovu prikazanih ulaznih podataka.

Sami proces projektiranja uvelike je ubrzan uz pomoć čarobnjaka (eng. Design Assistant) kojemu je potrebno definirati sve potrebne parametre (opremu, rute i označiti koje interesne točke koristimo) i daljnje projektiranje je u potpunosti automatizirano (slika 11. , slika 12.).

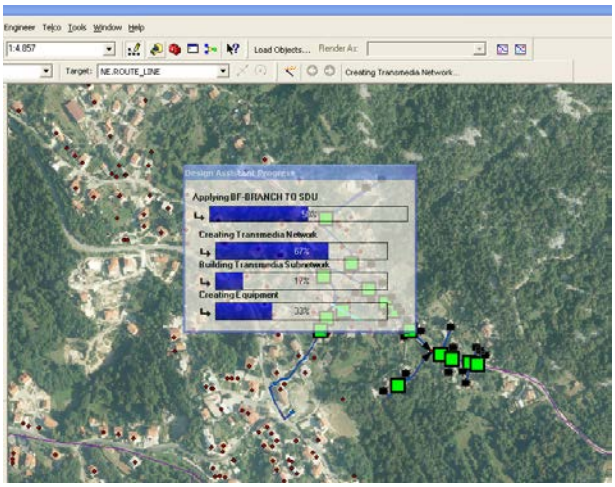


Slika 11. Jedan od prozora čarobnjaka za automatizaciju procesa (Design Assistant)

Time se značajno smanjuju potencijalne ljudske pogreške, povećava brzina samog projektiranja i pojednostavnjuje pronalazak optimalne varijante projekta.

Tijekom automatiziranog procesa projektiranja mreže na grafičkom sučelju sustava za projektiranje pratimo tijek procesa. U realnom vremenu prikazuje se izgenerirana topologija mreže (slika 11.) čiji se detalji u bilo kojem trenutku mogu izmjeniti i sa izmjenjenim detaljima nastaviti automatizirani proces.

Provjerom i usporedbom dobivenih tehničkih rješenja, projektant odabire optimalnu varijanu koja se onda završno obrađuje.



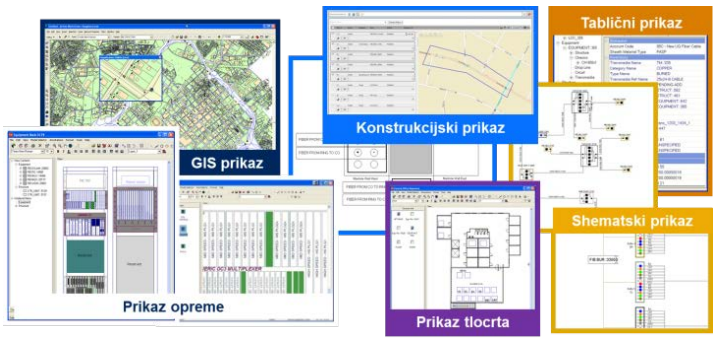
Slika 12. Dimenzioniranje svjetlovodne infrastrukture u ENE alatu

Konačni rezultat je optimalna topologija planirane mreže te kompletan troškovnik u skladu sa tehničkim rješenjem a što je prikazano slikom 13. .

Red. broj	Naziv produkta	Količina	Jedinica
1. Oprema			
1.1	Mikro cijev 1	5.533	m
1.2	Mikro cijev 4	1.544	m
1.3	Mikro cijev 7	746	m
1.4	Mikro cijev 12	1.182	m
1.5	Mikro cijev 19	1.094	m
1.6	Mikro cijev 24	3.377	m
1.7	Optičko vlakno (2 vlakna)	107.292	m
1.8	Optičko vlakno (8 vlakana)	5.062	m
1.9	Distribucijski ormar (vanjski)	1	komad
1.10	Spojnica za mikrocijevi (100 kom.)	1.570	komad
1.11	Račva za mikrocijevi	196	komad
1.12	Pristupni terminal (kućni)	193	komad
2. Građevinski radovi			
2.1	Instalacija mikro cijevi (veće)	5.475	m
2.2	Instalacija mikro cijevi (manje)	3.695	m
2.3	Iskop rupe	150	komad
2.4	DTK montažni zdenac	4	komad
2.5	Instalacija distr. ormara	1	komad
2.6	Instalacija prist. terminala	193	komad
2.7	Varenje opt. Vlakna	193	komad

Slika 13. Primjer troškovnika za jednu plansku ćeliju

Ovako dobiveno rješenje predstavlja optimalan idejni projekt i predstavlja kvalitetno ishodište za zadovoljenje brojnih potreba JLS (strateško planiranje i budžetiranje, priprema materijala za natječaje, katastar vodova, itd.).



Slika 14. Višedimenzionalno korištenje alata za projektiranje

Upravo korištenje GIS baziranog alata za projektiranje znatno ubrzava proces planiranja (slika 6.) i omogućava uspostavu kvalitetnog mehanizma (slika 14.) za

upravljanje životnim ciklusom mrežne infrastrukture (od planiranja preko izgradnje i eksploatacije te konačno održavanja).

V. ZAKLJUČAK

Prije neposrednog planiranja mreže potrebno je izraditi ažurnu GIS prostornu bazu podataka i uspostaviti mehanizme održavanja sustava. Time JLS dobiva kvalitetni sustav za podršku odlučivanju kojeg može koristiti ne samo u planiranju i izgradnji telekomunikacijske infrastrukture nego i u drugim projektima. Smjernice za implementaciju takvog sustava su:

1. Napraviti kvalitetni pregled stanja svih postojećih prostornih i alfanumeričkih podataka zasnovan na stručnom pristupu kontrole podataka (topologija, popunjenost atributa, jedinstvenost i usklađenost identifikatora).
2. Popisati potrebe korisnika za prostornim informacijama (informacijski proizvodi) prema trenutnim saznanjima, administrativnim i zakonskim obvezama i pravilima, te korištenjem metode intervjua dionika.
3. Izraditi model GIS baze podataka, u službenom koordinatnom sustavu HTRS96/TM, zasnovan na identificiranim potrebama korisnika.
4. Konvertirati postojeće podatke, provesti kontrolu kvalitete i unijeti podatke u GIS bazu.
5. Načiniti plan upravljanja podacima za određeno vremensko razdoblje, opisati procese prikupljanja i uređivanja podataka s jasnim ulogama pojedinih dionika.

Tako pripremljen GIS sustav omogućava da se uopće može pristupiti projektiranju mreža uz pomoć naprednih alata za projektiranje.

Time se značajno smanjuju troškovi te se olakšava dugoročna održivost cjelokupne komunalne infrastrukture.

LITERATURA

[1] The Impact of Broadband on the Economy, Research to Date and Policy Issues, April 2012, , www.itu.int/broadband

[2] ITU-T Rec. G.984 Gigabit-capable Passive Optical Networks (GPON), 2003

[3] Ericsson interna dokumentacija

[4] Roger Tomlinson, Thinking about GIS, ESRI Press 2003.

Upotreba alata za mjerenja na radijskom sučelju u nastavi mobilnih komunikacija

Leonard Novosel
Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva
Zavod za radiokomunikacije
Unska 3, 10 000 Zagreb, Hrvatska
leonard.novosel@fer.hr

Sažetak – Moderni su radijski komunikacijski sustavi složeni te zahtijevaju od studenata, budućih inženjera, opsežno poznavanje teorije. Nažalost, često puta se zanemaruje praktičan rad na stvarnim radijskim komunikacijskim mrežama što zbog objektivnih, a ponekad i subjektivnih razloga. Razvoj profesionalnih alata za mjerenje na radijskom sučelju u sustavima mobilnih komunikacija omogućuje razvijanje novih metoda u nastavi. Nove metode teže aktivnijem sudjelovanju studenata i usvajanju novih praktičnih znanja primjenjivih u profesionalnom telekomunikacijskom okruženju. Postojanje alata za mjerenja na radijskom sučelju koje svakodnevno koriste telekomunikacijski inženjeri za terenska mjerenja i optimizaciju odlična je prilika za korištenje tih alata u nastavnom procesu sa svrhom boljeg razumijevanja sustava mobilnih komunikacija. U ovom će se radu predstaviti korištenje jednog od dostupnih profesionalnih alata za mjerenje parametara radijskog sučelja Anite Nemo Handy s naglaskom na aktivnom sudjelovanju studenata u mjerenjima, analizi podataka te donošenju zaključaka vezanih uz performanse sustava mobilnih komunikacija.

Abstract - Modern wireless communication systems are complex and require from students, future engineers, extensive theoretical knowledge. Unfortunately, practical work on real wireless communication networks is often neglected because of objective, and sometimes subjective reasons. Development of professional mobile communications radio interface measurement tools allows the development of new teaching methods with active student participation. Students learn new practical knowledge applicable in professional telecommunication environment. The existence of radio interface measurement tools that telecommunication engineers use every day for field measurements and optimization is an excellent opportunity to reuse these tools in the educational process with the aim of better understanding the mobile communication systems. This paper will introduce the use of one of the available professional mobile communications radio interface measurement tools Anite's Nemo Handy with an emphasis on the active student participation in measurements, data analysis and drawing conclusions on the performance of mobile communications system.

I. UVOD

Današnja nastava mobilnih komunikacija zahtijeva od studenta da shvati mnoge apstraktne pojmove što predstavlja problem u njihovom vjernom i jednostavnom prikazu. Teorijski dio u nastavi mobilnih komunikacija dobro je obrađen i objašnjen dok je praktični dio najčešće zanemaren. Glavni razlog manjeg udjela praktičnog rada je cijena kao i sam format opreme koji ne dopušta korištenje izvan laboratorija, odnosno ne omogućuje studentima da samostalno provode mjerenja, analiziraju rezultate i izvode zaključke.

Studenti, odnosno budući inženjeri, tijekom studija ne dolaze u doticaj s mjernim uređajima i sustavima koje koriste profesionalni telekomunikacijski inženjeri te im je potrebno određeno vrijeme da prouče priručnike i odaberu način mjerenja pri prvom susretu s takvom opremom. Problem je i sama analiza rezultata i izvođenje zaključaka koje zahtijeva određeno predznanje i proučavanje opsežne literature kako bi se shvatilo međudjelovanje svih elemenata sustava mobilnih komunikacija [1]. Primjena analizatora spektra uobičajena je pomoć u nastavi mobilnih komunikacija [2], no javlja se problem mjerenja izvan laboratorija zbog veličine opreme. Korištenjem alata za terensko mjerenje na radijskom sučelju vrlo se jednostavno mogu izmjeriti i proučiti parametri mobilnih mreža. Mjerenjem na stvarnim sustavima proširuje i upotpunjuje se teorijsko znanje dobiveno kroz nastavu na predmetima koji su tematsko vezani uz sustave mobilnih komunikacija.

Ovaj rad će pokušati predstaviti neke primjere primjene alata za testiranje mreža mobilnih komunikacija u objašnjavaju fenomena koji se javljaju u stvarnim sustavima. Rad je organiziran na način da se u II. poglavlju uvode pojmovi koji se podučavaju u nastavi mobilnih komunikacija i objašnjavaju osnovni fenomeni koji bi se mogli provjeriti mjerenjem. U III. poglavlju objašnjavaju se mogućnosti Nemo Handy sustava te uvodi u način mjerenja pomoću alata Nemo Handy. U IV. poglavlju razmatra se upotreba mjerenja i mjerenih vrijednosti za objašnjavanje utjecaja radijskog kanala te povezuje s teorijskim znanjima koje bi studenti trebali posjedovati. Poglavljem V. zaključuje se rad.

II. NASTAVA IZ RADIJSKIH KOMUNIKACIJA

Tijekom studija studenti se podučavaju osnovama propagacije elektromagnetskih valova, utjecaju radijskog kanala na elektromagnetski val, osnovnom arhitekturom i načinom rada sustava mobilnih komunikacija. Osnovni cilj ovog teoretskog uvoda jest upoznati studente s mogućnostima i ograničenjima sustava mobilnih komunikacija. Detaljno se izučavaju sustavi mobilnih komunikacija koje studenti mogu svakodnevno sresti, a na

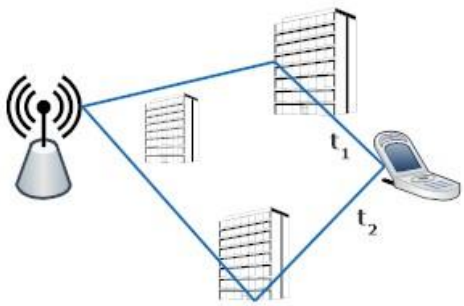
kojima će raditi kao budući inženjeri. Primjeri su sustavi druge i treće generacije mobilnih komunikacija (GSM, UMTS) te sustavi lokalnih radijskih mreža (IEEE 802.11).

Poseban osvrt daje se na ćelijski raspored u sustavima druge i treće generacije kroz teme kapaciteta ćelijskog sustava, raspodjele kanala te međudjelovanja susjednih ćelija na kvalitetu prijenosa. Zanimljiva je i tema utjecaja veličine ćelije na prostiranje elektromagnetskog vala kao i sam tip ćelije. Razmatraju se makro (1- 30 km), mikro (200 – 2000 m), te u posljednje vrijeme i nano i piko (4 – 200 m) ćelije kojima se povećava kapacitet ćelijskog sustava. Smanjenje opsega pokrivanja ćelije rezultira manjim kašnjenjem te povećanjem propusnosti, no povećava se interferencija među susjednim ćelijama kao i broj prijelaza mobilnog uređaja iz jedne ćelije u drugu.

Mehanizmi širenja elektromagnetskog val, a i njihov utjecaj na primljeni signal još je jedna od dobro obrađenih tema kroz pojmove prigušenja slobodnog prostora te proračuna prigušenja pomoću Frissovog zakona. Teme refleksije od tla i prepreka, ogiba, raspršenja vala na objektima koji su reda veličine valne duljine te apsorpcije vala pri prolazu kroz prepreke poučavaju se kroz primjere i simulacije kako bi se dobio vizualni prikaz prostiranja vala u takvom okruženju [3].

Posebno je zanimljiv pojam višestaznog širenja koji se obrađuje kroz teme sporo promjenjivog i brzo promjenjivog fedinga pomoću statističke analize primljenih signala te rezultantnog konstruktivnog ili destruktivnog zbrajanja signala na prijamu [3]. Sporo promjenjivi feding ili sjenjenje javlja se u slučaju postojanja jedne izravne zrake ili uz prisustvo slabe refleksije dok se brzo promjenjivi feding javlja u slučaju pojave mnogo zraka koje najčešće imaju slučajno raspodijeljene faze. Utjecaj brzo promjenjivog fedinga na primljeni signal očituje se kroz nagle i učestale propade primljenog signala dok se pojava sporo promjenjivog fedinga primjećuje kao spora promjena srednje vrijednosti primljenog signala [3].

Kroz teme ćelijskog rasporeda te višestaznog širenja elektromagnetskog vala studenti upoznaju utjecaj radijskog kanala na rasprostiranje valova i efekte na signal na prijamnoj strani. Utjecaji se sinteziraju i proučavaju kroz različite modele kanala koji su dobiveni mjerenjima te naknadnom obradom ili empirijski razradom osnovne teorije. Veličina, tip i raspored objekata u ćeliji utječu na elektromagnetski val, a taj se utjecaj očituje kroz promjene i kašnjenje signala na prijamu.



Slika 1. Višestazno širenje

Teorijski dio modela kanala dobro je razrađen i proučava se kroz sve stupnjeve studija, no praktični odnosno mjerni dio loše je zastupljen u nastavi. U laboratorijskim se uvjetima može vrlo jednostavno mjeriti signal na prijamu pomoću analizatora spektra [2], osciloskopa ili drugih specijaliziranih uređaja te vršiti potrebna obrada i analiza te izvesti zaključci, no u takvim se uvjetima ne mogu promotriti utjecaji različitih okolina na rasprostiranje vala te provesti identifikacija okoline. Terenska se mjerenja mogu provesti i provode se takvim mjernim uređajima no ona se redovito vrše pomoću automobila ili drugog prijevoznog sredstva što, zbog veličine opreme i potrebe za napajanjem, nije prihvatljivo za mjerenja koja bi trebali provoditi studenti u sklopu nastave.

Umjesto korištenjem analizatora, terenska mjerenja studenti mogu jednostavno provesti pomoću profesionalnih alata za mjerenje na radijskom sučelju kakve koriste telekomunikacijski inženjeri. Nedostatak tih alata jest nedovoljna vremenska razlučivost i gruba podjela u mjerenju snage, no za promatranje osnovnih fenomena propagacije u raznim okruženjima, odnosno modelima kanala, ta je razlučivost sasvim dovoljna.

Upoznavanjem mjerne opreme te programske podrške studenti proširuju svoje praktično znanje, a samim mjerenjima i obradom podataka dobivaju uvid u stvarne fenomene koji se odvijaju u radijskom kanalu. Analizom rezultata uz proučavanje literature i usporedbom s teorijom produbljuju svoje znanje o utjecaju kanala na signal na prijamu i kompleksnom međudjelovanju dijelova sustava. Jedan takav alat kojeg svakodnevno koriste telekomunikacijski inženjeri u terenskim mjerenjima, a može se koristiti i u nastavi, jest Nemo Handy tvrtke Anite.

III. MJERENJA POMOĆU NEMO HANDY SUSTAVA

A. Nemo Handy mjerni sustav

Nemo Handy je aplikacija s grafičkim korisničkim sučeljem koja sadrži više različitih stranica prilagođenih za različite zadatke koje uređaj može izvršavati, a instalira se na mobilni telefon. U paketu koji se koristi u ovom radu nalazi se mobilni telefon Nokia C7-00 sa Symbian operativnim sustavom. Navedeni mobilni uređaj je standardni mobilni telefon bez dodatnih funkcija dok mobilna aplikacija Nemo Handy pruža mogućnost pregleda signalizacijskih parametara koji se razmjenjuju između bazne stanice i mobilnog uređaja pri standardnom načinu rada. Mjerenja je moguće provoditi u aktivnom stanju (*dedicated mode*) ili stanju mirovanja (*idle mode*) [4].

Testiranje mobilnih mreža moguće je provesti korištenjem skripti kojima se definiraju aktivnosti mobilne stanice prilikom mjerenja, na primjer broj i trajanje poziva, slanje SMS i MMS poruka, količina prenesenih podataka, slanje elektroničke pošte te protokol za prijenos. Spajanjem mobilnog uređaja na baznu stanicu inicijalizira se spremanje svih dostupnih signalizacijskih podataka, a mobilni uređaj prelazi u aktivno stanje. Sam postupak mjerenja vrlo je jednostavan: na lokaciji predviđenoj za mjerenje pokreće se program Nemo Handy, a rezultati se

zapisuju u zasebnu datoteku koja se kasnije otvara i obrađuje u aplikaciji Nemo Outdoor.

Nemo Outdoor je računalna aplikacija koja se nalazi u istom programskom paketu kao i Nemo Handy. Njena svrha je analiza i pregled mjerenja u jednostavnom grafičkom okruženju koja su ostvarena mobilnim uređajem i Handy aplikacijom.

B. Mjerenja na mrežama druge generacije

Spajanjem na mrežu druge generacije (GSM) u aktivnom stanju osnovni su mjerni parametri razina prijamne snage (*RX level*) i kvaliteta veze (*RX quality*) te odnos signal/interferencija. Za prikaz prijamne snage koristi se njezina vrijednost u rasponu od -120 do -20 dBm (u praksi od -110 do -48 dBm), dok se kvaliteta prikazuje u razinama od 0 do 7. Kvaliteta signala odnosi se na broj krivo prenesenih bitova u 4 uzastopna okvira. Uz navedene podatke prikazani su i broj kanala, smjer prijenosa informacije, poruka i vrsta poruke pojedinog sloja te vrsta korištenog logičkog kanala. Ovime je omogućeno promatranje ponašanja mreže pri upućivanju poziva ili prijenosu informacija.

C. Mjerenja na mrežama treće generacije

Mjerni uređaj Nemo Handy omogućuje detaljna mjerenja na mrežama treće generacije kojima se dobiva uvid u rad zračnog sustava. Nabrojeni su samo neki parametri koji se mogu upotrijebiti za analizu i otkrivanje uvjeta u kanalu [4]:

- HSDPA propusnost (*HSDPA Throughput*) je parametar kojim pokazuje izmjerenu brzina prijenosa podataka (propusnost), kako na aplikacijskom (korisničkom) sloju, tako i na fizičkom MAC (*Medium Access Control*) sloju;
- Procjena snage primljenog signala (*Received Signal Strength Indication*, RSSI);
- Postotak pogrešaka u prijenosu na MAC sloju (*MAC Block Error Rate*) prikazuje parametar *BLER*, a dobiva se kao omjer pogrešno prenesenih blokova i ukupnog broja prenesenih blokova;
- Postotak ponovno poslanih blokova prvi put na MAC sloju (*MAC 1st retransmission rate*) je mjera ponovno prenesenih paketa dobiven omjerom ponovno prenesenih blokova i ukupnog broja blokova koji se promatra pri mjerenju;
- Omjer energije podimpulsa i spektralne gustoće šuma (E_c/N_0) jedan je od bitnijih parametara kvalitete, definira se kao omjer primljene snage korisnog signala (*Received Signal Code Power*, RSCP) i ukupno primljenog signala sumiranog u svim ćelijama (*Received Signal Strength Indication*, RSSI) [5];
- Procjena kvalitete kanala (*HSDPA Channel Quality Indication*) je broj u rasponu [0, 30] koji definira kvalitetu kanala, a samo mjerenje tog parametra ovisi o uređaju koji se koristi. Ako je *CQI* = 30, kvaliteta prijenosa u kanalu je najviša, dok je za *CQI* = 0 najniža [6];
- QoS parametri (*QoS Parameters*) su parametri kvalitete usluge koji su dogovoreni između mobilne stanice i bazne stanice prije samog početka prijenosa.

IV. KORIŠTENJE HANDY MJERNOG SUSTAVA U NASTAVI

Uvođenjem Nemo Handy sustava u nastavu prvenstveno se želi postići veća uključenost studenata u samostalnom istraživanju i mjerenjima. Na Zavodu za radiokomunikacije Fakulteta elektrotehnike i računarstva Sveučilišta u Zagrebu aktivno se koristi Anite Nemo mjerni alat u nastavi na nekoliko predmeta preddiplomskog i diplomskog studija. Uvođenjem mjernih alata u nastavu pokušava se studente potaknuti na samostalan rad. Samostalan rad je često stran studentima koji su navikli na učenje teorije i rješavanje zadataka iz pripremljene literature. Tijekom studija, na zasebnom predmetu ili u sklopu laboratorijskih vježbi, studentima je zadan neki inženjerski zadatak koji moraju samostalno riješiti. Studenti se pokaznim vježbama prvo upoznaju s mogućnostima mjernog sustava, osnovama korištenja i instalacijom potrebne programske podrške. Drugi korak je ustupanje mjernog sustava na korištenje studentima. Kao pripremni zadatak studenti su dužni napraviti okvirni plan mjerenja s planiranim lokacijama mjerenja, parametrima koje će mjeriti kao i očekivanim rezultatima mjerenja. Pripremni zadatak je važan zbog pripreme studenata za mjerenja kroz samostalno proučavanje korisničkih priručnika mjernog sustava te identifikaciji potrebnih parametara za mjerenje. Korisnički priručnici prikazuju mogućnosti sustava, a studenti odabiru koje mogućnosti, odnosno parametre za mjerenje, žele iskoristiti u rješavanju problema. Studentima se daje i literatura iz područja radijskih komunikacijskih sustava kako bi mogli proučiti teoriju na kojoj se baziraju parametri koje će mjeriti. Dodatna literatura pomaže i proširivanju i upotpunjavanju postojećih znanja iz radijskih komunikacijskih sustava uvođenjem i objašnjavanjem novih termina, a koji su važni za rad stvarnog sustava radijskih komunikacija. Proučavanjem literature i korisničkih priručnika dobiva se teorijsko znanje o međudjelovanju elemenata radijskog sustava kao i međusobnom utjecaju mjernih parametara. Planiranjem mjerenja primjenjuje se teorijsko znanje kako bi se dobili dosljedni mjerni rezultati te pokazalo razumijevanje utjecaja i ograničenja sustava.

Provođenjem planiranih mjerenja studenti se upoznaju s upotrebom mjernog sustava na terenu te načinom pohrane podataka. Mjerni se podatci mogu provjeriti odmah nakon mjerenja na samom mobilnom uređaju kako bi se osigurala točnost i dosljednost mjerenih vrijednosti. Sljedeći je korak obrada mjerenja pomoću dostupnih programskih alata. Nemo mjerni sustav dolazi s Nemo Outdoor aplikacijom za obradu i prikazivanje podataka. Pomoću aplikacije odabiru se parametri koji se odnose na rješenje zadanog problema, dok se ostali mogu prikazati informativno. Podatci se mogu usporedno prikazati u Outdoor aplikaciji ili pretvoriti u drugi format. Pretvorbom u drugi format podatci su dostupni za obradu u drugim programskim alatima poput Matlaba ili Mathematice. Korištenje tih alata posebno je korisno pri obradi velike količine podataka te upotrebe statističkih

funkcija za analizu. Studenti su dobro upoznati s tim alatima što im omogućava brzu obradu podataka i njihov kvalitetan prikaz.

Nakon obrade i analize podataka pristupa se izradi dokumentacije gdje se sinteziraju teorijska i praktična znanja. Studenti u izrađenoj dokumentaciji opisuju upotrebljene mjerne metode, prikazuju mjerene parametre, komentiraju rezultate te izvode zaključke. Zaključivanje je proces spajanja teorijskog znanja prikupljenog iz literature te uočavanja utjecaja koji su dobiveni kao rezultati analize mjerenih vrijednosti. Izradom dokumentacije studente se navodi na diskusiju kako bi razmjenjivali ideje i prepoznali uzroke dobivenih rezultata. U nastavku su navedeni neki primjeri mogućih zadataka koje bi studenti mogli riješiti mjerenjem pomoću Nemo sustava.

Pokretanjem korisničke skripte za mjerenje u aktivnom stanju započinje se postupak mjerenja i spremanja vrijednosti. Primjerice, moguće je mjeriti i spremati podatke o ukupnom kašnjenju (*Round-Trip Delay Time*, RTT). RTT je vrijeme koje je potrebno da primatelj primi signal (podatak) i potvrdu o primitku vrati pošiljaocu [4]. RTT je podatak koji pokazuje utjecaj svih elemenata sustava, tipa ćelije, smještaja mobilnog uređaja unutar ćelije kao i interferencije na ukupno kašnjenje i kvalitetu veze. Usporedbom mjerenih vrijednosti RTT i srednje izmjerene propusnosti na MAC sloju kao u tablici I., može se primijetiti trend smanjenja propusnosti s povećanjem kašnjenja. Mjerenjem i prikazom ovisnosti brzine prijenosa o kašnjenju, studenti mogu dobiti prikaz utjecaja svih povezanih elemenata sustava kao i raznih propagacijskih mehanizama na parametre veze.

Posebno je zanimljiva upotreba aplikacije Nemo Outdoor za prikaz nekoliko povezanih parametara što omogućuje bolje razumijevanje povezanosti fenomena u kanalu te posljedice na kvalitetu prijama. Na primjer, vrlo se lako može promatrati utjecaj promjene razine snage na kvalitetu kanala te brzinu prijenosa podataka kao na slici 2 gdje se na horizontalnoj osi nalazi točno vrijeme mjerenja, a na vertikalnoj vrijednosti u ovisnosti o odabranom parametru. Studenti mogu vrlo lako izmjeriti i prikazati usporedno više parametara te komentirati utjecaj raznih fenomena u kanalu na određeni parametar što omogućuje primjenu teorijskih znanja na stvarnim sustavima [7]. Mjerenjem se može dobiti ovisnost brzine prijenosa o indikaciji kvalitete (CQI) kanala te RSSI. Mjerenjem u aktivnom stanju te obradom u Nemo Outdoor aplikaciji dobiva se odnos brzine prijenosa u odnosnu na RSCP. RSCP se koristi za odluke o prekapčanju, proračunu prigušenja staze kao i mjera primljene snage [8].

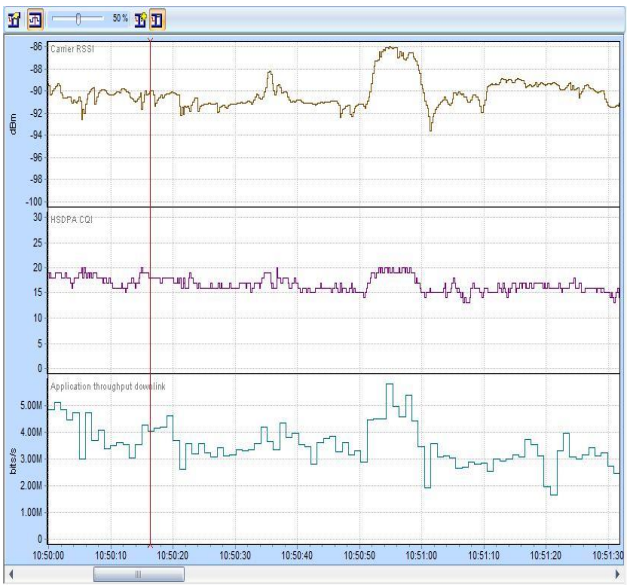
TABLICA I. OVISNOST PROPUSNOSTI O KAŠNJENJU

RTT [ms]	Propusnost na MAC sloju [Mbit/s]
128	4.18
144	4.52
265	4.03
278	4.23
293	3.69
309	3.16
318	3.17
334	2.40
384	3.25
414	2.48
500	2.06
547	1.23
749	1.82
4097	0.003

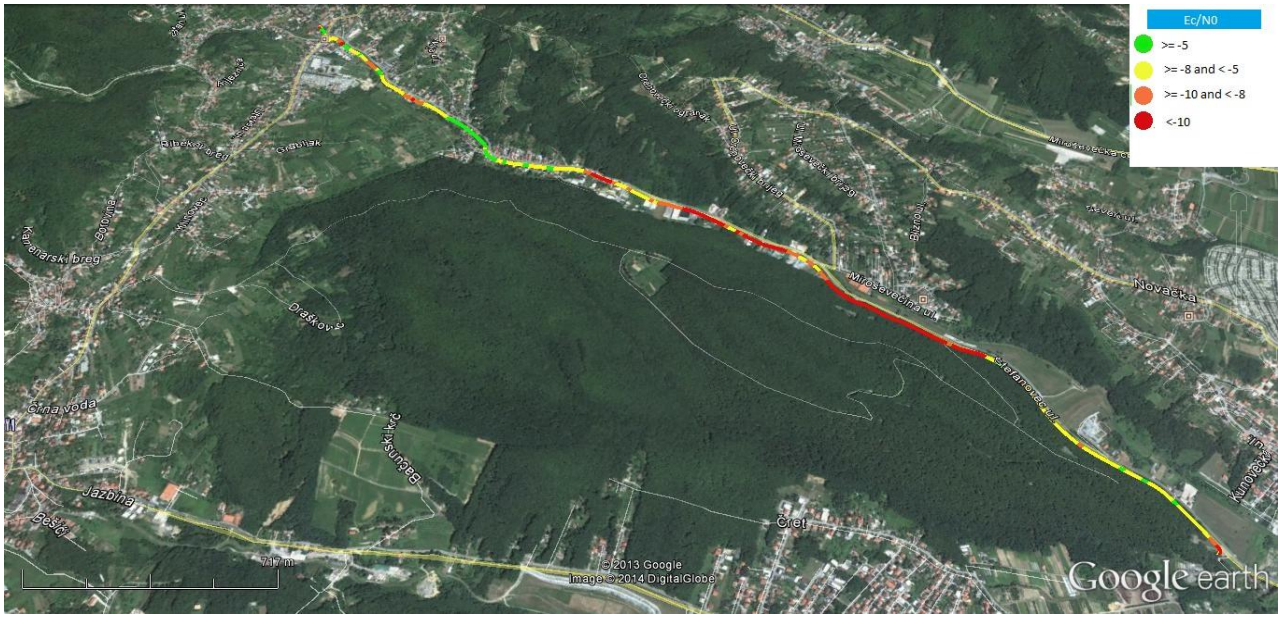
Iscrtavanjem izmjerenih vrijednosti E_c/N_0 na karti kao na slici 3. vrlo se jednostavno mogu uočiti promjene parametra u odnosu na razne objekte koji se pojave u radijskom kanalu kretanjem mobilnog uređaja. Mjerni uređaj mjeri i sprema svoju poziciju, brzinu kretanja i nadmorsku visinu pomoću GPS-a. Tako se na slici 3. vidi smanjenje E_c/N_0 prilikom prolaska mobilnog uređaja kroz brdovito područje uslijed sjenjenja prirodnim zaprekom. Općenito E_c/N_0 pokazuje pokrivanje i kapacitet ćelije jer uzima u obzir pokrivanje ćelije preko RSCP vrijednosti te razinu interferencije preko RSSI vrijednosti. E_c/N_0 može se iskazati preko RSCI i RSSI kao [8]:

$$\frac{E_c}{N_0} = \frac{RSCP}{RSSI} \tag{1}$$

Promjene E_c/N_0, zbog povezanosti s RSSI, pokazuju promjene razine šuma u mjernoj trasi kao na slici 3. Promjene šuma mogu se pojaviti zbog utjecaja okoline ili promjene opterećenja ćelije [8]. Studenti mogu izmjerene vrijednosti E_c/N_0, kao i izmjerena kašnjenja iskoristiti za identifikaciju raznih vrsta *fedinga* te proučiti utjecaj okoline na stvarni komunikacijski sustav kao i utjecaje opterećenja u raznim tipovima ćelija s različitom gustoćom korisnika. Zanimljiva je upotreba opcije zaključavanja Nemo Handy mobilnog uređaja na jednu baznu stanicu za mjerenje pokrivanja bazne stanice i promatranje utjecaja objekata u ćeliji na promjene izmjerenih vrijednosti.

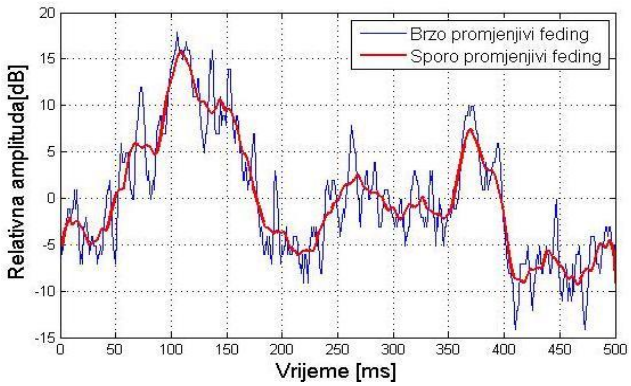


Slika 2. Usporedni prikaz parametara u Nemo Outdoor aplikaciji



Slika 3. Prikaz promjene E_c/N_0 na mjernoj trasi

Mjerenjem primljene snage signala na mreži druge generacije u stanju mirovanja u jednoj urbanoj makro ćeliji te naknadnom obradom može se dobiti prikaz brzo i sporo promjenjivog fedinga kao na slici 3. Rezultati mjerenja preuzeti su iz [9], dokumenta koji je nastao kroz studentski projekt diplomskog studija. Iako vrijeme uzorkovanja od 1 ms nema dovoljnu razlučivost za fino mjerenje brzog fedinga, on je ipak vidljiv kao nagle i značajne promjene amplitude signala na prijamu.



Slika 4. Brzo i sporo promjenjiv feding

V. ZAKLJUČAK

Cijena i kompleksnost opreme koja se koristi za mjerenja u bežičnim komunikacijama često puta je ograničavajući faktor pri samostalnom radu studenta na mjerenjima. Postoje mjerni alati kojima se koriste inženjeri u telekomunikacijama, a koji vrlo učinkovito mogu zamijeniti skupe i glomazne mjerne uređaje u terenskom mjerenju no uz cijenu smanjenje osjetljivosti mjerenja. Posebno je korisna opcija takvih alata mogućnost usporednog prikaza nekoliko izmjerenih parametara kako bi se povezalo znanje o utjecaju kanala na izmjerene vrijednosti.

Alati poput Nemo Handy mjernog sustava posebno su zanimljivi za razne studentske projekte kao i u nastavi upravo zbog jednostavnosti mjerne opreme i mogućnosti prikaza rezultata. Korištenjem takvih alata studenti imaju priliku praktično se upoznati s opremom i programskom podrškom koju će koristi kao telekomunikacijski inženjeri. Tijekom studija studenti dobiveno praktično iskustvo mogu vrlo lako povezati s teorijskom podlogom korištenjem literature ili kroz diskusiju sa svojim mentorima i profesorima.

LITERATURA

[1] B. Coll-Perales, M. Sepulcre, J. Gozalvez (2011), "Innovative mobile communications engineering teaching laboratory using professional network testing tools," EDULEARN11 Proceedings, pp. 5531-5535.

[2] L.M. Tomas, "Using the Spectrum Analyzer as an Educational Tool for Mobile Communications," Departamento de Senales, Sistemas y Radiocomunicaciones. Universidad Politecnica de Madrid, lipanj 2012.

[3] A.F. Molisch, Wireless Communications, 2nd Edition. Wiley-IEEE, veljača 2011.

[4] ANITE FINLAND, "Nemo Handy-S 3.50 Manual", veljača 2012.

[5] M. Sauter, From GSM to LTE: an introduction to mobile networks and mobile broadband. Wiley, veljača 2011.

[6] Universal Mobile Telecommunications System (UMTS) Physical layer procedures (FDD), Technical Specification. ETSI, ožujak 2007.

[7] ANITE FINLAND, "Nemo Outdoor Playback 6.1 Manual", studeni 2011.

[8] Jarno Niemelä, Jakub Borkowski and Jukka Lempiäinen: Using Idle Mode Ec/N0 Measurements for Network Plan Verification. in Proc. 8th IEEE International Symposium on Wireless Personal Multimedia Communications (WPMC'05), Aalborg, Denmark, rujan 2005.

[9] Mjerenje parametara bežičnih mreža pomoću alata Anite Nemo Handy-S, Tehnička dokumentacija, prosinac 2012.

NETWORKS & ENVIRONMENT

Reselection of base station in TETRA network

Saša Punčikar
Visoka škola za informacijske tehnologije, Zagreb, Croatia
spucikar@vsite.hr

Abstract – TETRA (Terrestrial Trunked Radio) is ETSI defined standard for professional digital mobile radio network. It is mission critical system and one of most important parameter is reliability. Mobile functionality is of great importance for end users and special attention should be given to reselection parameters. The article gives a detailed explanation of the process of reselection, it is explained what are slow reselect threshold and fast reselect threshold, list of adjacent base station that are sent from serving base station to terminal and their influence to smooth transition from serving cell to adjacent cell. Examples of parameters will be shown and practice guides and tools to discover problems of reselection.

I. INTRODUCTION

TETRA (Terrestrial Trunked Radio) is digital mobile system developed with primary function to operate for police department, fire brigades, ambulance, and military purposes but because of rough construction and reliable functionality it was adopted also by electricity and oil companies, airports, mines and it is used for many other applications. TETRA first goal was to offer voice communication and using TDMA, it was possible to deliver opportunity for 4 different users to communicate at same time on same radio channel. Talkgroups were defined and many talkgroups can passively connected on one base radio (one base radio use one radio channel) but only 4 at same time. Also possibility of deploying dynamic allocation of resources was great breakthrough comparing TETRA to analog system where one channel has been used for one group of users and no possibility of dynamic allocation of resources.

Data communication trough TETRA at first was using only SDM (Short Data Messages) trough control channel and without affecting voice communication. Available bitrates trough control channel are sufficient system for control and management for SCADA and similar systems.

Cell reselection of MS can influence both voice and data communication but since data traffic is much shorter and has implemented mechanism for retransmission reselection influence to data traffic is negligible. Also considering fact that in SCADA system probes are mainly mounted on one position and doesn't change position cell reselection is perform rarely, mostly because of equipment malfunction.

On the other hand Mobile Station (MS) primary used for voice communication are constantly changing positions and cell reselection happens very frequently, it should be done as smooth as possible. It is very important for reselection is done quickly; best possible result is for customer not to know about any this.

Special attention should be consider to take into count transmit power balance from BS and transmit power from TETRA terminal. In most TETRA system BS can provide 60-80 W (48-49 dBm) per each base radio that means TX power at antenna connector can be up to 20 W (43 dBm). Typical TX power of handheld terminal is 1W (30 dBm) and mobile terminal (intended to be mounted in vehicle) is 3W (35 dBm). TX power figures are depending on manufacturer but those are typical values that can be seen on practically all systems. Transmit power balance must be achieved to accomplish network quality and lowering power consumption.

Power disbalance of complete system can strongly influence cell reselection, especially if some base stations are set to maximum power because Mobile Station (BS) measuring *RSSI* and calculating *CX* parameter base on *RSSI* from Base Station.

II. RECEIVED SIGNAL STRENGTH

Important difference between TETRA system and most of other commercial mobile network is that in TETRA mobile subscriber unit initiate reselection of Base Station (BS). MS is constantly measuring receiving signal strength indicator (*RSSI*) from serving cell and from adjacent base stations. Using *RSSI* mobile station is calculating *CX* parameter which is basically fade margin for mobile station. *CX* figure defines amount of available additional attenuation before *RSSI* falls below minimum defined value.

For serving cell *C1* parameter is calculated and for adjacent site *C2* parameter is calculated but calculation is done using same formula:

$$CX = RSSI - Rx_Min_Access_level - max(0, Max_MS_Tx_Power - Power_MS) \quad (1)$$

Rx_Min_Access_level is minimum received signal strength value from BS measured on MS that is permitted MS to operate on. If *RSSI* decrease below this value MS will not be able to communicate on this BS. This parameter must be higher than RX sensitivity of base station (BS) and it is defined trough system software. In Motorola Dimetra system this parameter is configured in Zone Configuration Manager separately for each BS and can be used to decrease cell covering area and decrease number of MS that will be connected to BS.

Parameter *max(0, Max_MS_Tx_Power - Power_MS*) is needed to modify *CX* parameter depending on difference of maximum transmit power of MS that can be

used in system and transmit power of specific MS that is in use since two way communication must be achieved. Most manufactures of TETRA terminals produce handheld terminals with 1W (30 dBm) TX power and mobile terminals for installation in vehicles with 3W (35 dBm).

For example *Rx_Min_Access_level* is set to -110 dBm and measured *RSSI* is -82 dBm *C1* would be:

- -82 – (-110) -5= 23 dBm for handheld terminal and (2)
- -82 – (-110) -0= 28 dBm for mobile terminal. (3)

III. RESELECT PARAMETERS

Most important parameters that are determing reselection of BS are fast reselect threshold, slow reselect threshold, fast reselection hysteresis and slow reselection hysteresis:

A. *fast reselect threshold (FRT)* is the maximum level above C1 = '0' for a radio relinquishable link,

B. *fast reselect hysteresis (FRH)* is the hysteresis for a radio relinquishable link

C. slow *reselect threshold (SRT)* is the maximum level above fast reselect threshold for a radio improvable link *and*

D. *slow reselect hysteresis (SRH)* is the hysteresis for a radio improvable link.

Slow and fast hysteresis parameters are in use for defining hysteresis for reselection to occur while slow and fast reselect thresholds are defining minimal conditions for reselection to occur.

Fast reselect parameters are generally in use for relinquishable link. To make the site switching smoother, it is needed to increase the fast threshold and reduce the fast hysteresis parameters. With this set-up the site switching will occur earlier, more often and with less hysteresis.

For site switching in areas of weak coverage, fast hysteresis should be set to 2 dB or 4 dB and slow hysteresis always need to be set to greater value than the fast hysteresis.

Through the Zone Configuration Manager (ZCM) in Motorola TETRA system adjacent sites (eg. Adjacent base stations) must be defined for any base station BS. List of adjacent sites and other parameters (*FRT, SRT, FRH SRH*) are transmitted to mobile station (MS) trough Broadcast Network Channel (BNCH) and Broadcast Control Channel (BCCH). Since every base station should be working on different frequency channel Zone Configuration Manager is sending ID and frequency channel number of adjacent base stations to mobile station (MS). MS is measuring *RSSI* for serving cell and all of adjacent sites and calculating *CX* parameter. Any mistake in defining adjacent site can be critical in process of reselection.

For instance if near base station BS1 is not defined as adjacent site mobile station (MS) is not measuring its signal because it did not received information that this site exist in network. This is the reason why it is not possible to use BS1 if receive signal from active base station is lost. Mobile Station will try to connect to BS2 with lower receiving signal strength. If BS1 is configured as adjacent site for BS2 then it will connect to BS1. Every reselection need some time to take place and problems in communication will probably appear. This kind of behavior is very confusing for end users and most probably they will not be able to describe what is problem.

IV. SERVING CELL STATUSES

Five possible statuses for serving cell are defined through TETRA standard:

A. Exclusive - *RSSI* from serving cell is higher than

Fast Reselect Threshold + Slow Reselect Threshold above Fast + Start Monitoring Threshold Delta. (4)

In Exclusive mode MS is not scanning for adjacent sites since it has excellent signal from serving cell and this saving battery energy,

B. Superior mode - enable MS to start scanning for adjacent sites but it is not possible to start reselection process in this mode,

C. Usable – MS is scanning for adjacent sites and reselection process is possible in this mode but rarely happens because *RSSI* from serving cell is still high,

D. Improvable - MS is scanning for adjacent sites and reselection process is possible and

E. Relinquishable – *RSSI* is low and it is near *RX sensitivity* of MS, MS is scanning for adjacent sites and reselection will take place as soon as possible.

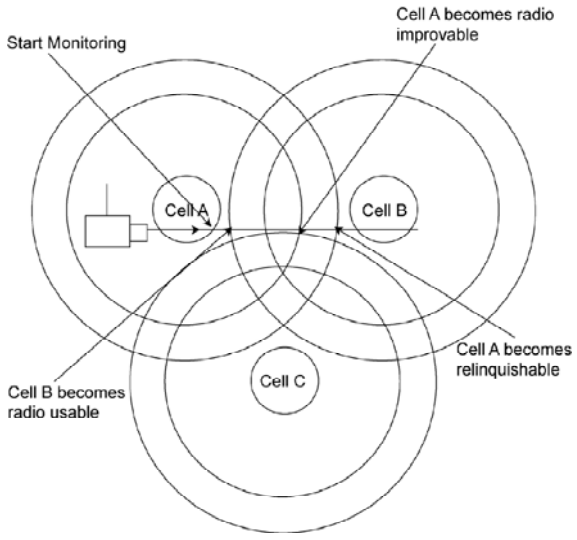


Figure 1. Statuses of serving and neighbor cell

V. NEIGBOR CELL STATUS

Three possible statuses for neighbor cell are defined through TETRA standard. Moving Mobile Station trough TETRA network is shown in figure 1 and boundaries different thresholds for each status.

Statuses for neighbor cell are:

A. Usable – reselection to Usable cell is not allowed if serving cell and neighbor cell have same service level since in this case there is no benefit of reselection. Reselection is possible only if usable cell provide higher service level (higher subscriber class, home location area, cell load or similar). Reselection to usable cell happens very rarely.

B. Improvable – *RSSI* of neighbor cell becomes better than *RSSI* of serving cell so MS is trying to register on neighbor cell. Following criteria must be accepted for reselection to take place:

- *RSSI* of neighbor cell must be for value of *SRH* higher then *RSSI* of serving cell BS and
- *RSSI* of serving cell must lower than *Rx_Min_Access_level* + *FRT* + *SRT*.

C. Relinquishable – *RSSI* of serving cell is lowering and MS is trying to register on neighbor cell as soon as possible. Following criteria must be accepted for reselection to improvable cell to take place:

- *RSSI* of neighbor cell must be for value of *FRH* higher then *RSSI* of serving cell BS
- *RSSI* of serving cell must lower than *Rx_Min_Access_level* + *FRT*

VI. EXAMPLES FOR RESELECTION TO NEIGHBOR CELL

Following parameter on ZCM are configured for all base stations:

Rx_Min_Access_level: -105 dBm

Slow reselect threshold (SRT): 8 dB

Fast reselect threshold (FRT): 12 dB

Slow reselect hysteresis (SRH): 6 dB

Fast reselect hysteresis (FRH): 10 dB

Example 1:

Serving cell *RSSI*: - 76 dBm

Neighbor cell *RSSI*: - 68 dBm

First criteria for neighbor cell to be improvable state is achieved (-68+6>-76 dBm) but second criteria is not achieved (-76>-105+12+8) and reselection will not take place.

Example 2:

Serving cell *RSSI*: - 86 dBm

Neighbor cell *RSSI*: - 78 dBm

First criteria for neighbor cell to be improvable state is achieved (-78+6>-86 dBm). Second criteria is also achieved (-86<-105+12+8) so reselection will take place since neighbor cell is defined as improvable.

VII. TOOLS FOR RESELECTION TEST

Mobile handheld devices and mobile terminals provide service menu (Figure 2.) where it is possible to see *RSSI* and calculated *C1* for serving cell. Also *RSSI* and *C1* can be shown for sites that are defined as adjacent sites for that serving cell.

Entering service menu for Motorola MTH800 is available using combination of four keys: *, #, MENU and RIGHT ARROW. Service menu for Sepura STP8000 is available using combination of four keys: *477.



Figure 2. Motorola MTH800 and Sepura STP8000 terminals in service menu

Basic information about *RSSI* and *CX* from serving cell and adjacent site that are available from service mode can be very useful for basic troubleshooting and it is possible to get information from nontechnical person. Service mode is implemented in all mobile station and it is free tool, network operator can easily guide end users to enter this menu. Trough Zone Configuration Manager (ZCM) it is possible to define profile for Mobile Station to constantly use same serving cell and to determine cell coverage. After determing of coverage for serving cell and for adjacent site together with information about *RSSI* reselection parameters can be set properly.

Motorola Scout is full solution for air interface monitoring and it can be used for testing of TETRA system reselection. Mobile terminal is connected to PC running Scout software using serial RS232 connection. Scout software tool can be used only using USB key with HASP software licensing.

Using Scout information that Mobile station (MS) is receiving trough Broadcast Network Channel (BNCH) and Broadcast Control Channel (BCCH) can be seen. Serving Cell Data window provide information about serving cell reselection parameters, operating channel *RSSI*, *C1* and status of serving cell.

Adjacent Sites Data window provide basic information about adjacent sites that are defined trough Zone Configuration Manager, *RSSI*, *CX* and operating Channel. Any mistake during adjacent site configuration in Zone Configuration manager will be seen on this display.

Scout software can obtain geographic coordinates from Mobile station or using external GPS receiver and it is possible to show on map position of terminal and position of serving cell and adjacent cells.

Microsoft MapPoint maps are used and if database with information about base stations in TETRA network is implemented also symbols and labels for Base Stations will be displayed. Distance to the sites can be shown and GPS track in different colors depending on *RSSI* measured at Mobile Station (MS). Special marks are placed when cell reselection occurs.

RSSI graph for serving cell and monitored adjacent cells is drawn that can provide useful information.

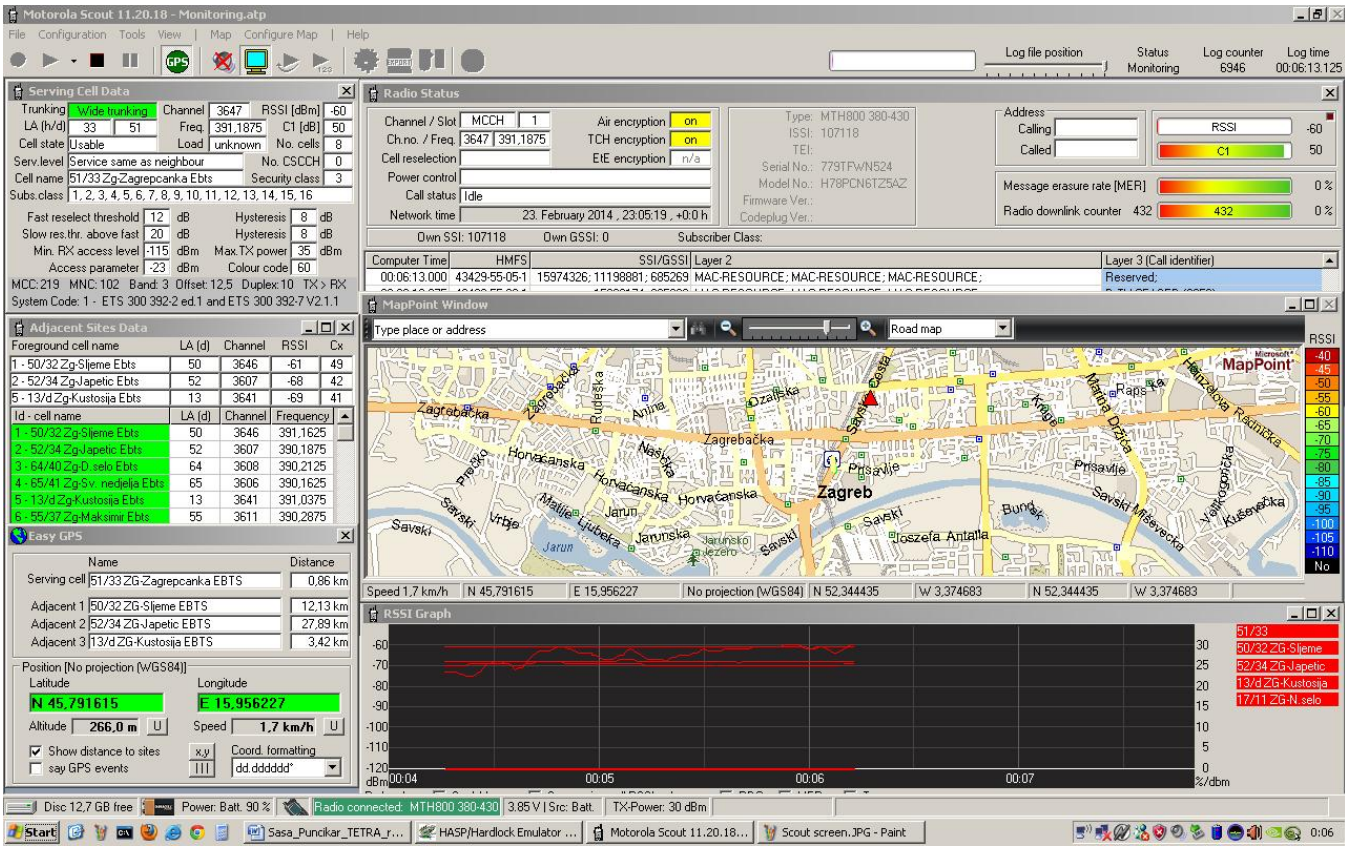


Figure 3. Motorola Scout software

Force channel menu delivers possibility to lock Mobile Station to specific radio channel (specific Base Station).

It is possible to lock MT to specific cell to determine coverage area for any specific cell much easier and much faster way then by Zone Configuration Manager.

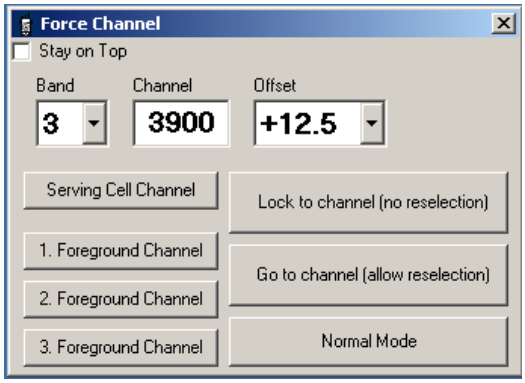


Figure 4. Force Channel menu in Scout software

VIII. CONCLUSION

When configuring TETRA system or adding additional base station in existing system special attention should be dedicated to reselection parameter and defining adjacent sites to enable smooth cell reselection.

After installation empirical field measurement should be done using appropriate tool to verify that all parameters are set properly. During exploitation period using same tools it is possible to determine possible coverage problems due to malfunctioning of base stations, antennas or other parts of system.

Most important parameters for reselection process, *Slow reselect threshold* (*SRT*), *Fast reselect threshold* (*FRT*), *Slow reselect hysteresis* (*SRH*) and *Fast reselect hysteresis* (*FRH*) should be set according to manufacturer specification. After installation and verification at field specific modification should be done for each Base Station since coverage and signal propagation depend on geographic profile and user density.

REFERENCES

[1] Motorola, "The Scout TETRA Air Interface Monitoring and Recording Software for Windows User Manual", September 2008;

[2] Motorola, "Dimetra IP System Documentation", July 2010;

[3] Motorola, "Dimetra IP 2004 Configuration Management", October 2005;

[4] Terrestrial Trunked Radio (TETRA); Voice plus Data (V+D); Designers' guide; Part 1: Overview, technical description and radio aspects, Retrieved February 14, 2014 from http://www.etsi.org/deliver/etsi_etr/300_399/30001/01_60/etr_30001e01p.pdf.

[5] Terrestrial Trunked Radio (TETRA); User Requirement Specification TETRA Release 2.1; Part 4: Air Interface Enhancements, Retrieved February 14, 2014 from http://www.etsi.org/deliver/etsi_tr/102000_102099/10202104/01.04.01_60/tr_10202104v010401p.pdf.

[6] Terrestrial Trunked Radio (TETRA); TETRA mobiles moving at high velocity, Retrieved February 14, 2014 from http://www.etsi.org/deliver/etsi_tr/102700_102799/102753/01.01.01_60/tr_102753v010101p.pdf.

Inventory management system for water supply network

O. Jukić, I. Heđi
Virovitica College, Virovitica, Republic of Croatia
{oliver.jukic | ivan.hedi}@vsmti.hr

Abstract - In this paper, inventory management system for water supply network is presented. Considering the existing experience in telecommunications network management area, we have made comparison between basic aspects of telecommunications network management and water supply network management. As a first step, we have detected the need for developing water supply network inventory system, covering configuration management functional area. The main goal of this system is to collect, process and store inventory data through graphical user interface, to be presented to network operating personnel. Related work is researched and briefly presented. After that, the inventory management system solution is presented. Complete network structure is modelled as information model in Data Warehouse. Graphical user interface, developed in web technology with Google maps as background, with basic knowledge integrated, is presented. Some basic functions of GUI (graphical user interface) are discussed, such as: logging of network maintenance and repair actions, network planning support and reporting. Finally, we have tried to define future steps for evolution of this inventory management system in complete water supply network solution. Hence, we have made basic discussion related to technical aspects of real-time data collecting in water supply networks, introducing performance and fault management functional areas in water supply network management.

Keywords: Water supply network, network management, leak detection

I. INTRODUCTION

A. Telecommunication and water supply network

Services provided by network operators rely on operator's network infrastructure. Network infrastructure capacity (read: "cost") should be optimized, but service level from the customer's point of view must not be degraded. Because of that, every network operator has some kind of centralized network and/or service operation and maintenance center (official names vary, such as "Operation and Maintenance Center", "Network/Service Operation Center", etc.), tracing current network status and triggering appropriate restoring actions in case of network problems.

Originally, above text was published in [3] within the context of telecommunications network. However, it can be applied to any other network type, such as water supply network. Indeed, there is a network, as a structured collection of nodes (vertices) joined by edges (links); there is a specific network infrastructure (pipes, valves, etc.), there is a service provided; there is measurable service quantity (cubic meters of water delivered); and finally, there is measurable service quality (i.e. chemical content of water, network down-time in seconds, average water pressure in specific network link, etc.).

TABLE I: COMPARISON OF NETWORK ELEMENTS IN TELECOMMUNICATION AND WATER SUPPLY NETWORK

Network elements and terms	
Telecommunication network	***Water supply network***
Optical link; airwave link; UTP link, etc.	Pipe
Router, switch, telephone exchange	i.e. T- valve
Firewall	Valve
Data flow (bit/s)	Water flow (l or m^3/s)
Information flow source – e.g. Video on demand server	Water flow source – e.g. water wellhead

In network theory, there are four general classes of networks: technological networks, social networks, information networks and biological networks. The division into classes is a useful one, since networks in the same class are often treated using similar techniques, ideas and algorithms [4]. Both telecommunication and water supply network belong to the class of technological networks. Terminology and brief network elements comparison is shown in table 1.

B. Inventory (configuration) management

In telecommunication management network concept, there are five main system management functional areas (SMFA). Those system management functional areas are:

- Configuration management
- Fault management
- Performance management

This paper is result of project "Knowledge-based water supply system management" sponsored by company Komrad d.o.o., Slatina

- Accounting management
- Security management

Configuration management is used to locate resources, and to keep track of the resources and their details. Fault management primarily covers detection, isolation and correction of any unusual network operational behavior. Performance management covers performance data collection, analysis, presentation and reporting. Accounting management covers the usage of resources, including charges for this usage. Finally, security management covers detection of security violation [2].

All of those areas are also of great importance in management of water supply networks. Basic functional area is configuration management, which is the basis for all other functional areas. Hence, we have started our research and development activities related to that area.

In water supply network almost all network elements are invisible, located in the ground. In most cases, network elements are not accessible through any communication infrastructure, and they are "software-free" elements. Hence, there is no possibility to implement self-detection or auto-discovery features. Hence, it is rather important to implement and maintain network configuration database. Water supply network operators usually have "database" as a collection of different type documents describing network, even in the electronic form. However, that is not enough for maintenance and management activities. There is very strong need for information system for network configuration management, which we call "inventory" management (in water supply network we use term "network inventory" for all elements building water supply network). As a first step in our research we have focused on related work and developing of our own water supply inventory management system.

C. *Related work*

Vienna Waterwork's approach to network maintenance and rehabilitation is described in [1]. Network information system of Vienna Waterwork is briefly presented. System is primarily used to provide assistance platform for planning process. For predicting pipe failures the proportional hazards model (PHM) is used.

GIS-based water supply information system for an urban area is described in [7]. The basic aim of that water supply information system is to serve as a decision support tool for Waterwork's management by helping maintain records of leaks and repairs.

Both systems are focused on inventory management, but currently without GPS coordinates. Real-time data integration is also not considered.

In [8] system using telemetry for obtaining real-time data from water-supply network is described. However, all data are propagated to SCADA application; inventory management is not considered.

All facts mentioned above have encouraged us to continue research and development activities leading to the integration of real-time collected data with network inventory system based on GPS coordinates.

II. INVENTORY MANAGEMENT SYSTEM ARCHITECTURE

A. *Database layer*

System core is located at database layer. There are three main databases: network inventory database, repairing action database and real-time performance database. Central database in the context of network inventory system is network inventory database. It can be assumed as configuration database. Inventory management system architecture for water supply network is shown on figure 1.

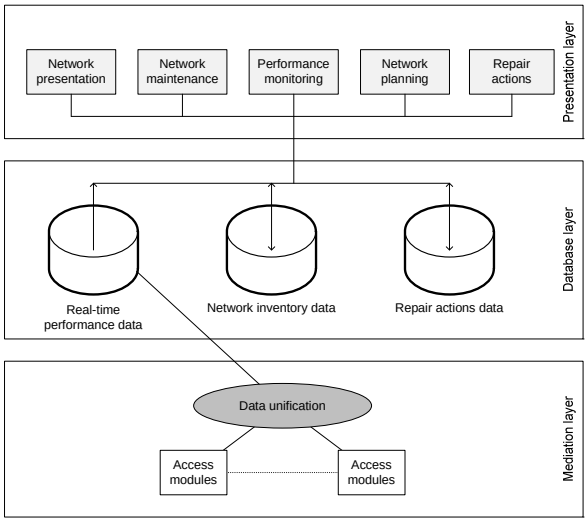


Figure 1. Inventory management system architecture

Since there are no auto-discovery or self-detection features in a "software-free" network, network inventory data must be imported into the database through graphical network presentation interface. Data collection process includes consolidation of all possible sources within the company, such as: network schemas, network descriptions, project plans, repairing logs, internal human knowledge, informal notes etc. Generally, there are two types of network elements: basic elements, such as valves or pipes, that cannot be divided into the smaller parts; compound elements, such as manhole, are compound of number of basic elements. UML class diagram shown on figure 2 presents one compound element containing basic elements such as pipes or valves.

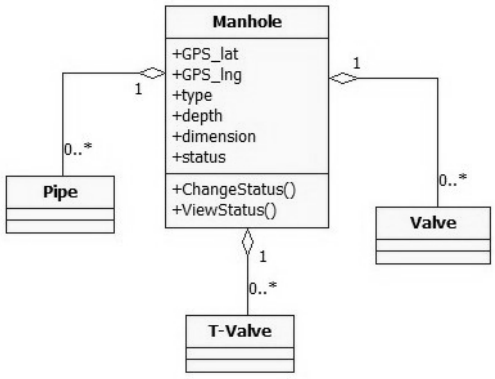


Figure 2. UML class diagram of network element types

GPS coordinates are a very important part of network inventory data model, since they allow network presentation on any GPS supporting platform (figure 3).

Network inventory database is the foundation for all other network-related data and network maintenance and management actions. Repair actions database represents some kind of "log" or "network diary", where all data about network repairing actions are stored and paired with an appropriate network element. Repairing actions data are usable for prediction of future pipe failures using appropriate prediction models. By knowing the number of failures in the past, it is possible to predict time of the next failure with a certain probability. Research and development of suitable models will be subject of the future work, while one possible approach (proportional hazards model - PHM) is described in [5]. All mentioned information are an important part of network planning process.

When water-supply network information model is stored in database, it can be used as foundation for performance management. Every single network element is allowed to become performance measurement point, where specific key performance indicators can be measured. All measurements should be collected in real-time manner from the water-supply network.



Figure 3. Network inventory data presentation on GPS supporting platform

B. *Mediation layer*

Main functions of mediation layer are:

- data collection from water-supply network
- unification of all data formats
- aggregation of collected data on low level, preparing it for storage in real-time performance data warehouse

All data are collected by specific applications called access modules. Every access module uses specific techniques and performs specific actions in order to obtain data from the water-supply network. For instance, data can be collected through any kind of communication protocol (usually using GPRS for data transport), from existing databases or industrial measurement solutions.

All collected measurements are stored into a real-time performance data warehouse. Every measurement is related to some specific performance measurement point, which is defined in a water-supply network information model. This is the basis for data unification. Namely, every measurement is tuple containing:

- performance measurement point (where was it measured?)
- performance measurement indicator name (what was measured?)
- performance measurement value (which value was measured?)

Following the data unification, aggregation of collected data can be performed. Data can be aggregated by any mathematical function, such as sum, average or count. Aggregation can be grouped by performance measurement points or performance measurement indicator names.

C. *Presentation layer*

Network presentation GUI is web interface located on the top of database layer. There are five main modules within network presentation layer:

- Network presentation
- Network maintenance
- Performance monitoring
- Network planning
- Repair actions

Network presentation is aimed to present data about network inventory to the user. However, it is possible to include some other data sources into the graphical presentation. For instance, some very important performance measurement values can be placed on the main screen. This graphical interface is layered, with zooming feature included. Furthermore, compound elements can be "drilled-down", showing its content (figure 4).

Performance monitoring graphical interface serves as "performance cockpit" where all relevant performance measurement indicators can be monitored. Also, some performance measurement configuration data can be set here.

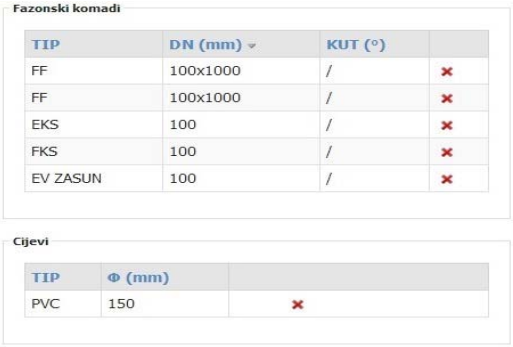


Figure 4. Content of the compound element

Network planning is used for performing different algorithms on network pipes and node elements, in order to predict failures in the future. Input of this process is number of repair actions per network element, which is then processed by algorithm. As a candidate for algorithm arises proportional hazards model (PHM) [6]. This part of the system is still under construction.

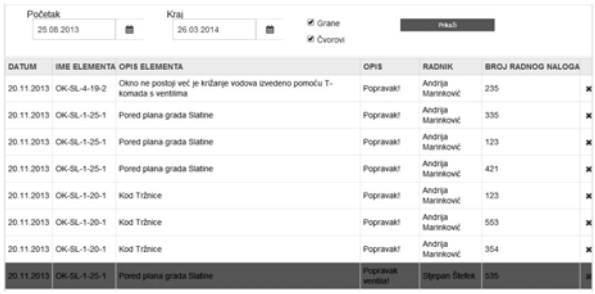


Figure 5. Repairing actions graphical user interface

In order to obtain data about past repairing actions, repair action GUI is developed (figure 5). This is where all data related to some specific repairing action are "logged". Activities during repair are shown on figure 6 using UML activity diagram.

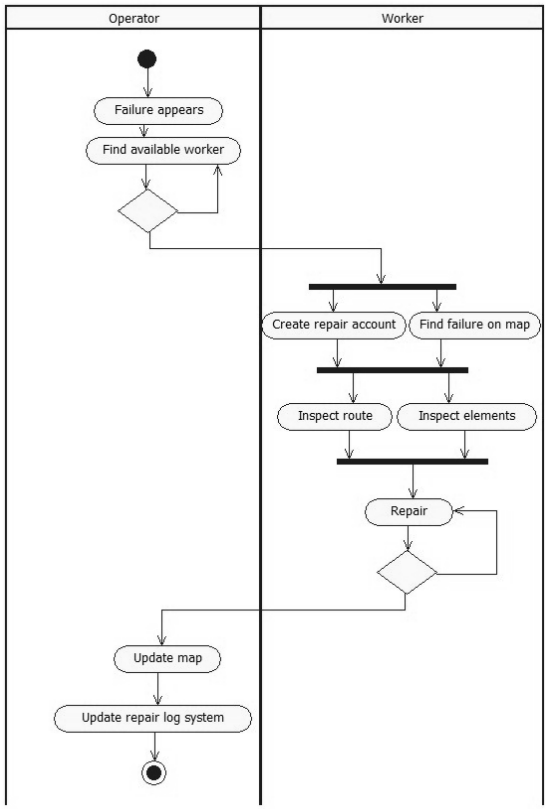


Figure 6. UML activity diagram for repair actions

Since every repair action is followed by number of official records, internal to organization, repairing actions database is typically connected to some internal data sources, for instance working orders database in some ERP (enterprise resource planning) system.

III. CONCLUSION

In this paper we have presented information system for inventory management in water-supply network. Based on existing experience in telecommunications management area, we have compared basic aspects of management in water-supply and telecommunication networks. Encouraged by the results, we have developed inventory management system for water-supply network.

Complete network structure is modelled as information model in Data Warehouse. Graphical user interface, developed in web technology with Google maps as background, with basic knowledge integrated, is presented.

The system was developed and launched during the project "Knowledge-based water supply management" sponsored by company "Komrad d.o.o.", Slatina. It is used during network management, maintenance and repairing process. Some of system components are still subject of research efforts; it will be subject of future work and will be presented in future papers. It is primarily related to real-time data collecting functionality, which will enable performance data integration into the network planning process and network presentation graphical user interface.

REFERENCES

[1] D. Kasses, D. Fuchs-Hanusch, "Vienna Waterwork's approach to network maintenance and rehabilitation" Water 21 – Magazine of the Internationa Water Association, United Kingdom, pp. 33-35, February 2013

[2] D.K.Udupa, "TMN – Telecommunications Management Network", McGraw-Hill Telecommunications, New York, 1999.

[3] O. Jukić and M. Kunštić, "Integrated view on telecommunication network status", Proceedings of the 34rd International Convention MIPRO 2011, pp. 141-145, MIPRO, Opatija, 2011.

[4] M.E.J. Newman, "Networks – An introduction", Oxford University Press, Oxford, 2010.

[5] D. Fuchs-Hanusch, B. Kronberger, F. Friedl, R. Scheucher, "Whole of Life Cost Calculations for Water Supply Pipes", Water Asset Management International, 8 (2), 19-24, 2011.

[6] D. R. Cox, "Regression Models and Life-Tables", Journal of the Royal Statistical Society. Series B (Methodological), Vol. 34, No. 2. (1972), pp.

[7] A.K. Dikshit, P.V. Krishnarao, "A GIS-based Water Supply Information System for an Urban Area", Managing Environmental Knowledge, EnviroInfo, Graz, 2006.

[8] M. Danijel, T. Lisak, G. Malčić, "Telemetrija i upravljanje vodoopskrbnim sustavom", Proceedings Vol. III MIPRO 2009, Computers in technical systems & Intelligent systems, pp. 163-167, MIPRO, Opatija, 2009.

Procjena vrijednosti parametra SAR u blizini GSM/UMTS baznih stanica

D. S. Šuka*, P. S. Medjedović**, M. I. Simić***
*Elektrotehnički fakultet Istočno Sarajevo,Republika Srpska, BiH
** Ministarstvo za ekonomske odnose i regionalnu saradnju u vladi RS, Republika Srpska, BiH
*** Elektrotehnički fakultet u Beogradu, Republika Srbija
darko.suka@etf.unssa.rs.ba
p.medjedovic@meoi.vladars.rs.net
mira@etf.rs

Sažetak - Dnevno izlaganje elektromagnetskim poljima GSM/UMTS baznih stanica, iz godine u godinu, povećava javnu zabrinutost o mogućim negativnim efektima po zdravlje ljudi koji žive i rade u njihovoj blizini, a koje danas predstavljaju najrasprostranjeniji izvor nejonizujućeg zračenja. Međunarodni standardi i smjernice ustanovljeni za ograničavanje izloženosti ljudi EM poljima podjeljeni su u dvije kategorije: Osnovna ograničenja (SAR - *Specific Absorption Rate*, indukovana gustina struje i indukovana gustina snage) i Referentni nivoi (jačina električnog polja, jačina magnetskog polja i gustina snage u slobodnom prostoru). Da bi se provjerilo slaganje sa obje ove kategorije, u radu su predstavljena mjerenja jačine el. polja i gustine snage, te su dobijeni rezultati iskorišteni za numeričku predikciju SAR. Mjerenja su izvršena u urbanoj zoni u periodu vršnog (*peak*) saobraćaja. Rezultati su upoređeni sa ICNIPR, IEEE i CENELEC standardima.

I. UVOD

Povećana upotreba mobilnih komunikacija dovela je do postavljanja GSM/UMTS baznih stanica, čak i u blizini kuća, škola, itd, odnosno u gusto naseljenim područjima (zone povećane osjetljivosti). Ovo podiže zabrinutost javnosti u pogledu sigurnosti stanovništva u odnosu na elektromagnetsko (EM) zračenje iz takvih izvora. Uradjene su mnoge studije o potencijalnim biološkim i termičkim efektima EM polja iz GSM/UMTS opsega, ali jasnog stava o mogućoj štetnosti za sada jos uvijek nema. Interakcija GSM/UMTS EM polja i ljudskog organizma treba da obuhvati više aspekata, [1]:

- „Materijal“ (ljudsko telo) ima veoma neobične elektromagnetske osobine i svojstva: električna permitivnost, električna provodljivost;
- Ove osobine nisu sasvim poznate i zavise od aktivnosti osobe;
- Ovaj „materijal“ je aktivan materijal na ćelijskom nivou;
- U većini slučajeva, problem je zapravo spregnut: termički efekat je jedan od glavnih efekta i u zavisnosti je od cirkulacije krvi u organizmu;
- Geometrija i orijentacija ljudskog tijela takođe utiču na apsorpciju elektromagnetske energije u tkivima;
- Raspodjela elektromagnetskog polja je veoma složena i mora se uzeti u obzir aproksimacija ravanskim talasima u dalekoj zoni zračenja.

S druge strane, nadležne organizacije su uspostavile smjernice za ograničavanje izlaganju EM poljima. Međunarodni standardi i smjernice su date u dvije kategorije: (a) Osnovne ograničenja - ograničenja o izlaganju vremenski promjenljivim EM poljima direktno zasnovanim na ustanovljenim zdravstvenim efektima (opsegu GSM/UMTS frekvencija fizička količina koja se koristi da se izračuna ovo ograničenje je SAR) i (b) Referentni nivoi - obezbjeđuju praktičnu procjenu izloženosti u cilju provjere da li su prekoračena osnovna ograničenja. Osnovni cilj provjere usklađenosti za vrijeme ljudske izloženosti EM polju je da data osnovna ograničenja budu zadovoljena. Iz tog razloga sve smjernice i preporučena ograničenja za izlaganje ljudi EM poljima u GSM/UMTS opsegu dati u pogledu SAR (*Specific Absorption Rate*) vrijednosti.

II. SPECIFIČNA BRZINA APSORPCIJE (SAR)

Polja iz opsega radiofrekvencija koja prodiru u tijelo opadaju sa porastom frekvencije. U cilju razumijevanja efekata koje ovo prodiranje može da ima na ljudsko tijelo, potrebno je utvrditi veličinu polja u različitim dijelovima tijela koji su izloženi polju. Ovo zahtijeva poznavanje električnih svojstava različitih vrsta tkiva. Kada se odrede ova svojstva, moguće je proračunati vektore E i H za svaki dio tijela koji je izložen određenoj vrsti izvora. SAR se definiše kao:

$$SAR = \frac{\sigma \left| E^2 \right|}{\rho_m} \left[\frac{\mathrm{W}}{\mathrm{kg}} \right] \qquad (1)$$

gdje je: σ - provodljivost tjelesnog tkiva, E - jačina el. polja u posmatranoj tački, ρ_m - gustina mase tkiva u toj tački. Pošto je SAR, promjena apsorbovane RF energije u jedinici vremena po jedinici mase, veoma teško i složeno mjeriti u biološkim tkivima, standardi dozvoljavaju upotrebu referentnih nivoa gustine snage (W/m^2) u slobodnom prostoru. IEEE standardi ustanovili su osnovna ograničenja za efektivne (*rms*) vrijednosti jačine električnog i magnetskog polja, E i H, tzv. maksimalnu dozvoljenu izloženost (MPE, *Maximum Permisible*

Exposure) i slično kao i ICNIRP standard definišu referentne granice za jačine polja u slobodnom prostoru. Zadovoljavanjem ovih ograničenja osigurano je poštovanje SAR vrijednosti. Dakle, umjesto kompleksnih SAR mjerenja, za usklađenost procjene, gore pomenuti standardi nam omogućavaju jednostavnije mjerenje jačine električnog polja (*rms*). Kao što je spomenuto u uvodu, vrijednost SAR-a zavisi od nekoliko faktora, [2,3]:

- Parametara incidentnog (upadnog) polja, kao što je npr. frekvencija, intenzitet, polarizacija i konfiguracija izvor – objekat (da li se radi o bliskom ili dalekom polju),
- karakteristika izloženog tijela, kao što su veličina tijela, interna i eksterna geometrija, kao i dielektrična svojstva raznih tkiva,
- uticaja ostalih objekata koji se nalaze u polju u blizini izloženog tijela.

Kada je vektor *E* paralelan sa osom koja prolazi kroz centar ljudskog tijela i kada se to tijelo nalazi u dalekom polju, SAR za cijelo tijelo dostiže svoj maksimum. Količina apsorbovane energije zavisi od mnogo faktora, uključujući i veličinu tijela. Za "čovjeka prosječne veličine", koji nije uzemljen, rezonantna apsorpciona frekvencija (frekvencija na kojoj se apsorbuje najviše energije) je 70 MHz, dok ista može varirati kod osoba nestandardnih dimenzija, viših, nižih, djece i osoba koje sjede. Kod osoba koje su uzemljene vrijednost ove frekvencije je manja dva puta. S druge strane, kada imamo uslove izlaganja koji odgovaraju bliskom polju, onda treba napomenuti da se ove frekvencije, kao i uopšte procjena SAR-a potpuno razlikuje. U ovom slučaju, prilikom izlaganja može doći do velikih vrijednosti lokalnog SAR-a, a ta vrijednost, kao i vrijednost SAR-a za cijelo tijelo je jako zavisna od rastojanja čovjeka od izvora visokih frekvencija. Iz (1) vidi se da vrijednost SAR-a zavisi od provodnosti i gustine tkiva. S obzirom na to da se provodnost mijenja u različitim dijelovima tijela, postoji i zavisnost SAR-a od položaja tijela, što je i jedan od razloga zbog kojeg se SAR nekada računa za cijelo tijelo, a nekada za lokalne dijelove tijela (kao što su glava, vrat, trup i ekstremiteti). Gustina tkiva je poprilično konstantna duž tijela, tako da promjena SAR-a ne zavisi od njene vrijednosti. Postojanje drugih objekata u polju može da utiče na samo polje preko intenziteta, polarizacije, i na taj način da promijeni vrijednosti za SAR. SAR se proračunava na osnovu eksperimentalnih podataka iz laboratorije ili određenim numeričkim metodama. To znači da prilikom provjere kompatibilnosti sa nekim graničnim nivoom iz standarda nisu česti slučajevi u kojima će se to provjeravati direktno iz vrijednosti SAR-a. Zbog toga su date tzv. Referentne vrijednosti koje su izražene nekim veličinama koje se lako mogu izmjeriti (npr. vektori *E* i *H*), a koje odgovaraju određenim vrijednostima SAR-a. Inače, SAR pored toga što se nalazi u standardima kao granična vrijednost, takođe se daje kao aproksimacija količine apsorbovane energije koja potiče od nekog uređaja u vidu podatka za taj uređaj, kao što je to slučaj npr. sa mobilnim telefonima. U odnosu na apsorpciju energije u tijelu EM polja se mogu podijeliti u četiri opsega, [2, 3]:

- Frekvencije od približno 100 kHz do vrijednosti malo manje od 20 MHz, na kojima apsorpcija u trupu opada brzo sa opadanjem frekvencije, a značajna apsorpcija može da se pojavi u vratu i nogama,
- frekvencije u opsegu od 20 MHz do 300 MHz, na kojima relativno visoke apsorpcije mogu da se pojave u cijelom tijelu, a ukoliko su dostignute rezonantne frekvencije mogu biti još veće,
- frekvencije u opsegu od 300 MHz do nekoliko GHz, kod kojih imamo lokalnu i neuniformnu apsorpciju i
- frekvencije preko 10 GHz, kod kojih se apsorpcija energije dešava primarno na površini tijela.

Koncept SAR ima smisla samo u opsegu od 100 kHz do 10 GHz, jer je u tom području dubina penetracije veća od 1 cm, [2,3]. Gustina indukovane struje je važan parametar na frekvencijama ispod 100 kHz, dok se na frekvencijama iznad 6 GHz energija apsorbuje površinski, pa je važna gustina upadne snage. Postoji opsežna literatura o procjeni srednje vrijednosti SAR za cijelo tijelo i prostornoj raspodjeli SAR za različite modele životinja, pa čak i za model čovjeka. Neke ranije procjene bazirale su se na prostim sferičnim ili elipsoidnim modelima, npr. Durnijev, [4]. Međutim, u novijim istraživanjima koriste se numeričke simulacije anatomski korektnijih modela odraslog čovjeka. SAR je ključni koncept u planiranju i analizi eksperimenata, kako onih *in vivo*, tako i *in vitro* i služi kao osnova savremenim standardima za zaštitu ljudi od RF/MT zračenja. Pri tom treba istaći da podjednaku važnost imaju i prosječna vrijednost za SAR za cijelo tijelo, kao i lokalni maksimumi prostorno usrednjene vrijednosti za SAR, [5].

III. PRORAČUN SAR VRIJEDNOSTI

U zoni dalekog polja izloženost opšte populacije zračenju baznih stanica može se smatrati kao izloženost ravanskim elektromagnetskim talasima. Na frekvenciji GSM 900 MHz, prosječna provodljivost ljudskog tijela je $\sigma = 0{,}97$ S/m, dok je relativna permitivnost $\varepsilon_r = 47{,}4$. Na frekvencijama GSM 1800 MHz, te UMTS 2100 MHz, ove vrijednosti su približno: ($\sigma = 1{,}35$ S/m, $\varepsilon_r = 45{,}21$) i ($\sigma = 1{,}5$ S/m, $\varepsilon_r = 45{,}2$), u skladu sa srednjom vrijednosti, računatom za cijelo tijelo prema [6]. Relativna magnetska permeabilnost ljudskog tijela, prema [7], približno iznosi $\mu_r = 0{,}99$, dok je prosječna gustina ljudskog tijela, prema [8], 1062 kg/m^3. Imajući u obzir gore navedene činjenice, SAR se može izračunati u funkciji incidentnog električnog ili magnetskog polja koristeći (1):

$$SAR = \frac{\sigma}{\rho} \cdot \frac{\mu\omega}{\sqrt{\sigma^2 + \varepsilon^2\omega^2}} \left(1+\gamma_r\right)^2 \left|H\right|^2, \qquad (2)$$

$$\gamma_r = \frac{2\left|\sqrt{\varepsilon'}\right|}{\left|\sqrt{\varepsilon'} + \sqrt{\varepsilon_0}\right|} - 1, \qquad (3)$$

$$\varepsilon' = \varepsilon + j\frac{\sigma}{\omega}, \qquad (4)$$

$$H = \frac{E}{Z_0}, \tag{5}$$

$$SAR = \frac{\sigma}{\rho} \cdot \frac{\mu\omega}{\sqrt{\sigma^2 + \varepsilon^2\omega^2}} (1+\gamma_r)^2 \frac{|E|^2}{Z_0^2}, \tag{6}$$

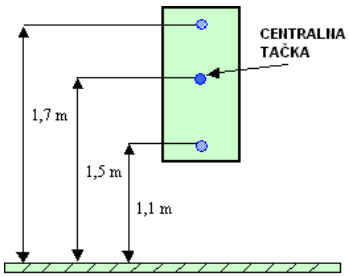
pri čemu su: γ_r – odgovarajući koeficijent refleksije, ε' – kompleksna permitivnost sredine, H i E – efektivne (*rms*) vrijednosti magnetskog i električnog polja, respektivno.

U ovom radu su predstavljeni rezultati mjerenja jačine el. polja i gustine snage u blizini bazne stanice u urbanom okruženju. Procjena izloženosti se vršila u cilju provjere nivoa izloženosti elektromagnetskom zračenju, te poređenju istih sa odgovarajućim standardima. Na osnovu dobijenih rezultata, koristeći izraz (6), prikazane su i SAR vrijednosti.

IV. MJERENJE JAČINE ELEKTRIČNOG POLJA I GUSTINE SNAGE U BLIZINI BAZNE STANICE

Precizno mjerenje jačine el. polja i gustine snage u složenom i dinamičnom okruženju je težak zadatak. To je uglavnom zbog refleksije, apsorpcije i interferencije EM talasa. Stoga, različita mjerenja mogu dovesti do vrlo različitih rezultata zbog ovakvih uslova [9]. Mjerenja EM zračenja baznih stanica (BS) mobilne telefonije izvršena u ovom radu su samo jedan od segmenata iz oblasti mjerenja zračenja radiofrekventnih EM polja, u skladu sa [10, 11]. U nastavku je ukratko opisan postupak mjerenja u skladu sa Slučajem 2 (skeniranje promjenljivog frekventnog opsega po određenim frekvencijama u dalekoj zoni zračenja) Pravila 37/2008. Ovaj tip mjerenja je najbolje raditi koristeći lagane, baterijski napajane analizatore spektra (AS). Analizator spektra bi trebao da ima mogućnost softverske kontrole, koja je neophodna zbog velike količine frekventnih i amplitudskih podataka koji se moraju prikupiti tokom ispitivanja, a da bi se zadržala konzistentnost rezultata dobijenih od većeg broja kompleta mjerne opreme kojom rukuje više ljudi. Softver treba da ima mogućnost programiranja faktora pojačanja antene i gubitaka na konektorima kablova. Ovo će dozvoliti sistemu za ispitivanje da koristi različite antene i kablove dozvoljavajući veći stepen prilagođavanja za svaki opseg. Na ovaj način mogucnost greške ljudskog faktora se svodi na minimum. Analizatori spektra će povremeno raditi u opstruirajućem RF okruženju (*hostile RF environments*). Stoga, dobar dinamički opseg i intermodulacione preformanse će biti od najveće važnosti za vjerodostojne i ponovljive rezultate. Antene za ispitivanje trebaju biti lagane i robusne, a trebaju se koristiti i napojni kablovi dobrog kvaliteta. Antene koje se preporučuju za korištenje su: (a) Magnetna petlja za HF područje; (b) Širokopojasna dipol antena ili (zatvorena) *log* periodic antena; (c) Bi-konična antena; (d) Direkciona antena za druge emisione tipove (korištenje ovih antenna se preporučuje kada postoji glavno zračenje, dok je sekundarno zračenje zanemarljivo); (e) Izotropna, trosna antena. Sva mjerna oprema treba biti kalibrisana (u skladu sa preporukama proizvođača ili prema postupcima za izvođenje kontrole kvaliteta). RF instalacije (kablovi, konektori, talasovodi...) se trebaju, svaki pojedinacno, odrediti i provjeriti prije upotrebe zbog mogućih mehaničkih oštećenja. Svaka promjena pojačanja faktora antene i gubici na kablovima trebaju se programirati u mjerni prijemnik/softver. Postupak mjerenja se treba odvijati u sljedećim koracima:

a) **Mjerna tačka.** Mjerenja treba obaviti u jednoj jedinstvenoj tački na visini od 1,5 m od nivoa zemlje (ili nivoa poda).



Slika 1. Centralna mjerna tačaka u zoni dalekog polja

b) **Frekventni opseg.** Mjerni postupak iz Slučaja 2 se koristi za frekventni opseg „9 kHz – 3 GHz".

c) **Podešavanje analizatora spektra (AS).** Usljed različitih RF izvora u radio spektru, mjerna širina opsega trebala bi biti kompromisna veličina. Unutar spektra postoji mješavina širokih/uskih, analognih/digitalnih i kontinualnih/diskontinualnih izvora zračenja. Preporučuje se za analizatore spektra da koriste sljedeću širinu opsega/vrijeme ispitivanja (konkretno za posmatrani opseg u Slučaju 2, 9 kHz – 3 GHz): RBW = 100 kHz, VBW = 300 kHz sa vremenom ispitivanja od 700 ms – 1 s ili prema preporukama proizvođača. Prag osjetljivosti od –40 dB (može do -70 dB) se odnosi na usvojeni nivo iz smjernica. Antena treba biti korištena u oba, horizontalnom i vertikalnom položaju. Takodje, trebaju biti korišteni "*Max-hold*" i "*Peak*" način rada (za detektor).

d) **Obrada podataka.** Imajući u vidu da se razmatra slučaj mjerenja u dalekoj zoni zračenja, mag. polje H ili gustina snage S mogu se, u odnosu na prostiranje u slobodnom prostoru (impedansa vazduha $Z_0 = 377\ \Omega$), prema poznatoj vrijednosti el. polja, izračunati prema (7):

$$S = EH = E^2 / Z_0 = H^2 Z_0. \tag{7}$$

Da bi se izračunala ukupna gustina snage u podesivom frekventnom opsegu, za AS korišten u ovom radu, neophodno je bilo znati faktor pojačanja antene u tom opsegu, prema [12]. Za ovu konverziju se koristi izraz (8):

$$S = \frac{10^{\left(\frac{p-G}{10}\right)}}{1000} \cdot \frac{4\pi}{\lambda^2} \tag{8}$$

gdje je S gustina snage (u W/m^2), p je izmjerena snaga (u dBm), G je faktor pojačanja antene u odnosu na izotropni radijator (u dBi) i λ je talasna dužina izvora (u metrima).

e) **Provjera koeficijenta ukupne izloženosti.** Za frekvencije iznad 100 kHz sljedeći zahtjevi bi

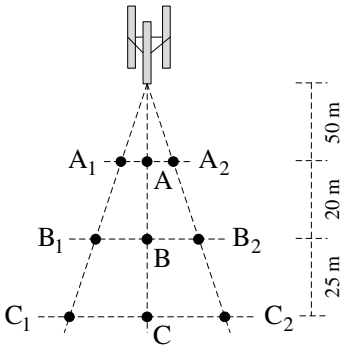
trebali da budu ispunjeni za vrijednosti el. i mag. polja, te srednju gustinu snage EM zračenja:

$$\sum_{i=100kHz}^{1MHz}\left(\frac{E_i}{c}\right)^2+\sum_{i>1MHz}^{300GHz}\left(\frac{E_i}{E_{L,i}}\right)^2\le 1 \qquad (9)$$

$$\sum_{j=100kHz}^{1MHz}\left(\frac{H_j}{d}\right)^2+\sum_{j>1MHz}^{300GHz}\left(\frac{H_j}{H_{L,j}}\right)^2\le 1 \qquad (10)$$

$$\sum_{i=1}^{N}\frac{S_i^{meas}}{S_i^{guid}}=\frac{S_1^{meas}}{S_1^{guid}}+\frac{S_2^{meas}}{S_2^{guid}}+\ldots+\frac{S_N^{meas}}{S_N^{guid}}<1. \qquad (11)$$

gdje su: E_i intenzitet električnog polja na frekvenciji i, $E_{L,i}$ referentni nivo električnog polja dat u Tabeli II u odnosu na ICNIRP standard, H_j intenzitet magnetskog polja na frekvenciji j, $H_{L,j}$ referentni nivo magnetskog polja prema [13], $c = 87/f$ V/m i $d = 0{,}73/f$ A/m za izlaganje opšte populacije, S_i^{meas} izmjerena gustina snage na frekvenciji i, te S_i^{guid} referentni nivo gustine snage dat Tabeli II u odnosu na ICNIRP standard.



Slika 2. Raspored mjernih tačaka u zoni dalekog polja



Slika 3. (a) Spectran HF-4040; (b) HyperLOG 7040, Aaronia



Slika 4. EME SPY 140, Satimo

Mjerenja u ovom radu su obavljena u više tačaka u pristupačnoj zoni dalekog polja na platou Elektrotehničkog fakulteta Istočno Sarajevo. Sve mjerne tačke su odabrane kao najbolji reprezenti raspodjele i maksimalnih nivoa EM polja, tako da je većina njih bila u osi glavnih snopova usmjerenih dijagrama zračenja antena. Slika 2 prikazuje raspored mjernih tačaka u odnosu na lokaciju bazne stanice smještene na krovu Elektrotehničkog fakulteta u Istošnom Sarajevu. U ovoj zoni, jačinu el. polja i srednju gustinu snage čine komponente upadnog polja (LOS – *Line of Sight*), te komponente polja nastale usljed refleksije, rasijavanja, itd. [1, 2]. U cilju preciznijih rezultata mjerenja istovremeno su korištena dva mjerna uredjaja u dalekoj zoni zračenja: (a) širokopojasni, frekventno-selektivni, programabilni analizator spektra Aaronia HF4040 (sa HyperLog 7040 antenom u opsegu 700 MHz - 4GHz) u sva tri ortogonalna smjera, u kombinaciji sa računarom uključujući softver *MCS Spectrum Analyzer* i (b) izotropski dozimetar Satimo EME Spy 140, koji je mjerio jačinu el. polja nezavisno od polarizacije emitovanih talasa, Slike 3 i 4, respektivno. Detaljnije specifikacije mjerne opreme date su u [12,14].

V. REZULTATI MJERENJA I SAR VRIJEDNOSTI

U Tabeli I prikazani su rezultati mjerenja jačine el. polja i gustine snage u opsegu frekvencija GSM900, GSM1800 i UMTS2100, u rasponu (SPAN) od 35 MHz, 75 MHz i 60 MHz, respektivno. Mjerenja su izvršena u periodu vršnog (*peak*) saobraćaja. Vremenski interval mjerenja u svakoj mjernoj tački iznosio je 6 minuta. U radu su prikazani i dijagrami promjene signala u vremenu iz UMTS opsega snimljeni pomoću analizatora spektra (AS) Spectran HF-4040 i EME SPY 140, Slike 5 i 6, respektivno. Dobijeno rezultantno el. polje pomoću AS Spectran HF-4040, Tabela I, izračunato je prema zbiru jačina komponenti vektora el. polja E na više zabilježenih frekvencija unutar posmatranog frekvencijskog opsega. Vrijednosti polja dozimetra EME SPY 140 predstavljaju usrednjene vrijednosti unutar posmatranog opsega. Podešavanja za RBW izabrana su po preporuci proizvođača, a u skladu sa izabranim opsegom i vrstom mjerenog signala. Kod svih rezultata dobijenih u programu *MCS Spectrum Analyzer*, Tabela I, u obzir je uzeta mjerna netačnost instrumenta od ±3 dB, pojačanje antene, kao i slabljenje kalibrisanog kabla. Cijelo vrijeme tokom mjerenja, u programu je bila korištena opcija „*Trace Maximum*“, jer bi bez nje procedura mjerenja bila neizvodljiva, kako zbog promjena amplitude EM talasa u vremenu, tako i zbog dugog vremena mjerenja. Ako izvor signala na određenoj frekvenciji nije bio operativan prilikom skeniranja tog opsega, onda on nije bio uključen u procjenu izlaganja. Slično tome, ako su signali bili emitovani povremeno ili sa promjenljivim snagama, samo su maksimalni nivoi snage, tokom vremena skeniranja, bili zabilježeni, pa je, za određene frekvencije signala, izlaganje moglo biti precjenjeno. Stoga, ovi rezultati predstavljaju maksimalno izlaganje, koje se povremeno dešava na mjestu mjerenja. U Tabeli I su prikazane i SAR vrijednosti izračunate primjenom izraza (6) za procjenu SAR u funkciji izmjerene jačine el. polja. Izračunate vrijednosti su date za visinu od 1,5 m od nivoa tla, prema Slici 1. Dobijeni rezultati su upoređeni sa referentnim nivoima [11, 13, 15], te pokazuju da su gustine snage EM

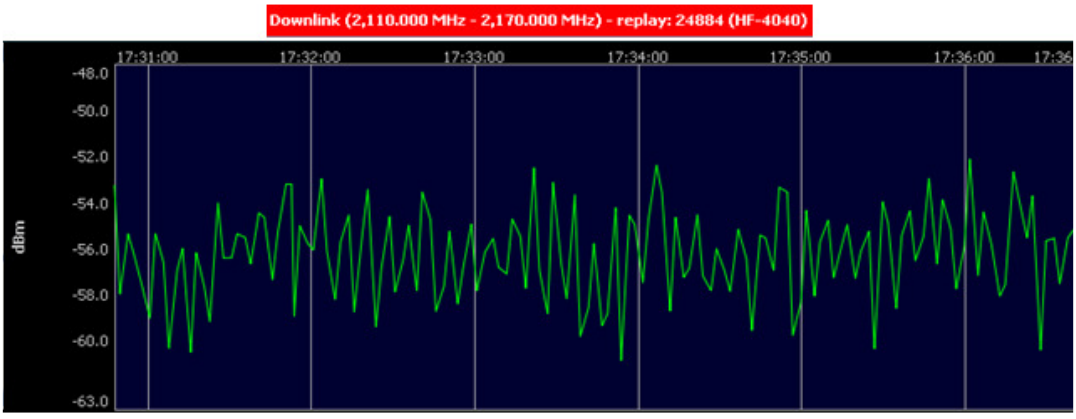
polja, u trenutku mjerenja, bile daleko ispod ovih vrijednosti, što je i u skladu sa [16]. Granične vrijednosti el. polja i gustine snage, te SAR vrijednosti, prikazane su i u Tabelama II i III, respektivno. Rezultati mjerenja pomoću dozimetra EME SPY 140 pokazuju odredjenu razliku u odnosu na vrijednosti izmjerene pomoću AS Spectran HF-4040, koja se mogu objasniti akumulacijom greške tačnosti Spectrana HF-4040 i dozimetra EME SPY 140. Drugi uzrok ovih razlika može biti vremenski aspekt. Ukupna gustina snage je veoma promjenljiva u vremenu. Takođe, prilikom mjerenja, bila je prisutna interferencija signala usljed smetnji od nekih drugih izvora (najčešće DECT telefona), što se ogleda u većem odstupanju izmjerene frekvencije u odnosu na centralnu ARFCN pojedinih kanala. Razlike su i u mogućnostima podešavanja samih parametara unutar mjernih uređaja (parametre za RBW i SPAN, širinu frekventnog opsega, kao i mjerenje jačine el. polja i gustine snage po kanalima u GSM/UMTS mreži nije moguće podešavati kod dozimetra EME SPY 140). Razlikuje se i vrijeme uzorkovanja signala tj. vrijeme prebrisavanja frekventnog opsega ova dva uređaja. Konačno, AS Spectran HF-4040, za razliku od dozimetra EME SPY 140, nema mogućnost izračunavanja ukupne gustine snage u posmatranom frekvencijskom opsegu.

TABELA I. REZULTATI MJERENJA ELEKTRIČNOG POLJA I GUSTINE SNAGE U GSM I UMTS OPSEGU

Koordinate MT		Opseg (MHz)	E_{HF4040} (V/m)	E_{EMESPY} (V/m)	Devijacija (V/m)	S_{HF4040} (μW/m^2)	S_{EMESPY} (μW/m^2)	SAR_{HF4040} (μW/kg)	SAR_{EMESPY} (μW/kg)	$SAR_{DEVIJACIJA}$ (μW/kg)
A	N 43^{0}49,377' E 018^{0}22,480'	925-960	0,291	0,605	-0,314	224,6	970,89	2,34	10,14	-7,8
		1805-1880	0,038	0,005	0,033	3,85	0,066	0,05	0,0009	0,0491
		2110-2170	0,040	0,061	-0,021	4,21	9,87	0,065	0,151	-0,086
A1	N 43^{0}49,377' E 018^{0}22,469'	925-960	0,208	0,292	-0,084	114,59	226,16	1,2	2,36	-1,16
		1805-1880	0,009	0,005	0,004	0,21	0,066	0,003	0,0009	0,0021
		2110-2170	0,083	0,084	-0,001	18,51	18,72	0,281	0,286	-0,005
A2	N 43^{0}49,381' E 018^{0}22,489'	925-960	0,104	0,142	-0,038	28,65	53,48	0,3	0,56	-0,26
		1805-1880	0,009	0,005	0,004	0,21	0,066	0,003	0,0009	0,0021
		2110-2170	0,023	0,031	-0,008	1,46	2,55	0,022	0,039	-0,017
B	N 43^{0}49,365' E 018^{0}22,489'	925-960	0,210	0,193	0,017	117,02	98,80	1,22	1,03	0,19
		1805-1880	0,022	0,005	0,017	1,32	0,066	0,018	0,0009	0,0171
		2110-2170	0,132	0,106	0,026	46,02	29,80	0,71	0,46	0,25
B1	N 43^{0}49,364' E 018^{0}22,482'	925-960	0,322	0,449	-0,127	275,3	534,75	2,87	5,58	-2,71
		1805-1880	0,010	0,005	0,005	0,26	0,066	0,004	0,0009	0,0031
		2110-2170	0,105	0,065	0,04	29,01	11,40	0,448	0,172	0,276
B2	N 43^{0}49,368' E 018^{0}22,497'	925-960	0,356	0,412	-0,056	337,69	450,23	3,51	4,7	-1,19
		1805-1880	0,02	0,005	0,015	1,05	0,066	0,015	0,0009	0,0141
		2110-2170	0,105	0,078	0,027	29,3	16,14	0,448	0,247	0,201
C	N 43^{0}49,354' E 018^{0}22,496'	925-960	0,162	0,215	-0,053	70,313	122,61	0,73	1,28	-0,55
		1805-1880	0,005	0,005	0	0,082	0,082	0,0009	0,089	-0,0881
		2110-2170	0,050	0,064	-0,014	5,8	10,86	0,102	0,167	-0,065
C1	N 43^{0}49,352' E 018^{0}22,485'	925-960	0,259	0,358	-0,099	179,045	339,96	1,86	3,55	-1,69
		1805-1880	0,010	0,005	0,005	0,264	0,066	0,004	0,0009	0,0031
		2110-2170	0,093	0,068	0,025	23,04	12,27	0,352	0,188	0,164
C2	N 43^{0}49,360' E 018^{0}22,508'	925-960	0,128	0,202	-0,074	43,58	108,23	0,45	1,13	-0,68
		1805-1880	0,013	0,005	0,008	0,42	0,066	0,006	0,0009	0,0051
		2110-2170	0,017	0,036	-0,019	0,796	3,44	0,012	0,053	-0,041

TABELA II. GRANIČNE VRIJEDNOSTI ELEKTRIČNOG POLJA I GUSTINE SNAGE PREMA ICNIRP, IEEE I CENELEC STANDARDU

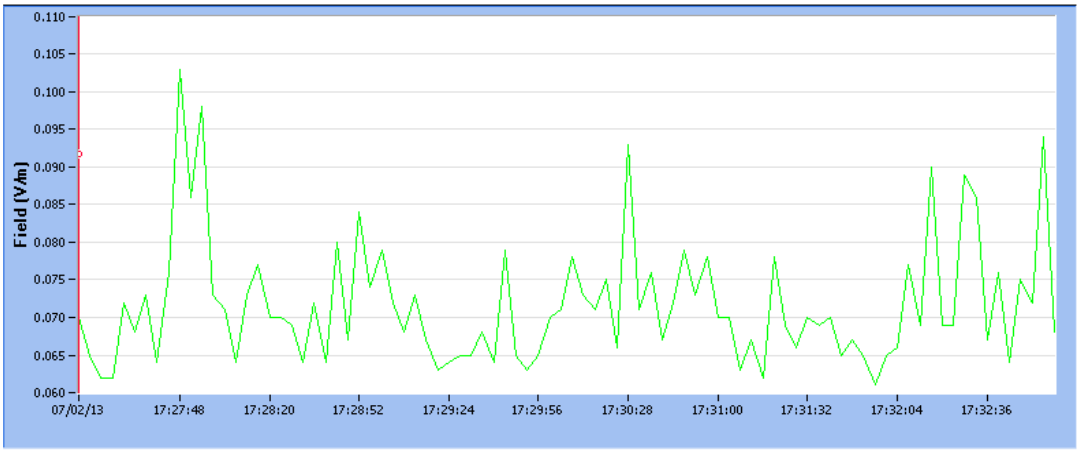
Standard	GSM 900 MHz		GSM 1800 MHz		UMTS 2100 MHz	
	E (V/m)	S (W/m^2)	E (V/m)	S (W/m^2)	E (V/m)	S (W/m^2)
ICNIRP	41	4,5	58	9	61	10
IEEE	41	4,5	58	9	61	10
CENELEC	41	4,5	58	9	61,4	10



Slika 5. Dijagram zračenja UMTS opsega u mjernoj tački A, AS HF-4040

TABELA III. VRIJEDNOSTI OSNOVNIH OGRANIČENJA SAR, ICNIRP

f	SAR cijelo tijelo (W/kg)	SAR glava i trup (W/kg)	SAR udovi (W/kg)	Gustina snage *S* (W/kg)
(9 – 100) kHz	–	–	–	–
(0,1 – 10) MHz	0,08	2	4	–
(0,01 – 10) GHz	0,08	2	4	–
(10 – 300) GHz	–	–	–	10



Slika 6. Dijagram zračenja UMTS opsega u mjernoj tački A, EME SPY140

VI. ZAKLJUČAK

BS mobilne telefonije su danas veoma interesantne sa aspekta procjene intenziteta EM zračenja. Razlog za to je što veliki procenat zračenja, kome je čovjek danas izložen, potiče upravo od njih. Ta procjena nije jednostavan zadatak, jer zavisi od relativnog položaja čovjeka i BS, karakteristika emitovanog signala, konfiguracije terena i antenskog sistema, stepena urbanizacije sredine, itd. U okviru mobilnih komunikacionih sistema, jedan od parametara koji se koristi za procjenu izloženosti ljudske populacije EM poljima je SAR. Korištenjem izraza za predikciju SAR mogu se odrediti indukovane SAR vrijednosti za osobe koje se nalaze u neposrednoj blizini BS. Analizirajući rezultate praktičnih mjerenja koja su izvršena na ograničenom broju mjesta u zoni u kojoj je uobičajno prisustvo opšte populacije, na visini od 1,5 m u odnosu na nivo tla, jačina el. polja i gustina snage koju su emitovale posmatrane antene GSM/UMTS opsega, te SAR vrijednosti u dalekoj zoni zračenja, nikada nisu prešle referentne nivoe po ICNIRP, IEEE i CENELEC standardu, kao ni prema važećem Pravilniku 37/2008 Regulatorne agencije za komunikacije BiH. Pošto su svi dijelovi ljudskog tijela izloženi EM zračenju, paznja je najvećim dijelom usmjerena na prosječnu izlozenost SAR cijelog tijela. Konačno, numerička predikcija SAR vrijednosti, takođe, zadovoljava osnovna ograničenja.

LITERATURA

[1] M. Ibrani-Pllana, L. Ahma, E. Hamiti, „Human exposure assessment in the vincinity of 900MHz GSM base station antenna“ International journal of communications, p. 57-61, Issue 2, Volume 1, 2007.

[2] Kemal Dervić: „Čovjek u bliskom radiofrekvencijskom polju”, Pljevlja 2008.

[3] Darko Šuka: „Analiza elektromagnetske emisije baznih stanica mobilne telefonije u urbanom okruženju“, Magistarski rad, Elektrotehnički fakultet Istočno Sarajevo, maj 2012.

[4] Durney CH, Massoudi H, Iskander MF.: „*Radiofrequency Radiation Dosimetry Handbook*“, 4th ed, Brooks AFB, TX: USAF School Aerosp. Med. 1986.

[5] Balzano Q, Chou C-K, Cicchetti R, Faraone, Tay R Y-S.: „An Efficient RF Exposure System with Precise Whole-Body Average SAR Determination for in vivo Animal Studies at 900MHz

[6] http://niremf.ifac.cnr.it/tissprop/htmlclie/htmlclie.htm#atsftag

[7] Nur A. Husni, Mohammad T. Islam, Mohammad R. I. Faruque and Norbahiah Misran: „Effects of electromagnetic absorbtion towards human head due to variation of its dielectric properties at 900, 1800 and 1900 MHz with different antenna substrates“, Progress In Electromagnetics Research, Vol. 138, 367{388, 2013}.

[8] Christopher M. Collinsa, Bei Yanga, Qing X. Yanga, Michael B. Smitha: „Numerical calculations of the static magnetic field in three-dimensional multi-tissue models of the human head“, Magnetic Resonance Imaging 20 (2002) 413–424.S. Miclaus, P. Bechet, Estimated and measured values of the radio-frequency radiation power density around cellular base stations, Rom.Journ.Phys, Vol.52, Nos.3-4 P.399-409. Bucharest, 2007.

[9] Pravilo 37/2008 o ograničavanju emisija EM zračenja, Regulatorna agencija za komunikacije BiH

[10] European Committee for Electrotehnical Standardization. www.cenelec.com

[11] Spectran HF4040, Spectrum analyzer instruction manuals, Aaronia AG, www.aaronia.de

[12] International Commission on Nonionizing Radiation Protection. www.icnirp.com

[13] Satimo EME Spy 140, Instruction manuals, www.satimo.com

[14] European Committee for Electrotechnical Standardization, www.cenelec.eu

[15] Pravilo 37/2008 o ograničavanju emisija EM zračenja, Regulatorna agencija za komunikacije BiH

[16] S. M. Mann, D. S. Addison, R. P. Blackwell, and M. Khalid, “Personal dosimetry of RF Radiation,” Heatlth Protection Agency (Mobile Telecommunications and Health Research Programme), Chilton, Didcot, Oxfordshire OX11 0RQ, Tech. Rep. ISBN 0 85951 568 0, 2005. [Online]. Available: http://www.mthr.org.uk

[17] C.A. Coget, and J. Fondu: “Design of an RF Dosimeter”, Katholieke Universiteit Leuven, June 2008.

Utjecaj LTE tehnologije na razine elektromagnetskih polja u mobilnim mrežama Hrvatskog Telekoma

Ružica Gajić *, Ksenija Janaček*
* Hrvatski Telekom, Roberta Frangeša Mihanovića 9, Zagreb, Hrvatska
ruzica.gajic@t.ht.hr; ksenija.janacek@t.ht.hr

U području pokretnih komunikacija u Hrvatskoj su prisutni izvori elektromagnetskog polja tehnologija GSM, UMTS, DCS i LTE. Mreža pokretnih komunikacija tehnologije LTE je u komercijalni rad u Hrvatskoj puštena u ožujku 2013. godine, od strane dvaju telekomunikacijskih operatera. S obzirom da su elektromagnetska polja emitirana od strane visokofrekvencijskih izvora u mobilnim mrežama stalni predmet interesa javnosti, u radu se razmatraju izmjerene razine elektromagnetskih polja i njihova usporedba s dozvoljenim razinama. Analizirane su izmjerene vrijednosti električnog polja u mobilnim mrežama Hrvatskog Telekoma: 882 mjerenja izvora tehnologije GSM, 987 mjerenja izvora tehnologije UMTS, 130 mjerenja izvora tehnologije DCS i 22 mjerenja izvora tehnologije LTE. Utvrđeno je da su sve izmjerene vrijednosti unutar razina propisanih Pravilnikom o zaštiti od elektromagnetskih polja, kao i istodobni utjecaj više polja koja djeluju na istom mjestu.

I. UVOD

Mreža pokretnih komunikacija tehnologije LTE (eng. Long Term Evolution) je u komercijalni rad u Hrvatskoj puštena u ožujku 2013. godine, od strane dvaju telekomunikacijskih operatera. Uz LTE, u Hrvatskoj su prisutne i ostale tehnologije: GSM (eng. Global System for Mobile Communications), UMTS (eng. Universal Mobile Telecommunications System) i DCS (eng. Digital Cellular System). Razloga za uvođenje sustava LTE je više: rast usluga koje zahtijevaju pristup širokopojasnom mobilnom pristupu Internetu, kapacitivna ograničenja sustava UMTS iz tehnologije 3G (tehnologija treće generacije, eng. 3G, 3rd generation), povećanje učinkovitosti korištenih radijskih resursa. Neminovno je da su u mobilnim mrežama tehnologije LTE postignute veće brzine prijenosa podataka te da će već sama ta činjenica omogućiti kvalitetnije korištenje usluga, kao i razvoj novih usluga.

S druge strane se postavlja pitanje razina elektromagnetskih polja emitiranih u mobilnim mrežama i njihov utjecaj na zdravlje. Javnost je vrlo senzibilizirana utjecajem elektromagnetskih polja a uočava se i manjak objektivnih činjenica, bilo o poznavanju zakonodavnog okvira, bilo o stvarnim razinama elektromagnetskih polja koje se javljaju u praksi. Još jedna tehnologija mobilnih komunikacija postavlja u prvi plan sljedeće teme: Zadovoljavaju li razine elektromagnetskih polja tehnologije LTE zakonom propisane granice? Jesu li zadovoljeni uvjeti istovremenog rada polja više frekvencija? Kolike su razine elektromagnetskih polja emitiranih od već uvedenih tehnologija GSM, DCS i UMTS?

S obzirom da su elektromagnetska polja emitirana od strane visokofrekvencijskih izvora u mobilnim mrežama stalni predmet interesa javnosti, u radu se razmatraju izmjerene razine elektromagnetskih polja i njihova usporedba s dozvoljenim razinama. Mjerenja su obavljale isključivo tvrtke koje posjeduju ovlaštenje Ministarstva zdravlja, sukladno odredbama Pravilnika o zaštiti od elektromagnetskih polja. U Republici Hrvatskoj su dozvoljene razine referentnih veličina definirane u Pravilniku o zaštiti od elektromagnetskih polja. Cilj rada je utvrditi zadovoljavaju li izmjerene referentne veličine granične razine propisane u Republici Hrvatskoj, u slučaju rada jednog izvora te u slučaju istodobnog rada više izvora elektromagnetskih polja. Analizirane su izmjerene vrijednosti električnog polja, referentne veličine u mobilnim mrežama Hrvatskog Telekoma. Razmatrane su bazne postaje tehnologija GSM, DCS, UMTS i LTE. Izmjerene vrijednosti u okolini izvora elektromagnetskih polja su uspoređene sa propisanim graničnim razinama iz Pravilnika o zaštiti od elektromagnetskog polja.

II. REGULATIVA U VEZI ZAŠTITE OD ELEKTROMAGNETSKIH POLJA U REPUBLICI HRVATSKOJ

U Republici Hrvatskoj na snazi je Pravilnik o zaštiti od elektromagnetskih polja, izdan od strane Ministarstva zdravlja [1]. Pravilnik opisuje temeljna i referentna ograničenja. Prema članku 2. stavak 3 Pravilnika o zaštiti od elektromagnetskih polja, zbog mogućih znanstveno još nepotvrđenih štetnih zdravstvenih utjecaja elektromagnetskih polja, primjenom načela predostrožnosti, granične razine referentnih razina su smanjene u odnosu na pripadna temeljna ograničenja.

Kako se iz jedne referentne veličine mogu računski izračunati sve ostale referentne veličine, u radu se analiziraju rezultati mjerenja električnog polja, koje pripada u referenta ograničenja, i to isključivo u području povećane osjetljivosti, radi interesa javnosti. Područja povećane osjetljivosti su zgrade namijenjene boravku ljudi, čestice na kojima su izgrađene zgrade stambene namjene, škole, vrtići, rodilišta, bolnice, smještajni turistički objekti, dječja igrališta te površine neizgrađenih

čestica koje su prema urbanističkom planu namijenjene navedenoj uporabi. Referentna ograničenja su značajna u praktičnoj primjeni jer se odnose na mjerljive veličine. Ostale referentne veličine su gustoća snage, jakost magnetskog polje i gustoća magnetskog toka. U uvjetima mjerenja u slobodnom prostoru, u dalekoj zoni izvora, gustoća snage iznosi:

$$S = E \cdot H \tag{1}$$

Ovdje su E i H efektivne vrijednosti električnog i magnetskog polja. U slobodnom prostoru, u dalekoj zoni izvora, iznosi električnog i magnetskog polja povezani su izrazom:

$$\eta = \frac{E}{H} \tag{2}$$

gdje je η valna impedancija slobodnog prostora i iznosi 377 Ω. Iz navedena dva izraza, gustoća snage u dalekom polju može se izraziti i kao:

$$S = \frac{E^2}{\eta} = H^2 \cdot \eta \tag{3}$$

Na temelju ove jednadžbe je vidljivo da se mjerenjem jedne referentne veličine mogu dobiti i ostale veličine pa je analiza, zbog obujma podataka, napravljena samo za električno polje.

Tablica I. podrobnije opisuje granične razine referentnih veličina u području povećane osjetljivosti.

U odnosu na preporuke Međunarodne komisije za zaštitu od neionizirajućeg zračenja (ICNIRP – International Commission on Non-Ionizing Radiation Protection) [2], na referentne veličine u navedenim frekvencijskim područjima primijenjen je faktor umanjenja koji za područje povećane osjetljivosti iznosi 2,5 – 6,25, ovisno o referentnoj veličini. Za jakost električnog polja primijenjen je faktor umanjenja od 2,5. Ovu činjenicu je potrebno naglasiti radi eventualnih usporedbi rezultata ovog rada s međunarodnim preporukama.

TABLICA I. GRANIČNE REFERENTNE VELIČINE ZA PODRUČJE POVEĆANE OSJETLJIVOSTI PREMA PRAVILNIKU O ZAŠTITI OD ELEKTROMAGNETSKIH POLJA

Referentna veličina	***Pravilnik o zaštiti od elektromagnetskih polja***
Jakost električnog polja E (V/m)	400-2000 MHz: 0,55 $f^{1/2}$
	2-300 GHz: 24,4
Jakost magnetskog polja H (A/m)	400-2000 MHz: 0,00148 $f^{1/2}$
	2-300 GHz: 0,064
Gustoća magnetskog toka B (μT)	400-2000 MHz: 0,00184 $f^{1/2}$
	2-300 GHz: 0,08
Gustoća snage (ekvivalentnog ravnog vala) Sekv (W/m2)	400-2000 MHz: f/1250
	2-300 GHz: 1,6

Pravilnik o zaštiti od elektromagnetskih polja definira i pojam značajnog izvora. Sukladno članku 2. Pravilnika o zaštiti od elektromagnetskih polja, značajan izvor elektromagnetskog polja pojedine frekvencije jest onaj stacionarni izvor čije elektromagnetsko polje u području povećane osjetljivosti, ili u području profesionalne izloženosti, doseže barem 10% iznosa granične razine zadane za tu frekvenciju. U radu se analizira broj značajnih izvora prema tehnologiji.

Pravilnik o zaštiti od elektromagnetskih polja definira i uvjete koji moraju biti zadovoljeni u slučaju istodobnog (tzv. kolociranog) djelovanja izvora elektromagnetskih polja više frekvencija. Na takvim lokacijama svaki od izvora elektromagnetskog polja mora udovoljavati graničnim razinama, a dodatno treba provjeriti i njihov zajednički učinak prema zahtjevima članka 13. Pravilnika o zaštiti od elektromagnetskih polja, prema izrazu:

$$\sum_{>1MHz}^{300GHz} \left(\frac{E_f}{E_{g,f}} \right)^2 \leq 1 \tag{4}$$

gdje je :

E_f = efektivna vrijednost jakosti električnog polja u V/m na frekvenciji f

$E_{g,f}$ = efektivna vrijednost jakosti granične razine električnog polja u V/m na frekvenciji f

Ovaj uvjet postaje posebno interesantan s razvojem raznih tehnologija u području pokretnih komunikacija, jer je na jednoj lokaciji danas moguće imati izvore elektromagnetskih polja iz područja tehnologija GSM, UMTS, DSC i LTE.

III. ANALIZA MJERNIH IZVJEŠĆA

U radu su korištena mjerna izvješća koja su izradile isključivo tvrtke ovlaštene za provođenje mjerenja od strane Ministarstva zdravlja, radi objektivnosti. Razmatrano je isključivo područje povećane osjetljivosti, zbog strožih granica i izloženosti građana. Pregled tehnologija i broja izvršenih mjerenja prema tehnologijama dan je u Tablici II.

TABLICA II. BROJ IZMJERENIH IZVORA PREMA TEHNOLOGIJAMA

Mjereni izvori po sustavima	***Broj izmjerenih izvora***
GSM 900 MHz	882
DCS 1800 MHz	130
UMTS 900 MHz	553
UMTS 2100 MHz	434
LTE 800 MHz	8
LTE 1800 MHz	14

Izvješća su obrađena na način da su pregledom izvješća za pojedinu lokaciju (baznu postaju) izdvojeni rezultati najvećih izmjerenih vrijednosti jakosti elektromagnetskog polja u pojedinim točkama u području povećane osjetljivosti. Za potrebe obrade, razmatrani su rezultati s uračunatim mjernim nesigurnostima pojedinog izvođača mjerenja na kojeg se rezultati odnose. Obrada rezultata obavlja se na način da se analizira dobivena vrijednost jakosti električnog polja za svaku pojedinačnu frekvenciju u području pojedinog sustava (GSM, DCS, UMTS, LTE) usporedbom s dozvoljenim vrijednostima za danu frekvenciju prema Pravilniku o zaštiti od elektromagnetskih polja.

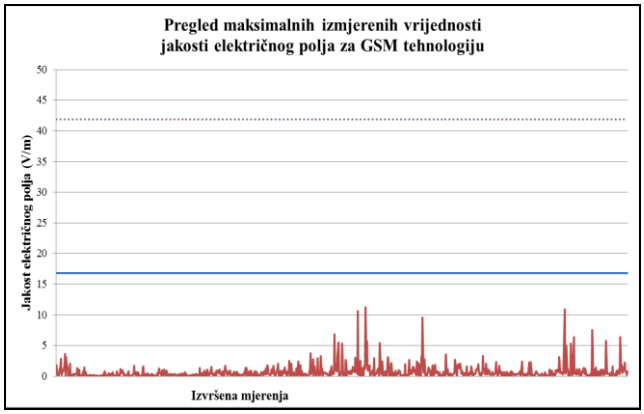
Glavni cilj rada je utvrditi udovoljavaju li postavljeni izvori elektromagnetskog polja ograničenja iz Pravilnika o zaštiti od elektromagnetskih polja za područje povećane osjetljivosti. Razmatrano je električno polje (referentna veličina), a prema Pravilniku, dozvoljena vrijednost električnog polja izražena u V/m u području povećane osjetljivosti definirana je u odnosu na frekvencije te iznosi:

- u rasponu od 400-2000 MHz: 0,55 $f^{1/2}$
- u rasponu od 2-300 GHz: 24,4

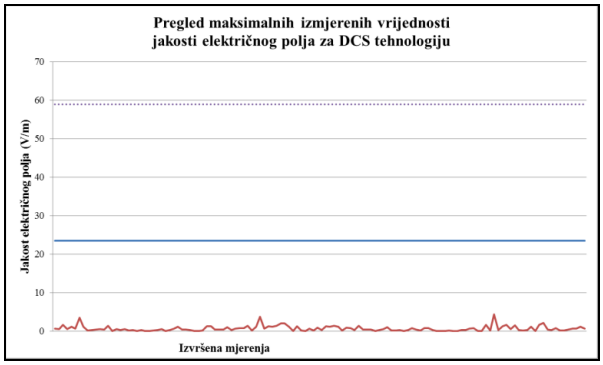
Upravo zbog frekvencijske ovisnosti, analize su provedene za svaku tehnologiju zasebno. U grafovima je napravljena analiza mjernih rezultata, koji su uspoređeni s maksimalnom dozvoljenom vrijednosti prema preporuci ICNIRP-a te maksimalnom dozvoljenom vrijednosti prema Pravilniku o zaštiti od elektromagnetskih polja.

Na slici 1. prikazani su rezultati mjerenja električnog polja tehnologije GSM i uspoređeni s dozvoljenim razinama. Isprekidana linija označava dozvoljenu razinu električnog polja prema preporuci ICNIRP-a, dok puna linija označava dozvoljenu razinu električnog polja prema Pravilniku o zaštiti od elektromagnetskih polja, za frekvencije na kojima je obavljeno mjerenje. Crvenom bojom je označena vrijednost stvarno izmjerenog električnog polja u ukupno 882 lokacije tj. mjerne točke s najvišom vrijednošću električnog polja u pojedinom mjerenju. Najviša izmjerena vrijednost električnog polja iznosi 11,253 V/m, što iznosi 67,06 % od dozvoljene vrijednosti.

Na slici 2. prikazani su rezultati mjerenja električnog polja tehnologije DCS i uspoređeni s dozvoljenim razinama. Isprekidana linija označava dozvoljenu razinu



Slika 1. Pregled maksimalnih izmjerenih vrijednosti jakosti električnog polja za tehnologiju GSM

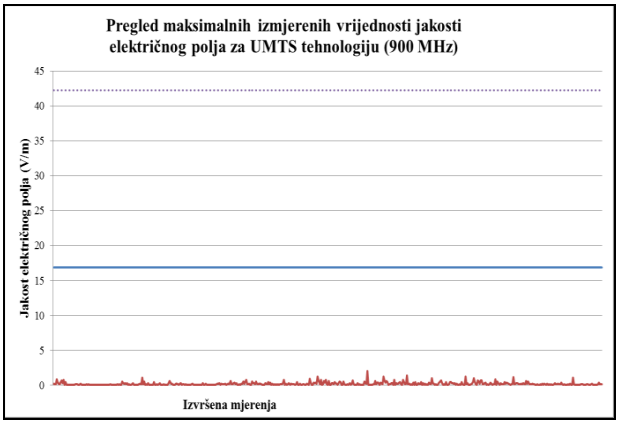


Slika 2. Pregled maksimalnih izmjerenih vrijednosti jakosti električnog polja za tehnologiju DCS

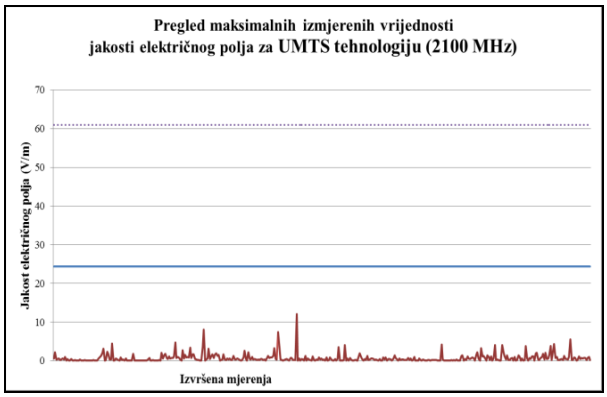
električnog polja prema preporuci ICNIRP-a, dok puna linija označava dozvoljenu razinu električnog polja prema Pravilniku o zaštiti od elektromagnetskih polja, za frekvencije na kojima je obavljeno mjerenje. Crvenom bojom je označena vrijednost stvarno izmjerenog električnog polja u ukupno 130 lokacija tj. mjernih točaka s najvišom vrijednošću električnog polja u pojedinom mjerenju. Najviša izmjerena vrijednost električnog polja iznosi 4,44 V/m, što iznosi 18,85 % od dozvoljene vrijednosti.

Na slici 3. prikazani su rezultati mjerenja električnog polja tehnologije UMTS 900 i uspoređeni s dozvoljenim razinama. Isprekidana linija označava dozvoljenu razinu električnog polja prema preporuci ICNIRP-a, dok puna linija označava dozvoljenu razinu električnog polja prema Pravilniku o zaštiti od elektromagnetskih polja, za frekvencije na kojima je obavljeno mjerenje. Crvenom bojom je označena vrijednost stvarno izmjerenog električnog polja u ukupno 553 lokacija tj. mjernih točaka s najvišom vrijednošću električnog polja u pojedinom mjerenju. Najviša izmjerena vrijednost električnog polja iznosi 2,06 V/m, što iznosi 12,19 % od dozvoljene vrijednosti.

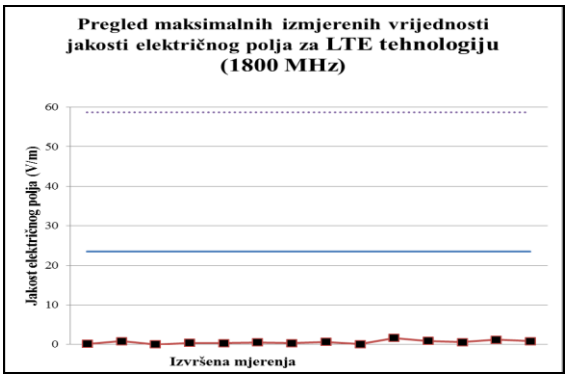
Na slici 4. prikazani su rezultati mjerenja električnog polja tehnologije UMTS 2100 i uspoređeni s dozvoljenim razinama. Isprekidana linija označava dozvoljenu razinu električnog polja prema preporuci ICNIRP-a, dok puna linija označava dozvoljenu razinu električnog polja prema



Slika 3. Pregled maksimalnih izmjerenih vrijednosti jakosti električnog polja za tehnologiju UMTS 900



Slika 4. Pregled maksimalnih izmjerenih vrijednosti jakosti električnog polja za tehnologiju UMTS 2100

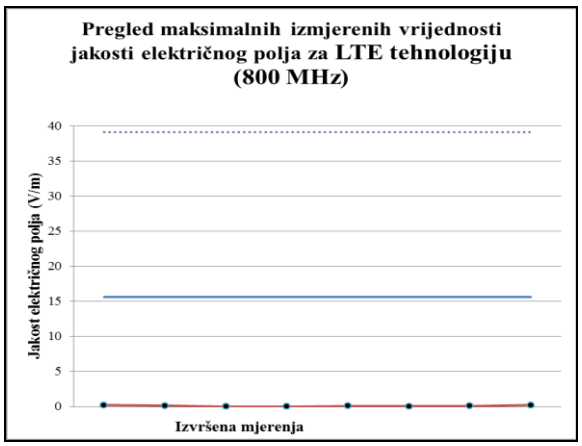


Slika 6. Pregled maksimalnih izmjerenih vrijednosti jakosti električnog za tehnologiju LTE 1800

Pravilniku o zaštiti od elektromagnetskih polja, za frekvencije na kojima je obavljeno mjerenje. Crvenom bojom je označena vrijednost stvarno izmjerenog električnog polja u ukupno 434 lokacija tj. mjernih točaka s najvišom vrijednošću električnog polja u pojedinom mjerenju. Najviša izmjerena vrijednost električnog polja iznosi 12,14 V/m, što iznosi 49,77 % od dozvoljene vrijednosti.

Na slici 5. prikazani su rezultati mjerenja električnog polja tehnologije LTE 800 i uspoređeni s dozvoljenim razinama. Isprekidana linija označava dozvoljenu razinu električnog polja prema preporuci ICNIRP-a, dok puna linija označava dozvoljenu razinu električnog polja prema Pravilniku o zaštiti od elektromagnetskih polja, za frekvencije na kojima je obavljeno mjerenje. Crvenom bojom je označena vrijednost stvarno izmjerenog električnog polja u ukupno 8 lokacija tj. mjernih točaka s najvišom vrijednošću električnog polja u pojedinom mjerenju. Najviša izmjerena vrijednost električnog polja iznosi 0,25 V/m, što iznosi 1,57 % od dozvoljene vrijednosti.

Na slici 6. prikazani su rezultati mjerenja električnog polja tehnologije LTE 1800 i uspoređeni s dozvoljenim razinama. Isprekidana linija označava dozvoljenu razinu električnog polja prema preporuci ICNIRP-a, dok puna linija označava dozvoljenu razinu električnog polja prema



Slika 5. Pregled maksimalnih izmjerenih vrijednosti jakosti električnog polja za tehnologiju LTE 800

Pravilniku o zaštiti od elektromagnetskih polja, za frekvencije na kojima je obavljeno mjerenje. Crvenom bojom je označena vrijednost stvarno izmjerenog električnog polja u ukupno 14 lokacija tj. mjernih točaka s najvišom vrijednošću električnog polja u pojedinom mjerenju. Najviša izmjerena vrijednost električnog polja iznosi 1,6 V/m, što iznosi 6,81 % od dozvoljene vrijednosti.

Analiza rezultata mjerenja pokazala je da izmjerene razine elektromagnetskog polja u svim mjernim točkama udovoljavaju graničnim razinama za područja povećane osjetljivosti prema zahtjevima Pravilnika o zaštiti od elektromagnetskih polja. Vidljivo je da se izmjerene razine kreću u velikom rasponu, što je i razumljivo s obzirom da su obavljena mjerenja sustava različitih snaga, na različitim visinama i udaljenostima. Važno je istaknuti da su u svim slučajevima izmjerene vrijednosti jakosti polja unutar dozvoljenih vrijednosti.

I prethodno izvršena analiza razina polja za 1778 izvora tehnologija GSM, UMTS i DCS te njihova usporedba s dozvoljenim razinama [4] pokazala je isti rezultat, tj. da su sve izmjerene vrijednosti, jednako kao i istodobni (kolocirani) utjecaj više polja koja djeluju na istom mjestu, unutar razina propisanih Pravilnikom o zaštiti od elektromagnetskih polja.

IV. ZNAČAJNI IZVORI

Pravilnik o zaštiti od elektromagnetskih polja definira pojam značajnog izvora. Sukladno članku 2. Pravilnika o zaštiti od elektromagnetskih polja, značajan izvor elektromagnetskog polja pojedine frekvencije jest onaj stacionarni izvor čije elektromagnetsko polje u području povećane osjetljivosti, ili u području profesionalne izloženosti, doseže barem 10% iznosa granične razine zadane za tu frekvenciju.

Na temelju analize mjerenja, utvrđen je broj značajnih izvora u mjernom uzorku (Tablica III.).

Na temelju analize je vidljivo da je postotak značajnih izvora mali, pogotovo u tehnologijama UMTS i DCS, čak i uz strogo postavljenu granicu za značajne izvore (10% od dozvoljene vrijednosti). Očekivano je najveći broj značajnih izvora u tehnologiji GSM.

TABLICA III. BROJ ZNAČAJNIH IZVORA U MJERNOM UZORKU

Mjereni izvori po sustavima	*Svi izmjereni izvori*	*Broj značajnih izvora*	*Postotak značajnih izvora*
GSM 900 MHz	882	77	8,73%
DCS 1800 MHz	130	3	2,31%
UMTS 900 MHz	553	1	0,18%
UMTS 2100 MHz	434	24	5,53%
LTE 800 MHz	8	0	0,00%
LTE 1800 MHz	14	0	0,00%

Značajnih izvora u tehnologiji LTE nema, što se može objasniti malim brojem postavljenih i izmjerenih izvora ove tehnologije.

V. KOLOCIRANI IZVORI

S obzirom da su gotovo na svim lokacijama Hrvatskog Telekoma postavljeni kolocirani izvori, obavljena je i analiza učinka zajedničkog rada sustava GSM, DCS, UMTS i LTE na lokacijama na kojima je instalirana tehnologija LTE. Svi rezultati pokazuju da je uvjet Pravilnika o zaštiti od elektromagnetskih polja vezan za zajednički učinak svih kombinacija sustava Hrvatskog Telekoma ispunjen na svim izmjerenim lokacijama.

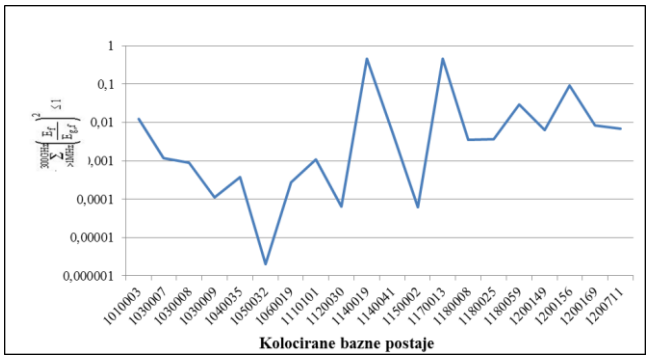
Na slici 7. dan je primjer izračuna za nekoliko kolociranih baznih postaja sustava GSM, DCS, UMTS i LTE. Ordinata pokazuje vrijednost izračuna uvjeta Pravilnika o zaštiti od elektromagnetskog polja prema (4), a na apscisi su dane identifikacijske oznake kolociranih baznih postaja.

Ukoliko bi za neki izvor izmjerena vrijednost električnog polja bila veća od granične vrijednosti dozvoljene prema Pravilniku o zaštiti od elektromagnetskih polja (što ovdje nije slučaj), brojnik u (4) bio bi veći od nazivnika, te bi njihov kvadrat a time i ukupni zbroj svih članova niza bio veći od 1, čime uvjet iz Pravilnika ne bi bio ispunjen. Nadalje, budući da je za sve tehnologije većina izmjerenih vrijednosti unutar 10% granične vrijednosti dozvoljene Pravilnikom (svega je 5,2% značajnih izvora u ukupnom mjernom uzorku za sve sustave), tako su i kvadrati odnosa izmjerene i granične dozvoljene vrijednosti brojke još manjih vrijednosti (tj. bliže nuli). Iz tog su razloga vrijednosti na slici 7. prikazane u logaritamskom mjerilu.

Vidljivo je da su izračunate vrijednosti manje od vrijednosti dozvoljene prema zahtjevima Pravilnika o zaštiti od elektromagnetskih polja. Kolociranost s ostalim tehnologijama koje se nalaze u istom frekvencijskom spektru nije promatrana.

VI. ZAKLJUČAK

U području pokretnih komunikacija u Hrvatskoj su prisutni izvori elektromagnetskog polja tehnologija GSM,



Slika 7. Grafički prikaz rezultata kolociranih sustava GSM, DCS, UMTS i LTE

UMTS, DCS i LTE. Pravilnik o zaštiti od elektromagnetskih polja, koji se primjenjuje u Republici Hrvatskoj, postavlja granice referentnih veličina za frekvencijska područja u kojima emitiraju izvori elektromagnetskih polja u mobilnim komunikacijama (400-2000 MHz, 2-300 GHz). S obzirom na vrlo stroge granice, koje su određene visokim faktorom umanjenja u odnosu na preporuku Međunarodne komisije za zaštitu od neionizirajućeg zračenja, u radu se analiziraju rezultati mjerenja referentnih veličina u okolini izvora elektromagnetskog polja u mobilnim mrežama Hrvatskog Telekoma za tehnologije GSM, DCS, UMTS i LTE. Posebno le razmatran istodobni rad svih izvora elektromagnetskih polja na pojedinim mjestima (istodobni rad izvora GSM, UMTS, DCS i LTE tehnologija).

Nakon analize 882 izvora sustava GSM, 987 izvora sustava UMTS (od toga 434 za UMTS 2100 MHz i 553 za UMTS 900 MHz), 130 izvora sustava DCS i 22 izvora sustava LTE (od toga 8 za LTE 800 MHz i 14 za LTE 1800 MHz), utvrđeno je da su svi izmjereni rezultati unutar granica propisanih Pravilnikom o zaštiti od elektromagnetskih polja.

Na lokacijama na kojima dolazi do istovremenog rada više izvora elektromagnetskih polja različitih frekvencija izvršena je analiza zadovoljenja uvjeta iz Pravilnika o zaštiti od elektromagnetskih polja (njihov zajednički učinak). Promatrano je 20 lokacija na kojima istodobno djeluju izvori tehnologija GSM, DCS, UMTS i LTE. Analizom je utvrđeno da sve lokacije udovoljavaju zahtjevima članka 13. Pravilnika o zaštiti od elektromagnetskih polja koji se odnosi na zajednički učinak više izvora koji istodobno rade.

LITERATURA

[1] Pravilnik o zaštiti od elektromagnetskih polja, NN 98/11, 2011.

[2] Guidelines for Limiting Exposure to Time-Varying Electric, Magnetic, and Electromagnetic Fields (up to 300 GHz). Health Physics 74 (4): 494-522; 1998.

[3] S. Gros i R. Gajić, "Analiza rezultata mjerenja elektromagnetskih polja u mobilnim mrežama Hrvatskog Telekoma", MIPRO, Opatija, svibanj 2013.

NETWORKS PERFORMANCE

Incremental-Relaying Cooperative-Networks using Dual Transmit Diversity and Decode and Forward Relaying Scheme with Best Relay Selection.

Ali M Hayajneh, Mahmoud Khodeir, Mamoun F. Al-Mistarihi
amhayajneh11@eng.just.edu.jo, makhodeir@just.edu.jo, mistarihi@just.edu.jo

***Abstract*— A new incremental-relaying cooperative-diversity technique with a simple dual transmit diversity, decode-and-forward and best relay selection scheme is presented. Closed-form expressions for the the bit error rate, the signal to noise ratio outage probability, and average achievable rate are derived using binary phase shift keying modulation scheme over independent non-identical Rayleigh fading channels. The results show that the simple space time block coding scheme with the incremental-relaying cooperative diversity and best relay selection outperforms the regular incremental fixed cooperative diversity and saves resources.**

***Index Terms*—Rayleigh fading, adaptive relaying, incremental-relaying, maximal ratio combining, error performance, outage probability, space time block coding.**

I. INTRODUCTION

In the last few years the incremental-relaying cooperative-diversity networks has become one promising technique which is proposed widely in wireless communication systems. The motivation in the technique is the better utilization of the resources and the great saving in power consumption for relaying. This can be achieved by allowing the relay to operate in very specific cases (i.e., when the received *signal to noise ratio* (SNR) at the destination is below a pre-defined threshold value). Here, the relay should send another copy of the transmitted signal (i.e., already saved at the relay) to the receiver. The authors in [1] and [2] discussed the performance of the regular cooperative-diversity networks, where they derived a closed-form expressions for both the average *bit error rate* (BER) and SNR *outage probability* (P_{out}). Next, the authors in [3] derived a closed-form expressions for the same metrics for incremental-relaying cooperative-networks based on the results of [1] and [2]. The results were shown assuming a *single input single output* (SISO) transmission model is used and exploiting both *amplify-and-forward* (AF) and *decode-and-forward* (DF) relaying schemes. In particular, the results showed that the regular incremental cooperative-diversity achieved superior channel utilization as compared with the regular cooperative-diversity. For more generalization on the incremental relaying technique, the authors in [4] derived a closed-form expressions for both the BER and SNR outage probability assuming a simple transmit diversity based on the simple *space time block coding* (STBC) for Alamouti and DF relaying scheme [5].

Finally, the motivation of this study is to capture the effect of using the adaptive best relay selection with the incremental cooperative-diversity utilizing a simple dual transmit diversity technique. This will add more diversity orders to the achieved virtual diversity gain in [3]. Thus, the transmitter in the proposed model is equipped with two antennas with a minimum required separation to achieve independent fading channels and transmitting due to the simple transmit diversity scheme [5], [6]. Fig. 1 below shows the configuration of the antennas and the geometry of the network, where the *relay* (R) is between the *source* (S) and the *destination* (D) assuming asymmetric network geometry (i.e., variable relay location). Moreover, the DF relaying scheme (i.e., regenerative relays) is used, where the transmitted signals from the source are fully decoded then retransmitted to the destination when the received SNR is below the pre-defined threshold as a response to a request from the destination. Here, using the simple transmit diversity does not require a perfect *channel state information at transmitter* (CSIT).

Overall, the rest of this paper is organized as follows. In Section II, a closed-form expression for the average BER of the incremental cooperative-diversity is derived, where a BPSK modulation scheme is assumed over independent non-identical Rayleigh fading channels with a dual transmit diversity. Then a closed-form expression for the SNR outage probability analysis is presented in Section III. Next, Section IV shows the average achievable rate performance. Furthermore, Section V analytical and numerical results analysis. Finally, Section VI presents the conclusion and future work.

II. SYSTEM MODEL

A new scheme of relaying in incremental-relaying is assumed, where the network has multiple relays to assist in the relaying phase by selecting the relay that has the best performance as shown in Figure 1. Here, we have *L* relays cooperating in the network and all the relays are assumed to operate using DF scheme (i.e., re-generative relaying). By utilizing the dual transmit diversity, then every relay must has the ability to decouple the Alamouti received signals, decode

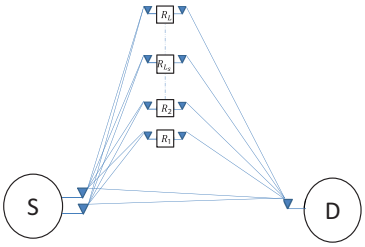


Fig. 1. Illustration of the incremental-relaying cooperative-diversity networks using dual transmit diversity and best relay selection.

and re-encode these signals, and then re-transmit them to the destination. Moreover, the destination will request a signal retransmission from the best relay (i.e., selected among the L relays) if the received SNR coming from the direct path $S \rightarrow D$ is below the pre-defined SNR threshold. Due to this request, the selected relay will retransmit the encoded signals to the destination. This will add virtual diversity gain to the gain achieved in the aforementioned relaying schemes.

In general, the received signal at any of the cooperating relays, which is originally coming from the source can be expressed as follows:

$$y_{S,R,i} = h_{S,R,i}\sqrt{E_s/2}\,x(t) + n_i \quad , i = 1, 2, 3, ..., L \quad (1)$$

where L is the number of the cooperating relays, $y_{S,R,i}$ is the received signal at the relay from the source, $h_{S,R,i}$ is the source relay channel coefficient, $x(t)$ is the transmitted symbol from the source with unity energy, and n_i is the AWGN terms. Then the received signal coming from any relay at the destination can be written as follows:

$$y_{R,D,i} = h_{R,D,i}\sqrt{E_s}x_r(t) + n_i \quad , i = 1, 2, 3, ..., L \quad (2)$$

where $y_{R,D,i}$ is the received signal at the destination coming from the *i-th* relay, $h_{R,D,i}$ is the relay destination channel coefficient, $x_r(t)$ is the retransmitted signal from the relay.

Actually, the selection of the best relay is a routing protocol problem. In particular, in order to get the total instantaneous received SNR from the best relay, one can get the SNR coming from the best relay as follows:

$$\gamma^*_{R,D} = \underset{i=1,2,3,...,L}{max}\{\gamma_{R,D,i}\} \quad (3)$$

where $\gamma^*_{R,D}$ denote the received SNR coming from the best relay at the destination. Finally, all the the channel coefficients are assumed to have non-identical Rayleigh fading channels.

Due to best relays selection, the knowledge of ordered statistics of the received signals is a must. In particular, ordered statistics means that the the routing protocol has the capability to order the received signals realizations due to the backward channels due to their SNR values. Here, ordered statistics deal with the PDFs of the received signals. The authors in [7],[8], and [9] developed unified approaches to determine the PDF of the ordered statistics of independent and identically distributed random variables with a common PDF and CDF. Here the protocol that is utilized ranks the realizations beginning from the one with the largest received SNR to the one with the smallest SNR (i.e., $\gamma_{1:L} \geq \gamma_{2:L} \geq ... \geq \gamma_{L-1:L} \geq \gamma_{L:L}$). Furthermore, the relay with the best SNR value is denoted by $\gamma_{1:L}$ and for simplicity assume $\gamma^*_{R,D} = \gamma_{1:L} = \gamma^*$. Therefore, to write the PDF of $\gamma^*_{R,D}$ assuming that all the received SNR PDFs from the relays have a common mean, one can write,$f_{\gamma^*_{R,D}}(\gamma) = 2Lf_{\gamma_{R,D}}(\gamma)[1-F_{\gamma_{R,D}}(\gamma)][2F_{\gamma_{R,D}}(\gamma) - (F_{\gamma_{R,D}}(\gamma))^2]^{L-1}$ For Rayleigh fading channels, then the PDF of $\gamma_{R,D,i}$ follows the exponential distribution as follows:

$$f_{\gamma_{R,D,i}}(\gamma) = \frac{1}{\overline{\gamma}_{R,D,i}} e^{-\frac{\gamma}{\overline{\gamma}_{R,D,i}}} \quad (4)$$

Next, with the common mean assumption:

$$f_{\gamma_{R,D}}(\gamma) = \frac{1}{\overline{\gamma}_{R,D}} e^{-\frac{\gamma}{\overline{\gamma}_{R,D}}} \quad (5)$$

Now, the the received signal from the best relay has the following PDF [7]:

$$f_{\gamma^*_{R,D}}(\gamma) = \sum_{i=1}^{L}(-1)^{i-1}\binom{L}{i}\frac{2i}{\overline{\gamma}_{R,D}} e^{-i\frac{2\gamma}{\overline{\gamma}_{R,D}}} \quad (6)$$

Moreover, in order to capture the effect of the path loss on the performance of the proposed model, one can define the statistical average operators as in [10] such that $\mathbf{E}\left(h^2_{S,D}\right) = 1, \mathbf{E}\left(h^2_{S,R,i}\right) = \left(\frac{d_{S,D}}{d_{S,R}}\right)^{\alpha}, \mathbf{E}\left(h^2_{R,D,i}\right) = \left(\frac{d_{S,D}}{d_{R,D}}\right)^{\alpha}$, where $\mathbf{E}(\cdot)$ is the statistical average operator, α is the path loss, and $d_{k,m}$ is the distance between terminals k and m

III. Error Performance Analysis

In order to analyze the error performance of the proposed system, one needs to determine the error probability of the combined signal at the receiver using the following [3]:

$$\begin{aligned} P(e) &= P_r(\gamma_{S,D,\Sigma} \leq \gamma_o) \times P_{div}(e) \\ &+ (1 - P_r(\gamma_{S,D,\Sigma} \leq \gamma_o)) \times P_{direct}(e) \end{aligned} \quad (7)$$

However, the total SNR due to the received signal at the receivers is the sum of the SNR from the signal coming from the direct path, and the SNR due to the signal received from the best relay. Thus, the total SNR for the total received signal can be written as follows:

$$\gamma_{tot} = \gamma^*_{R,D} + \gamma_{S,D,\Sigma} = Z \quad (8)$$

Moreover, the PDF of the total received signal can be written as the convolution of the two random variables $f_{\gamma^*_{R,D}}(\gamma)$ and $f_{\gamma_{S,D,\Sigma}}(\gamma)$ which can be written as follows [11]:

$$\begin{aligned} f_{\gamma_{tot}}(z) = \frac{1}{\overline{\gamma}^2_{S,D}} \int_0^{\infty} \sum_{i=1}^{L}(-1)^{i-1}\binom{L}{i}\lambda_i \\ \times e^{-\lambda_i\gamma}(z-\gamma)e^{-\frac{z-\gamma}{\overline{\gamma}_{S,D}}}d\gamma \end{aligned} \quad (9)$$

where $\lambda_i = \frac{2i}{\overline{\gamma}_{R,D}}$. Then, after some manipulations, one can write the above equation as in Eqn. (10), which can be further simplified to Eqn. (11). Finally, to find the average BER assuming a BPSK, one must find ,$P_{div}(e)$, which is the error due to the combined signal at the destination as follows:

$$P_{div}(e) = P^*_{S,R,\Sigma}(e) \times P_x(e) + (1 - P^*_{S,R,\Sigma}(e)) \times P_{tot}(e) \quad (12)$$

$$f_{\gamma_{tot}}(z)=\begin{cases}\dfrac{1}{\overline{\gamma}^2_{S,D}}\displaystyle\int_0^z\sum_{i=1}^{L}(-1)^{i-1}\binom{L}{i}\lambda_i e^{-\lambda_i\gamma}(z-\gamma)e^{-\frac{z-\gamma}{\overline{\gamma}_{S,D}}}d\gamma & \text{if } z\le\gamma_o.\\ \dfrac{1}{\overline{\gamma}^2_{S,D}}\displaystyle\int_{z-\gamma_o}^z\sum_{i=1}^{L}(-1)^{i-1}\binom{L}{i}\lambda_i e^{-\lambda_i\gamma}(z-\gamma)e^{-\frac{z-\gamma}{\overline{\gamma}_{S,D}}}d\gamma & \text{if } z>\gamma_o.\end{cases} \tag{10}$$

$$f_{\gamma_{tot}}(z)=\begin{cases}\dfrac{1}{\overline{\gamma}^2_{S,D}}\displaystyle\sum_{i=1}^{L}\frac{(-1)^{i-1}\binom{L}{i}\lambda_i}{\left(\lambda_i\overline{\gamma}_{S,D}-1\right)^2}\\ \times\left[\overline{\gamma}^2_{S,D}e^{-\lambda_i z}+\overline{\gamma}^2_{S,D}\lambda_i z e^{-\frac{z}{\overline{\gamma}_{S,D}}}-\overline{\gamma}^2_{S,D}e^{-\frac{z}{\overline{\gamma}_{S,D}}}-\overline{\gamma}_{S,D}z e^{-\frac{z}{\overline{\gamma}_{S,D}}}\right] & \text{if } z\le\gamma_o.\\ \dfrac{1}{\overline{\gamma}^2_{S,D}}\displaystyle\sum_{i=1}^{L}\frac{(-1)^{i-1}\binom{L}{i}\lambda_i}{\left(\lambda_i\overline{\gamma}_{S,D}-1\right)^2}\times\left[\overline{\gamma}^2_{S,D}e^{\lambda_i\gamma_o}\lambda_i\gamma_o e^{-\frac{\gamma_o}{\overline{\gamma}_{S,D}}}e^{-\lambda_i z}\right.\\ \left.+\overline{\gamma}^2_{S,D}e^{-\lambda_i z}-\overline{\gamma}^2_{S,D}e^{\lambda_i\gamma_o}e^{-\frac{\gamma_o}{\overline{\gamma}_{S,D}}}e^{-\lambda_i z}-\overline{\gamma}_{S,D}e^{\lambda_i\gamma_o}\gamma_o e^{-\frac{\gamma_o}{\overline{\gamma}_{S,D}}}e^{-\lambda_i z}\right] & \text{if } z>\gamma_o.\end{cases} \tag{11}$$

where $P_{tot}(e)$ is the BER in the combined signal (i.e., the direct path signal and the best relay signal), and $P^*_{S,R,\Sigma}(e)$ is the probability of a desision error at the best relay. Thus, to complete the study of the best relays selection with the OSTBC scheme, one can write $P_{tot}(e)$ as follows:

$$P_{tot}\left(e\right)=\frac{1}{2}\int_0^\infty f_{\gamma_{tot}}\left(z/\gamma_{S,D,\Sigma}\le\gamma_o\right)erfc(\sqrt{z})dz \tag{13}$$

Next, the conditional PDF of γ_{tot} has the following distribution:

$$f_{\gamma_{tot}}\left(z/\gamma_{S,D,\Sigma}\le\gamma_o\right)=\frac{f_{\gamma_{tot}}\left(z\right)}{1-\left(\frac{\gamma_o}{\overline{\gamma}_{S,D}}+1\right)exp\left(-\frac{\gamma_o}{\overline{\gamma}_{S,D}}\right)} \tag{14}$$

Thus, by substituting Eqn. (11) in the above equation one can find a closed-form expression for $f_{\gamma_{tot}}\left(z/\gamma_{S,D,\Sigma}\le\gamma_o\right)$ Here, after solving Eqn. (13) and after some manipulations one can write $P_{tot}(e)$ as follows:

$$\begin{aligned}P_{tot}(e)=&\frac{1}{2\overline{\gamma}^2_{S,D}\left(1-\left(\frac{\gamma_o}{\overline{\gamma}_{S,D}}+1\right)exp\left(-\frac{\gamma_o}{\overline{\gamma}_{S,D}}\right)\right)}\\ &\times\sum_{i=1}^{L}\frac{(-1)^{i-1}\binom{L}{i}\lambda_i}{\left(\lambda_i\overline{\gamma}_{S,D}-1\right)^2}\left[\Theta(0,\gamma_o\overline{\gamma}^2_{S,D},\lambda_i)\right.\\ &+\Phi(0,\gamma_o,\overline{\gamma}^2_{S,D}\lambda_i,\frac{1}{\overline{\gamma}_{S,D}})-\Theta(0,\gamma_o,\overline{\gamma}^2_{S,D},\frac{z}{\overline{\gamma}_{S,D}})\\ &\left.-\Phi(0,\gamma_o,\overline{\gamma}_{S,D},\frac{z}{\overline{\gamma}_{S,D}})+\Theta(\gamma_o,\infty,\eta,\lambda_i)\right]\end{aligned} \tag{15}$$

where $\eta=\overline{\gamma}^2_{S,D}+\overline{\gamma}^2_{S,D}e^{\lambda_i\gamma_o}\lambda_i\gamma_o e^{-\frac{\gamma_o}{\overline{\gamma}_{S,D}}}-\overline{\gamma}^2_{S,D}e^{\lambda_i\gamma_o}e^{-\frac{\gamma_o}{\overline{\gamma}_{S,D}}}-\overline{\gamma}_{S,D}e^{\lambda_i\gamma_o}\gamma_o e^{-\frac{\gamma_o}{\overline{\gamma}_{S,D}}}$, and $\Phi(A,B,\omega,\lambda)$ and $\Theta(A,B,\omega,\lambda)$ are auxiliary functions that are defined in the appendix (A).

Now, assuming the dual transmit diversity with BPSK modulation scheme is used under Rayleigh fading channels, the expression for the probability of decision error at the best relay in decoding the Alamouti signals,$P^*_{S,R,\Sigma}(e)$, can be written as follows:

$$\begin{aligned}P^*_{S,R,\Sigma}(e)&=\int_0^\infty f_{S,R,\Sigma}(\gamma)\times P\left(e/\gamma\right)d\gamma\\ &=\frac{\left[1-\left(1+\frac{3}{2\overline{\gamma}_{\mathrm{S,R}}}\right)^{-3/2}\sqrt{1+\frac{1}{\overline{\gamma}_{\mathrm{S,R}}}}\right]}{2}\end{aligned} \tag{16}$$

Finally, by substituting the above equation and Eqn. (13) in Eqn. (12) and substituting the result in Eqn. (7) one can find a closed-form expression for the total average BER for the incremental relaying with best relay selection and simple STBC scheme.

IV. Outage Probability

In this section, the performance of the SNR outage probability is analyzed as extension to the analysis of error probabilities. Here, the total SNR outage probability for the total received signals from the best relay assuming the DF scheme is employed. For the whole bunch of the relays, one can write the total probability of outage as follows [3]:

$$\begin{aligned}P_{out}&=P^*_r\left(\gamma_{S,R,\Sigma}\le\gamma_o\right)P_r\left(\gamma_{S,D,\Sigma}\le\gamma_o\right)\\ &+\left(1-P^*_r\left(\gamma_{S,R,\Sigma}\le\gamma_o\right)\right)P_r\left(\gamma_{tot}\le\gamma_o\right)\end{aligned} \tag{17}$$

where $P^*_r\left(\gamma_{S,R,\Sigma}\le\gamma_o\right)$ is the outage probability at the best relay. $P_r\left(\gamma_{S,D,\Sigma}\le\gamma_o\right)$ is the outage probability if the direct path is used only, and $P_r\left(\gamma_{tot}\le\gamma_o\right)$ is the outage probability when the direct path signal and the relayed signal are combined at the destination. In particular, Eqn. (17) is valid for evaluating the outage probability for the regular incremental relaying with the OSTBC scheme derived in [4].

Furthermore, by using the previous definitions, to evaluate Eqn. (17), one can write $P_r^*\left(\gamma_{S,R,\Sigma} \leq \gamma_o\right)$ as follows:

$$\begin{aligned} P_r^*\left(\gamma_{S,R,\Sigma} \leq \gamma_o\right) &= \int_0^{\gamma_o} \frac{1}{\overline{\gamma}_{S,R}^2} \gamma e^{-\frac{\gamma}{\overline{\gamma}_{S,R}}} d\gamma \\ &= 1 - \left(1 + \frac{\gamma_o}{\overline{\gamma}_{S,R}}\right) e^{-\frac{\gamma_o}{\overline{\gamma}_{S,R}}} \end{aligned} \quad (18)$$

As a result, the SNR outage probability at the combined signal at the destination can be written as in Eqn. (19). Finally, by substituting Eqns. (19) and (18) in Eqn. (17) one can find a closed form expression for the SNR outage probability for the incremental relaying with the best relay selection and simple STBC scheme.

V. Average Achievable Rate

Now, the average achievable rate in the Shannon sense is analyzed in order to evaluate the system performance under which errors are recoverable. In particular, to find the total average achievable rate for the proposed best relay selection scheme, one can write the following:

$$\begin{aligned} C_{avg} &= P_r\left(\gamma_{S,D,\Sigma} \leq \gamma_o\right) \times C_{tot} \\ &+ \left(1 - P_r(\gamma_{S,D,\Sigma} \leq \gamma_o)\right) \times C_{direct} \end{aligned} \quad (20)$$

where C_{tot} is the average achievable rate for the combined signal from the best selected relay and the direct path transmission and C_{direct} is the average achievable rate achieved by the direct path only (i.e., the relay has no role in transmission). In particular, to evaluate the above equation, one have to write C_{direct} as follows:

$$\begin{aligned} C_{direct} &= BW \int_0^{\infty} log_2(1+\gamma) \\ &\times f_{\gamma_{S,D,\Sigma}}(\gamma/\gamma_{S,D,\Sigma} > \gamma_o\ ) d\gamma \end{aligned} \quad (21)$$

Next, after some manipulations on can write C_{direct} as in Eqn. (22), where $E_1(\bullet)$ is the exponential integral. Now, to complete Eqn. (20), one can write the average achievable rate for the combined signals, C_{tot}, as follows:

$$C_{tot} = \frac{BW}{2} \int_0^{\infty} log_2(1+z) \times f_{\gamma_{tot}}(z|\gamma_{\mathrm{S,D},\Sigma} \leq \gamma_o\ ) dz \quad (23)$$

Now, after some manipulation for the above integral one can write C_{tot} as follows: weher $\Phi_c(A,B,\omega,\lambda)$ and $\Theta_c(A,B,\omega,\lambda)$ are auxiliary functions that are defined in the appendix (B). Finally, by substituting Eqn. (24) and (22) in Eqn. (20) one can find the closed-form expression for the total average achievable rate for the incremental best relay selection when the simple OSTBC scheme is used.

VI. Numerical Results

In this section the numerical results for the incremental best relay selection are presented. In particular, average BER, SNR outage probability, and the average achievable rate versus the transmitted signal (E_s/N_o dB) are calculated. Furthermore, in order to capture the effect of the number of relays and the geometry of the network in the sense of the distance from S and R, many figures are shown to view the relationships between the performance metrics and different arguments of the whole system.

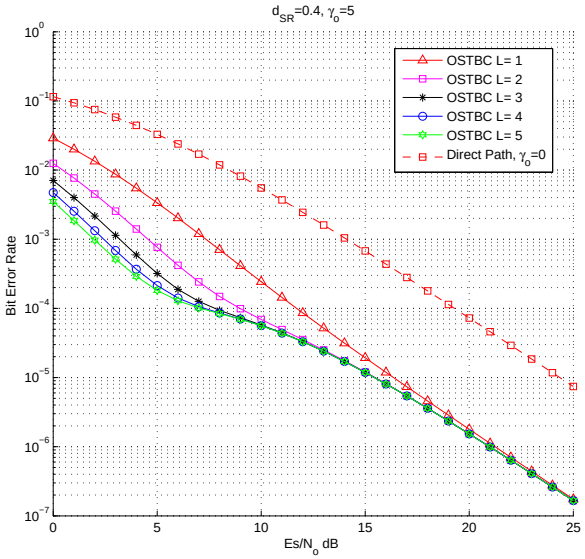


Fig. 2. Average BER versus the SNR of the transmitted signal ($E_s/N_o\,dB$) for the best relay selection with $\gamma_o = 5$.

Now, Figure 2 shows the BER versus the transmitted signal (E_s/N_o dB) for the proposed model and for multiple number of relays. Here, assuming that the relative distance between the source and the relay is denoted by $d = d_{S,R} = 0.4$ which is the optimal relative distance to achieve the best BER performance and the SNR threshold is denoted by γ_o. Moreover, the path loss exponent, α, is assumed to be 3. Obviously, the incremental best selection relaying with dual transmit diversity outperforms both of the fixed incremental relaying scheme as well as the one-hop transmission (i.e., $\gamma_o \to 0$). For example, for $d_{SR} = 0.4$, $\gamma_o = 8$ dB and L=2 relays, the incremental best relay selection with a simple OSTBC requires a 4 dB less SNR than that for the fixed incremental relaying (i.e., L=1) scheme to achieve the same value of the $BER = 10^{-6}$. However, the incremental best relay selection outperforms the fixed relaying schemes for low values of (E_s/N_o) and this depends on the value of γ_o and the proposed model will require more assistant from the relays. In fact, involving more relays in the relaying phase will add more diversity order and the BER will be enhanced to some extent. Hence, selection amongst more than 3 relays will not add a perceptible enhancement (i.e., this resembles the conventional selection mechanisms in the one-hop receive diversity). Furthermore, Figure 3 shows the P_{out} versus the SNR for the proposed model. The results show that the incremental best relay selection outperforms the fixed incremental relaying and achieves higher virtual diversity array gain and transmit diversity gain, as well as giving better SNR outage probability performance. In particular, for $d_{SR} = .4$, $\gamma_o = 5$ dB and $L = 2$ relays the incremental best relay selection with a simple OSTBC requires 6 dB less SNR than that for the fixed incremental relaying scheme to achieve the same outage probability (i.e., $P_{out} = 10^{-6}$). Moreover,

$$
\begin{aligned}
P_r\left(\gamma_{tot} \le \gamma_o\right) &= \int_0^{\gamma_o} f_{\gamma_{tot}}(\gamma) d\gamma \\
&= \frac{1}{\overline{\gamma}_{S,D}^2} \sum_{i=1}^{L} \frac{(-1)^{i-1}\binom{L}{i}\lambda_i}{\left(\lambda_i \overline{\gamma}_{S,D} - 1\right)^2} \\
&\quad \times \Bigg[\left(\overline{\gamma}_{S,D}^2 \gamma_o - \overline{\gamma}_{S,D}^3 \lambda_i \gamma_o - \overline{\gamma}_{S,D}^4 \lambda_i + 2\,\overline{\gamma}_{S,D}^3\right) e^{-\frac{\gamma_o}{\overline{\gamma}_{S,D}}} \\
&\qquad - \frac{\overline{\gamma}_{S,D}^2}{\lambda_i e^{\lambda_i \gamma_o}} + \frac{\overline{\gamma}_{S,D}^2}{\lambda_i} + \overline{\gamma}_{S,D}^4 \lambda_i - 2\,\overline{\gamma}_{S,D}^3 \Bigg]
\end{aligned} \tag{19}
$$

$$
\begin{aligned}
C_{direct} = \frac{BW}{ln(2)\left(1 + \frac{\gamma_o}{\overline{\gamma}_{S,D}}\right) exp\left(-\frac{\gamma_o}{\overline{\gamma}_{S,D}}\right)} \Bigg[& \ln\left(\gamma_o + 1\right) e^{\frac{-\gamma_o}{\overline{\gamma}_{S,D}}} + \frac{\gamma_o}{\overline{\gamma}_{S,D}} \ln\left(\gamma_o + 1\right) e^{\frac{-\gamma_o}{\overline{\gamma}_{S,D}}} \\
& + e^{\frac{1}{\overline{\gamma}_{S,D}}} E_1\left(\frac{\gamma_o}{\overline{\gamma}_{S,D}} + {\overline{\gamma}_{S,D}}^{-1}\right) + e^{\frac{-\gamma_o}{\overline{\gamma}_{S,D}}} - \frac{e^{{\overline{\gamma}_{S,D}}^{-1}}}{\overline{\gamma}_{S,D}} E_1\left(\frac{\gamma_o}{\overline{\gamma}_{S,D}} + {\overline{\gamma}_{S,D}}^{-1}\right) \Bigg]
\end{aligned} \tag{22}
$$

$$
\begin{aligned}
C_{tot} = & \frac{BW}{2\overline{\gamma}_{S,D}^2\left(1 - \left(\frac{\gamma_o}{\overline{\gamma}_{S,D}} + 1\right) exp\left(-\frac{\gamma_o}{\overline{\gamma}_{S,D}}\right)\right)} \\
& \times \sum_{i=1}^{L} \frac{(-1)^{i-1}\binom{L}{i}\lambda_i}{\left(\lambda_i \overline{\gamma}_{S,D} - 1\right)^2} \Bigg[\Theta_c(0, \gamma_o \overline{\gamma}_{S,D}^2, \lambda_i) + \Phi_c(0, \gamma_o, \overline{\gamma}_{S,D}^2 \lambda_i, \frac{1}{\overline{\gamma}_{S,D}}) \\
& - \Theta_c(0, \gamma_o, \overline{\gamma}_{S,D}^2, \frac{z}{\overline{\gamma}_{S,D}}) - \Phi_c(0, \gamma_o, \overline{\gamma}_{S,D}, \frac{z}{\overline{\gamma}_{S,D}}) + \Theta_c(\gamma_o, \infty, \eta, \lambda_i) \Bigg]
\end{aligned} \tag{24}
$$

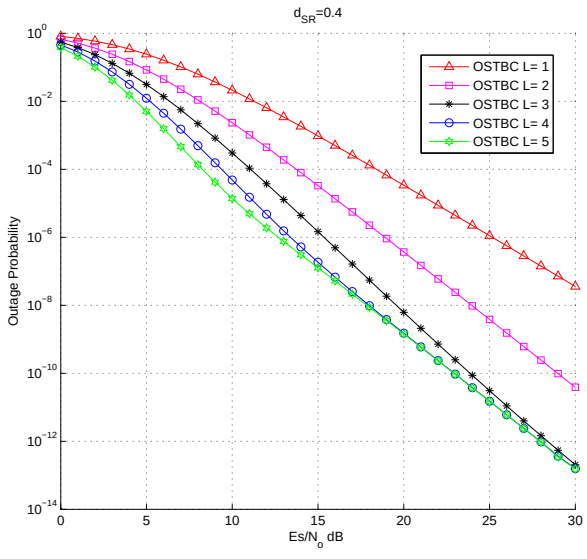


Fig. 3. Outage probability versus the SNR of the transmitted signal (E_s/N_o) for the best relay selection with $\gamma_o = 5$ dB.

one can notice that by increasing, L the outage probability, P_{out}, will be decreased and a better overall performance will be achieved. Furthermore, increasing L will result in an improvement in the outage performance to some extent whereas a slight improvement will be achieved when $L > 3$.

Finally, Figure 4 shows the performance of the average achievable rate versus the SNR of the transmitted signal for both the incremental cooperative diversity with best relay selection utilizing dual transmit diversity and the regular cooperative diversity OSTBC for different number of the participating relay nodes. The result shows that utilizing the OSTBC scheme with the incremental relaying best relay selection scheme saves channel resources (i.e., spectral efficiency improved) in the sense of the average channel capacity as compared to the regular cooperative scheme of relaying (i.e., $\gamma_o \to \infty$). However, increasing the number of the involved relaying nodes yields to a slight channel utilization (i.e., slight increase in the overall average achievable rate). However, a remarkable improvement in the the sense of the BER and the SNR outage probability with approximately the same average achievable rate performance.

VII. Conclusion

In paper, an incremental best relay cooperative diversity with a dual transmit diversity is studied. Closed-form expressions for the average bit error rate, average achievable rate, and signal to noise ratio outage probability are derived. Here, a

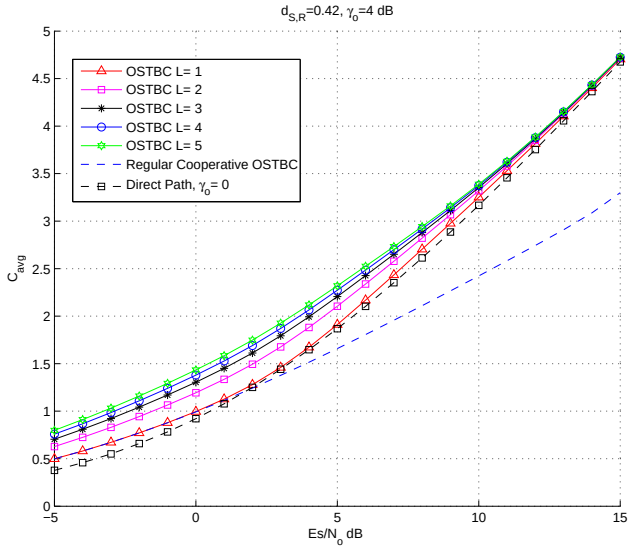


Fig. 4. Average achievable rate versus the SNR of the transmitted signal (E_s/N_o) for best relay selection with $(d = 0.42)$.

binary phase shift keying modulation scheme over independent and identically distributed Rayleigh fading channels is used. The results show that the best relay selection scheme with multiple input single output incremental-relaying outperforms the regular incremental cooperative diversity and can achieve full diversity orders with the same transmitted energy, which is the sum energy of the two transmit antenna.

APPENDIX A

For simplicity and to reduce the number of terms used in the integrals, the following auxiliary functions $\Phi(A, B, \omega, \lambda)$ and $\Theta(A, B, \omega, \lambda)$ are defined as follows:

$$\Phi(\ A\ , B, \omega, \lambda) = \int_A^B \omega\, z e^{-\lambda\, z} erfc\left(\sqrt{z}\right) dz$$
$$= \frac{\omega}{\sqrt{\pi}\lambda^2\,(1+\lambda)^{3/2}}\Bigg[\left(-3/2\,\lambda - 1\right)\sqrt{\pi} erf\left(\sqrt{1+\lambda}\sqrt{z}\right)$$
$$+ \sqrt{1+\lambda}\Bigg(e^{-\lambda\, z}\sqrt{\pi}\left(-1 + erf\left(\sqrt{z}\right)\right)(1+\lambda)\,(\lambda\, z + 1)\Bigg)$$
$$+ \lambda\sqrt{z} e^{-(1+\lambda)z}\Bigg]_A^B \qquad \text{(A-1)}$$

$$\Theta\ (A, B, \omega, \lambda) = \int_A^B \omega e^{-\lambda\, z} erfc\left(\sqrt{z}\right) dz$$
$$= \omega\left[-\frac{e^{-\lambda\, z}}{\lambda} + \frac{erf\left(\sqrt{z}\right)}{\lambda e^{\lambda\, z}} - \frac{erf\left(\sqrt{1+\lambda}\sqrt{z}\right)}{\lambda\sqrt{1+\lambda}}\right]_A^B \qquad \text{(A-2)}$$

APPENDIX B

For the average achievable rate evaluation, the integrals involving the *log* function are widely appearing. Next, we introduced some auxiliary functions used in this paper.

$$\Phi_c(A, B, \omega, \lambda) = \int_A^B \omega\, z e^{-\lambda\, z} log_2\,(z+1)\, dz$$
$$= \Bigg[-\frac{\omega\ (1+\lambda\, z)\, \mathrm{e}^{-\lambda\, z}\ln(1+z)}{\ln(2)\,\lambda^2}$$
$$- \frac{\omega\, \mathrm{e}^{-\lambda\, z}}{\ln(2)\,\lambda^2} + \frac{\omega\, \mathrm{e}^{\lambda} E_1\,(\lambda + \lambda\, z)}{\ln(2)\,\lambda}$$
$$- \frac{\omega\, \mathrm{e}^{\lambda} E_1\,(\lambda + \lambda\, z)}{\ln(2)\,\lambda^2}\Bigg]_A^B \qquad \text{(A-3)}$$

$$\Theta_c(A, B, \omega, \lambda) = \int_A^B \omega e^{-\lambda\, z} log_2\,(z+1)\, dz$$
$$= \Bigg[-\frac{\omega\, \mathrm{e}^{-\lambda\, z}\ln(1+z)}{\ln(2)\,\lambda} - \frac{\omega\, \mathrm{e}^{\lambda} E_1\,(\lambda + \lambda\, z)}{\ln(2)\,\lambda}\Bigg]_A^B \qquad \text{(A-4)}$$

where $E_1(.)$ is the exponential integral.

REFERENCES

[1] F. D. Boyer J. and Y. H, "Multi-hop diversity in wireless relaying channels," *IEEE Trans. Commun*, vol. 52, no. 10, pp. 1820–1830, 2004.

[2] F. D. Boyer J and Y. H, "A theoretical characterization of the multi-hop wireless communication channel with diversity," in *Proc. Global Telecommunications Conf*, (San Antonio, TX), pp. 841–845, November 2001.

[3] A. M. Ikki S.S., "Performance analysis of incremental-relaying cooperative-diversity networks over rayleigh fading channels communications," *IET*, vol. 5, pp. 337–349, Feb 2011.

[4] A. M. Hayajneh and M. Khodier., "Incremental-relaying cooperative-networks using dual transmit diversity over rayleigh fading channels," in *Proc. Global Engineering, Science and Technology Conference*, (Dubai, UAE), Apr 2013.

[5] S. Alamouti, "A simple transmit diversity technique for wireless communications," *IEEE J. Select. Areas Commun.*, pp. 1451–1458, Oct 1998.

[6] A. Goldsmith, *Wireless Communications*. Cambridge University Press, 1st edn ed., 2005.

[7] A. Adinoyi, Y. Fan, H. Yanikomeroglu, H. Poor, and F. Al-Shaalan, "Performance of selection relaying and cooperative diversity," *Wireless Communications, IEEE Transactions on*, vol. 8, no. 12, pp. 5790–5795, 2009.

[8] H. Yang and M. Alouini, *Order Statistics in Wireless Communications: Diversity, Adaptation, and Scheduling in MIMO and OFDM Systems*. Cambridge University Press, 2011.

[9] L. Dai, B. Gui, and L. J. Cimini, "Selective relaying in ofdm multihop cooperative networks," in *Wireless Communications and Networking Conference, 2007. WCNC 2007. IEEE*, pp. 963–968, IEEE, 2007.

[10] J. N. Laneman and G. W. Wornell, "Distributed space-time-coded protocols for exploiting cooperative diversity in wireless networks," *Information Theory, IEEE Transactions on*, vol. 49, no. 10, pp. 2415–2425, 2003.

[11] A. Papoulis and S. Pillai, *Probability, random variables, and stochastic processes*. McGraw-Hill electrical and electronic engineering series, McGraw-Hill, 2002.

Closed-form Expression for Outage Probability in Relay-Based Cooperative Diversity Systems over Nakagami Fading Channels with Interference

Rami Mohaisen[1], Mamoun F. Al-Mistarihi[1], and Ashraf Sharaqa[2]

(1) Jordan University of Science and Technology/ Electrical Engineering Department, Irbid, Jordan
rami.fayez@hotmail.com, mistarihi@just.edu.jo

(2) WorleyParsons Arabia Ltd / Communication and Security Projects Department, Al-Khobar, Saudi Arabia
elengashraf@yahoo.com

Abstract - **Incremental relaying, which is a form of the adaptive cooperative wireless networks, has been recently introduced to compensate for the channel resources utilization deficiencies associated with the fixed relaying schemes. Incremental relaying operates by exploiting a feedback from the destination to both the source and the relay that inform them about usability of the source-destination direct channel and by this, the channel resources will be saved since the relay will be used only when needed. This paper discusses the performance, namely the outage probability, of the decode-and-forward incremental relaying scheme over Nakagami-m fading channel and incorporates the effects of the multiple L interferers near the destination that will degrade the overall system performance. Numerical results are provided to validate our derivations.**

Keywords: Cooperative Diversity, Incremental Relaying, Fixed Relaying, Outage Probability, Signal to Noise Ratio, Signal to Interference Ratio, Decode-and-Forward, Maximal Ratio Combiner (MRC).

I. INTRODUCTION

The newly adopted form of the cooperative diversity is deployed by placing an intermediate relays between the source and the destination in order to relay the source signal to the destination. And by this, the spatial diversity gain will be achieved at the destination without placing the actual multiple antennas on both the source and the destination as in the Multiple Input Multiple Output systems (MIMO) [1].

Cooperative wireless networks is further splitted into regular or fixed relaying schemes and adaptive relaying schemes, and their functions is dependent, among many other factor, on the usability of the channel resources allocations between the source and the relay. In other words, in the fixed cooperative schemes the relay(s) is/are used to relay the source signal in a deterministic manner without taking into account the status of the direct channel from the source to the destination which in turn will lead to an improper utilization of the channel resources [2-3]. These schemes have been thoroughly analyzed and discussed as in [4-5] where a closed form expressions were obtained for the amplify-and-forward fixed relaying systems over Rayleigh fading channels.

A new form of the cooperative wireless networks is introduced in order to compensate for the deficiencies associated with the fixed cooperative schemes. The adaptive wireless networks are implemented in an opportunistic manner. In other words, the relay path(s) will be used only when needed and more precisely, when the direct path is in outage, by this, the channel resources will be saved effectively and properly used [6].

The incremental relaying scheme is a form of the adaptive cooperative networks in which a feedback from the destination is sent to both the relay and the source in each time slot to inform them about the usability of the direct channel. For instance if the direct transmission was successful, then a single bit will be sent from the destination informing the source and the relay about the successfulness of the transmission of the first time slot and requesting a new signal from the source within the second time slot. Otherwise, a negative feedback will inform the relay to resend the source signal within second time slot after which the destination will combine the two received signal from both the source and the relay by the maximal ration combiner (MRC) [7]. This system is thoroughly analyzed as in [8] where a closed form expression of the outage probability and the BER were obtained for the decode-and-forward incremental relaying over Rayleigh fading channels with no interferers introduced in the system.

In this paper, a closed form expression for the outage probability of the Decode-and-Forward Incremental Relaying over Nakagami-m fading channels with the presence of the multiple interferers near the destination that cause a Co-channel interference is derived. Numerical results are provided to show the effects on the system performance when changing the Nakagami-m order, the number of the interferers near the destination and the minimal threshold γ_o, below which the destination would request the assistance of the relay in sending the source signal.

This paper is further presented as follows; Section II provides the system model. Section III provides the performance analysis and derived tight closed form expression of outage probability. Results and discussion are presented in section IV. We concluded this paper in section V.

II. SYSTEM MODEL

Fig.1 shows the incremental relaying system where the source and the relay communicate with the destination over a channel whose channel coefficients, $h_{S,D}$, $h_{S,R}$ and $h_{R,D}$ which follow the Nakagami-m distribution. The noise terms at the destination are all assumed to be additive white Gaussian noise (AWGN) with zero mean and a unity variance N_0. The multiple access technique used to send the source signal can be any orthogonal access technique such as Frequency Division Multiple Access (FDMA) or the Time Division Multiple Access (TDMA), in our system, we adopted the TDMA. All the communicating nodes are equipped with only one single antenna. The channels from the interferers to the destination are assumed to be identically and independently distributed Rayleigh random variables. If the relay path is used, then upon the reception of the two signals from the source and the relay, the destination will combine these two signals by the maximal ratio combiner (MRC).

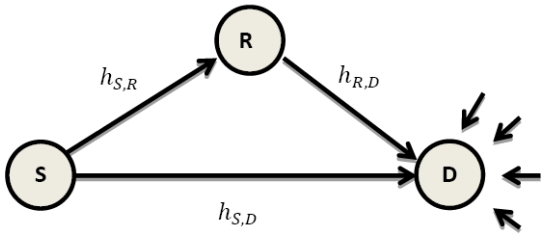


Fig. 1 Cooperative Diversity Network with Multiple L Interferers Located near the Destination.

Upon the reception of the source signal within the first time slot by the destination, the destination will decide whether or not the relay should participate in the transmission of the source signal by sending a feedback to both the source and the relay informing them about the usefulness of the direct path. For instance, if the direct path is in outage then the destination will send a negative feedback requesting the relay to send the source signal to the destination within the second time slot. Otherwise, a positive feedback will be sent to inform the source to send a new signal within the second time slot.

The received signals at the destination from both the source and the relay within the first time slot are as follows:

$$y_{S,D}(t_1) = h_{S,D}\sqrt{E_s}\, x(t_1) + n_1(t_1) + I(t_1) \tag{1}$$

$$y_{S,R}(t_1) = h_{S,R}\sqrt{E_s}\, x(t_1) + n_2(t_1) \tag{2}$$

where $x(t)$ is the source signal and E_s is the energy of the source signal. The interferers effects near the destination within the first time slot is represented by $I(t_1)$. $n_i(t_1)$ where $i \in \{1,2\}$, represent the noise effects at the destination from the source and the relay within the first time slot.

If the relay participates in the transmission, then the received signal at the destination within the second time slot will be:

$$y_{R,D}(t_2) = h_{R,D}\sqrt{E_s}x_r(t_2) + n_3(t_2) + I(t_2) \tag{3}$$

where $x_r(t_2)$ is the regenerated signal at the relay, $I(t_2)$ represents the effects of the Co-channel interference near the destination in the second time slot.

The quality assessment measure of the direct path by the destination is determined by comparing the signal to noise ratio (SNR) of the direct path against γ_o which is the minimal SNR beyond which the destination can receive the source signal without the help of the relay which is also an application dependent parameter. An important observation is that as γ_o increases, the relay path will be used more often which in turn will enhance the spatial diversity benefits in the expense of the deficiency of using the channel resource properly since the relay path will be used more often.

III. PERFORMANCE ANALYSIS

The corner stones in the derivations of the performance metrics of any communication system are by providing the probability density function (PDF) and the cumulative distribution function (CDF) of both the signal to interference ratio (SIR) at the destination from the direct and the relay paths. Since the effect of the noise at the destination is not comparable to the Co-channel interference, we will bypass the noise effects by working only on the SIR at the destination where interference limited environment is a common practice in the literature.

In the following subsections we will derive the PDFs and the CDFs of the SIRs ($\eta_{S,D}$, $\eta_{R,D}$) from both the source and the relay to the destination and the SNR ($\gamma_{S,R}$) from the source to the relay respectively.

A. Probability Density Function of SIR's $\boldsymbol{\eta_{S,D}}$ and $\boldsymbol{\eta_{R,D}}$

The PDF of the SIR ($\boldsymbol{\eta_{S,D}}$) at the destination from the source path is given as follows:

$$f_{\eta_{S,D}}(\eta_{S,D}) = \zeta_{S,D}\left(\frac{(\eta_{S,D})^{m-1}}{(\eta_{S,D}+\lambda_{S,D})^{L+m}}\right) \tag{4}$$

where $\zeta_{S,D} = \frac{(L+m-1)!}{m^L\Gamma(m)(L-1)!}(m\,\lambda_{S,D})^L$, $\lambda_{S,D} = \frac{\overline{\gamma_{S,D}}}{m\overline{\gamma_C}}$

The PDF of the SIR **($\boldsymbol{\eta_{R,D}}$)** at the destination from the relay path is given as follows**:**

$$f_{\eta_{R,D}}(\eta_{R,D}) = \zeta_{R,D}\left(\frac{(\eta_{R,D})^{m-1}}{(\eta_{R,D}+\lambda_{R,D})^{L+m}}\right) \tag{5}$$

where $\zeta_{R,D} = \frac{(L+m-1)!}{m^L\Gamma(m)(L-1)!}(m\,\lambda_{R,D})^L$, $\lambda_{R,D} = \frac{\overline{\gamma_{R,D}}}{m\overline{\gamma_C}}$

Proof:

Given that the SNR of the direct path $\gamma_{S,D}$ and the relay path $\gamma_{R,D}$ at the destination from both the source and the relay, respectively, follow the Nakagami-m distribution and the fading channels from the interferers to the destination ($\gamma_b = \sum_1^L \gamma_k$) is assumed to be a sum of identically and independently distributed exponential random variables. The PDF of the SIR ($\boldsymbol{\eta_{S,D}}$) can be calculated as [10]:

$$f_{\eta_{S,D}}(\eta_{S,D}) = \int_0^\infty \gamma_b f_{\gamma_{S,D}}(\gamma_b \eta_{S,D}) f_{\gamma_b}(\gamma_b) d\gamma_b \quad (6)$$

And the PDF of the SIR $(\boldsymbol{\eta_{R,D}})$ can be calculated as:

$$f_{\eta_{R,D}}(\eta_{R,D}) = \int_0^\infty \gamma_b f_{\gamma_{R,D}}(\gamma_b \eta_{R,D}) f_{\gamma_b}(\gamma_b) d\gamma_b \quad (7)$$

The PDFs of $\gamma_{S,D}$, $\gamma_{R,D}$ and γ_b are denoted as [9]:

$$f_{\gamma_{S,D}}(\gamma_{S,D}) = \frac{m^m {\gamma_{S,D}}^{m-1}}{\overline{\gamma_{S,D}}^m \Gamma(m)} exp\left(-\frac{m\gamma_{S,D}}{\overline{\gamma_{S,D}}}\right), \gamma_{S,D} > 0 \quad (8)$$

$$f_{\gamma_{R,D}}(\gamma_{R,D}) = \frac{m^m {\gamma_{R,D}}^{m-1}}{\overline{\gamma_{R,D}}^m \Gamma(m)} exp\left(-\frac{m\gamma_{R,D}}{\overline{\gamma_{R,D}}}\right), \gamma_{R,D} > 0 \quad (9)$$

$$f_{\gamma_b}(\gamma_b) = \frac{{\gamma_b}^{L-1}}{\overline{\gamma_c}^L (L-1)!} \exp\left(\frac{-\gamma_b}{\overline{\gamma_c}}\right) \quad (10)$$

where $\overline{\gamma_c} = \frac{\mathcal{E}}{N_o} E({\alpha_k}^2)$ is the average SNR per channel which is assumed to be independent and identical for all the channels, and $\mathcal{E}$ is the signal energy and:

$$\gamma_b = \frac{\mathcal{E}}{N_o} \sum_{k=1}^{L} {\alpha_k}^2 = \sum_{k=1}^{L} \gamma_k \quad (11)$$

γ_k is being the instantaneous SNR of the $k-th$ channel and L is the number of the interfering channels near the destination. And since the fading channels are mutually and statistically independent, then $\gamma_k's$ are statistically independent.

Solving the integrals in (6) and (7) to get the compact form of the SIR's PDF at the destination from the direct path as in (4) and from the relay path as in (5) using [11,(3.351.1)].

B. The Cumulative Distribution Function of SIR's $\boldsymbol{\eta_{S,D}, \eta_{R,D}}$

The CDF of the SIR $(\boldsymbol{\eta_{S,D}})$ at the destination from the source path is given as follows:

$$F_{\eta_{S,D}}(\eta_{S,D}) = \frac{\zeta_{S,D}}{(\lambda_{S,D})^{L+m}} \frac{{\eta_{S,D}}^m}{m} \, {}_2F_1\left(L+m, m; 1+m; -\frac{\eta_{S,D}}{\lambda_{S,D}}\right) \quad (12)$$

The CDF of the SIR $(\boldsymbol{\eta_{R,D}})$ at the destination from the relay path is given as follows:

$$F_{\eta_{R,D}}(\eta_{R,D}) = \frac{\zeta_{R,D}}{(\lambda_{R,D})^{L+m}} \frac{{\eta_{R,D}}^m}{m} \, {}_2F_1\left(L+m, m; 1+m; -\frac{\eta_{R,D}}{\lambda_{R,D}}\right) \quad (13)$$

where ${}_2F_1(a, b; c; z)$ is the Gaussian Hyper geometric function [11].

Proof:

The CDF of $\eta_{S,D}$ can be calculated as [10]:

$$F_{\eta_{S,D}}(\eta_{S,D}) = \int_0^{\eta_{S,D}} f_{\eta_{S,D}}(y) dy \quad (14)$$

where $f_{\eta_{S,D}}(y)$ is given as in (4).

The CDF of $\eta_{R,D}$ is found by integrating the PDF of $\eta_{R,D}$ given in (5) as follows:

$$F_{\eta_{R,D}}(\eta_{R,D}) = \int_0^{\eta_{R,D}} f_{\eta_{R,D}}(y) dy \quad (15)$$

Solving the integrals in (14) and (15) using [11,(2.117.2)], the compact form of the CDF of $\eta_{S,D}$ will be as in (12) and that of $\eta_{R,D}$ will be as in (13).

The CDF of the SNR $(\boldsymbol{\gamma_{S,R}})$ at the relay from the source path is given as follows:

$$F_{\gamma_{S,R}}(\gamma_{S,R}) = \frac{\gamma\left(m, m\frac{\gamma_{S,R}}{\overline{\gamma_{S,R}}}\right)}{\Gamma(m)} \quad (16)$$

where $\gamma(a, b)$ is the Incomplete Gamma function [11] and $\Gamma(m)$ is the gamma function.

Proof:

The CDF of $\boldsymbol{\gamma_{S,R}}$ is found by integrating the PDF of $\boldsymbol{\gamma_{S,R}}$ as follows [10]:

$$F_{\gamma_{S,R}}(\gamma_{S,R}) = \int_0^{\gamma_{S,R}} f_{\gamma_{S,R}}(y) dy \quad (17)$$

where $f_{\gamma_{S,R}}(y)$ is given as [9]:

$$f_{\gamma_{S,R}}(\gamma_{S,R}) = \frac{m^m {\gamma_{S,R}}^{m-1}}{\overline{\gamma_{S,R}}^m \Gamma(m)} exp\left(-\frac{m\gamma_{S,R}}{\overline{\gamma_{S,R}}}\right), \gamma_{S,R} \geq 0 \quad (18)$$

Solving the integral in (17) using [11,(3.381.1)], the compact form of the CDF of $\boldsymbol{\gamma_{S,R}}$ will be as in (16).

C. Outage Probability Analysis

The outage probability events of the incremental relaying are characterized in two events; basically, when the direct path is in outage, the destination will inform the relay to send the source signal in an attempt to enhance the reception of the signal by combining them using the MRC. The second event when the SNR of the MRC's two combined signals at the destination fall below the pre-specified threshold γ_o.

Representing the two cases of the decode-and-forward incremental relaying outage probability events, p_{out} is given as:

$$p_{out} = P_r(\eta_{S,D} \leq \gamma_o) \times P_r[\min(\gamma_{S,R}, \eta_{S,D} + \eta_{R,D} \leq \gamma_O | \eta_{S,D} \leq \gamma_o)]$$

$$p_{out} = P_r(\eta_{S,D} \leq \gamma_o)\left[1 - P_r(\eta_{S,D} + \eta_{R,D} > \gamma_o | \eta_{S,D} \leq \gamma_o)\right] \times P_r(\gamma_{S,R} > \gamma_o | \eta_{S,D} \leq \gamma_o)$$

$$p_{out} = P_r(\eta_{S,D} \leq \gamma_o)\left[1 - \frac{P_r(\eta_{S,D} \leq \gamma_o) - P_r(\eta_{S,D} + \eta_{R,D} \leq \gamma_o)}{P_r(\eta_{S,D} \leq \gamma_o)} P_r(\gamma_{S,R} > \gamma_o)\right] \quad (19)$$

It is worth to mention that the operator min(•|•) encompasses the two outage events, in other words, the first argument within the operator is the outage at the relay whereas the second one represents the outage at the destination .

Further simplifications on the above expression yields:

$$p_{out} = P_r(\gamma_{S,R} \leq \gamma_o) P_r(\eta_{S,D} \leq \gamma_o) + P_r(\gamma_{S,R} > \gamma_o) P_r(\eta_{S,D} + \eta_{R,D} \leq \gamma_o) \quad (20)$$

$$p_{out} \approx \frac{\zeta_{S,D}}{m(\lambda_{S,D})^{L+m}} \frac{\gamma\left(m, m\frac{\gamma_o}{\gamma_{S,R}}\right)}{\Gamma(m)} (\gamma_o)^m \; {}_2F_1\left(L+m, m; 1+m; -\frac{\gamma_o}{\lambda_{S,D}}\right) + \frac{\zeta_{R,D}}{m(\lambda_{R,D})^{L+m}} \left[1 - \frac{\gamma\left(m, m\frac{\gamma_o}{\gamma_{S,R}}\right)}{\Gamma(m)}\right] (\gamma_o)^m \; {}_2F_1\left(L+m, m; 1+m; -\frac{\gamma_o}{\lambda_{R,D}}\right) \quad (21)$$

When using the MRC, the value of $\eta_{S,D}$ is not comparable to $\eta_{R,D}$ since we used the relay path in the first place to compensate for the deficiency of the direct transmission. And hence, it will be intuitive to omit $\eta_{S,D}$ from the argument to get the approximated outage probability expression as given in (21).

The detailed mathematical derivations of (4), (5), (12), (13), (16), and (21) will be submitted to a journal for publication.

IV. NUMERICAL RESULTS

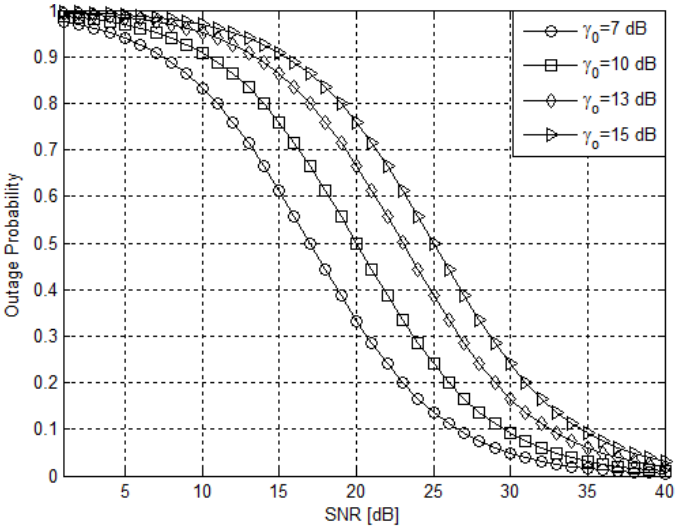


Fig. 2 Outage probability for different values of γ_o, (γ_o=7 dB, 10 dB, 13 dB and 15 dB), the Nakagmi-m order (m= 1) and only one interferer near the destination.

Fig.2 depicts the system performance when there is only one interferer and the Nakagmi-m order is one (m=1) and various values of γ_o. From the figure, one can see that as we increase γ_o, the system performance in terms of the oytage probability will be degraded. As an example, for a system with (γ_o =15 dB) to have the same performance as that with (γ_o=7 dB), an increment of about 8 dB is required given that the outage probability is 0.5.

Fig.3 shows the effects of changing the Nakagni-m order on the outage probability performance when there is only one interferer near the destination and (γ_o= 10 dB). As can be seen from the figure, as the Nakagami-m order increases, the system performance will be enhanced. For example, for an outage probability of 0.1, an increment of about 5 dB is required for a system with (m=1) to have the same performance as that with (m=3).

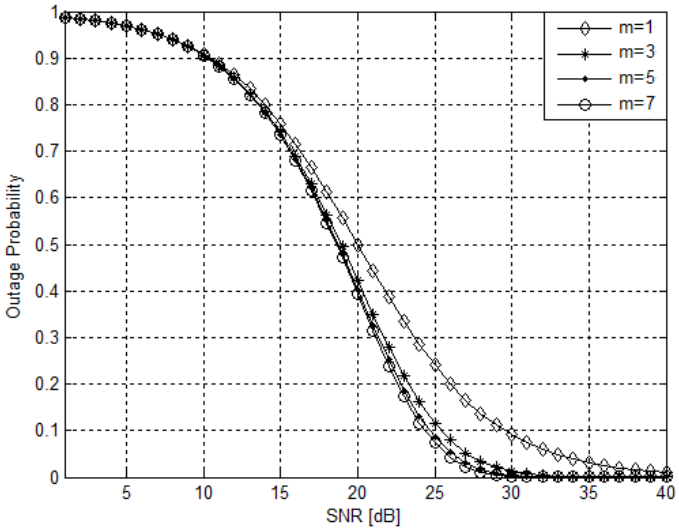


Fig. 3 Outage Probability for different values of the Nakagmi-m order (m=1,3,5,7), (γ_o=10dB) and only one interferer.

Fig.4 combines the impact of changing γ_o and the Co-channel effects caused by the interferers near the destination on the system outage probability. As shown from the figure, the impact of increasing γ_o on the system performance is more apparent than changing the number of interferers. For example, a system with (γ_o=10 dB) and (L=2+3) will behave much better than a system with (γ_o=15 dB) and (L=1).

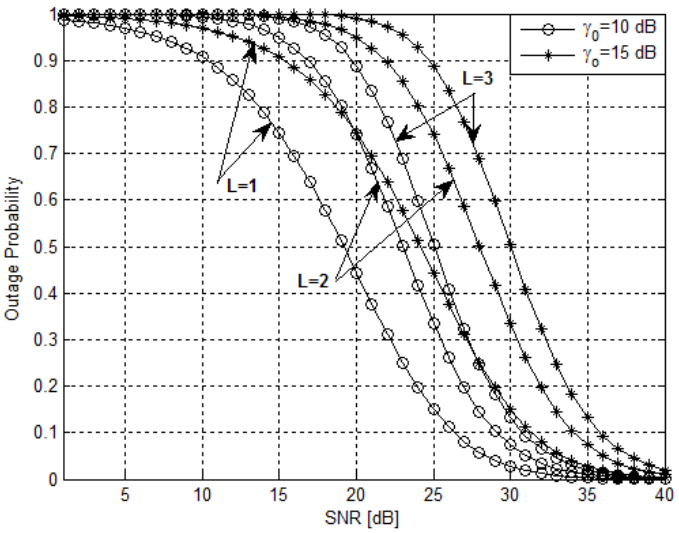


Fig. 4 Outage Probability for different numbers of interferers (L=1,2,3), different values of γ_o (γ_o=10 dB, 15 dB) and the Nakagami-m order (m=1).

Fig.5 depicts the outage probability behavior when changing the number of interferers and the Nakagami-m order for a system with γ_o=13 dB. and this is obvious especially in the high SNR region. From the figure, as we increase the Nakagami-m order and decrease the number of interferers a considerable enhancement will be achieved especially in the high SNR region. Also, the effects of changing the Nakagami-m order on the system performance is more influential on the outage probability than changing the number of the interferers. For example, given that (L=1), then for a system with (m=1) to behave as that with (m=4), an increment of about 5 dB is required to have the same performance for an outage probability of 0.1. To illustrate the effect of the interference, for an outage probability of 0.2 and Nakagami-m order of one (m=1), an increment of 5 dB is required for a system with (L=3) to behave the same as the one with (L=1).

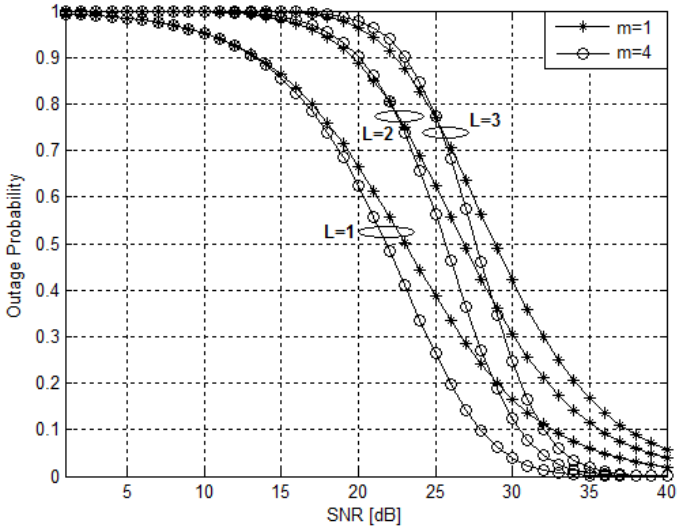


Fig. 5 Outage Probability for different numbers of interferers (L=1,2,3), different values of the Nakagami-m order (m=1,4) when (γ_o=13 dB).

V. CONCLUSION

In this paper, we analyzed the effects of the Co-channel interference on the decode-and-forward incremental relaying scheme over Nakagmi-m fading channel caused by the multiple L interferers near the destination which in turn will degrade the overall system performance. Closed form expression of the outage probability is derived. It is apparent that as we increase the Nakagmi-m order and decrease the number of the interferers near the destination and γ_o, the system will get a considerable enhancement regarding the outage probability performance.

REFERENCES

[1] J. N. Laneman, D. N. C. Tse, and G. W. Wornell, "Cooperative diversity in wireless networks: efficient protocols and outage behavior," *IEEETrans. Inf. Theory*, vol. 50, Dec. 2004.

[2] M. O. Hasna and M.-S. Alouini, "End-to-end performance of transmission systems with relays over Rayleigh-fading channels," *IEEE Trans. Wireless Commun.*, vol. 2, no. 6, pp. 1126-1131, Nov. 2003.

[3] P. A. Anghel and M. Kaveh, "Exact symbol error probability of a cooperative network in a Rayleigh-fading environment," *IEEE Trans. Wireless Commun.*, vol. 3, pp. 1416-1421, Sep. 2004.

[4] N. C. Beaulieu and J. Hu, "A closed-form expression for the outage probability of decode-and-forward relaying in dissimilar Rayleigh fading channels," *IEEE Commun. Letters*, vol. 10, pp.813-815, Dec. 2006.

[5] H. A. Suraweera, P. J. Smith, and J. Armstrong, "Outage probability of cooperative relay networks in Nakagami-m fading channels," IEEE Commun. Letters, vol. 10, pp.834-836, Dec. 2006.

[6] S. Ikki and M. H. Ahmed, "Performance of multiple relay cooperative diversity systems with best relay selection over Rayleigh fading channels," *EURASIP J. Advances Signal Process.*, vol. 2008.

[7] S. Ikki and M. H. Ahmed, "Performance analysis of cooperative diversity wireless networks over Nakagami-m fading channel," *IEEE Commun. Lett.*, vol. 11, no. 4, pp. 334-336, Apr. 2007.

[8] S. Ikki and M. H. Ahmad, ' Performance Analysis of Decode-and-Forward Incremental Relaying Cooperative-Diversity Networks over Rayleigh Fading Channels' *Proc. IEEE Technology Conference-Spring* (VTC-Spring'09), Barcelona , Spain, Apr.2009.

[9] M. K. Simon and M.-S. Alouini, *Digital Communication over Fading Channels*, John Wiley and Sons, New York, NY, USA, 2000.

[10] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, 4th ed. New York: McGraw-Hill, 2002.

[11] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, 7th ed. Academic Press, 2007.

On practical solutions of AQM queues with large buffers

Pawel Mrozowski and Andrzej Chydzinski
Faculty of Automatic Control, Electronics and Computer Science, Silesian University of Technology, Gliwice, Poland
andrzej.chydzinski@polsl.pl

***Abstract*—In a router equipped with the active queue management (AQM), the arriving packets may be rejected even if the buffer is not full. This is supposed to reduce the queue size, while maintaining a high link utilization. The simplest approach to AQM is rejecting arriving packets with the probability that is a function of the queue length. The recently developed mathematical models of such systems present numerical difficulties that are not trivial to overcome, especially in systems of large sizes. In this paper we first overview the analytical solutions of these models, which contain both Markovian (Poisson) and non-Markovian (general renewal) traffic models. Then we discuss the computational techniques that enable solving and designing of such AQM queues with buffers for thousands of packets. Finally, we demonstrate computational examples with such buffers.**

***Index Terms*—active queue management, Internet routers, dropping function, numerical solutions**

I. Introduction

It is known that FIFO tail-drop queues used for packets in Internet routers have several disadvantages. In particular, they induce large queueing delays, flow synchronization and unfairness between flows, to name only three of them. When these (and other) problems with FIFO tail-drop queues were noticed, the active queue management (AQM) for Internet routers was proposed. The idea is that the router can drop incoming packets even if the buffer is not full yet, thus preventing the queue from building up. Additionally, if random dropping is involved, the flow synchronization is reduced and the fairness improved. The first proposed AQM algorithms were the Random Early Detection gateway (RED) [1] and its modifications e.g. [2], [3], [4], [5], [6]. To date many other AQM algorithms were proposed, including the well-known solutions like REM [7], PI [8], [9], BLUE [10], GREEN [11], AVQ [12], and newest propositions [13], [14], [15], [16].

The simplest approach to AQM is rejecting arriving packets with the probability that is a function of the queue size observed upon packet arrival. The function assigning the dropping probability to the queue size is called a *dropping function*. Intuitively, the dropping function should be monotonic (non-decreasing). In RED for instance, the dropping function is linear.

Recently, two important models of the queueing systems with dropping functions have been solved analytically, namely the model with Poisson arrivals [17], and the model with exponential service distribution, [18]. These models are equivalent to the classic $M/G/1/N$ and $G/M/1/N$ finite-buffer queues, but with added dropping function mechanism.

Unfortunately, for the reasons discussed in details in Section V, the analytical results presented in [17], [18] cannot be used directly to obtain numerical results when the buffer size is large. In fact, they are even hard to use for moderate buffers. (In papers [17], [18] only examples with very small buffers are presented – for 10 and 6 packets, respectively).

Naturally, in practice we need solutions for much larger buffers. The routers' manufacturers use the bandwidth-delay product when dimensioning the buffers [19]. Typically, the buffer is supposed to accommodate 100ms (or 200ms) of the traffic. For instance, if the output link capacity is 1Gb/s and the average packet size equals 1000 bytes, then the buffer for this link should accommodate 12500 packets (1Gb/s $\cdot$ 100ms). The active queue management should reduce the average queue (and the necessary buffer size) by an order of magnitude. Thus in practice we have to solve models with buffers for 1000-2000 packets.

Therefore, the purpose of this paper is to present and discuss computational techniques that enable solving and designing of AQM queues with buffers for a few thousands of packets, basing on the analytical results of [17], [18].

The remaining part of the paper is structured as follows. In Section II, the formal description of the queueing model with the dropping function is recalled. Then, in Sections III and IV, the analytical solutions of the model with renewal (G) and Poisson (M) arrivals are reviewed, respectively. In Section V, the computational issues connected with large buffers are reviewed first. Then, for each of them a proposition of the solution is presented. In Section VI, several numerical examples with buffers for 2000 packets and different traffic and service types are presented and discussed. Finally, remarks concluding the paper are gathered in Section VII.

II. Queueing model

The following model of the queue of packets at the routers output interface will be considered.

Packets arrive to the queue according to a renewal process with the interarrival time distribution given by distribution function $F(t)$. The distribution of the service (transmission) time of a packet is given by distribution function $G(t)$. The packet buffer (waiting room) has size b (including the service position). Therefore, a packet incoming when already b packets

are present in the buffer is rejected. Additionally, even if the buffer is not full yet, every arriving packet can be dropped with probability d_n, where n is the queue size observed upon this packet arrival. The function d_n is called a dropping function and is not further specified - it can assume any form. By $\mathbb{P}$ we will denote the probability, by $X(t)$ the queue size at time t (including the service position), by X_k – the queue size left by the k-th departing packet, while by $f(s)$ – the Laplace transform of the interarrival time distribution, i.e:

$$f(s) = \int_0^\infty e^{-st} dF(t).$$

Finally, we denote:

$$P_m(t,n) = \mathbb{P}(X(t) = m | X(0) = n),$$

and

$$P_m = \lim_{t\to\infty} \mathbb{P}(X(t) = m).$$

In the following two sections we will recall the analytical results, which allow computation of P_m, obtained in [17], [18].

III. General arrivals and exponential service

In this section we assume that the service time is exponentially distributed with parameter μ, i.e.

$$G(t) = 1 - e^{-\mu t}, \tag{1}$$

while the interarrival time distribution can assume any form. Denote

$$p_m(s,n) = \int_0^\infty e^{-st} P_m(t,n) dt, \tag{2}$$

$$\mathbf{p}_m(s) = \left[p_m(s,0), p_m(s,1), ..., p_m(s,b)\right]^T. \tag{3}$$

In [18] it was proven that:

$$\mathbf{p}_m(s) = \mathbf{A}^{-1}(s)\mathbf{B}_m(s), \tag{4}$$

where $\mathbf{B}_m(s)$ is a vector of size $b+1$ defined as:

$$\mathbf{B}_m(s) = \left[I(m=0)\frac{1-f(s)}{s}, -\gamma_m(s,1), \ldots, -\gamma_m(s,b)\right]^T, \tag{5}$$

with

$$\gamma_m(s,n) = \int_0^\infty e^{-(s+\mu)t}\big(1-F(t)\big)\left[I(n \geq m \geq 1)\frac{(\mu t)^{n-m}}{(n-m)!} + I(m=0)\sum_{k=n}^{\infty}\frac{(\mu t)^k}{k!}\right] dt, \tag{6}$$

$$I(A) = \begin{cases} 1, & \text{if } A \text{ is true,} \\ 0, & \text{otherwise,} \end{cases} \tag{7}$$

while $\mathbf{A}(s)$ is a $(b+1)\times(b+1)$ matrix with the following elements:

$$a_{0,0}(s) = 1 - d_0 f(s), \tag{8}$$
$$a_{0,1}(s) = (d_0 - 1) f(s), \tag{9}$$
$$a_{0,j}(s) = 0, \quad 2 \leq j \leq b, \tag{10}$$
$$a_{i,0}(s) = d_0 \sum_{k=i}^{\infty} \alpha_k(s), \;\; 1 \leq i \leq b, \tag{11}$$
$$a_{i,1}(s) = d_1 \alpha_{i-1}(s) + (1-d_0)\sum_{k=i}^{\infty}\alpha_k(s) - I(i=1), \quad 1 \leq i \leq b, \tag{12}$$
$$a_{i,j}(s) = (1-d_{j-1})\alpha_{i-j+1}(s) + d_j \alpha_{i-j}(s) - I(i=j), \quad 1 \leq i \leq b, \;\; 2 \leq j \leq \min\{i, b-1\}, \tag{13}$$
$$a_{i,i+1}(s) = (1-d_i)\alpha_0(s), \;\; 1 \leq i \leq b-1, \tag{14}$$
$$a_{b,b}(s) = (1-d_{b-1})\alpha_1(s) + f(s+\mu) - 1, \tag{15}$$
$$a_{i,j}(s) = 0, \quad \text{otherwise,} \tag{16}$$

and

$$\alpha_k(s) = \int_0^\infty e^{-(s+\mu)t}\frac{(\mu t)^k}{k!} dF(t). \tag{17}$$

IV. Poisson arrivals and general service

In this section we assume that the interarrival time distribution is exponential with parameter λ, i.e.

$$F(t) = 1 - e^{-\lambda t}, \tag{18}$$

while the service time distribution may assume any form.

In [17] it was proven that the steady-state distribution of the queue size has the form:

$$P_b = 1 - \frac{\sum_{i=0}^{b-1}\pi_i/(1-d_i)}{\pi_0/(1-d_0)+\rho}, \tag{19}$$

$$P_m = \frac{\pi_m/(1-d_m)}{\pi_0/(1-d_0)+\rho}, \quad 0 \leq m \leq b-1, \tag{20}$$

where π_m is the distribution of the queue length left behind by departing packets, i.e.:

$$\pi_m = \lim_{n\to\infty}\mathbb{P}(X_n = m), \tag{21}$$

while ρ is the offered load:

$$\rho = \lambda \int_0^\infty t dG(t). \tag{22}$$

The numbers π_m can be computed by means of the system of equations:

$$\begin{cases} \pi_k = \sum_{j=0}^{b-1}\pi_j p_{j,k}, & 0 \leq k \leq b-1, \\ \sum_{j=0}^{b-1}\pi_j = 1, \end{cases} \tag{23}$$

with

$$p_{j,k} = \begin{cases} a_{1,k}, & \text{if} \quad j=0, \; 0 \leq k \leq b–1, \\ a_{j,k-j+1} & \text{if} \quad 1 \leq j \leq b–1, \; j-1 \leq k \leq b–1, \\ 0 & \text{otherwise,} \end{cases} \tag{24}$$

$$a_{n,k} = \int_0^\infty Q_{n,k}(u) dF(u), \tag{25}$$

where the function $Q_{n,k}(u)$ denotes the conditional probability that the first departure time is later than u and k packets are accepted in $(0, u]$, assuming $X(0) = n$. The values of $Q_{n,k}(u)$ can be obtained using its Laplace transform, i.e.:

$$q_{n,k}(s) = \int_0^\infty e^{-su} Q_{n,k}(u) du. \tag{26}$$

In [17] it was proven that

$$q_{n,k}(s) = \frac{\prod_{i=0}^{k-1} \lambda(1 - d(n+i))}{\prod_{i=0}^{k}(s + \lambda(1 - d(n+i)))}, \qquad n \geq 0, \quad k \geq 0. \tag{27}$$

V. Practical solutions

Although the analytical results presented in sections III and IV can be used in unaltered form to obtain numerical results, such approach may result in extensive, time-consuming computations, even in cases with moderate buffer sizes.

The numerical difficulties have several sources.

Firstly, the results are presented in terms of the Laplace transform. As demonstrated in [17], transform (27) is a rational function which can be inverted symbolically. In practice, however, it is achievable only for small values of b, therefore a numerical inversion have to be used. Formula (4) does not need an inversion, as the limit $\lim_{s\to 0+} s\mathbf{p}_m(s)$ can be used instead, but requires a proper choice of s for the limiting distribution – too small s may produce large errors connected with range overflows during computation. Secondly, there are integrals (e.g. in (6), (17)) that have to be computed either symbolically or numerically. Thirdly, infinite summations appear in formulas (6), (11) and (12). Finally, in (4) an inversion of a matrix of size $(b+1) \times (b+1)$ is needed. For a very large b this also may be computationally demanding.

In the remaining part of this section we address all the aforementioned issues.

A. Infinite sums

The infinite sums in formulas (6), (11) and (12) can be easily replaced by finite sums using the identity $e^x = \sum_{k=0}^\infty x^k/k!$. In particular, from (6) we obtain:

$$\gamma_m(s,n) = \int_0^\infty e^{-(s+\mu)t}\big(1 - F(t)\big)\left[I(n \geq m \geq 1) \frac{(\mu t)^{n-m}}{(n-m)!} + I(m=0)\left(e^{\mu k} - \sum_{k=0}^{n-1} \frac{(\mu t)^k}{k!}\right)\right] dt, \tag{28}$$

while (11) and (12) give:

$$a_{i,0}(s) = d_0\big[f(s) - \sum_{k=0}^{i-1} \alpha_k(s)\big], \;\; 1 \leq i \leq b, \tag{29}$$

$$a_{i,1}(s) = d_1 \alpha_{i-1}(s) + (1 - d_0)\big[f(s) - \sum_{k=0}^{i-1} \alpha_k(s)\big] - I(i=1), \quad 1 \leq i \leq b. \tag{30}$$

B. Integrals

The integrals in (6) and (17) can be computed symbolically for many types of distribution $F(t)$. Below we present ready-to-use results for four popular distributions.

1) Exponential distribution with parameter λ. We have

$$F(t) = 1 - e^{-\lambda t}, \qquad f(s) = \frac{\lambda}{s+\lambda}. \tag{31}$$

In this case, from (17) and (28) we obtain:

$$\alpha_k(s) = \frac{\lambda \mu^k}{(s + \lambda + \mu)^{k+1}}, \tag{32}$$

$$\gamma_n(s,m) = I(n \geq m \geq 1) \frac{\mu^{n-m}}{(s+\lambda+\mu)^{n-m+1}} + I(m=0)\left(\frac{1}{s+\lambda} - \sum_{k=0}^{n-1} \frac{\mu^k}{(s+\lambda+\mu)^{k+1}}\right). \tag{33}$$

2) Erlang distribution with parameters K and λ. We have

$$F(t) = 1 - \sum_{i=0}^{K-1} \frac{(\lambda t)^i}{i!} e^{-\lambda t}, \quad f(s) = \left(\frac{\lambda}{s+\lambda}\right)^K. \tag{34}$$

From (17) and (28) we obtain:

$$\alpha_k(s) = \frac{\lambda^K \mu^k (k+K-1)!}{(s+\lambda+\mu)^{k+K} k!(K-1)!}, \tag{35}$$

$$\gamma_n(s,m) = I(n \geq m \geq 1) \sum_{i=0}^{K-1} \frac{\lambda^i \mu^{m-n}(i+n-m)!}{(s+\lambda+\mu)^{n-m+1+i} i!(n-m)!} + I(m=0)\left(\sum_{i=0}^{K-1} \frac{\lambda^i}{(s+\lambda)^{i+1}} - \sum_{k=0}^{n-1}\sum_{i=0}^{K-1} \frac{\lambda^i \mu^k (i+k)!}{(s+\lambda+\mu)^{i+k+1} i! k!}\right). \tag{36}$$

3) Uniform distribution on interval $(0, v)$. We have

$$F(t) = \frac{1}{v} t, \qquad f(s) = \frac{1 - e^{-sv}}{sv}. \tag{37}$$

Now from (17) and (28) we obtain:

$$\alpha_k(s) = \mu^k \frac{k! - \Gamma(k+1, v(s+\mu))}{v k! (s+\mu)^{k+1}}, \tag{38}$$

$$\gamma_n(s,m) = I(n \geq m \geq 1)\frac{\mu^{n-m}(s+\mu)^{m-n-2}}{v(n-m)!}$$
$$\cdot \Big[-(n-m+1)!$$
$$+ v(s+\mu)\big((n-m)! - \Gamma(n-m+1, v(s+\mu))\big)$$
$$+ \Gamma(n-m+2, v(s+\mu))\Big] + I(m=0)\Big[\frac{e^{-sv}+sv-1}{s^2v}$$
$$- \sum_{k=0}^{n-1} \frac{\mu^k(s+\mu)^{-2-k}}{vk!}\Big(-(k+1)! + v(s+\mu)\big(k!$$
$$- \Gamma(k+1, v(s+\mu))\big) + \Gamma(k+2, v(s+\mu))\Big)\Big], \quad (39)$$

where $\Gamma(k,t)$ denotes the incomplete gamma function.

4) Constant interarrival time equal to c. W have

$$f(s) = e^{-cs}, \quad (40)$$

and

$$\alpha_k(s) = e^{-(s+\mu)c}\frac{(\mu c)^k}{k!}, \quad (41)$$

$$\gamma_n(s,m) = I(n \geq m \geq 1)\Big[\frac{\mu^{n-m}(s+\mu)^{m-n-1}}{(n-m)!}$$
$$\cdot \big((n-m)! - \Gamma(n-m+1, c(s+\mu))\big)\Big]$$
$$+ I(m=0)\Big[\frac{1-e^{-cs}}{s} - \sum_{k=0}^{n-1}\frac{\mu^k(s+\mu)^{-k-1}}{k!}\big(k!$$
$$- \Gamma(k+1, c(s+\mu))\big)\Big]. \quad (42)$$

C. Choice of s and n

In the case of the exponential service time, the steady-state queue size distribution can be computed from (4) using the well-known properties of the Laplace transform around $s = 0$. Namely, we have

$$P_m = \lim_{s \to 0+} sp_m(s,n), \quad (43)$$

for any n. Therefore, in numerical calculations we have to choose a proper, small value of s. Theoretically, the smaller s, the better. However, if s is very small, $p_m(s,n)$ and other intermediate results may become large, what may cause numerical errors, depending on the precision of numbers used in computations. We observed that in typical computations satisfactory results are obtained for $s = 10^{-10}$.

As for the choice of the initial queue size n (i.e. the element of vector (3)), theoretically any n is good. However, in typical cases the average queue size in the steady state is somewhere between 0 and b rather than close to 0 or b. Therefore, in order to exploit a quicker convergence rate to the steady state in (43), it is reasonable to choose e.g. $n = \frac{b}{2}$ rather than 0 or b.

D. Transform inversion

In the case of the exponential interarrival time, the Laplace transform given in (27) has to be inverted. When the buffer is large, this has to be done numerically. For this purpose, in [20] we checked several methods for numerical transform inversion, including the methods by Gaver and Stehfest [21], Talbot [22], Zakian [23], Weeks [24], Piessens [25], Dubner and Abate [26], Abate and Choudhury and Whitt [27]. We came to the conclusion that the Zakian method offers the best tradeoff between the computation time and accuracy of the results. This method is very fast, while in typical cases offering not worse accuracy than the most computationally demanding method [27] (see [28], [29] for examples of usage of the method).

If $H(t)$ is an original function and $h(s)$ is its Laplace transform, then the Zakian inversion formula has the form:

$$H(t) \approx \frac{2}{t}\sum_{k=0}^{n} \mathrm{Re}\big[\omega_k h\left(\frac{\alpha_k}{t}\right)\big]. \quad (44)$$

where n, α_k, ω_k are specially chosen parameters (α_k, ω_k are complex numbers). Commonly used values are $n = 4$ and the following values of α_k and ω_k.

k	α_k	ω_k
0	12.83767675 + i 1.666063445	-36902.08210 + i 196990.4257
1	12.22613209 + i 5.012718792	61277.02524 - i 95408.62551
2	10.93430308 + i 8.409673116	-28916.56288 + i 18169.18531
3	8.776434715 + i 11.92185389	4655.361138 - i 1.901528642
4	5.225453361 + i 15.72952905	-118.7414011 - i 141.3036911

E. Matrix inversion

In (4) we have to invert a matrix of size $(b+1) \times (b+1)$. In the case of a buffer for a few thousands of packets, this can be easily done, even using an average PC and the classic (Gaussian) method of matrix inversion, whose time complexity is of $O(b^3)$. For larger buffers, advanced methods of matrix inversion can be considered, for instance the Coppersmith & Winograd method, whose time complexity is of $O(b^{2.376})$.

VI. Examples

To perform sample numerical calculations, we assumed that the router output link of 1Gb/s capacity serves the arriving traffic of 1.1Gb/s bitrate. The buffer size was $b = 2000$ and the following dropping function was used:

$$d_n = \begin{cases} 0, & \text{if} \quad n \leq 500, \\ 0.0001201602n - 0.06008011, & \text{if} \quad 500 < n < 2000, \\ 1, & \text{if} \quad n \geq 2000. \end{cases}$$

As we can see, this is a linear dropping function of the RED type [1]. This particular form of the dropping function has been chosen so that the average queue size would be roughly 1250 packets.

The following six examples, with different types of traffic and service time distribution were computed. (All the distributions were scaled so that the average interarrival time is 1.)

1) The interarrival time is exponential, $\lambda = 1$, the service time is exponential, $\mu = 0.9090909$.

TABLE I
QUEUEING CHARACTERISTICS IN EXAMPLE 1)

	computation	simulation
mean queue size	1256.56	1256.83
std. dev. queue size	86.99	86.95
$\mathbb{P}(1000 \leq queue < 1200)$	0.252	0.252
$\mathbb{P}(1200 \leq queue < 1300)$	0.432	0.432
$\mathbb{P}(1300 \leq queue < 1500)$	0.311	0.311

2) The interarrival time is exponential, $\lambda = 1$, the service time is constant and equal to 1.1.
3) The interarrival time has Erlang distribution with parameters $K = 3$ and $\lambda = 3$, the service time is exponential, $\mu = 0.9090909$.
4) The interarrival time has Erlang distribution with parameters $K = 9$ and $\lambda = 9$, the service time is exponential, $\mu = 0.9090909$.
5) The interarrival time is constant and equal to 1 (CBR traffic), the service is exponential, $\mu = 0.9090909$.
6) The interarrival time is uniformly distributed on interval $(0, 2)$, the service is exponential, $\mu = 0.9090909$.

The computation results are presented in Tables I-VI. Each table contains also results of a long-run simulation (about 10^9 packets) of the queueing system, performed using OMNeT++ simulator, [30]. For computation we used an Intel Quad Core i7 (3GHz) machine with hyperthreading and an application written in Java 1.7 64bit. In most cases calculations were done using double precision numbers, with exceptions for Examples 6 and 5. These two examples caused range overflow, when double precision numbers were used. Therefore, in these examples the BigInteger and BigDecimal classes from standard Java math library and from Java JAS Project library [31] were used.

As we can see, in all the examples the computed queueing characteristics agree pretty well (though not perfectly) with the simulation results. This indicates that P_m's are overestimated a little for small m. To further illustrate this, in Figs. 1 and 2 we present distributions of the queue size in Example 2) and Example 6), respectively. In Fig. 1 the computed distribution agrees very well with the simulated one. However, in Fig. 2 the two distributions are apart for queue sizes below 1100. As both figures are plotted in logarithmic scale, the errors concern rather small probabilities and do not have a deep impact on the practical applicability of the computational method (in the worst case, they change the average queue size by 0.2 percent only). At this moment it is unclear what is the reason of these errors.

Finally, in Tab. VII the computation times of the considered examples are gathered. As we can see, they vary from a few minutes to a few hours, so they are not prohibitive for practical purposes. The significantly longer computation times of Examples 5 and 6 are connected with the usage of the BigInteger and BigDecimal classes during computation.

TABLE II
QUEUEING CHARACTERISTICS IN EXAMPLE 2)

	computation	simulation
mean queue size	1256.44	1257.37
std. dev. queue size	62.42	61.75
$\mathbb{P}(1000 \leq queue < 1200)$	0.177	0.175
$\mathbb{P}(1200 \leq queue < 1300)$	0.580	0.580
$\mathbb{P}(1300 \leq queue < 1500)$	0.243	0.245

TABLE III
QUEUEING CHARACTERISTICS IN EXAMPLE 3)

	computation	simulation
mean queue size	1256.87	1257.03
std. dev. queue size	72.61	72.74
$\mathbb{P}(1000 \leq queue < 1200)$	0.213	0.213
$\mathbb{P}(1200 \leq queue < 1300)$	0.506	0.505
$\mathbb{P}(1300 \leq queue < 1500)$	0.280	0.280

TABLE IV
QUEUEING CHARACTERISTICS IN EXAMPLE 4)

	computation	simulation
mean queue size	1256.97	1257.48
std. dev. queue size	67.15	67.35
$\mathbb{P}(1000 \leq queue < 1200)$	0.185	0.187
$\mathbb{P}(1200 \leq queue < 1300)$	0.539	0.540
$\mathbb{P}(1300 \leq queue < 1500)$	0.266	0.263

TABLE V
QUEUEING CHARACTERISTICS IN EXAMPLE 5)

	computation	simulation
mean queue size	1257.02	1257.62
std. dev. queue size	64.23	64.34
$\mathbb{P}(1000 \leq queue < 1200)$	0.183	0.181
$\mathbb{P}(1200 \leq queue < 1300)$	0.559	0.560
$\mathbb{P}(1300 \leq queue < 1500)$	0.257	0.259

TABLE VI
QUEUEING CHARACTERISTICS IN EXAMPLE 6)

	computation	simulation
mean queue size	1254.74	1257.84
std. dev. queue size	74.35	72.49
$\mathbb{P}(1000 \leq queue < 1200)$	0.223	0.212
$\mathbb{P}(1200 \leq queue < 1300)$	0.502	0.506
$\mathbb{P}(1300 \leq queue < 1500)$	0.273	0.281

TABLE VII
COMPUTATION TIMES OF NUMERICAL THE PRESENTED EXAMPLES

	computation time [s]
Example 1)	141
Example 2)	947
Example 3)	1732
Example 4)	1807
Example 5)	10740
Example 6)	6150

VII. CONCLUSIONS

We presented computational techniques that enable designing of queueing systems with dropping functions with buffers for thousands of packets. As demonstrated with a buffer for

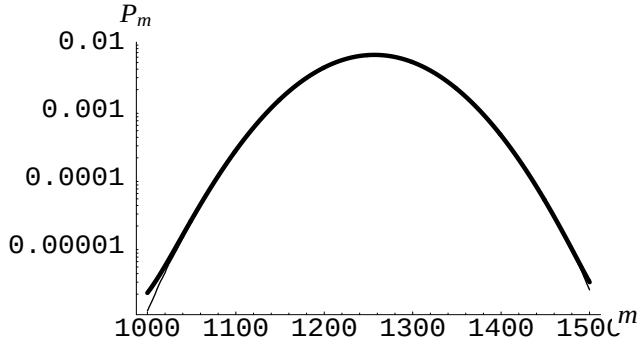


Fig. 1. Steady-state distribution of the queue size in Example 2). Thick curve - computation, thin curve - simulation.

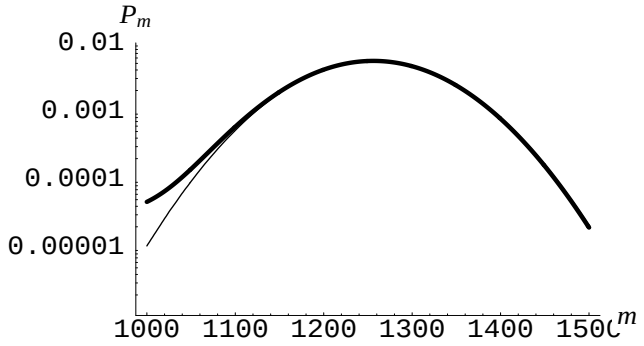


Fig. 2. Steady-state distribution of the queue size in Example 6). Thick curve - computation, thin curve - simulation.

2000 packets, the proposed techniques allow for obtaining results on an average PC within a few minutes to a few hours, with the accuracy good enough for engineering purposes.

Acknowledgement

This work was partially supported by the Polish National Science Centre under Grant No. N N516 479240.

REFERENCES

[1] S. Floyd and V. Jacobson. Random early detection gateways for congestion avoidance, IEEE/ACM Transactions on Networking, vol. 1, pp. 397–413, (1993).

[2] S. Floyd, R. Gummadi, S. Shenker et al. Adaptive RED: An algorithm for increasing the robustness of RED active queue management, Preprint, available at http://www.icir.org/floyd/papers.html, (2001).

[3] V. Rosolen, O. Bonaventure and G. Leduc. A RED discard strategy for ATM networks and its performance evaluation with TCP/IP traffic. SIGCOMM Comput. Commun. Rev. 29, 3, Jul. (1999).

[4] C. Wang, J. Liu, B. Li, K. Sohraby and T. Hou. LRED: A Robust and Responsive AQM Algorithm Using Packet Loss Ratio Measurement. IEEE Transactions on Parallel and Distributed Systems, v.18 n.1, pp. 29–43, (2007).

[5] B. Hariri, N. Sadati. NN-RED: An AQM mechanism based on neural networks. Electronics Letters. Volume: 43, Issue: 19, On page(s): 1053–1055, (2007).

[6] J. V. Chen, F.-C. Chen, J. M. Tarn and D. C. Yen. Improving network congestion: A RED-based fuzzy PID approach, Comp. Stand. Inter., vol. 34, no. 5, pp. 426-438, Sep. (2012).

[7] S. Athuraliya, S. H. Low, V. H. Li and Qinghe Yin. REM: active queue management, IEEE Network. On page(s): 48-53, Volume: 15, Issue: 3, May (2001).

[8] C. V. Hollot, V. Misra, D. Towsley and Weibo Gong. Analysis and design of controllers for AQM routers supporting TCP flows, IEEE Transactions on Automatic Control. On page(s): 945–959, Volume: 47, Issue: 6, Jun (2002).

[9] W. Wu, Y. Ren, and X. Shan. A self-configuring PI controller for active queue management. In Asia-Pacific Conference on Communications (APCC), Japan, (2001).

[10] W. Feng, K. G. Shin, D. D. Kandlur and D. Saha. The BLUE active queue management algorithms. IEEE/ACM Transactions on Networking. Page(s): 513 – 528, Volume: 10, Issue: 4, August (2002).

[11] B. Wydrowski and M. Zukerman. GREEN: an active queue management algorithm for a self managed Internet, in: Proceeding of IEEE International Conference on Communications ICC2002, vol. 4, pp. 2368-2372, April, (2002).

[12] S. S. Kunniyur and R. Srikant. An adaptive virtual queue (AVQ) algorithm for active queue management. IEEE/ACM Transactions on Networking. Page(s): 286–299, Volume: 12, Issue: 2, April (2004).

[13] Z. Na, Q. Guo, Z. Gao, J. Zhen and C. Wang C. A novel adaptive traffic prediction AQM algorithm Journal, Telecommunication Systems, Volume 49, Issue 1, pp 149-160, (2012).

[14] M. H. Suzer, K.-D. Kang and C. Basaran. Active queue management via event-driven feedback control, Computer Communications, vol. 35, no. 4, pp. 517-529, Feb. (2012).

[15] N. Farzaneh, R. Monsefi, M. H. Yaghmaee and A. H. Mohajerzadeh. A novel congestion control protocol with AQM support for IP-based networks, Telecommunication Systems, Volume 52, Issue 1, pp 229-244, (2013).

[16] G. Kahe, A. H. Jahangir and B. Ebrahimi. AQM controller design for TCP networks based on a new control strategy, Telecommunication Systems, DOI 10.1007/s11235-013-9859-y, August (2013).

[17] A. Chydzinski, L. Chrost, Analysis of AQM queues with queue-size based packet dropping, International Journal of Applied Mathematics and Computer Science 21:3, 567-577, (2011).

[18] W. Kempa. On non-stationary queue-size distribution in a finite-buffer queue controlled by a dropping function. Proc. International Conference on Informatics'13 (INFORMATICS 2013), Spisska Nova Ves, Slovakia, November 5-7, pp. 67–72, (2013).

[19] A. Dhamdhere, H. Jiang, and C. Dovrolis, Buffer sizing for congested internet links, INFOCOM 2005. 24th Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings IEEE, vol. 2, pp. 1072–1083 vol. 2, March (2005).

[20] P. Mrozowski, A. Chydzinski, Performance evaluation of routers with the dropping-function queueing, accepted for Computer Networks Conference, Brunow, June (2014).

[21] H. Stehfest, Numerical inversion of Laplace transforms. Communications of the ACM, 13:47–49, (1970).

[22] A. Talbot, The Accurate Numerical Inversion of Laplace Transforms, J. Inst. Maths. Applics. 23, 97–120, (1979).

[23] V. Zakian. Numerical inversion of Laplace transform. Electronic Letters 5:120–121, (1969).

[24] W. T. Weeks, Numerical Inversion of Laplace Transforms Using Laguerre Functions, JACM 13:419–426, (1966).

[25] R. Piessens, A New Numerical Method for the Inversion of the Laplace Transform, J. Inst. Maths. Applics. 10:185–192, (1972).

[26] H. Dubner, J. Abate, Numerical Inversion of Laplace Transforms by Relating them to the Finite Fourier Cosinus Transform, Journ. ACM 15, 115-123, (1968).

[27] J. Abate, G. L. Choudhury and W. Whitt. An introduction to numerical transform inversion and its application to probability models. Computational Probability, pp. 257–323, W. Grassman (ed.), Kluwer, Boston, (2000).

[28] A. Chydzinski, Transient Analysis of the MMPP/G/1/K Queue. Telecommunication Systems 32:4, pp. 247–262, (2006).

[29] A. Chydzinski, Duration of the buffer overflow period in a batch arrival queue. Performance Evaluation 63:4-5, pp. 493–508, (2006).

[30] http://www.omnetpp.org

[31] http://krum.rz.uni-mannheim.de/jas/

C E

International Conference on

COMPUTERS IN EDUCATION

Steering Committee

Chairs: Marina Čičin-Šain, University of Rijeka, Croatia
Jadranka Sunde, Defence Science and Technology Organisation, Salisbury, Australia
Jaak Henno, Tallinn University of Technology, Estonia
Hannu Jaakkola, Tampere University of Technology, Finland
Ivanka Sluganović, I. Tehnička škola Tesla, Zagreb, Croatia

Members: Michael E. Auer, Carinthia Tech Institute Villach, Austria
Ivan Gerlič, University of Ljubljana, Slovenia
Marija Marinović, University of Rijeka, Croatia
Slavomir Vukmirović, University of Rijeka, Croatia
Željka Požgaj, University of Zagreb, Croatia

INVITED PAPERS

Children as co-creators of video stories: mobile videos for learning

Jari Multisilta, Hannele Niemi
University of Helsinki, Cicero Learning, Helsinki, Finland
jari.multisilta@helsinki.fi, hannele.niemi@helsinki.fi

Abstract – In this paper we describe the MoViE mobile video storytelling platform, related Global Sharing Pedagogy framework and present examples of their use in learning and education. MoViE, the Mobile Video Experience platform, is a research platform for studying how learners create mobile video stories and how they share and learn with mobile social media. MoViE supports collaborative video creation, as people can use video clips of the same event made by others in their own mobile video stories because they are sharing the same experience. Currently, MoViE and GSP have been used in several international research projects, and there are over 1200 users and 1500 videos from 5 countries. In this paper, the research results and experiences from MoViE and GSP as tools for cross-cultural co-creation of learning related videos are presented.

I. INTRODUCTION

The use of videos on the internet has grown significantly in the last few years. For example, Khan Academy has a large collection of educational videos, especially on STEM subjects, available for free on the internet. It is also relatively easy to produce technically high-quality videos almost by anybody since the most new smartphones have high-quality media capture features. The growth of Internet video coupled with trends in social networking and participatory media, provide exciting new opportunities for video storytelling to become an integral component of distributed creative learning environments.

A growing number of education and research communities have started to use video in learning activities in new ways. According [1], the uses of "video are expanding from teachers simply showing videos to students to approaches where learners interact with, create, or comment on video resources as part of their knowledge-building activities."

The use of internet videos can be classified into three types, namely A) consuming videos from the internet, B) producing videos and uploading them to the internet, or C) remixing and enriching existing internet videos and their metadata.

Although the most popular videos in the internet are related to entertainment, web videos can have many educational uses. Khan Academy is an example of a type A web video service that has a large collection of educational videos, especially on STEM subjects. Khan Academy supports the traditional recorded lecture model. According to [2], "Khan Academy is a collection of video lectures that give demonstrations of mechanical processes." Much debate has taken place about the pedagogical model behind Khan Academy [2, 3, 4]. The main criticism is that Khan Academy is not supporting a constructivist model where learners actively create knowledge using activities that support knowledge construction. Our mobile video storytelling platform MoViE supports co-creation of mobile video stories and its use for learning can be designed and analyzed using the Global Sharing Pedagogy (GPS) model.

This paper reviews the use of video storytelling for learning. After the review the paper describes briefly the GSP model and the MoViE platform. Two examples of using GSP and MoViE will be presented. Finally, we discuss the lessons learned, and outline the need for the future research.

II. BACKGROUND

A. Video storytelling for learning

Instead of passive watching, the creation of video stories by the learners themselves seems to be a more efficient way of using video in learning. Mobile video storytelling is considered to be a new form of user participation and communication [5]. According to [5], "the ability to have constant access through mobile devices allows a new way of doing cinematographic narratives that can enhance the participants' experience in a significant way."

Isomursu et al. [6] summarized four ways that video clips can be used in a design process: the co-authoring of video materials, video as reflective material, the collaborative designing and editing of a video documentary, and collaborative sense-making. These four steps can also be applied to the educational uses of video storytelling.

Digital stories can be fictional or non-fictional. Non-fictional stories may originate from the author's own experience or document a specific event. Video stories can be scripted before shooting, or they can be ad-hoc stories. In many cases, digital video stories require editing or remixing. The life cycle of a web video is typically difficult to define. In many cases it may be impossible to control the video after it has been uploaded to a service. However, some of the recent internet video services support videos that can have a very short life. For example in Snapchat, the author of the video can set a life time of just a few seconds for a video. The video would be viewable only for that period, and afterwards be deleted permanently from the service.

Video stories can be interactive [7] or generated in real-time with scripting [8]. Storytelling platforms can also support automatic story creation [9, 10]. Multisilta and his research group created a mobile social media service for community created videos [9, 11]. MoViE has been used for example at Kasavuori Dream School in Kauniainen, Finland [12] and with university students at Tampere University of Technology, Finland [13].

Digital stories have also been used to preserve, reflect on, and share the life experiences of people who do not have access to personal computers and who are living in rural areas in South Africa [14]. Based on the classification in [6] it can be said that the video in this case is used as reflective material. Digital story creation may also be able to provide a means of exploring self-identity through the sharing and group construction of digital video stories [8].

Mobile video storytelling has also been used in user experience studies [6]. The users were testing a location-based mobile application, but collecting the data by observing users in real life contexts proved to be challenging. The solution was to recruit test persons in pairs, where one used the location application and the other recorded user experience material using their mobile phone camera. According to [6], by using mobile video, researchers received "richer emotional material and more versatile usage situations than with traditional observation methods". In this case the video were used as a way to collaborative design and edit of a video documentary.

The users' video clips were also used in the research interview as a starting point for a deeper discussion of the test situation. In this case, the video clips were used as a tool to document the situation. In addition, some test users not only documented the test situations, but "created, acted out and recorded mini dramas and mini comedies around use situations to make their point clear" [6]. Finally, the video stories were found to be an important tool in design and in communicating about design between stakeholders in the design process.

In designing digital storytelling systems, ethnographic studies have been used to understand the storytelling mechanisms and story structures [7, 14, 15].

B. Global Sharing Pedagogy

Niemi and Multisilta [16, 17, 18] have developed a Global Sharing Pedagogy (GSP) framework for designing and analyzing the learning activities and the learning process involving co-creation of video stories. The GSP is based on socio-cultural theories of learning. In GSP learning is seen as a result of dialogical interactions between people, substances, and artefacts [19, 20, 21, 22]. The primary objective is to strengthen student engagement in learning and mediate them to become active learners and knowledge creators. Engagement is seen as a motivational component that consists of the emotional states of students, such as joy and fun experienced in learning, as well as qualities that are typical in self-regulated learning. It includes commitment to learning tasks and a willingness to make efforts to achieve an objective [23, 24]. The GSP framework has four components, that are are called the mediators of learning: 1) learner-driven knowledge creation and skills development, 2) collaboration, 3) networking, and 4) digital media competencies and literacies. The GSP model has been applied to mobile video storytelling projects described below in Chapter III.

C. MoViE - Mobile Video Experience platform

MoViE (Mobile Video Experience) platform is used to create and share mobile videos. MoVIE enables users to record video clips using their mobile devices (phones, tablets, digital cameras, etc.), upload videos to the MoViE web site, and create video stories using all the clips they and their collaborators have taken [9, 25, 11, 12]. Existing video clips can also be remixed to new video stories. The content of the videos is not limited by the MoViE. Teachers and students can create videos that fit their needs and support learning both in and outside the classroom.

In the classroom, the teacher initiates the video project by providing the guidelines for using MoViE and agreeing the topic of the project with the learners. After that, the learners record short video clips of the topic and upload the clips to MoViE. They can record video clips at school or outside school on their free time. After uploading their video to MoViE, the users can add a description, tags and comments to the video. Next, the learners are asked to watch the videos others have produced and add comments to the videos. MoViE includes a commenting feature where the comments appear as a list after the video. Commenting videos and responding to the comments is an important phase where students give feedback, ask questions and answer questions others have asked. It is also possible to add links to other web pages or to other MoViE videos. Finally, the video clips, comments, and links represent the structure of the knowledge that the learners have collaboratively processed.

III. CASE EXAMPLES

A. Cross-cultural video sharing

In the Boundless Classroom project mobile video storytelling was used for experimenting with cross-cultural video sharing among schools in Finland, Greece and California in 2012-2013. The researchers requited several classes from these countries on voluntary basis. The researchers also facilitated the communication between the participating schools and teachers. They had the freedom to choose whatever topic they wanted for their mobile video storytelling project. In addition, they could do the project either by themselves or in collaboration with another class in another country. The teachers were provided training on using the MoViE platform in their teaching before the classroom projects started.

Students in Finland (n = 192), Greece (n = 111), and California (n = 16) created video stories and shared them with their peers locally and, in some cases, globally with their partner class in another country. The purpose was that each class would have an international partner class with whom they could share their videos and receive comments and feedback from their international peers. However, not all teachers and classes had the resources to

collaborate internationally; since working with videos was a new classroom activity to many teachers, they preferred to collaborate locally. The students who collaborated with a class in another country were excited and motivated to share their video stories with them.

The project took from one to six weeks depending on how it was integrated with other activities in schools. In the beginning of the classroom video projects, the students were brainstorming the topic of their video project and drafted a manuscript for their video. Often they had to seek additional knowledge on their topic from different sources. They recorded small captures on the theme of they story and uploaded them to the MoViE platform. A joint video repository was created for all group members.

Students produced their videos on a variety of themes, in some classes closely related to curriculum, in some classes more freely. The aim was that students would work actively, create knowledge collaboratively, learn to use digital devices and environments as tools for learning, and also become 'networked' while making connections with peers in other countries.

The main results from the Boundless classroom experiment indicate that students can become highly engaged through digital storytelling. The engagement seems to combine both an emotional aspect like fun and a commitment with hard work. During the project the students learned several 21st century skills when producing their own video stories. In addition, group work and co-creation were important for the motivation and enthusiasm. The MoViE platform itself supported both engagement and learning outcomes [16, 17, 18].

B. Video Inquiry Method in STEM

Video Inquiry Project [26] was a joint research project with Stanford University and University of Helsinki in 2013-2014. Learning "how to see" scientific and mathematical aspects of real world phenomena and raise real questions that foster engagement in inquiries is a key education problem for the STEM disciplines (Science, Technology, Engineering and Mathematics). The main research question in the VIP project was: what if learning across formal and informal environments could be bridged and enhanced by learners using mobile smartphones for capturing videos of STEM-rich worldly events and phenomena, raising questions about them that can anchor sense-making inquiries and discussions in classrooms and online forums?

The pilot phase was realized in fall 2013 and spring 2014 both in Finland and in California. In Finland, two 5th grade classes from an urban school and two classes from a rural area school participated to the pilot. In these classes there were around 80 learners. In the urban school learners used iPads that were offered by the school, whereas in the other school learners' own devices like tablets, smartphones, mobile phones, video cameras, or system cameras were used. An ability to shoot videos was the only technical requirement for those devices. In California there were two classes of students, total number of 59 sixth grade students (age 11–12). During the pilot phase students were learning to use the MoViE platform, get used to do video recordings with mobile devices and get familiar with online commenting. The topics of the videos were related to motion, air and water (Finland) and dissolving (California).

The study was based on both design research and intervention research in which the goal is to research how pupils' own mobile videos affect on learning and how they affect on interest/engagement in natural sciences.

The two pilot schools in Finland differed very much from each other. Pupils in the rural area school concentrated better on learning and seemed to be more interested about MoViE and learning itself. Interestingly, also the teachers in these two schools acted in the similar way: the teachers in the rural area school were more interested and involved in the project, whereas in the urban area schools the teachers usually withdrew from the activities during the editing lessons.

In California, both groups of the students were keen to use the tablets. Group A was more technologically intuitive regarding the use of the tablets. They were able to ask many technical questions and understand how tablets can be used in various situations. The students from Group B were intuitive toward the technology at the beginning, but more verbal instruction and explanation were needed in order to conduct the session. Toward the end of the seventh session, however, we found both groups of students were able to film the experiment, save the video, and upload it to the MoViE.

There were some technical problems that may have had an influence on the eagerness of the students to carry out the learning tasks involving video editing. The students both in the urban area school in Finland and in California expressed frustration because of the slowness of MoViE server while uploading and remixing their videos. Other problems included forgetting how to upload in certain situations, the length of the film, the sign-in passwords, and remembering how to search for their own video with the software.

Although the students were able to see the potential of using the videos and tablets as a way to develop their scientific exploration, the technical issues delayed their further engagement with the topic. Especially in California, due to the long uploading time, the students were not able to look at other videos or write comments during the classroom time — the inquiry process on the video was not able to take place. Despite the technical issues during the upload, we found that the students were able to start asking question as they conducted and filmed the experiment.

As a part of the design research, in Finland we also tested different teaching approaches for digital storytelling. Firstly, the task for the students was to remix together videos recorded during the lessons. They were also asked to add some comments to the videos others had uploaded to the service. However, because the students did not seem to progress enough, the researchers and the teachers came up special frame story for supporting the storytelling [26]. The purpose of the frame story was to give students an idea to follow in remixing their video clips. In the next phase at the rural area school the process of storytelling was supported by giving the students

detailed instructions on storyboarding and recording. This resulted in stories that resemble more natural learning process than in the first school. In these video stories learning natural sciences can be seen e.g. in the progress of videos, presenting of them, and also in the comments added to the videos.

IV. CONLUSION

Based on our studies, working with mobile videos both in a classroom and outside the classroom can be seen engaging. The engagement seems to combine both an emotional aspect like fun and a commitment with hard work. In addition, group work, co-creation and the MoViE platform were important factors for motivating students to learn and work with video stories.

The lousy user experience can however significantly lower the engagement. Students are used to very high quality commercial web services, and the same response time and quality is expected from the services the school is providing. Commenting adds an engaging feature to the learning process. It was also possible to add the engagement by creating a framework story to help students to create their videos. Despite the technical issues during the pilot projects, we found that the students were able to start asking question as they conducted and filmed the experiment.

Many videos (the outcomes of the pilot projects) had an important message aiming to influence other viewers' opinions (in other countries) in topics like recycling, water, homelessness, and animal testing. The student's own voice could clearly be seen in the videos. Based on our data, we think that the creation of mobile video stories and sharing support the learning of global awareness and 21st Century Skills, such as problem solving, collaboration, networking, and cultural understanding.

ACKNOWLEDGMENT

The author would like to thank Tekes – the Finnish Funding Agency for Innovation, Academy of Finland, and all the partners in the Finnable 2020 –project and the SAVI Innovations in Education –programme for supporting the research described in this paper. The author would also like to thank the Stanford University H-STAR Institute for an inspiring environment while being there as a Visiting Scholar.

REFERENCES

[1] Pea, R., & Lindgren, R. (2008). Video collaboratories for research and education: an analysis of collaboration design patterns. IEEE Transactions on Learning Technologies, 1(4), October-December, 2008.

[2] Talbert, R. (2012). The trouble with Khan Academy. The Chronicle. http://chronicle.com/blognetwork/castingoutnines/2012/07/03/the-trouble-with-khan-academy/.

[3] Thompson, C. (2011). How Khan Academy is changing the rules of education. Wired Digital, July 15.

[4] Prensky, M. (2011). Khan Academy. Available from http://www.marcprensky.com/writing/Prensky-Khan_Academy-EdTech-Jul-Aug2011.pdf

[5] Correia, N., Alves, N., Correia, H., Romero, L., Morgado, C., Soares, L., Cunha, J., Romão, T., Dias, A. E., Jorge, J. (2005) InStory: a system for mobile information access, storytelling, and gaming activities in physical spaces. ACE 2005, Valencia, Spain.

[6] Isomursu, M., Kuutti, K., Väinämö, S. (2004) Experience clip: method for user participation and evaluation of mobile concepts. Proceedings of the Participatory Design Conference 2004, Toronto, Canada. ACM.

[7] Ladeira, I., Marsden, G., and Green, L. (2011) Designing interactive storytelling: a virtual environment for personal experience narratives. In P. Campos et al. (Eds.): INTERACT 2011, Part II, LNCS 6947, pp. 430–437. IFIP International Federation for Information Processing.

[8] Vaucelle C., Davenport G. (2004). An open-ended tool to compose movies for cross-cultural digital storytelling: Textable Movie. Digital Culture & Heritage (ICHIM 04), 31.8 – 2.9, Berlin, Germany.

[9] Multisilta, J., Mäenpää, M. (2008) Mobile video stories. In Proceedings of the 3rd international Conference on Digital interactive Media in Entertainment and Arts (Athens, Greece, September 10 - 12, 2008). DIMEA '08, vol. 349. ACM, New York, NY, 401-406.

[10] Zsombori, V., Frantzis, M., Guimaraes, R. L., Ursu, M. F., Cesar, P., Kegel, I., Craigie, R., & Bulterman, D. C. A. (2011). Automatic generation of video narratives from shared UGC. Proceedings of the 22nd ACM conference on Hypertext and hypermedia (pp. 325–334). ACM New York, NY, USA

[11] Multisilta, J., Suominen, M., & Östman, S. (2012). A platform for mobile social media and video sharing. *International Journal of Arts and Technology*, Vol 5 No 1. pp. 53-72

[12] Tuomi, P. & Multisilta, J. (2012) Comparative study on use of mobile videos in elementary and middle school. *International Journal of Computer Information Systems and Industrial Management Applications*. Vol. 4. 255-266

[13] Kiili, K., Multisilta, J., Suominen, M., & Ketamo, H. (2010). Learning experiences of mobile social media. *Int. J. Mobile Learning and Organisation*, Vol. 4, No. 4, pp.346–359.

[14] Bidwell, N. J., Reitmaier, T., Marsden, G., & Hansen, S. (2010). Designing with mobile digital storytelling in rural africa. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 1593–1602). ACM New York, NY, USA.

[15] Engström, A., Esbjörnsson, M., & Juhlin, O. (2008). Mobile collaborative live video mixing. Proceedings of the 10th international conference on Human computer interaction with mobile devices and services (pp. 157–166). ACM New York, NY, USA.

[16] Niemi, H. & Multisilta, J (2014). Global is becoming everywhere – Global Sharing Pedagogy. In Niemi, H. & Multisilta, J. (eds.) Finnish innovations & technologies in schools: a guide towards new ecosystems of learning. Sense Publishers, Rotterdam: The Nederlands. (unpublished)

[17] Niemi, H. & Multisilta, J. (2014) Digital story telling promoting 21st century skills and students' engagement. Int J. Technology, Pedagogy and Education. (unpublished)

[18] Niemi, H. & Multisilta, J. (2014). Towards Global Sharing Pedagogy. In Niemi, H., Multisilta, J. & Löfgren, E. (eds.) Crossing boundaries for learning - through technology and human efforts. Cicero Learning, University of Helsinki, Helsinki:Finland. (unpublished)

[19] Pea, R. D. (2004). The social and technological dimensions of scaffolding and related theoretical concepts for learning, education, and human activity. *The Journal of the Learning Sciences, 13*(3), 423–451.

[20] Cole, M., & Cigagas, X. E. (2010). Culture and cognition. In M. H. Bornstein (Ed.), *Handbook of cultural developmental science* (pp. 127–142). New York, NY: Psychology Press.

[21] Säljö, R. (2012). Schooling and spaces for learning: Cultural dynamics and student participation and agency. In E. Hjörne, G. van der Aalsvoort, & G. Abreu (Eds.), *Learning, social interaction and diversity - Exploring school practices* (pp. 9–14). Rotterdam: Sense.

[22] Hakkarainen, K., Paavola, S., Kangas, K., & Seitamaa-Hakkarainen, P. (2013). Sociocultural perspectives on collaborative learning: Towards collaborative knowledge creation. In C. E. Hmelo-Silver, A. M. O'Donnell, C. Chan, C. A. Chinn (Eds.), *The international handbook of collaborative learning* (pp. 57–73). New York, NY: Routledge.

[23] Pintrich, P. R. (2000). The role of motivation in self-regulated learning. In P. R. Pintrich & P. Ruohotie (Eds.), *Conative constructions and self-regulated learning* (pp. 51–66). Hämeenlinna, Finland: RCVE.

[24] Pintrich, P. R., & McKeachie, W. J. (2000). A Framework for conceptualizing student motivation and self-regulated learning in the college classroom. In P. R. Pintrich & P. Ruohotie (Eds.), *Conative constructions and self-regulated learning* (pp. 31–50). Hämeenlinna, Finland: RCVE.

[25] Multisilta, J., Perttula, A., Suominen, M., & Koivisto, A. (2010). Mobile video sharing: Documentation tools for working communities. In Proceedings of the 8th International Interactive Conference on Interactive TV & Video (pp. 31–38). New York, NY: ACM.

[26] Multisilta, J., Kallunki, V., Ojalainen, J., Penttilä, J., Liu, A., Edouard, K. & Pea, E. (2014). Engagement in inquiry based learning with mobile videos. NFSUN 2014 Helsinki, Finland. (unpublished).

Children's safety on the Internet - development of the school curriculum

Lidija Kralj*
* Osnovna škola Veliki Bukovec, Veliki Bukovec, Croatia
lidija.kralj@skole.hr

Abstract - Generations that are currently receiving their education and those yet to come should get prepared for everyday interaction with information and communication technology. European Parliament and the Council of the European Union included digital competence in key competencies which each person needs to possess in order to adapt to the rapidly changing world. Their definition of digital competence, along with knowledge and skills, include critical attitude toward the responsible use of ICT. At the moment, Croatia does not have a national strategy for the internet safety and not even the minimum of children's knowledge about appropriate and safe use of internet is obligatory in school curriculum.

OŠ Veliki Bukovec together with partners OŠ Popovača, OŠ "Mladost", OŠ "Gripe" and OŠ "Mato Lovrak", is running a 16-month European Union-funded project " Children's safety on the Internet " developing new school curriculum area for children's safety on the Internet for students aged 7 -14, their parents, teachers and local community. Curriculum will consist of pedagogical-didactical model, acceptable use policies, multimedia resources, textbooks and guides. The project aims to improve students' digital competences and encourage children to assume responsibility for their own safety with a focus on empowerment, emphasizing responsible behavior and digital citizenship and to raise student teacher, parents and general public awareness and understanding of issues relating to children's safety online in synergy with the EU policies.

I. INTRODUCTION

Generations that are currently receiving their education and those yet to come should get prepared for everyday interaction with information and communication technology (ICT). This was recognized by the European Parliament and the Council of the European Union [1] which included digital competence in key competencies which each person needs to possess in order to adapt to the rapidly changing world. Their definition of digital competence, along with knowledge and skills, include critical attitude toward the responsible use of ICT.

Currently, students in compulsory schools in Croatia have a chance to obtain ICT related skills and knowledge only if they choose an elective subject - Informatics in grades 5 to 8, or if similar extra-curriculum activity is enabled for grades 1 to 4 [2]. Therefore not even the minimum of children's knowledge about appropriate and safe use of internet is obligatory at the time and none of Croatian schools has developed an acceptable use policy regarding the use of ICT and mobile devices in schools.

Education Sector Development Plan 2005-2010 [3] considers equipping people with the skills to use information and communication technology in everyday life as one of the prerequisites for active involvement in an information-based society. In National Curriculum Framework [4] the use of ICT is transversal theme the aim of which is to develop information literacy skills as part of competences for lifelong learning.

All children should reach minimum standard and gain experience and practices regarding risky and safe use of the Internet and new online technologies. Children should be prepared to use computers and the Internet, get advice appropriate to their age, with emphasis on visibility and public availability of their personal information on the Internet, various nasty, malicious and harmful effects which happen in online communication while encouraging creativity and lifelong learning through the exploration and learning through the Internet.

According to the latest UNICEF's survey in Croatia [5], 85% of children have Internet access at home, their favorite activities on the Internet are seeking entertainment, communication with friends, and use of social networking sites while using the internet for research and learning are less frequent. The same survey showed that 34% of children experienced some form of electronic violence and that usual teachers' reaction when faced with such cases is asking for advice from an expert. Therefore, teachers need further professional development in that area to prevent electronic violence and successfully deal with children's safety on the Internet.

EU Kids Online research "Risks and safety on the internet" encompassing 25 countries in 2011 [6] showed that the average age of first internet use is seven, that 93% of children aged 9 - 16 use internet daily or several times a week and that 55% of them have public profiles with personal information published on social networks. 12% of European children say that they have been bothered or upset by something on the internet. The research findings suggest that digital skills training needs continued emphasis and updating in terms of training, safety features and applications operation to ensure that all children reach a minimum basic standard and to prevent digitally isolated and unskilled children. All policy actors have the responsibility to ensure greater availability of age-

appropriate positive content for children, especially in small language communities such as Croatia.

EU Safer Internet programme recommend using programs that include children, parents and teachers to develop effective models of support for children using the internet for learning and research. Such model of targeted initiatives is especially important for countries in Eastern Europe, where there is a major difference between child and parental knowledge of the capabilities of the Internet.

Those demonstrate the importance of incorporating courses and learning objects about children's safety on the Internet into the school curriculum of compulsory education in Croatia.

II. PROJECT AIMS

Project Children's safety on the Internet is a 16-month European Union-funded project developed by OŠ Veliki Bukovec, together with partners OŠ Popovača, OŠ "Mladost", OŠ "Gripe" and OŠ "Mato Lovrak", Croatia, financed from grant scheme Further development and implementation of the Croatian Qualifications Framework (Europeaid/131254/ M/ACT/HR). The project is taking place in Croatia from 19th August 2013 till 19th December 2014. Project associates include Education and Teacher training agency (Agencija za odgoj i obrazovanje), Croatian academic and Research net (Hrvatska akademska i istraživačka mreža CARNet), Croatian personal data protection agency (Agencija za zaštitu osobnih podataka), Association "Partners in learning (udruga "Suradnici u učenju"), municipality of Veliki Bukovec (Općina Veliki Bukovec), municipality of Mali Bukovec (Općina Mali Bukovec), Brodsko-posavska county, City and Univesity library Osijek, City of Nova Gradiška and Tourism office Nova Gradiška.

The project has been designed after consultations with the Education and Teacher Training Agency, public institution responsible for teachers' training, CARNet and the NGO Suradnici u učenju as Safer Internet Committee for Croatia. Partner schools get support from their Counties who consider this project to be a great potential for county development. All consulted organizations highly recommend this project and will be involved in project realization and dissemination.

The project objectives are:

- to develop, according to principles of the Croatian Qualifications Framework (CROQF), learning outcome-based school curriculum area for children's safety on the Internet which will enable holistic approach involving students, parents and teachers on the same goal;
- to develop and implement appropriate pedagogical and didactical model for student centered learning which will make the most of teachers' potentials and new technology strength;
- to improve primary school teachers' educational skills and expertise so they could apply the new methodologies for student centered learning and principles of CROQF in their work;
- to improve students' digital competences and encourage children to assume responsibility for their own safety as much as possible with focus on empowerment, emphasizing responsible behavior and digital citizenship;
- to raise students, teachers, parents and general public awareness and understanding of issues relating to the children's safety online in synergy with the EU policies.

Project facilitates cooperation and experience exchange as well as sharing best practice between Croatian and EU schools on issues relating to children's safety online and thus ensure European added value.

Furthermore, project promotes gender equality and equal opportunities through carefully chosen characters and their roles in all forms of learning materials. Digital materials have options for modification for use by disabled people. Project also gives equal opportunities to all students to reach standard learning outcomes in area of safe, legal and ethical behavior on the Internet and to prevent digitally isolated and unskilled children.

III. PROJECT ACTIVITIES

The project is going to last for 16 months with activities organized in four phases during which the school capacity for implementation of student centered learning will be improved and modernized.

First 6 months were the Research phase - pedagogical team of teachers individually and during 2 study visits (Belgium and United Kingdom) explored the existing school curriculum and examples of good practice in the area of online child safety in EU. First study visit was Insafe and European Schoolnet, which works in the area of Internet safety and has continuous experience in professional development for teachers. Second study visit was to the United Kingdom, to schools which have perennial experience in development of school curriculum and implementation of topics about e-safety in teaching ICT and other subjects and acceptable use policies implemented.

The pedagogical team was trained in 2 workshops (Plitvička jezera and Đurđevac) related to e-safety, EU policies and strategies, curriculum development, determining learning outcomes and learner-centered educational methods and then prepared a report and planned strategies for curriculum development.

Next 6 months is the Development phase - all teachers are trained in 5 workshops where colleagues who have attended study visits and workshops will share their experience and knowledge (cascade training). Pedagogical team will work on development and publishing of school curriculum area (compound of 4 courses & sets of materials for students, teachers and parents) for children's safety on the Internet with specified learning outcomes according to the principles of CROQ. During this phase all partner schools are equipped with mobile learning environment with modern tablets, interactive board and wireless access to internet.

Two months of Testing and improvement phase follows - developed school curriculum area will be tested, evaluated and improved according to students, parents and teachers feedbacks. Research regarding students digital competences and parental knowledge of the capabilities of the Internet will be carried out as part of the pilot project. New versions of curriculum and learning object will be published and made freely available.

Last 2 months will be the Dissemination phase - devoted entirely to dissemination of project results, reporting and evaluation. Final conference and exhibition of students' works resulting from the pilot project will be held.

IV. CURRICULUM

School curriculum will consist of four courses addressing one of four age groups in primary schools (students aged 7-8, 9-10, 11-12 and 13-14). That level of curriculum structure is chosen to enable the creation of age appropriate contents and methods in balance with available resources for this project.

One set of teaching/learning materials would include textbooks for students, multimedia resources, teachers' guides and guides for parents. Learning objects within teaching/learning materials will be created in form of text, hypertext, pictures, animated stories, videos, audios, computer games, social games, coloring pages, photos, interactive quizzes, learning quests. Colorful characters have been created and will be used in all learning resources so curriculum will have unique and recognizable multimedia representations. Special attention will be paid to the flexibility of learning objects so students with disabilities can use them more easily. Every course will be described in terms of learning outcomes, as well as knowledge, skills and competences. That description will include theoretical knowledge, practical and technical skills and social competences needed for responsible, safe and appropriate use of the Internet. Evaluation methods, associating self-assessment, feedback mechanisms and procedures for improvement will be planned in each course. Acceptable Use Policy for ICT in schools will be created.

School curriculum area for children's safety on the Internet is vertically adjusted in five units: information, communication, content creation, safety and problem solving as recommended in EU Framework for Developing and Understanding Digital Competence in Europe [7]. Some of the topics covered in curriculum are; personal data protection, e-mail phishing and scams, netiquette, online communication and collaboration, risks on social networks, responsible use of mobile devices, sharing and authoring rights, identity theft, digital footprints, e-portfolios and online presence, evaluation of information on the Internet, how to protect computer and family, prevention of cyberbullying.

School curriculum would be tested and evaluated by two sides - schools which take part in project and external advisors. To be sure that school curriculum area is well planned and all resources well created, pilot testing will be performed with 220 students and 220 parents - 2 classes from each participating schools.

After feedback and recommendations, school curriculum would be adjusted and prepared for publication and further implementation

V. IMPACT

The proposed action is directly targeting the following target groups: teachers, students and parents.

Teachers in Croatian primary schools have the basic level of ICT competencies but usually poor knowledge about capabilities of the Internet, possible risks, safety measures and prevention of electronic violence. They have only general information about CROQF, learning outcomes and student-oriented approaches. During the project 15 teachers attended 2 study visits, 25 teachers participated in 2 five-day workshops, and 218 teachers (all teachers and principals from partner schools) will be trained in the area of e-safety. The project will be disseminated among approximately 2500 teachers participating in 50 County Teacher Councils in different subjects in 5 counties where partner schools are situated.

Today, students grow up surrounded by a variety of digital and electronic devices and online tools. They often misuse those devices and embrace potentially risky behavior on the Internet. Students need guidance in how to behave in a safe, legal and ethical way and protect themselves on the Internet. 220 students aged 7-14 will be involved in a pilot project of implementation of developed school curriculum area. During the pilot testing participating students will create at least one practical work in which they present the knowledge they have gained, and give advice to peers about responsible, appropriate and safe use of Internet. Students' works will be showcased on the project web and at the final conference as an exhibition of works resulting from the implementation of new curriculum. Approximately 2700 students will be involved in awareness raising campaign and school curriculum implementation during this and next school year.

Parents often have poor knowledge about opportunities and risks related to the use of the Internet and children usually know more than parents. Therefore parents' awareness about the internet safety issues have to be raised and they need to be educated in order to support their children. 220 parents of the students involved in the pilot project will also take part in the project and give feedback. Approximately 2700 parents will be reached after the project results are implemented in all classes of the partner schools.

Local community and general public will be positively influenced through raising awareness about internet safety. To accompany school curriculum and establish stronger connection between educational institutions and local community, awareness raising campaign was created. Awareness raising campaign includes workshops, exhibitions, lectures, interactive round tables, articles, radio and TV shows targeting on local, regional and national level. Project publicity and dissemination

activities will be synchronized with that campaign so every event will serve both causes - raising awareness and project dissemination.

VI. CONCLUSION

At the moment, Croatia does not have a national strategy towards the internet safety so our project could be used as a reference or case study for future prevention programme or national strategy and all project results may easily be available for future use on the national level.

School curriculum with four courses and all associated teaching/learning materials will be published in digital form online and made freely available on project web page **petzanet.HR** for all interested schools.

The long-term project objective is to reach every single primary school in Croatia through different national seminars and online publication. Awareness raising campaign will set model to help other schools in establishing connection between school and local community on common goal of protecting children on the Internet.

REFERENCES

[1] Recommendation of the European Parliament and of the Council on key competences for lifelong learning, 2006.

[2] "Nastavni plan i program za osnovnu školu," Ministarstvo znanosti, obrazovanja i športa, Zagreb 2006.

[3] "Education Sector Development Plan 2005-2010",Ministry of Science, Education and Sport, Zagreb, 2005.

[4] "Nacionalni okvirni kurikulum", Ministarstvo znanosti, obrazovanja i športa, Zagreb 2011.

[5] J. Pregrad, et. al."Iskustva i stavovi djece, roditelja i učitelja prema elektroničkim medijima", Ured UNICEF-a za Hrvatsku, 2011.

[6] S. Livingstone, L. Haddon, A. Görzig. and K. Ólafsson, "Risks and safety on the internet: The perspective of European children", Full Findings. LSE, London: EU Kids Online, 2011.

[7] A. Ferrari, "DIGCOMP: A Framework for Developing and Understanding Digital Competence in Europe", Luxembourg: Publications Office of the European Union, 2013.

PAPERS

Introducing e-CRM into academic libraries: exploration of needs and possibilities

Anita Papić* and Marija Primorac**
* Faculty of humanities and social sciences/Department of information sciences, Osijek, Croatia
** Faculty of humanities and social sciences/Department of information sciences, Osijek, Croatia
apapic@ffos.hr, mprimorac@ffos.hr

Abstract - This paper gives theoretical insight into importance of introducing customer relationship management (CRM) into libraries. CRM is defined as system or software for storage or analysis of data, but also as strategy oriented to customer and change of organizational culture. One of reasons for CRM implementation is that companies sometimes more care for their products and services than for customers. E-CRM can help libraries to sophisticate their e-services which is an imperative for libraries as leading educational institutions. Namely, libraries have mission to gather, process, store and make access to information and that's why they must make their e-services better to attract new users and to keep existing users. Working hypothesis in this paper is that level of readiness of Croatian academic libraries for e-CRM is low. According to research results this hypothesis is confirmed, namely discrepancy between possibilities and needs for introducing e-CRM into academic libraries is rather high.

I. INTRODUCTION

Libraries in 21. century are characterized by constant changes due to development of information and communication technology. Expanding libraries' e - services and their sophistication becomes an imperative in libraries and a way how libraries can stay leading educational institutions. The concept of customer relationship management (CRM) is exactly that business system or tool which can help libraries to keep existing users and attract new users.

CRM system is very important option in libraries. Namely, libraries as educational institutions have mission to gather, process, store and make access to information and that's why they must make their services better and attract new customers as well as keep existing users. CRM implementation can contribute to development of better libraries' e-services according to many studies as for example is research conducted in academic library in Taiwan [10]. Namely, the main advantage of CRM implementation in libraries can be seen in their efficiency [6].

Customer relationship management (CRM) is business strategy which includes selection and customer relationship management with a goal to optimize their long term loyalty. The term CRM is used from early 60-ies. Namely, the acronym CRM was defined as customer relationship management but also customer relationship marketing [3]. Theodore Levitt, an expert in marketing, stated that the main goal in business is to recruit and keep the customer. He interceded for fact that companies should see the whole business process which consists of discovering, creating, challenging and fulfilling customer needs. In addition to that, recent marketing studies have shown that CRM software implementation would be increased for 700% from that time [4]. CRM can be observed as continuation of so called precise marketing which implies art of merging product or service with customer's need to provide safe sale. CRM also implies use of contact centers for cohesive and unique view of customer but also driver of investments in data warehouse systems. CRM is defined as technology, system or software package for storage of data or analysis of data, but also as strategy oriented to customer and change of organizational culture [1]. One of reasons for this kind of strategy is that companies sometimes more care for their products and services than for customers. It should be accentuated that CRM has its roots in relationship marketing which goal was to conquer new customers. Namely, it differs from today's marketing which tends to keep existing customers [5]. The birth of CRM was strongly influenced by high concurrence among companies in all fields. That's why CRM motto could be "know your customer and you will know what he/she buys" [9].

CRM systems are extremely growing field of investments into IT solutions. CRM systems differ in a function which system covers and which solution would choose a particular company depends upon its needs and business strategy. The fact is that implementation of typical CRM lasts minimally two years and that in 60 - 70% cases completes with failure [9]. The main reason for CRM system failure lies in a fact that companies' top management don't understand the meaning of CRM. Lots of researches shown that a lot of people think that CRM software manages customer relationships instead that it is an instrument which helps in customer relationship management [9]. Namely, whole CRM solution consists of three parts: analytical, collaborative and operative. Operative and analytical parts are fundamental elements of system. Operative system is in charge for everyday communication with customers and analytical system, the brain of whole CRM system, is in charge for analysis of data gathered from operative part and for generating strategies and adequate decisions based on results of analysis. On the contrary, collaborative system is based on external organization namely customer sees that system in communication with company. Collaborative

system encompasses all CRM functions that provide interactions with customers, partners and suppliers. In this way operative data are generated and these data are the base for analytical part of CRM which gathers, stores, processes, extracts, interprets and informs about customers. That's why analytical part can be seen as business intelligence of CRM system unlike operative part which gives options for data exchange and integration with information systems [9]. CRM implementation is a very demanding process. An imperative for successful CRM implementation lies in clear communication between CRM strategy at all organizational levels and employers involvement into its implementation. Some problems which occur before CRM implementation are: identification of customers' needs, understanding customers, giving feedback etc. [8]. CRM system does not provide quick success, namely long period of time is necessary for CRM implementation and to make customer happy which is a safe way for keeping customer's loyalty.

For the purpose of conducting content analysis of Croatian academic libraries' web sites in regards to exploring needs and possibilities for introducing e-CRM the working hypothesis is stated. Namely, working hypothesis in this paper is that level of readiness of Croatian academic libraries for e-CRM implementation is low.

II. METHOD

For the purpose of conducting research in this paper scientific method of content analysis was selected. Content analysis is flexible research method and it can be qualitative or quantitative but also within it can be used combination of these two methods. Namely, in this paper was used quantitative content analysis of text within Croatian academic libraries' web sites according the list in appendix.

Content analysis is research method which objectively and systematically describes written, verbal or visual communication [2]. Krippendorff describes content analysis as research method for derivation of repeatable and valid conclusions from text within its context [7]. Content analysis helps to determine presence of certain words, phrases or concepts within text or collections of texts. Texts can be from books, speeches, interviews, essays, historical documents, scientific articles, pictures, web pages etc.

Content analysis can be implemented to any written or other form of human communication and that's why in recent decade it is often used in marketing, cultural studies, journalism, sociology, politics, phycology, cognitive science, economy, information sciences etc.

III. INTERPRETATION OF RESULTS

Croatian academic libraries' web sites according to the list in appendix were analysed in purpose to explore needs and possibilities for introducing e-CRM. E-options considered as preconditions for introducing e-CRM into academic libraries were grouped into following eight categories obtained by the method of content analysis: (1) communication with users, (2) design of user-friendly system, (3) information about library and public relationships, (4) affective engineering, (5) promoting content of library, (6) expanding and enhancing e-services, (7) superior service quality and (8) contacts with potential future users.

A. Communication with users

Fig. 1 shows that majority of libraries have e-mail (73) and newsletters (50) but only few of them have surveys options (8) and webinars (5).

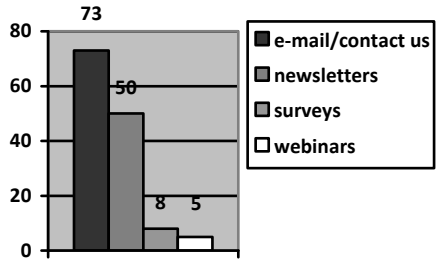


Fig. 1. Communication with users

B. Design of user-friendly system

The majority of libraries (50) have drop-down menus as part of design of user-friendly interface and almost half of libraries take care about presenting legal matters on their web sites (34) as it is shown at Fig. 2. Instructions for use of web site (9) and site map (13) are rare options. Only four libraries have an option FAQ (frequently asked questions).

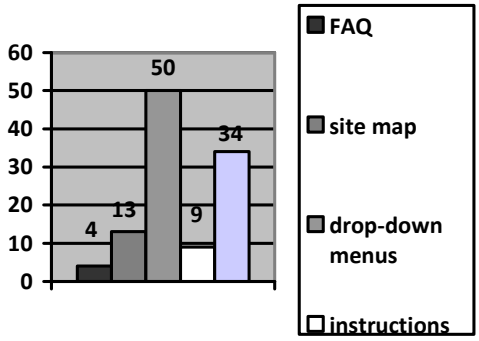


Fig. 2. Design of user-friendly system

C. Information about library and public relationships

Almost all libraries' web sites consist of information about library (65), library location (60), history of library (55) and annual reports give 21 libraries shown on Fig. 3.

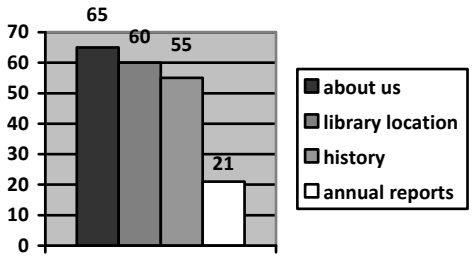


Fig. 3. Information about library and public relationships

D. Affective engineering

In regards to libraries' affective engineering, Fig. 4 shows that 46 libraries offer useful textual materials and materials for download (20). Charitable activities (6) and ideas for products' use (3) like for example *Bookcrossing* are very rare.

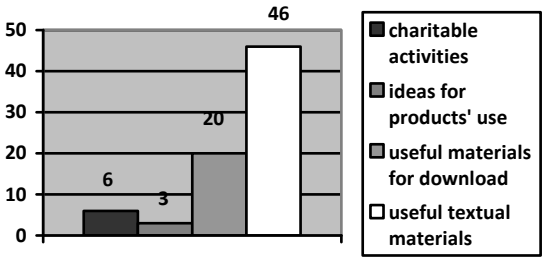


Fig. 4. Affective engineering

E. Promoting content of library

Fig. 5 shows that 64 libraries promote their content through catalogs and 26 libraries points out prices of their services.

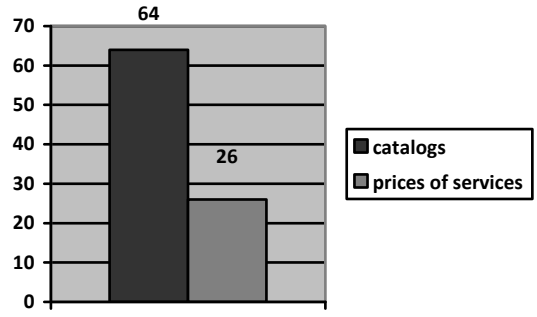


Fig. 5. Promoting content of library

F. Expanding and enhancing e-services

Fig. 6 shows some of ways through which libraries try to expand and enhance their e-services like for example 63 libraries offer educational material dealing different databases, online dictionaries etc. Great number of libraries offer information about e-services (45), innovative and exclusive e-services (30), users' forum (25) and few of them e-entertainment (8).

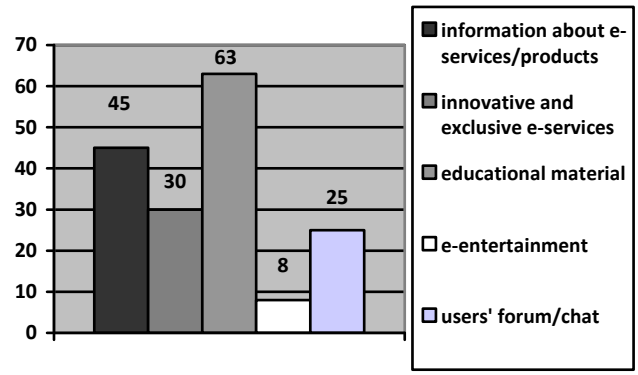


Fig. 6. Expanding and enhancing e-services

G. Superior service quality

Very small number of libraries have attractive visual design of users' interface (19), remind users about new content/services (18), seek for feedback (13) and conduct e-surveys about users' satisfaction (3) which is shown on Fig. 7.

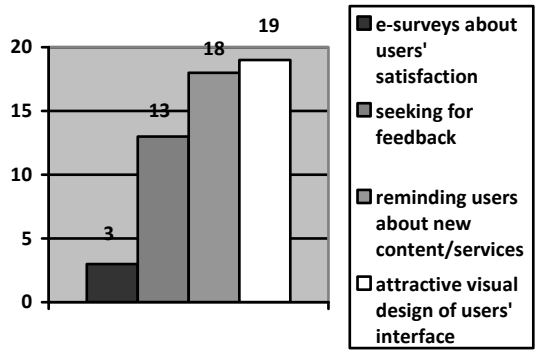


Fig. 7. Superior service quality

H. Contacts with potential future users

Fig. 8 shows that only 10 libraries try to contact with potential future users through noncommercial e-content or access to users outside the traditional library. Five libraries try to recruit users with new needs.

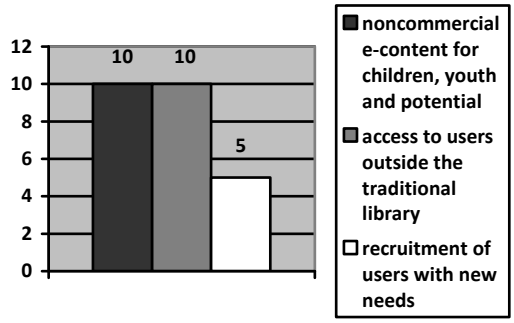


Fig. 8. Contacts with potential future users

IV. CONCLUSION

CRM is defined as system or software for storage or analysis of data, but also as strategy oriented to customer and change of organizational culture. One of reasons for CRM implementation is that companies sometimes more care for their products and services than for customers. E-CRM can help libraries to sophisticate their e-services which is an imperative for libraries as leading educational institutions. Namely, libraries have mission to gather, process, store and make access to information and that's why they must make their e-services better to attract new users and to keep existing users. Working hypothesis in this paper was that level of readiness of Croatian academic libraries for e-CRM is low.

Croatian academic libraries' web sites according to the list in appendix were analyzed in purpose to explore needs and possibilities for introducing e-CRM. E-options considered as preconditions for introducing e-CRM into academic libraries were grouped into following eight categories obtained by the method of content analysis: (1) communication with users, (2) design of user-friendly system, (3) information about library and public relationships, (4) affective engineering, (5) promoting content of library, (6) expanding and enhancing e-services, (7) superior service quality and (8) contacts with potential future users.

According to research results the hypothesis is confirmed, namely discrepancy between possibilities and needs for introducing e-CRM into academic libraries is rather high.

REFERENCES

[1] R. Baran, R. Galka and D. Strunk, Customer Relationship Management. USA: Thomson Learning, 2008.

[2] B. Barelson, Content analysis in communication research. New York : Free Press, 1952.

[3] F. Buttle, Customer Relationship Management: concepts and technologies. Hungary: Elsevier Ltd, 2009.

[4] S. Das, Customer Relationship Management. New Delhi: Excel Books, 2007.

[5] B. Dukić and M. Meler, Upravljanje odnosima od potrošača do klijenta (CRM). Osijek: Ekonomski fakultet, 2007.

[6] R. Jamali, [et al.],"Customer relationship management in electronic environment," Electronic Library, vol. 31, pp. 119-130, 2013. http://www.emeraldinsight.com/journals.htm?articleid=17077496 (2013-12-12)

[7] K. Krippendorff, Content analysis: an introduction to its methodology. Thousand Oaks : London : New Delhi: Sage Publications, 2004.

[8] S. R. Kumar, Customer Relationship Management. New Delhi : New Age International, 2003.

[9] G. Raab, [et al.], Customer Relationship Management: a global perspective. England: Gower, 2008.

[10] M. Y. Wang, "Introducing CRM into an academic library," Library Managment, vol. 22, pp. 281-291, 2007. http://www.emeraldinsight.com/journals.htm?articleid=1617026&show=abstract (2013-12-12)

Appendix. List of Croatian academic libraries

1.	Nacionalna i sveučilišna knjižnica Zagreb	http://www.nsk.hr/
2.	Sveučilišna knjižnica Pula	http://www.skpu.hr
3.	Sveučilišna knjižnica Rijeka	http://www.svkri.uniri.hr
4.	Sveučilišna knjižnica Dubrovnik - knjižnica za tehničke i biotehničke znanosti	http://www.unidu.hr/odjeli.php?idizbornik=40
5.	Sveučilišna knjižnica Dubrovnik - knjižnica Odjela za ekonomiju i poslovnu ekonomiju	http://www.unidu.hr/odjeli.php?idizbornik=41
6.	Agronomski fakultet Zagreb - Središnja agronomska knjižnica	http://www.agr.unizg.hr/cro/ustrojstvo/cak/
7.	Akademija dramskih umjetnosti Zagreb	http://www.adu.hr/
8.	American College of Management and Technology, Dubrovnik	http://www.acmt.hr/
9.	Ekonomski fakultet, Rijeka	http://www.efri.uniri.hr
10.	Ekonomski fakultet, Zagreb	http://www.efzg.hr
11.	Fakultet ekonomije i turizma Dr. Mijo Mirković, Pula	http://oet.unipu.hr/
12.	Fakultet elektrotehnike i računarstva, Zagreb	http://www.fer.hr/knjiznica
13.	Fakultet kemijskog inženjerstva i tehnologije, Zagreb	http://knjiznice.szi.hr/?libid=62&task=nsl
14.	Fakultet organizacije i informatike, Varaždin	http://www.foi.unizg.hr/

15.	Fakultet političkih znanosti, Zagreb	http://www.fakultet.fpzg.hr/
16.	Fakultet strojarstva i brodogradnje, Zagreb	http://www.fsb.hr/library/
17.	Fakultet za turistički i hotelski menadžment, Opatija	http://www.fthm.uniri.hr
18.	Farmaceutsko-biokemijski fakultet, Zagreb	http://www.pharma.hr
19.	Filozofski fakultet Družbe Isusove - knjižnica "Juraj Habdelić", Zagreb	http://www.ffdi.unizg.hr/knjiznica
20.	Filozofski fakultet, Rijeka	http://www.ffri.uniri.hr
21.	Filozofski fakultet, Zagreb	http://knjiznica.ffzg.hr/
22.	Geotehnički fakultet, Varaždin	http://www.gfv.hr/hr/knjiznica.html
23.	Građevinski fakultet, Zagreb	http://info.grad.hr/gf/index.asp?pid=1530
24.	Grafički fakultet, Zagreb	http://knjiznica.grf.hr/
25.	Hrvatski studiji, Zagreb	http://www.hrstud.hr/
26.	Katolički bogoslovni fakultet, Zagreb	http://www.kbf.hr/stranica.aspx?pageID=58
27.	"Andrija Štampar" - Medicinski fakultet, Škola narodnog zdravlja, Zagreb	http://knjiznice.szi.hr/index.php?libid=10&task=nsl
28.	Medicinski fakultet, Zagreb - Središnja medicinska knjižnica	http://smk.mef.hr/
29.	Medicinski fakultet, Rijeka - Središnja medicinska knjižnica	http://www.medri.uniri.hr/knjiznica/
30.	Metalurški fakultet, Sisak	http://www.simet.hr
31.	Pomorski fakultet, Rijeka	http://www.pfri.uniri.hr/knjiznica/
32.	Pravni fakultet, Rijeka	http://www.pravri.hr/knjiznica/index.html
33.	Pravni fakultet, Zagreb	http://www.pravo.hr/biblioteka
34.	Prehrambeno-biotehnološki fakultet, Zagreb	http://www.pbf.hr/hr/sluzbe/knjiznica
35.	Prirodoslovno-matematički fakultet, Zagreb	http://www.pmf.unizg.hr/
36.	Središnja geografska knjižnica Zagreb	http://www.geog.pmf.hr
37.	Središnja kemijska knjižnica Zagreb	http://www.pmf.unizg.hr/chem/skk
38.	Središnja biološka knjižnica Zagreb	http://hirc.botanic.hr/sbk/sbk-home.html
39.	Središnja geofizička knjižnica Zagreb	http://www.gfz.hr/
40.	Središnja geološka knjižnica Zagreb	http://knjiznice.szi.hr/?libid=6&task=nsl
41.	Središnja knjižnica za fiziku Zagreb	http://www.knjiznica.phy.hr/
42.	Središnja matematička knjižnica Zagreb	http://www.math.hr/Default.aspx?sec=329
43.	Rudarsko-geološko-naftni fakultet, Zagreb	http://www.rgn.hr/Default.aspx?UCID=243
44.	Stomatološki fakultet, Zagreb	http://www.sfzg.unizg.hr/
45.	Strojarski fakultet, Slavonski Brod	http://www.sfsb.hr/nastava/knjiznica/
46.	Šumarski fakultet, Zagreb	http://www.sumfak.hr/

47.	Tehnički fakultet, Rijeka	http://www.riteh.uniri.hr/zav_katd_sluz/knjiznica/index.html
48.	Tekstilno-tehnološki fakultet, Zagreb	http://www.ttf.hr/index.php?str=11
49.	Veterinarski fakultet, Zagreb	http://www.vef.hr/lib/?adnce=j&m=ff00f7c1f85b
50.	Učiteljski fakultet, Čakovec	http://www.ufzg.unizg.hr/cakovec/
51.	Učiteljski fakultet, Gospić	http://www.unizd.hr/nstgospic/Home/tabid/3432/Default.aspx
52.	Ekonomski fakultet Osijek	http://www.efos.hr
53.	Elektrotehnički fakultet Osijek	http://www.etfos.hr/?fakultet&Sluzbe&Knjiznica
54.	Filozofski fakultet Osijek	http://web.ffos.hr/knjiznica/
55.	Građevinski fakultet Osijek	http://knjiznice.szi.hr/?libid=46&task=nsl
56.	Katolički bogoslovni fakultet Đakovo	http://www.djkbf.unios.hr/
57.	Medicinski fakultet Osijek	http://knjiznice.szi.hr/?libid=4&task=nsl
58.	Poljoprivredni fakultet Osijek	http://knjiznice.szi.hr/?libid=28
59.	Pravni fakultet Osijek	http://zakon.pravos.hr/knjiznica
60.	Prehrambeno-tehnološki fakultet Osijek	http://www.ptfos.hr/
61.	Strojarski fakultet Slavonski Brod	http://knjiznice.szi.hr/?libid=47&task=nsl
62.	Odjel za matematiku Osijek	http://www.mathos.hr/hr/knjiznica
63.	Umjetnička akademija u Osijeku	URL:http://www.uaos.hr/
64.	Učiteljski fakultet u Osijeku	http://www.ufos.unios.hr/modules/news/
65.	Gradska i sveučilišna knjižnica Osijek	http://knjiga.gskos.hr/
66.	Veleučilište Lavoslav Ružička u Vukovaru	http://www.vevu.hr/
67.	Veleučilište u Požegi	http://www.vup.hr/
68.	Sveučilište u Zadru – Sveučilišna knjižnica	http://www.unizd.hr/Default.aspx?alias=www.unizd.hr/knjiznica
69.	Ekonomski fakultet Split	http://www.efst.hr/content.php?k=biblioteka&p=49
70.	Fakultet elektrotehnike, strojarstva i brodogradnje Split	http://www.fesb.hr/
71.	Fakultet prirodoslovno-matematičkih znanosti i odgojnih područja Split	http://www.pmfst.hr/knjiznica/index_pmfst.html
72.	Filozofski fakultet Split	http://www.ffst.hr/index.php?option=com_content&view=category&layout=blog&id=153&Itemid=325&lang=hr
73.	Građevinsko-arhitektonski fakultet Split	http://www.gradst.hr/Knjiznica/Novosti/tabid/1217/Default.aspx
74.	Kemijsko-tehnološki fakultet Split	http://www.ktf-split.hr/bib/index.html
75.	Medicinski fakultet Split	http://knjiznice.szi.hr/index.php?libid=33&task=nsl
76.	Pomorski fakultet Split	http://www.pfst.hr/old/?a=knjiznica
77.	Pravni fakultet Split	http://www.pravst.hr/fakultet.php?p=0&s=107

Virtual Machines in Education

Vladimir Mateljan, Vedran Juričić, Marko Moguljak
Faculty of humanities and social sciences, Zagreb, Croatia
Department of information and communication sciences
vladimir.mateljan@gmail.com, vedran.juricic@gmail.com, marko.moguljak@gmail.com

Virtualization finds wide application in information and communication technology. Some of these applications can be used in education, to simplify the administration and teaching performance while saving money. In the last few years, computer technology has progressed sufficiently thus facilitate the development of virtual machines that provide a completely new method of using computers and their resources.

The virtual machine is a software implementation of a machine that executes programs like a physical machine. There are several types of virtualization, span from partial to complete, and most of them could find a useful application in education. Setting up test environments, virtualization of classroom for a particular course, remote connection, and use of existing computer without any hardware upgrades, faster backup, and easier maintenance are some of the benefits of implementing virtual machines in education.

I. INTRODUCTION

Information and communication technologies have significantly changed and improved business practices, thus created a process in which companies stimulate the further development of these technologies. Among these trends virtualization finds wide application, and the requirements of information technology and the market situation are creating the tendency for further development.

II. VIRTUALIZATION

Although the concept of virtualization is relatively new, scientists have used virtual machines for decades on mainframe computers. Only in the last few years computer technology has advanced enough for virtualization to have wider application. Some of these applications could be useful in education, to simplify administration and teaching performance while saving money.

The term "virtualization" traces its roots to 1960's mainframes, during which it was a method of logically dividing the mainframe's resources for different applications. The usual goal of virtualization is to centralize administrative tasks while improving scalability and overall hardware-resource utilization.

A. Virtual computer

A virtual computer is a logical representation of a computer in software. By decoupling the physical hardware from the operating system, virtualization provides more operational flexibility and increases the utilization rate of the underlying physical hardware. ("Wikipedia, 2014)

Although virtualization is implemented primarily in software, many modern microprocessors now include hardware features explicitly designed to improve the efficiency of the virtualization process. In a traditional physical computer, one instance of the operating system supports one or more application programs. In a virtualization environment, a single physical computer runs software that abstracts the physical computer's resources so that they may be shared between multiple "virtual computers." Each virtual computer may be running a different operating system from all of the other virtual machines on the physical machine. A crash or other program error on any of the virtual machines leaves all of the other virtual machines unaffected. (IBM Corporation, 2007)

III. VIRTUALIZATION IN EDUCATION

Virtualization technology comes in several different varieties. It represents an abstraction from physical resources. All uses of virtualization are based on this concept. There are three major types of virtualization; server virtualization, desktop (client) virtualization and storage virtualization. All three major types can be useful in education by improving operational efficiency while reducing hardware and space requirements, energy consumption, and maintenance. It can also provide a layer of security because the virtualization applications run independently of each other so that a problem with one server can be confined to that server, and will not spread to others.

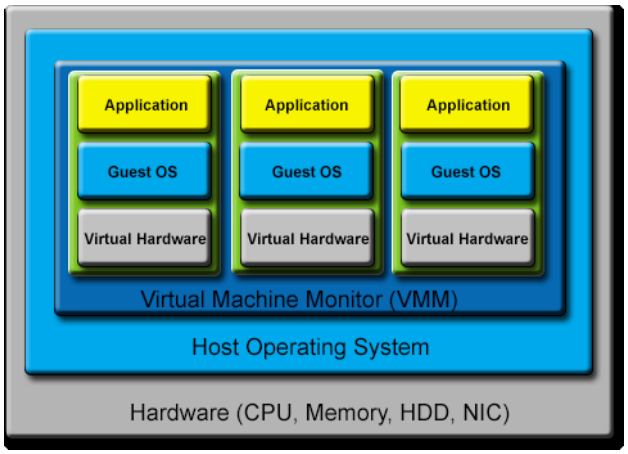


Figure 1. Virtual Computer diagram
Source: National instruments "Virtualization Technology Under the Hood", 2009,

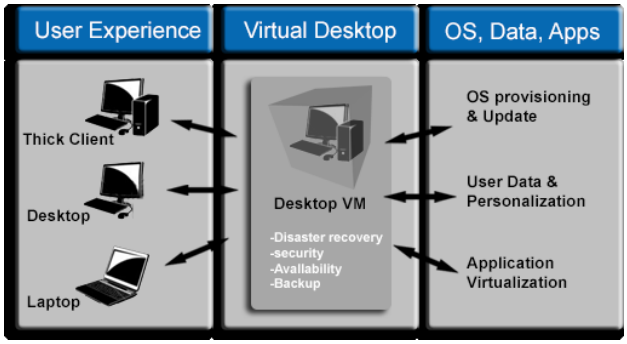


Figure 2. Desktop Virtualization
Source: Larry Dignan, "VMware: Desktop virtualization tipping point in 2011?" 2010.

A. *Desktop virtualization*

Desktop virtualization is the concept of separating the logical desktop from the physical machine. This type of virtualization moves the operating system and the applications it supports from the student desktop to a virtual computer running in a server. Each user can customize the desktop environment depending on the extent permitted by school policies.

This approach provides great advantage over working on a single assigned PC. When the student moves to a different location, such as from the classroom to a lab or library, the student's entire workspace appears on whichever workstation or device the student is currently using. If workstation hardware fails, the student's entire workspace is instantly available on any other workstation on the network. Software licenses for applications the student uses may be available for use by other students when the student's virtual desktop is not actually in use.

B. *Server virtualization*

Server virtualization lets one server do the job of multiple servers by sharing the resources of a single server across multiple environments. Each server typically serves one function (i.e., mail server, file server, Internet server, etc.), with each server using only a fraction of its true processing power, server virtualization breaks through the "one application, one server" barrier and facilitates the consolidation of numerous servers into one physical server. This equates to less physical servers required, and 70 to 80 percent or higher utilization of existing hardware as opposed to the previous 10 to 15 percent. (Bill Johonnesson, 2010,)

If the resource requirements of one of the server-based products running in a virtual server grows, because of increased usage for example, moving that virtual server to a different physical server with more available resources is as simple as copying a file. This ease of replication of virtual servers also means that it's simple to maintain snapshots of virtual servers as file back-ups and quickly restore complex systems to operation in the event of physical server hardware failures. This increased resilience results in more efficient use of existing physical server resources, lower operating costs, reduced power consumption and greater overall reliability. (IBM Corporation, 2007)

C. *Storage virtualization*

Storage virtualization technology separates virtual storage partitions from physical storage devices (storage devices that hold, spin, read and write magnetic or optical disks such as CD, DVD, or even a hard disk drive, etc.).

The virtualization software presents users with a logical storage area for storing their data, and itself handles the process of mapping it to the actual physical location. Virtualizing storage gives administrators greater control over storage resources and allocation of storage space for applications and users that need it most. Data migration is also made easier through storage virtualization. Data can be moved to another storage device without disrupting computer operations. It's also easier to expand storage space when the physical location is full.

IV. THE BENEFITS OF VIRTUALIZATION FOR *EDUCATION*

A. *Hardware life span savings*

Most student workstation computers are acquired over an extended period. Most schools are faced with the dilemma of whether to spend scarce technology budget money to expand the number of student workstations toward one per student or to replace the oldest machines with newer models that will better support current operating systems and desktop applications. Desktop virtualization provides a way to increase the useful life of existing older PCs while at the same time reducing the cost-per-seat of new student workstations. This is done by running virtual desktop software on older PCs and purchasing less expensive thin client workstations rather than full desktop computers to add additional workstation seats to the school. Even older PCs that cannot run current versions of new operating systems and resource demanding desktop applications can do so when used as thin clients running virtualized desktop software. There are no specific minimum requirements, the PC only needs

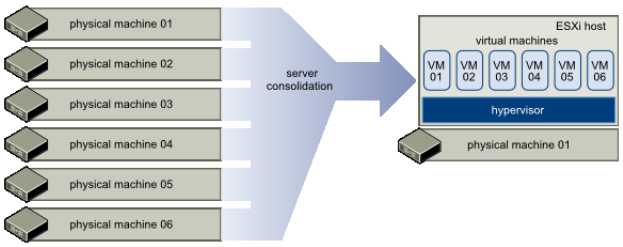


Figure 3. Virtual Server
Source: WMware, "Aspects of Virtualization" 2014.

to be able to run the application for virtual desktop software, which access the resources on the host computer. Working performance depends on many different factors, most important are the number of client computers, server's hardware resources, network bandwidth, type of virtualization, and virtualization software. (IBM Corporation, 2007)

B. *Hardware efficiency*

Any virtual machine can be run on any server in the data center because the virtualization layer abstracts the resources of the underlying computer and presents them in a standardized way to the virtual machine's operating system and applications. Older, smaller servers will support fewer virtual machines than newer, larger servers. However, the management interfaces to all of the virtual machines will be uniform and will result in lower administration costs.

Virtualization provides a way to deliver a uniform student experience for desktop workstations regardless of PC age. A lab equipped with PCs with different processors and memory will provide a uniform student user experience because the virtualized desktop environment exists on the network servers rather than individual PCs. The students' personalized workspaces will follow them from classroom to lab to library, and, where connectivity is available, the workspaces will even follow students home. (IBM Corporation, 2007)

C. *Remote connectivity*

As connected devices are becoming more ubiquitous, schools will be faced with the problem of supporting not only their devices, but also a growing range of devices owned by students. More and more students now own personal network capable devices. Some, such as cell phones, smart phones, tablets, laptops, etc. students already bring to school. Others, such as game consoles and personal computers, could be used to extend the school day by providing access to curriculum content from home. Virtualization technologies provide a way to abstract the workspace environment in such a way as to be able to present educationally valid content on a broad range of devices (IBM Corporation, 2007)

The additional network security provided by virtualization makes it possible for school networks to allow connection of student-owned devices to the school network without compromising the security, which is essential to maintain the reliability of the network. With client virtualization, it is no longer a problem that all of the devices on the network are not similar. It is possible to deliver curriculum content to students regardless of where they are or what sort of device they are using.

D. *Licencing utilization*

With virtual desktop workstations, purchase of software licenses is needed only for the peak instantaneous usage of any given product. Because licensed software runs in virtual PC clients on the server rather than on individual machines, it's easier to control. Limiting usage of a particular program to the number of licenses on hand becomes a manageable task. Just as virtual servers maximize utilization of physical server resources, virtual PC clients maximize utilization of software licenses. (IBM Corporation, 2007)

E. *Security improvement*

In a virtual desktop environment, software, data and most of the hardware elements that comprise the desktop workstation are in the data center and not on the desktop. This makes virtual PC clients inherently more secure since all of the data resides on servers in a secure data center rather than spread over desktops throughout the school. The inherent security provided by virtualizing client machines lets you open the network to outside connections without compromising security. (IBM Corporation, 2007)

Allowing students to connect from home has a beneficial effect on student performance, the network security issues and lack of standardized equipment in homes has made this difficult until now. Client virtualization addresses both issues. Students connecting to virtual PC clients on the school server can use a broad range of workstation devices ranging from game consoles to the family PC. The school network is protected because the virtual client connection interface is inherently secure

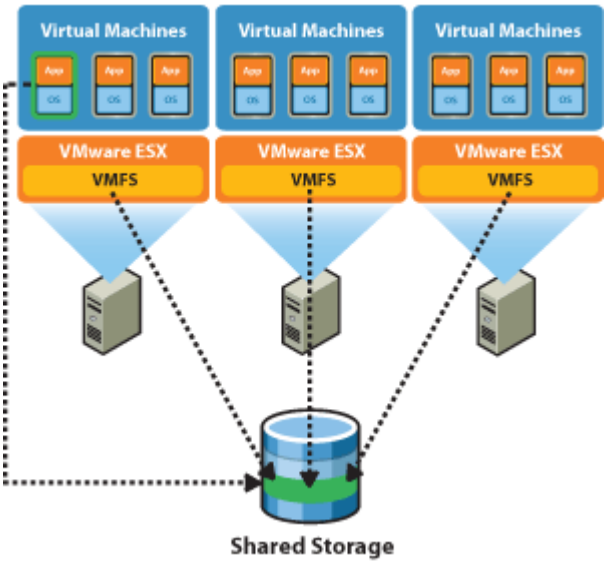


Figure 4. Storage Virtualization, VMFS_diagram
source: Eric Horschman, Virtual Reality, 2008.

and protects the school network from malicious content on the students' physical devices. (IBM Corporation, 2007)

F. *Easy lab enviorment settup*

Labs are set up to provide environments where the students can learn and solve problems. Entire rooms are often allocated for this purpose and contain different software and hardware as well as a number of machines. These labs are usually difficult to manage successfully. The different courses and the number of machines they might need lead to conflicting requirements. Shortages of available space and finances are a concern and limits the number of students in a lab and the amount of available equipment. Virtual lab can solve most of this problems and improve user experience. A significant savings can be realized when the whole virtual lab runs on a single machine. This machine can stand in the college's server room and won't take up much space. There is no need for student access to the main server in order to recover from problems. The use possibilities of the virtual laboratory are practically unlimited; it can be a base for any studies, course or training program. It is assumed that projects of the laboratory are fully multimedial. The participants of this laboratory can e.g. investigate training experiments from the definite field of material engineering, ask questions, pass tests, make contact with lecturers and the different users of the laboratory, and participate in development of its design and content. (IBM Corporation, 2007)

V. VIRTUAL MACHINES IN DIFFERENT ENVIORMENT

Most of the educational institutions in Croatia use Windows operating system as a main platform for classes. Although windows has wide support for applications, it is not always the best solution, and another operating system is required to perform tasks.

Virtualization offers a way of dealing with that problem in a simple manner, by installing a virtual machine software that contains operating system different from the one on a physical computer. There are many virtual machine software to choose from, and it depends on a computer hardware and installed software how well



Figure 5. Configuration

it works. To test virtual machines in different environments, a research needs to be done.

A. *The research*

Virtual machine used in research is VirtualBox, chosen because it is free, and is most frequently used among freeware. Operating systems used are; Windows7 ultimate and Linux mint 15. On a personal computer (see configuration reference figure 5.) are installed Windows7, and VirtualBox with Linux mint 15, which are given 1024MB RAM and 25GB of hard disk space. Benchmark tool used for testing is Geekbench 3, which measures integer performance, floating point performance and memory performance The process is repeated three more times in all other combinations of Windows and Linux as host OS and virtual machines, with same parameters. Results are then compared and ranked.

B. *VirtualBox*

Oracle VM VirtualBox is a virtualization software package for x86 and AMD64/Intel64-based computers from Oracle Corporation as part of its family of virtualization products. It was created by innotek GmbH, purchased in 2008 by Sun Microsystems, and now developed by Oracle. It is installed on an existing host operating system as an application; this host application allows additional guest operating systems, each known as a Guest OS, to be loaded and run, each with its own virtual environment.

C. *Geekbench 3*

Geekbench 3 is a cross platform benchmark tool. Scores are calibrated using a Mac mini (Mid 2011) with an Intel Core i5-2520M @ 2.50 GHz processor as a baseline with a score of 2,500 points. Higher scores are better, with double the score indicating double the performance. Geekbench 3 uses a number of different tests, or workloads, to measure performance. The workloads are divided into four different sections:" (Primate Labs, 2013,)

- "Integer performance: Integer workloads measure the integer instruction performance of your computer by performing processor-intensive tasks that make heavy use of integer instructions. All software makes heavy use of integer instructions, meaning a high integer scores indicates good overall performance." (Primate Labs, 2013)

- "Floating point performance: Floating point workloads measure floating point performance by performing a variety of processor-intensive tasks that make heavy use of floating-point operations. While almost all software makes use of floating point instructions, floating point performance is

TABLE I. TEST SCORES

OVERALL INTEGER PERFORMANCE				
	OS: Windows VM: Linux	OS: Linux VM: Linux	OS: Windows VM: Windows	OS: Linux VM: Linux
Single-core	1306	1099	1167	1099
Multi-core	1294	1106	1154	1106
OVERALL FLOATING POINT PERFORMANCE				
	OS: Windows VM: Linux	OS: Linux VM: Linux	OS: Windows VM: Windows	OS: Linux VM: Linux
Single-core	1237	1221	1154	1221
Multi-core	1231	1229	1161	1229
OVERALL MEMORY PERFORMANCE				
	OS: Windows VM: Linux	OS: Linux VM: Linux	OS: Windows VM: Windows	OS: Linux VM: Linux
Single-core	619	733	572	733
Multi-core	258	724	99	724

- especially important in video games, digital content creation, and high-performance computing applications." (Primate Labs, 2013)

- "Memory performance: Memory workloads measure memory bandwidth. Geekbench 3 uses tests based on the STREAM benchmarks developed by John D. McCalpin. Software working with large amounts of data (e.g., digital content creation) relies on good memory bandwidth performance to keep the processor busy." (Primate Labs, 2013)

D. *Results*

Tests has shown that there is a difference in performance in every combination of host OS and virtual OS. As shown in overall score (see Figure 6.) even single-core and multi-core scores differs in a single test. From a user experience these differences are negligible and virtual machine can work properly in any given combination. This is simple and free solution that offers many possibilities, and it is educator's choice of which of the operating system will be used for their class.

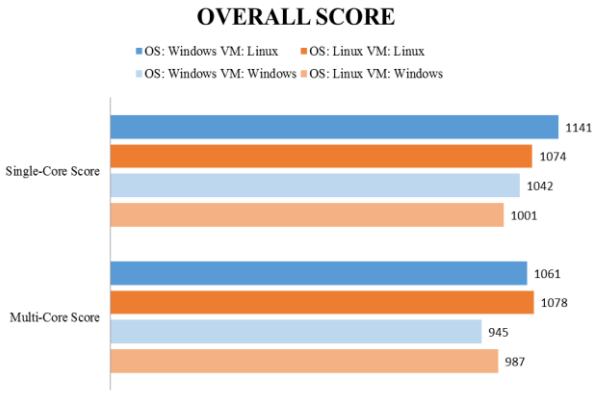


Figure 6. Overall Score

VI. SUMMARY

Virtualization is a powerful technology trend that is happening now. The trend is already underway and the outlook is that it will grow very rapidly. Educators should examine the opportunity very closely and see when and how they can start taking advantage of this capability. Virtualization provides better experience for students and teachers at lower cost and with greater security and reliability.

LITERATURE

[1] Chlajared, "The different types of virtualization", April 23, 2010 http://losangelescomputerhelp.com/2010/04/the-different-types-of-virtualization/

[2] Wikipedia "Virtualization" 2014., http://en.wikipedia.org/wiki/Virtualization

[3] IBM „Virtualization in education", 2007 http://www-07.ibm.com/solutions/in/education/download/Virtualization%20in%20Education.pdf

[4] What Are the Different Types of Virtualization Technology? http://www.wisegeek.com/what-are-the-different-types-of-virtualization-technology.htm#did-you-know

[5] Larry Dignan, "VMware: Desktop virtualization tipping point in 2011?" July, 2010, http://www.zdnet.com/blog/btl/vmware-desktop-virtualization-tipping-point-in-2011/37025

[6] "Interpreting Geekbench 3 Scores", 2013, http://support.primatelabs.com/kb/geekbench/interpreting-geekbench-3-scores

[7] Wikipedia "Hardware virtualization", 2014. http://en.wikipedia.org/wiki/Hardware_virtualization

[8] National instruments "Virtualization Technology Under the Hood", 2009, http://www.ni.com/white-paper/8709/en/

[9] Wmware "Aspects of Virtualization" 2014. http://pubs.vmware.com/vsphere-50/index.jsp?topic=%2Fcom.vmware.vsphere.introduction.doc_50%2FGUID-7EE617A2-4A10-424F-BAE2-56CA6692A93F.html

[10] http://www.zdnet.com/blog/btl/vmware-desktop-virtualization-tipping-point-in-2011/37025

[11] Eric Horschman, Virtual Reality, June 2008, https://blogs.vmware.com/virtualreality/2008/06

[12] Bill Johonnesson, The Different Types of Virtualization, 2010. *http://bucarotechelp.com/computers/winadmin/89050501.asp*

Open^{n}

Hannu Jaakkola[*], Timo Mäkinen[*], Jaak Henno[**] and Jukka Mäkelä[***]
[*] Tampere University of Technology/Pori Department, Pori, Finland
[**] Tallinn University of Technology/Information Technology, Tallinn, Estonia
[***] University of Lapland/Method Sciences, Rovaniemi, Finland
hannu.jaakkola@tut.fi; jaak@cc.ttu.ee; jukka.makela@ulapland.fi

Abstract - One of the key issues in the current information society is the easy and open availability of required resources. The concept of "openness" covers several resource categories, such as data, software sources, services, innovation and education. Open Data is seen as a promising source of new business. The concept of open innovation is based on the idea of distributing "useless" innovations for beneficial use in a new context. The role of open education has grown quickly at different education levels – especially university level. Open social networks are important platforms both for information distribution and as a source of information. There is a bidirectional influence between the availability of open resources and applications based on them: On the one hand, new applications create trends that demand open availability of resources, while on the other hand, the availability of resources creates opportunities for new applications. From the technology perspective, important change factors include mobility, big data technologies, digitalization, and context recognition. At the beginning of our paper we build a "big picture" of the important phenomena in the modern information society; open resources are seen in the role of enabler. National (Finland) and regional (AVARAS Project) initiatives are used to concretize the overall discussion.

.

I. INTRODUCTION

A. Towards an open information society

The importance of information as a factor of The importance of information as a factor of production has grown and information itself has a high commercial value. The Internet transfers enormous amounts of data in a reasonably open format. The skills for acquiring, marshalling and analyzing raw data to build utilizable knowledge and its visualization has a rapidly increasing value. Traditionally, these skills are used inside closed organizations like companies. *Business intelligence (BI*) is a technology that transforms raw data into knowledge to help decision makers in business management. BI software tools are able to handle large amounts of data, in most cases produced by the information systems of the company. However, nowadays *external data sources* are becoming more and more important sources of raw data – either combined with the company's internal data or also as such. The Internet makes enormous amounts of data sources available to all Internet users. Simultaneously it provides the means for *tracking users' Internet activities*; elements of which are scattered in an unmanaged way in several repositories. The term *Big Data* is connected to processing very large (usually public) data sets. The goal is – as in BI – to collect, organize, structure and publish individual data chunks in a form applicable for the users. The technologies used both by BI and Big Data tools apply a wide variety of Knowledge Engineering (KE) technologies. One important new source of data for analysis is Open Data; it is worth differentiating this term from *openly available data*. *Open data* refers to the idea that certain data should be freely available for use and re-use [1]. In general discussion, data sources are most often connected to the data produced, collected or paid for by public bodies. Examples of such data sources are geographical information (map data), statistics and weather data. However, an increasing amount of data is produced by autonomous sources (machines), like CCTV cameras, real time sensors, microphones, RFI readers, and information sensing mobile devices. Openly available data has a wider scope – it covers all data that can be accessed without meaningful restrictions, e.g. Facebook stream, public web pages, Twitter tweets, RSS feeds.

B. ICT Based Incremental and Radical Changes, New Paradigms and Phenomena

There are accelerating changes in current (modern) societies. In innovation theory, the changes are classified into four categories: incremental innovations, radical innovations, changes in technological systems, and changes in techno-economic paradigms [2]. This approach is applied in Figure 1 to classify some current phenomena.

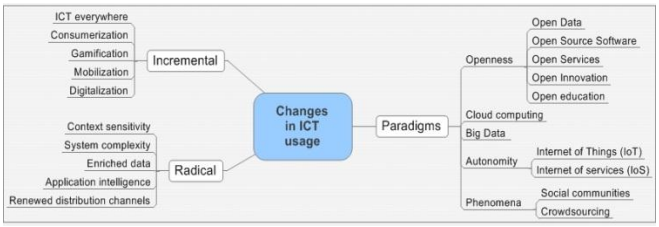


Figure 1. Changes in ICT usage Openness in figure above

Although the topic of this paper is focused on the role of open data – or openness at a general level – we wish to start by building a big picture of the detectable information technology phenomena having openness – especially open data as a driving force. We are trying to clarify the importance of the changes by using the categorization derived from innovation theory. *Incremental changes* appear continuously in existing products and services (continuing the existing trend). *Radical changes* appear when new research findings are applied in products to transfer their properties or

performance to a new step or cycle (movement to a new trend). *Changes in technological systems* are caused by combinations of several incremental and radical innovations in societal and organizational systems. *Changes in paradigms* are revolutionary and lead to pervasive effects throughout the target system under discussion (in our case the information society). The classification is used (in Fig. 1) in a quite broad way, which provides a means for criticism. Our aim is, however, to point out the importance of some existing phenomena and dependencies between them. The detailed aspects included in Figure 1 are discussed in Sections 2 and 3 of this paper.

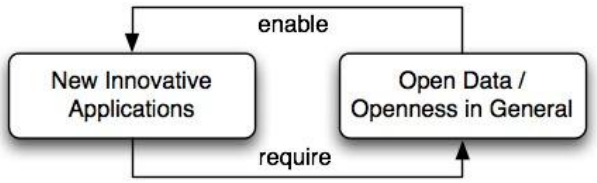


Figure 2. Bidirectional dependency between Openness and Innovative Applications

There is a bidirectional influence between the "paradigms" in the role of "enablers" on the right side of the figure and the phenomena on the left side in the role of incremental and radical changes (Figure 2): On the one hand, the changes recognized create demand for information resources, and on the other hand the availability of these resources creates a platform for new kinds of applications and use of information (Figure 2).

C. Structure of the paper

The aim of this paper is to point out the importance of open access to resources – especially to open data – as one of the driving forces of the current society and changes in it; the manifestations of changes are new services and applications and improvements in them. Sections 2 and 3 of this paper discuss the topics introduced in Fig. 1 more extensively. The aim is to introduce the key concepts related to openness and their role as an enabler of currently recognizable phenomena. Section 4 returns to the role of Open and Openly available Data as a key resource of improved and innovative products / services. It includes a general analysis and reports some experiences of Open Data in Finland. Section 5 covers a short introduction to the role of Open Data in education as one of its important and fast growing application areas. Section 6 introduces our regional Open Data project AVARAS (AVoimista tietovARAnnoista liiketoimintaa Satakuntaan –New Business for the Region of Satakunta From Open Data) by introducing its goals set and some preliminary results. Section 7 concludes the paper.

II. PARADIGMS AS ENABLERS

There is a common trend towards *cloud*-based service structures that provide a wide variety of scalable computing resources through the Internet [3;4]. For the software business, cloud architecture provides a standardized platform for application development and distribution. From the user point of view, cloud-based software solutions (SaaS – Software as a Service) provide easy location and device independent access to applications – data and services – in a symmetrical (device and platform independent) way.

The ability to apply openly available data resources is based on two factors: the ability to find the applicable resources (data mining) and the technology to apply intelligent data handling methods [5] *Big data* (discussed above) technologies are under active development; new algorithms and tools are being developed to provide data with high relevance in fulfilling the goals of the data mining requests [6]. The components of data mining technologies cover e.g. access technologies, methods used to understand the meaning of data items, methods for combining different data items to create understanding (knowledge), and technologies to publish and visualize the findings. In their study of the "Digital Universe," IDC [7] reported fast growth in Big Data: the amount of data applicable for open analytics is expected to grow from 130 EB (ExaBytes = 10^{18} Bytes) in 2005 to 40,000 EB in 2020. According to the report, the utilization level of available data is reasonably low: in 2013 only 0.5% of data was utilized in analysis, whereas 25% would be utilizable (33% in 2020). Additionally, the proportion of data that requires protection (because of privacy issues) is growing faster than the digital universe (from 30% to 45% of all data in ten years); only half of the data that needs protection is protected. Security and privacy are important issues when processing open data[8].

The source of data is also changing dramatically in the period covered by the report's analysis: 11% of data in 2005 was produced by *autonomous devices*; in 2020 this percentage is expected to grow to the level of 20%. Consequently, Big Data technologies have the opportunity to combine data from traditional data sources into real time data produced by cameras, sensors and other technical devices connected to the Internet. The term *Internet of Things (IoT)* refers to a collaborative structure in which the Internet or an Internet-like structure provides a platform for these connections. In addition to the source of data, IoT provides a means for the collaborative work of separate devices. In the service context, the term *Internet of Services (IoS)* is used to describe the service activity based on collaborating independent services. In connection with IoT, data technologies are used to find and analyze "sensor" type data that is combined with data from other open sources in an intelligent way.

At the human level, there is a great deal of evidence of the importance and power of collaborating individuals as members of social networks – *social communities*. Facebook provides a social networking platform for 1.4 billion users, Twitter for 646 million and LinkedIn for 261 million users [9; 10; 11]. These are only examples of the best-known collaboration platforms, which also have different usage profiles. LinkedIn has become an important source of data in recruiting, and Facebook and Twitter in experience-sharing and social communication. Simultaneously, the *data streams* of all these networks are important sources of data mined by big data technologies (social analytics) for commercial and profiling purposes. Active social network members are easy targets for social gesture building – the value of a network user is the value of its social gesture. The traffic is monitored by the service providers (like Facebook) for "quality

improvement purposes"; the reality is more or less related to the improvement of their business. In addition, this kind of open Internet traffic is also vulnerable to social analytics made by other parties in an uncontrolled way.

A special form of social networking is *crowdsourcing*, in which tasks are distributed to an undefined public group of people (crowd) for execution. There are different variations of the basic idea, e.g. *crowdfunding*, collective creativity, cloud labor, collective knowledge, community building, and open innovation [12]. There are several examples of successful crowdsourcing services: Duolingo (http://www.duolingo.com/) is a foreign language learning portal that is simultaneously used for crowd translations of texts as a commercial service; Kickstarter (http://www.kickstarter.com/) is a portal for the crowdfunding of innovations, which is aimed at providing a platform for startups to find funding for their product development.

Common to all of the above-mentioned "paradigms" is *openness* – openly available (data) resources provide a means for their successful implementation. *Open source software* combines in a way both the ideas of crowd labor and open innovation (software development in collaboration with loosely connected teams) and free (taking into account the licensing rules) exploitation of the final product without license fees. *Open Innovation* refers to the idea that the (internal) innovations (of organizations) are provided for open use, as well as existing (external) innovations being utilized beneficially. The idea is opposite to the tight licensing policy and careful protection of Intellectual Property Rights (IPR).

Open Education implements the idea of the open innovation principle in education. *Massive Open Online Courses (MOOC)* are provided to gain access to educational courses. *Open services* refer to service interfaces that provide a means for collaborative service implementation.

Open Data is freely available for everyone to use and republish without restrictions from copyright, patents, or other control mechanisms. The "Open Data" movement originates in the USA, but also the European Union (2003; Amendment 2013) was active as early as 2003 by creating a directive that encourages member countries to open public data sources for free business use; several member countries have passed laws or regulations to follow this recommendation.

III. FROM PARADIGMS TO CHANGES

As discussed above, the open availability of data and the ability to use the data in a beneficial way is one of the key factors of success in the utilization of Information and Communication Technology (ICT). ICT is more or less invisibly embedded in daily activities, processes, and products (*ICT everywhere*). Its utilization is bidirectional: on the one hand, it is used as an interface to services and, on the other hand, it is an important source of information.

Mobilization (the growing importance of mobile devices) provides a means for time and location independent service usage in a symmetrical way (the same service for different platforms and devices, scalable according to the network quality and type of user interface). The functionality (a wide variety of inbuilt sensors) of mobile devices provides a means for a growing amount of versatile data to be used by applications based on data transferred in a network. Mobile devices also have a richer context than workstations connected to a fixed network. In addition to the mobile terminal role, these devices also have the role of *information provider*. Their capability to collect user and usage dependent information is high. As a network connected device, a mobile terminal also has the capability to combine this information with other (openly available) information in the network as well as to send it to application developers (for analysis purposes) and interested external sources (e.g. intelligence agencies). The location and usage profile as well as the mobile terminal content are of high value, especially when connected to other user profile information (e.g. in a Facebook profile, Twitter profile, e-shopping profile).

Digitalization refers to the automatization of applications, digital marketing, and e-commerce – in general to technologies that are based on the handling of digital content. It is worth noting that even in situations where the original content is analog (like a video camera stream), in many cases it is transferred into and handled in digital format. An interesting application of digitalization is the virtual currency Bitcoin (https://bitcoin.org/en/) initiated in 2008 by Satoshi Nakamoto (a pseudonym). It uses peer-to-peer technology with no central authorities. It manages transactions and the issuing of Bitcoins is carried out collectively by the network. Bitcoin is also an example of an open-source product; its design is public, nobody owns or controls it. At the moment (February 2014) the amount of Bitcoin currency is 12 million and the p2p protocol allows its growth up to the level of 21 million (which will be reached in the next few years). Its current exchange rate is 1 Bitcoin to 700 EUR; the lowest rate has been 35 and this virtual currency suffers from fast changes.

Context sensitivity – the ability of an application to adapt to the context of its usage – enables the enrichment of applications. Although in connection with mobile terminals the term "context sensitivity" is most often seen as a synonym for the term "location sensitivity", the new applications are able to handle versatile context data in an organized way. *Lifeline* provides a *time* dimension for versatile context data items covering information items about *location*, *user* (e.g. health data, movement), *physical situation* (temperature, air pressure), *social environment*, *user preferences*, *presence and status*, etc. taking into account the capabilities built in to the device used. Context sensitivity is also the main source for growing *application intelligence* – the application's ability to create better quality knowledge based on complex handling and combining of data resources.

Adding "intelligence" to the data itself - *Enriched data* - provides new opportunities to use the same data for different purposes and to support more intelligent use of the data. In practice it is a question of combining different data formats in one visible / manageable data item. Examples of enriched data are embedded www-links in images, (reading the image automatically directs the device to the web page having the embedded address), QR codes directing the reader to the real information

connected to the code, embedded meta information in images, and applications used to extend the original information (e.g. taking a photo of an advertisement in a newspaper opens information related to it in a browser (see e.g. http://shortcut.karkimedia.fi).

The terms consumerization and gamification describe the characteristics of modern applications from several points of view. *Consumerization* is the product and service designs approach, in which the individual consumer (user value) is the primary driver. Such phenomena like the growing importance of social networks in marketing (you have to go where the consumers are) and BYOD (Bring Your Own Device) are typical manifestations of this, as well as the transfer of application development from experts to enthusiasts (crowd; a large share of mobile applications are developed by "non-professionals"). The users are also used to buying applications from platform providers' software online stores instead of special software shops or companies, at reasonably low prices. This *overhaul/restructuring of distribution channels* and the growing importance of cloud-based applications is changing the ICT business dramatically. *Gamification* describes users' expectations of usability; the same user experience and attractive look-and-feel that is available in games software is expected from serious applications. Consumerization and gamification can also be interpreted as an *attitude*: users tend to use ICT as a consumer product instead of a special product – as a part of everyday life, without any threats.

System complexity growth is actually a paraxial concept. Modern applications are designed for use in reasonably low performance devices (like mobile phones and tablets), but in a symmetrical way also in more powerful equipment. This application scalability can be implemented by distributing some part of the processing needed in (more) powerful (cloud) servers. As a consequence, complex monolithic systems are being replaced with less complex collaborating components. From the software development point of view, the complexity of a single component has decreased, but collaboration has increased the complexity of interfacing solutions. In general, we can also see that to an increasing extent applications are no longer stand-alone but a part of a complex environment. Barry Boehm [15] has described this using the term "*complex system of systems*" indicating that the current information systems are collaborating, independent systems.

IV. OPEN DATA

A. Sources of Open and Openly Available Data

Sections 2 and 3 painted the big picture related to the role of openly available data covering different points of view. This "big picture" covers both *Open Data* in its traditional meaning and *Openly Available Data*. The latter covers such data sources that are connected to a public network and (due to a lack of or insufficient security) provide a means for monitoring its content. The latter sources have a growing value in Big Data analytics. These include data sources opened purposely for public use (like the road cameras of some weather stations) but also such sources that are available but without the data owners' permission (like data collected by mobile phones, social network discussion streams, e-shopping data, private communication, etc.). Somewhere between these two types is open www-data, which is easy to use for data mining and analytics using text and image analysis. The rest of this paper concentrates on open data in line with the definition of the European Union [1;13; 14; 16]: "Open data refers to the idea that certain data should be freely available for use and re-use."

Open data is seen as a meaningful source of new business, especially for small and medium-sized companies; the importance of public sector information - government data - is considered important. Examples are geographical information (maps), statistics, weather data, data from publicly funded research projects, and digitized books. The Commission lists the following reasons to open governmental data[1]:

- Public data has significant potential for re-use in new products and services. Overall economic gains from opening up this resource could amount to € 40 billion a year in the EU direct business volume (and € 140 billion indirect).
- Addressing societal challenges – having more data openly available will help us discover new and innovative solutions;
- Achieving efficiency gains through sharing data inside and between public administrations;
- Fostering participation of citizens in political and social life and increasing transparency of government.

The Gartner group has positioned Open Data on the Emerging Technologies Hype Slope [17] in the category of potential breakthrough in 2-5 years, locating it in the curve position of the "Trough of Disillusionment" – passing through the Valley of Death!

It is worth noting that, in addition to governmental data, several private bodies (e.g. companies, private individuals, and private associations) and non-governmental public bodies (like cities, regional administration, publicly owned companies) have valuable information that would be beneficial for open use. Examples include applications using bus timetables, flight radar, city maps, and measurements from public sensors.

B. Challenges related to open data

In spite of the positive expected future of open data, there are several related challenges:

- problems in the capability of users and application developers to recognize the opportunities related to open data;
- unfamiliar business idea – how to use data to create new business;
- lack of technical readiness to use data sources because of complex data format or interfaces;

- inability to find suitable data sources for application purposes;
- insufficient availability of regional data sources for developing applications for local services;
- unclear licensing: in spite of openness there are rules limiting the rights of usage (as in open source software);
- legal issues – rules and limitations to republish data that is available in public sources (e.g. web pages) are unclear when processing open data [18] or when securely accessing it [19];
- technical limitations related to data publishing platforms;
- guarantee of maintenance and updatedness of data is missing;
- variation in data quality and reliability, etc.

In this paper we have no opportunity to handle these topics in detail; therefore we will satisfy ourselves with giving a short list of the challenges and reference to our future work on the AVARAS project introduced in Section 6 of this paper.

C. Open Data in Finland

At national level, Open Data has been noticed in several countries – we use Finland as our example. The Finnish government has followed the EU directives and regulated laws that demand that governmental data be opened for public use. In spite of having a positive attitude on the topic itself earlier, the real activities took place in the 2010s. Here are some important milestones worth recognizing:

- The first steps in the area of open data were taken by the National Land Survey of Finland in 2012 (geographical and map data) and the Finnish Meteorological Institute in 2013 (weather data); at the same time, Statistics Finland opened its statistical data.
- The report of the Ministry of Employment and Economy (TEM) [20] "Analysis of success paths for the future" reports the ability to apply information technology and information resources as one of the future success factors; in the Finnish economy this means the transfer from successful ICT producer role to the role of innovative adapter of data resources as a source of new business.
- At the same time, and based on the above-mentioned report, the Government of Finland published an "Open Data Programme" (http://www.vm.fi/vm/fi/05_hankkeet/02381_avoin_tieto/index.jsp) for the period 2013-2015; this program includes a list of operative actions to promote and coordinate open data related activities.

The Government Internet service portal *Suomi.fi* includes a special section dedicated to Open Data sources (http://www.suomi.fi/suomifi/tyohuone/yhteiset_palvelut/avoin_data/).

Some regions were faster than the government. Helsinki (the capital of Finland) started its project *Helsinki Region Infoshare* (http://www.hri.fi/fi/hri-projekti/) as early as 2010, and after the project the local open data portal has been developed and maintained as a normal part of the operative activities of the organizations involved. The *Tampere region* (http://www.hermia.fi/opendatatre/) was one of the pioneers in opening regional data sources for free use. Our project AVARAS ("Open Satakunta") also belongs to the group of pioneers with the goal of supporting application developers to find and apply open data sources in their business, encouraging developers to develop regional services in the form of open data based applications, and encouraging data owners to open their data sources for open access. One of the manifestations of open data related activities is the annual "Open Data Contest" Apps4Finland (http://www.apps4finland.fi/) that has been organized at national level annually since 2010. There is also a regional contest in the Pirkanmaa - Tampere region (http://apps4pirkanmaa.fi/). The contests provide a nice overview of the application base and also the types of data sources used in the applications. An analysis is discussed in Section 6 of this paper.

V. OPEN EDUCATION

Openly available education in the form of MOOCs is a fast growing opportunity to organize educational activities at different levels. At the moment this is seen as opportunity with big potential, but also drawbacks. In most cases it is a question of the availability of single courses for wide distribution and their official acceptance as a part of the curriculum; there are also cases where a full university degree is built on MOOCs as a "low-budget" alternative to more expensive traditional implementations of the curriculum. The popularity of MOOCs is explained by easy availability and in most cases also free access to lectures by the "world's leading specialists" available everywhere on the Internet. However, the motivation to participate in courses varies: sometimes it relates to the need to pass the course as part of the curriculum, sometimes it is at the level of general interest only. There are no reliable statistics available of the participants' behavioral patterns, but there are some case studies that report very low percentages of students passing the final examinations compared to the number of registrations, which is in the hundreds of thousands.

The most commonly known and globally used MOOC platforms are Coursera, Udacity, EdX, NovoEd and Khan Academy; in addition there are national level MOOC platforms, like that of the University of Helsinki (http://mooc.cs.helsinki.fi/). *Coursera* (https://www.coursera.org/) has its roots in Stanford University. It is a for-profit educational technology company, currently having 2.7 million users and partnership with 62 universities. Its offering covers courses in physics, engineering, humanities, medicine, biology, social sciences, mathematics, business, computer science, and other areas. Like Coursera, Udacity (https://www.udacity.com/) is also a for-profit educational organization having its roots in the free computer science classes offered in 2011 through Stanford University. Udacity has 25 (as of 30 September 2013) active Stanford University courses (source:

http://en.wikipedia.org/wiki/Udacity). EdX (EdX https://www.edx.org/) was founded by the Massachusetts Institute of Technology and Harvard University in May 2012. Its purpose was to provide university-level courses in several disciplines for worldwide distribution at no charge. It also conducts research in the area of learning. EdX has nearly 1.6 million users (http://en.wikipedia.org/wiki/EdX). Khan Academy (https://www.khanacademy.org/) is a non-profit educational website created in 2006 by educator Salman Khan, who has close connections to MIT and Harvard Business School. Its mission is to provide 'a free world-class education for anyone anywhere' for free, forever. Khan Academy provides thousands of educational resources, including a personalized learning dashboard, over 100,000 exercise problems, and over 4000 micro lectures via video tutorials stored on YouTube. The scope of study topics covers mathematics, history, healthcare, medicine, finance, physics, general chemistry, biology, astronomy, economics, cosmology, organic chemistry, American civics, art history, macroeconomics, microeconomics, and computer science. (http://en.wikipedia.org/wiki/Khan_Academy). NovoEd (https://novoed.com/) is another Stanford university spin-off in MOOC supply. There are also two recently new efforts in the area of open MOOC platforms: Stanford University has developed OpenEdX (http://code.edx.org/) and Google mooc.org (http://mooc.org/) as a OSS solution for establishing own MOOC platform.

MOOC based education provides the opportunity to reorganize (part of) education at all levels. The traditional approach in the form of distance learning has already been accepted as a part of the education system; videoconferencing, video recordings of lectures, and self-study based network courses have been widely adopted by educational institutes. What is new in MOOC? Maybe nothing other than wide geographical coverage and a platform that supports network based learning and an idea of a new type of learning. For partner organizations, MOOCs provide an alternative to self-organized courses. Instead of lecturing on the same topic that is available as a MOOC, the user institutes would use the same resources in student support activities at individual or small group level. This topic has been discussed in more detail by Henno et al. in their conference papers [21; 22].

Open Data – in general - also provides opportunities for use as a study material for teaching, studying and research work. The added value compared to the raw data available in www-pages is the opportunity to handle it in analysis, to combine and compare several sources of data and to illustrate the results of analysis.

VI. Regional Initiative to Reveal the Opportunities of Open Data

In autumn 2013, the Pori Department of Tampere University of Technology initiated a regional project to reveal the opportunities offered by open data. The one-and-a-half-year project is conducted in co-operation with the City of Pori and eight SME-sized software development companies from the Satakunta region in western Finland. The role of the partners is to direct and give guidance to project activities forming the interest group of the project. The acronym of the initiative, AVARAS, comes from its Finnish name "Avoimista tietoVARAnnoista liiketoimintaa Satakuntaan" which is "New Business for the Region of Satakunta from Open Data" in English. The project is funded by the European Union Regional Fund (ERDF), as part of the Satakunta Regional Centre for Economic Development, Traffic and Environment funding program.

The overall goal of the AVARAS project is to provide software companies with information on how to apply open data sources in their businesses. The project is expected to answer the following three questions: 1) what kind of business opportunities do open data offer? 2) What sort of open data sources exist? 3) How can open data be utilized technically? The questions were the basis for the three main work packages of the project: WP1 Open Data in Business, WP2 Meta-portal and Reviews, and WP3 Applications and Technologies. An additional target of the project is to encourage the public organizations of the region to open their data sources for open public use.

The first work package of the project (WP1) is composed of questionnaire studies, public seminars, and workshops with the interest group. It also includes tasks clarifying open data in context and the phenomenon of openness in general. This paper is one of the outcomes of the tasks. The first questionnaire-based study targeted the business and public sector in Satakunta. 166 respondents received the web-based survey, the results of which are being analyzed. The aim is to find out how well the basics of open data are known, and how its opportunities in business are seen. At first glance, it seems that awareness of the topics is not very high among the respondents.

The second work package (WP2) is intended to discover relevant sources of open data, resulting in an open data portal (www.tut.fi/avaras), which is called a meta-portal, because, instead of listing individual open data sources, the aim is to offer links to such portals. Thus, the outcome of the work package is for the most part a portal of open data portals. The portal is used also to deliver all the results of the project.

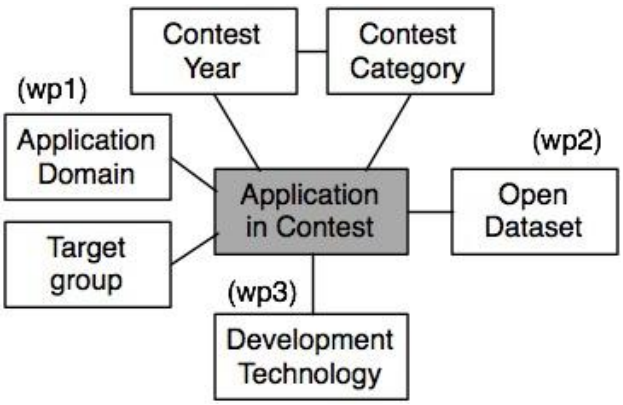


Figure 3. The dimensions to be used in the analysis and synthesis of applications, and their connections to the work packets of the project

The third work package (WP3) a) builds pilot applications using open data and b) examines existing software. During the development of the pilots, the technologies and tools needed in application development

are evaluated. The pilot applications will be openly available via the portal for public use. Their purpose is especially to bring out interfacing technologies and interface specifications for re-use in new applications. One of the pilots is discussed in [23]. The input of task group (b) is applications which have participated in the open data contests in Finland. There are nearly 400 contest applications from the period 2010-2013 to examine. The result of the study will also contribute to the other two work packages. The analysis and synthesis will be done from several dimensions as depicted in Fig 3. In addition to *development technologies*, there are dimensions for *application domains*, *target groups* (WP1), and *open datasets* (WP2).

In addition to the three initial work packages, two further topics for review have come to the fore during the project: a) Opportunities for companies to employ open education offered by several universities in the form of Massive Open Online Courses (MOOC), and b) NoSQL databases as a means to deliver open data. From the point of view of data models, there are several approaches for NoSQL solutions, such as key value, document, columnar, and graph. This sort of data models seems to be more flexible and straightforward to use than traditional relations. Commonly, the systems also offer e.g. ready-made REST interfaces and inherent replication. The steering committee of the project has just confirmed the inclusion of these topics in the work packages.

VII. CONLUSION

The paper discusses topics related to Open and Openly available Data at a general level. The approach is based on the hypothesis that open resources (data) are enablers for new innovative applications and new innovative applications create a demand for the availability of new resources. We have analyzed some visible phenomena in the modern information society; some of these are seen to have an "incremental" character and some to have more importance in the role of radical changes. Availability and opening of resources for public use is seen as a new paradigm in ICT; it provides new opportunities that will cause remarkable changes. In addition to the opportunities, there are also threats which are discussed briefly in the paper. Open Education – the fast growing availability of MOOCs is an opportunity to change education, or it can also be seen as a threat to the current state. Our case report (related to one European country) lists activities that have manifested the beneficial use of Open Data as a consequence of the European Union Directive to encourage its member countries to open the data resources owned by governmental organizations for free business use by application developers. On this continuum, a regional initiative (AVARAS) is introduced. The project is still in its "embryonic" phase and the results of the studies related to it are not available yet. The aim is to continue this introductory paper with more detailed reports of our project results at a later date.

ACKNOWLEDGMENT

The AVARAS project is funded by the European Union Regional Fund (ERDF), as part of the Satakunta Regional Centre for Economic Development, Traffic and Environment funding program.

REFERENCES

[1] European Union (2014), Digital Agenda for Europe: A Europe 2020 Initiative. Open Data. http://ec.europa.eu/digital-agenda/en/open-data-0 (retrieved January 18th, 2014).

[2] Freeman, C. and Perez, C. (1988). Structural Crises of Adjustment, Business Cycles and Investment Behavior. Technical Change and Economic Theory. Dodi, G., Freeman, C., Nelson, R., Silverberg, G. and Soete, L. (eds.). London, Pinter Publishers.

[3] Juric, M.B., Rozman, I., Brumen, B., Colnaric, M., & Hericko, M. (2006). Comparison of performance of Web services, WS-Security, RMI, and RMI-SSL. Journal of Systems and Software, 79(5), 689-700

[4] Juric, M.B., Sasa, A., Brumen, B., & Rozman, I. (2009). WSDL and UDDI extensions for version support in web services. Journal of Systems and Software, 82(8), 1326-1343

[5] Brumen, B., Hölbl, M., Harej Pulko, K., Welzer, T., Heričko, M., Jurič, M.B., & Jaakkola, H. (2012). Learning Process Termination Criteria. Informatica (Lithuan.), 23(4), 521-536

[6] Zorman, M., Pohorec, S., & Brumen, B. (2010). Opening the Knowledge Tombs - Web Based Text Mining as Approach for Re-evaluation of Machine Learning Rules. Lect Notes Comput Sc (Advances in Databases and Information Systems), 6295, 533-542.

[7] IDC (2012), The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East. http://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf (retrieved January 18th, 2014).

[8] Brumen, B., Golob, I., Welzer, T., Rozman, I., Družovec, M., & Jaakkola, H. (2003). An Algorithm for Protecting Knowledge Discovery Data. Informatica, 14(3), 227-288.

[9] Internet World Stats (2014). Internet World Statistics, Usage and Population. http://www.internetworldstats.com/, (Retrieved January 24th, 2014).

[10] Statisticbrain (2014), Social networking statistics. http://www.socialbakers.com/linkedin-statistics/ (Retrieved January 24th, 2014).

[11] Socialbakers (2014), Linkedin Statistics. http://www.socialbakers.com/linkedin-statistics/ (Retrieved January 24th, 2014).

[12] Crowdsourcing (2014), Crowdsourcing.org, The Industry website. http://www.crowdsourcing.org/ (Retrieved January 24th, 2014).

[13] European Union (2003), DIRECTIVE 2003/98/EC of the European Parliament and of the Council of 17 November 2003 on the re-use of public sector information. Official Journal of the European Union. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2003:345:0090:0096:EN:PDF. (retrieved January 18th, 2014).

[14] European Union (2013), DIRECTIVE 2013/37/EU of the European Parliament and of the Council of 26 June 2013 amending Directive 2003/98/EC on the re-use of public sector information. Official Journal of the European Union. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2013:175:0001:0008:EN:PDF. (retrieved January 18th, 2014).

[15] Boehm, B. (2006). A view of 20th and 21st Century Software Engineering. Proceedings of the 28th International Conference on Software Engineering (ICSE). Shanghai, China, ACM SIGSOFT and IEEE TCSE.

[16] European Union (2011), Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the R. Open data - An engine for innovation, growth and transparent governance. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2011:0882:FIN:EN:PDF. (Retrieved January 27th, 2014).

[17] Gartner Group (2010), Emerging Technologies Hype Cycle 2010.

http://www.gartner.com/it/content/1395600/1395613/august_4_whats_hot_hype_2010_jfenn.pdf. (Retrieved January 29, 2014).

[18] Brumen, B., Heričko, M., Rozman, I., & Hölbl, M. (2013). Security Analysis and Improvements to the PsychoPass Method. Journal of Medical Internet Research, 15(8), e161.DOI 10.2196/jmir.2366; PMID: 23942458

[19] Brumen, B., Heričko, M., Sevčnikar, A., Završnik, J., & Hölbl, M. (2013a). Outsourcing Medical Data Analyses: Can Technology Overcome Legal, Privacy, and Confidentiality Issues? J Med Internet Res, 15(12), e283. DOI 10.2196/jmir.2471.

[20] TEM (2013), ICT 2015 työryhmä, 21 Polkua kitkattomaan Suomeen. ICT 2015-työryhmän raportti. Työ- ja elinkeinoministeriön julkaisuja, Innovaatio, 4/2013. http://www.tem.fi/files/35440/TEMjul_4_2013_web.pdf. (Retrieved January 29, 2014).

[21] Henno, J., Jaakkola, H., Mäkelä, J. and Brumen, B. (2013). Will universities and university teachers become extinct in our bright online future? MIPRO 2013. Opatija, Croatia, Mipro and IEEE. ISBN 978-953-233-074-8; ISBN 978-953-233-075-5.

[22] Henno J., Jaakkola H., Mäkelä J., Brumen B. (2014), From Learning to E-learning to M-learning to C-learning to ... ?. MIPRO 2014. Opatija, Croatia Mipro and IEEE. Submitted to Mipro 2014..

[23] Rantanen P., Sillberg P. (2014), Event Calendar for Open Data Sources. Submitted to Mipro 2014.

From Learning to E-learning to M-learning to C-learning to ... ?

J. Henno*, H. Jaakkola** and J. Mäkelä***
* Tallinn University of Technology, Estonia
** Tampere University of Technology, Pori
*** University of Lapland, Rovaniemi, Finland
jaak@cc.ttu.ee

Abstract – Need for Higher Education is growing rapidly - by 2020 two thirds of all jobs will require postsecondary education [1]. The current higher education system cannot satisfy this need in many fields, e.g. Gartner predicts, that already in 2015 only a third of jobs for big data analysts can be filled [2].

The rapid development of Information and Communication Technology (ITC, IT) has changed many aspects of teaching and learning. Educational materials are moving into cloud(s) and degrees are given on-line ('Master's Degree Online Directory' lists over 90,000 programs [3]), old-fashioned Colleges and Universities with large campus, staff and high costs are getting more and more competition from numerous educational entrepreneurs, who start massive open online courses (MOOC) with millions of users/students from all around the world; they establish new on-line universities where costs of obtaining a degree are marginal compared to 'brick-and-mortal' universities and which are open to nearly everyone independent of user geographical location. Until now MOOC have not yet proved their value - less than 4% of enrollees complete the courses [4] and they do not receive (yet) university diploma. But the World most prestigious technical university, MIT, has already set up a certification system - students who can master the subjects taught in MOOC receive an official certificate of completion [5]. Although these certificates will not carry the weight of a traditional MIT diploma (they will be issued by started by MIT online learning initiative called MITx), but the word 'MIT' alone carries a weight.

When speaking about possibilities of IT in education most speakers/writers are excitedly discussing technology - servers, communication lines, bandwidth, security (preferably all in thick cloud of technical jargon). Nearly nobody presents practical experience, suggestions and/or theories about how to teach with this new technology in order best to utilize its possibilities. The 'education in the cloud' mainly consists of video clips trying to repeat the classroom experience: recorded before the old-fashioned blackboard (only nowadays it is white), where professor is presenting the same lecture what he presented in lecture hall, only now lectures are cut into 5-10 min video clips - the 'mobile approach'. Is this why the old lecture halls are still needed?

Rapidly increasing costs of University Education, increasing competition of educational entrepreneurs and constant increase of population of working students have made clear, that both sides of University education - teaching/learning and accrediting, giving diplomas need changes. University Education should more take into account students informal and non-formal learning, their participation in professional work - both in organization of study and in certification.

In the following are considered several aspects of current situation, presented practices which authors use when teaching IT students and examined possible changes in universities teaching and certification methods.

I. INTRODUCTION

Advancements in technology are constantly reshaping our teaching/learning/training practices.

In last 10..15 years, the biggest influence for educational practice have had computers and Internet Technologies (IT) and their role is increasing, especially for delivering training. The impact of IT to education was characterized in developed countries already in 15 years from now as the 'e-Learning Revolution' [6],[7]. Currently the E-learning revolution has started or is starting even in rural Africa [8], while developed countries have already embraced the M-learning revolution [9], our classrooms have become cloudy [10], [11] and some are moving classrooms into boxes ('the classroom in a box') [12], [13].

Technology and buzz-words around it have changed rapidly. But disparity between what goes into higher education and what many get out of it are not always encouraging.

Greek writer Aristophanes complained already more than 2,000 years ago in his comedy 'The Clouds' [14], that universities ('Thinkeries with quacks') left students heavily indebted and without practical skills. Clouds, the 'patron goddesses of thinkers and other layabouts' arrives in the comedy singing majestically and promising with teachings from professor 'Inferior Argument' earthly 'summum bonum' and a life of ease and pleasure, typical of men who know how to talk their way out of trouble.

Headings indicate, that little has changed during the last 2000 years, except that youth less and less believes that professors can lead them to life of ease and pleasure : "*Student Loan Problems: One Third Of Millennials Regret Going To College*" [15], "*Graduate data reveal England's lost and indebted generation* " [16], "*Today's College Graduates: In Debt and Unable to Find a Job*" [17].

The forerunner of the computer revolution in education has been USA - very innovative country, the very birthplace of computers, PC, the iPad and also initiator of application of computers in education

[18],[19]. Creativity and innovation have always been in USA highly valued [20]. But recent research indicates, that creativity of American children is on decline [21].

Has something gone wrong with rapid advance from learning to e-leaning to m-learning to c-learning (c- cloud, [22]) to ... - and if, then what ?

II. LEARNING

Education is the most complex area of production. Schools take in young human beings and try to develop their minds - create an 'added value' - and charge either society or directly the customer, student. But at the same time those minds, their ways of life are also rapidly changing and these changes happen quicker than changes in big systems - in universities, in the whole European Higher Education Area [23].

Learning occurs constantly, “from cradle to grave”. In 1996 the Organization for Economic Co-operation and Development (OECD) member states agreed to develop lifelong learning strategy, which includes formal, non-formal and informal learning [24], [25].

The formal learning is highly structured and implemented by specialized organizations - schools. Formal learning is intentional, schools and learners have explicit objectives covering knowledge, skills and competences, what learner should get.

Informal learning is learning from life experience - language acquisition, accepting cultural norms and social behavior. This is happening all the time, everywhere and is not intentional nor organized.

With development of society everything becomes more organized, thus also learning which happens not in schools tends to be organized. An example are Massive Open On-Line Courses (MOOC), which are rapidly gaining popularity, e.g. the Khan Academy [26] provides every month online lessons for more than 10 million visitors from about 200 countries. Organized 'outside school' learning is by OECD terminology non-formal learning. This kind of learning is very accessible - lessons/videos can be followed even on iPhone, iPad or Android phones [27]. While the first MOOC (Udacity, Codeacademy, Khan Academy) were focused on math, computer programming and computer science, currently MOOC are providing very broad selection of courses and new providers appear constantly. Even the World Economic Forum (WEF) announced at its last meeting about launch of its Forum Academy, where the huge network of industry leaders and opinion-makers of WEF will make their expertise and visions available to the global audience [28].

All these 'educational entrepreneurs' are directly moving into universities playground - several already provide accreditations. Currently they are mostly just for participation. But this is comparable with universities - participating, 'seat time' is in universities also often considered as something that gives students 'deeper understanding' (whatever this is), thus as a base for passing a course, although a correct mark were "Great Listener - constantly gave a nod of assent!")

Many university apparatchik's are doing everything in their power to make sure that universities continue to function the way they have. The notable difference is steady increase in the number of universities non-teaching staff, e.g. in the flagship Stanford University, USA - 3-5% every year [29]) - the first 'Thinkerie with quacks', Socrates, managed to shepherd his disciples to 'earthly summum bonum' alone.

And the academic staff should produce more and more 'bubbls' ('bubblications' - many are very soft inside).

Universities consider the formal learning as the only 'real' learning and want to continue in traditions of the Bologna university dating back to 1088. But students learn more and more in spheres of non-formal, informal learning, acquire skills and appreciation in many (professional) social networks. Universities should find a way to consider all these tendencies.

III. STUDENTS

Today's students are different from what they were five years ago and in the next year they will be even more different. Young people accept changes, especially new technology, much easier than their peers in universities. Today's students are multitasking and collaborative; they're constantly connected and use multiple devices for being connected. They're highly visual and experiential; they're action- and just-in-time-oriented.

They value highly their time, since most of them work, often full-time.

Students are worldwide increasingly participating in work life, e.g. the U.S. Census report indicated that 71 percent of college undergraduates and 82 percent of graduate students were working in 2011 [30]. In Estonia and in Finland, more than two-thirds of students are employed. The official statistics possibly do not indicate the real percentage, since e.g. for an IT student it is very easy to work (e.g. create www-sites) from home or even from computer class at university and not to indicate (for taxation) all sources of income. When the first author asked students in Feb 2014 his group (83 IT specialty first -year students) in the opening lecture, all students indicated, that they work. Even the academic time-tables have been already for several years adjusted for working students - some weekdays are free, but on some days lectures/exercises end at 21.00

Research results indicate, that working does not reduce academic achievements [31],[32], even contrary, working students are more organized, value their time and manage to achieve good results in their studies. This clearly shows in their behavior in lectures and exercises - they never just sit there, they are very active and want to get maximum what teacher can provide. Besides Facebook and Google+ accounts they have accounts in LinkedIn, StackOverflow, Quora, Scribd and in many other professional, topic-oriented social networks.

If they want to solve a problem, learn something, then they ask in on-line professional forums. Some are already answering professional questions posed by others and thus earning the appreciation. And for courses which they just want to pass, they share their essays, presentations,

laboratory exercises, answers to test questions etc in online databases like "Ucionica.hr je banka znanja na internetu" [33] or [34] (the Estonian version).

In these professional social networks they learn a lot and their knowledge in some areas connected with their work exceeds that of their professors in the university.

The generations X (the “lost” generation, some of them still study) and Y (Millennials) [35],[36] grow with Internet and have developed different techniques for solving their problems:

- *googling* .The first thing they do in computer class - they start Google and whatever problem with e.g. programming languages syntax - they google, and usually get necessary piece of knowledge already from the Google's page of results, without opening any page from there. This is a very useful habit and allows them quickly start using new applications - nowadays everything changes very quickly and it is pointless to try to memorize all UI/syntax details.
- *hacking*. They read program code much quicker than older generations (their peers) and grasp instantly the general structure, the information flow of a program - what is imported/input, what is exported/output and understanding of the structure allows them make modifications without even reading the whole text of the program - another very useful technique in our quickly changing world

Role of every kind of non-formal and informal learning and non-academic techniques is increasing and universities have not yet found adequate responses. As a result the millennials increasingly question the value of university education. A PhD students of the first author complained after successful defense, that in employment interviews several reviewers suggested him to remove mention of his PhD degree from his Curriculum Vitae - he looked 'over-educated'. His interests and practice before entering doctorate was management of IT projects. But there were no IT projects waiting for managers - all IT projects were invented (with financial help from the Ministry of Economy or crowd sourcing) to be managed by similarly desperate project managers (and most failed).

IV. Can Technology Replace Teachers ?

Education is a rather expensive sector in every government budget. In Finland expenses for education during the last ten years have been over 13%; in Estonia - around 12% [37]. So it is understandable, that governments want to reduce this item of expenses, suggesting use of more technology in schools, which should make teachers more effective or replace them.

Attempts of using technology for educating masses have long history. Already in 1920's started in USA educational radio broadcasts; later appeared educational TV, the PLATO - Programmed Logic for Automatic Teaching Operations (Automatic Teaching!) and currently MOOC. But all these ventures show very low course completion rates, e.g. the study from University of Pennsylvania revealed, that in average MOOC courses complete only 4% of users [38].

The number of on-line educational materials is growing exponentially. Most of them are short (5-10min) videos.

Most of these videos try to repeat the classroom experience and have been created in traditional style of university lectures - a 'speaking head'- style video, where professor reads chapters from his own textbook, occasionally jotting some hard to understand handwritten characters/figures/schemas on the blackboard (modern professors have already whiteboard).

Video is far less interactive than printed book. With book user can slip pages, look for references, skip uninteresting sections, with video one has to follow everything with video's speed. This breaks the fundamental principle of UI (User Interface) design: user, not the computer, should always have control. Videos cannot have links to outside materials and user can not copy-paste formulae or data from a video.

Video clips exploit very little of human's audio-visual capabilities. Bandwidth of human eye's is ca 10 Mbps (million bits per second) [39], ears - 1.4 Mbps, i.e. we receive every second ca 11.4 Mb of information. Bandwidth of educational video clips e.g. in YouTube is on average only 4500 kbps (everyone can check - right-click on clip and select 'Stats for nerds'), thus these clips engage only a quarter of our input capacity. But a weaker signal creates also weaker perceptions [40].



Figure 1. Sensory (visual+audio) signals are in classroom lecture many times stronger than in educational video from WWW.

Video lectures are for many more difficult to comprehend than the actual textbook which has been used, if there are formulae or data which user should write down, there are references to topics which are not covered in this particular lecture - from video user can not copy-paste and video does not support links. The widely used by universities simultaneous reading the slide text with voice ('speaking head') is useful only for users with reading disabilities, for other users it makes text comprehension more difficult [41], although some research claims (for different age group) the opposite [42].



Figure 2. 'Speaking head' reading the text of 20 slides. User should watch for 10min 47sec, in order to get the data and formulae, user

should pause the video and manually wrote everything down (no copy-paste, no links).

A good on-line study material should be self-contained - it is easy to open another paragraph from book, but rather troublesome to find another video clip explaining missing issues. And the aim should be to explain the topic, not to show the professor.

However, some video clips are really excellent - self-contained, with content menu and using the best features of digital media - animations, sound/voice; examples are e.g. the Introduction to Modern Information Theory [43] from Khan Academy or the winners of the Medical College Admission Test (MCAT) Video Competition, a collaboration between Khan Academy, the Robert Wood Johnson Foundation, and the Association of American Medical Colleges AAMC [44]. As this last example shows, students can make better educational materials than professors - students know, what their fellow students need.

There are many educational tutorials on WWW about nearly everything. They have been created without considering traditions of university lecture, but considering audience - to whom this may be useful - and maximally utilizing latest technical possibilities of modern browsers, e.g. allow to demonstrate interaction of different parts of www-page - html code, JavaScript, CSS [45]. And IT students already use these for learning far more than university lectures - lectures are for many only for getting diploma.

V. Best Practices

The technology is not the lesson presented by a live teacher. The competition in (higher) education is all the time becoming stronger and under the pressure of it, all teachers are seeking better practices in their teaching. The following are suggestions following from our everyday practices.

A. Activate, engage students

Getting students and professors together in a lecture room is an effort which has a cost - both students and professor's time, the lecture room etc. This time should not be wasted to make students passive listeners of oracle, who is (often) reading passages from his own book and who (often) has never tried to implement his recipes in some real business enterprise.

The advancement in communication technology have made such sessions of passive listening utterly senseless. Students could listen/watch these sermons online. The precious time when the whole group and professor are face-to-face should be used actively.

Many studies have shown, that learning is effective only when students are actively engaged in the learning process. Thus the first what lecturer should do is *Engage, activate students, make them active participants, make a lecture/practice active dialog instead of one-sided sermon!*

B. The Flipped classroom

University education is a rare (probably only) field of social activity which has nearly not changed from the times when first universities appeared in 11th..12th centuries - many of them are still active [46].

Proud of longevity and academic traditions professors continue to deliver lectures nearly the same way as did their predecessors during centuries: top-down, deductively - first the lecture explaining theory and then some examples and exercises illustrating the theory. This is the test-oriented method of teaching - tell students exactly what students need to know for the test, nothing more, nothing less.

In order to acquire knowledge and not simply to pass tests gives the inductive, problem-based learning, discovery learning, inquiry learning much better results than deductive [47]. Learning should start from a case, a problem, a story what students can understand. This is the Scahnk's story-based teaching method [48].

An often used tradition is to assign at the end of a lecture readings for home. But usually teachers discover at the start of the next lecture, that nobody has done their readings. Thus instead of continuous crying "How to get them to Do the Readings ?" maybe it is time to ask, whether this tradition should be changed. The essential knowledge, which earlier was assigned as home readings after a lection, should be discussed in class, but in order the discussion to be successful, students should first do some exercises from on-line tutorials, where are assigned on-line materials which present theory.

C. Use on-line materials and video clips

Teachers should use the best examples of the vast and rapidly growing pool of freely (still!) available materials. It has no sense to invent the bicycle or to sit a long evening trying to compile an animation for your next lecture, which will run in couple of seconds. There are already lot of ready material on-line - use it!

D. Hackers style

The most influential figures in modern computing world - Bill Gates, Steve Jobbs, Mark Zuckerberg and others where college drop-outs. Instead of theories they wanted to 'see how this thing works'.

Understanding the general, 'top-down' structure, flow of information in a program or a system is a very valuable ability, which allows quickly to understand new systems and programs (and currently they appear nearly every day). Many students already possess this ability - they see instantly, where is a closing bracket of a block or function, which lines have syntax errors etc. For instance, when the first author showed in class program lines which caused the recent Apple's SSL/TLS bug [49], several students quickly indicated the erroneous line. Many university teachers consider such 'top-down' style as 'non-serious' and even dangerous, but rapid increase of new systems and technologies makes this approach inevitable everyone want first to see how it works, how it can be used and only (if it can be used) - what are the details. The quick advancement of HTML5 is due just to its openness - everyone can investigate, 'how this works'.

E. Give students more choise

In his commencement speech Steve Jobbs described his decision to drop out from college: "... it was one of the

best decisions I ever made. The minute I dropped out I could stop taking the required classes that didn't interest me, and begin dropping in on the ones that looked interesting."

Students, especially working students, are well aware of developments in their field, since companies follow all the time, what is happening in their field - this determines their success or failure. They should have more freedom in selecting courses, for visiting or not visiting lectures and presenting their projects for credits. The rigid division to semesters and sessions is very questionable, e.g. why can teacher assess the achievements of a student only in pre-determined time, during the examination session? The marks go to university's database (the Studies Information System); this system runs like all server databases 7/24, so should also be accessible 7/24. Some universities are already experimenting with more flexible practices, e.g. the Rio Salado College in Arizona allows for a semester 48 start times in a year [50].

VI. Example Case

Some of relevant practices are illustrated in the following with a brief overview of first author's exercise class in theme "3D communication" in spring 2014. Participants were first year bachelor students, this was the semester's first exercise and the exercise was the first step for creating 3D web shop window, where customers could freely manipulate 3D objects.

A. Story

In 2008 appeared in Apple on-line store for iOS applications an app, which showed on screen a glowing gem with some lines of verse ("I'm rich...); the application was described as "a work of art with no hidden function at all" (i.e. just a picture), prize - 999.99 USD. During the first day eight people bought it, then Apple removed the app [51].

The story noticeably enlivened students and they started to create an enormous (1m x 1m x 1m) 3D diamond, which they could expose rotating on their www-sites using the html5/WebGL (the result: [52]).

B. Using on-line tutorials

The 3d model was created with Sketchup; students followed the YouTube tutorial [53].

C. Gamification

The indicated tutorial (and the following steps) could be improved, thus students were encouraged to propose improvements. The 3D model was improved with creating an additional upper bevel with author's initials (not part of on-line tutorial).



Figure 3. Upper bevel and author's initials

Those, who could propose improvements to the tutorial, could implement the additional bevel and remove unnecessary hidden surfaces earned credits - +-points, which will be considered in course final assessment.

D. Hacking

The 3D- model was transformed in Collada format to html5 document. For this was quickly (3-5 minutes) introduced the Javascript WebGl package three.js [54]. The whole package is > 200 MB, therefore students copied to assigned to them hard disk area only some files, first the examples/webgl_loader_collada.html and all files needed to run them in browser (they had to investigate the Javascript import clauses and create a corresponding directory structure). When the monster from webgl_loader_collada.html become alive, they had to create a directory parallel with monster's directory, export there from Sketchup their precious stone in Collada format and replace the monster's import with import of their Collada object. For additional points they had first to replace rotation of the world with rotation of the stone only; better students could also add camera control hacking the file , examples/misc_controls_orbit.html.

After 90 min (end of the class) all students succeeded in creating the rotating stone in html5 page. Those, who did not understand everything, had to consider at home everything from the www-tutorial (the 'flipped' classroom).

VII. Future of Technologies

It is very difficult to predict - changes are all the time accelerating. However, most technologies have up to now followed the Gartner's Technology Hype development stadiums, which illustrates utilization of emerging technologies [55]:

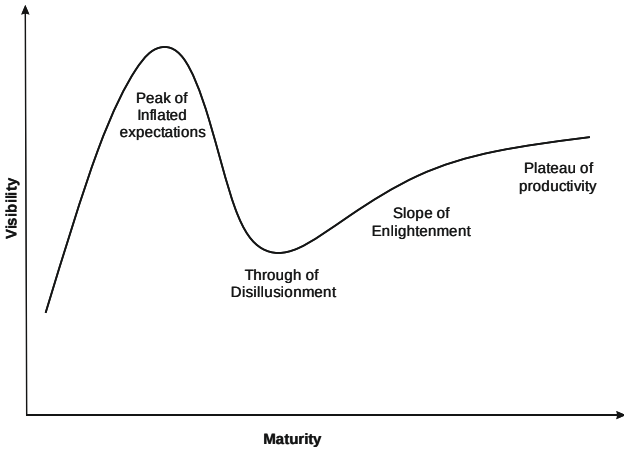


Figure 4. The Technology Hype curve (Gartner [55])

Some of education-connected technologies are currently in the following stadiums(for the full list check [56]):

- **On the Rise**:

  Open Microcredits - free or low-cost way to demonstrate competency in some area; some number of them could be used to obtain a certificate or diploma [57], [58];

  Student Retention CRM - relationship management software, which helps to identify at-risk students

before retention and predicts and reports student behavior to faculty members and administrators; replacing a drop-out student costs twice;

Affective Computing - procedures to influence our emotions and other affective phenomena (could become a part of the previous);

- **At the Peak** (beginning to drift down):

MOOC - the initial enthusiasm around MOOC is cooling down [59], since on university level participants of MOOC courses (even with on-line mentors to interact with students, 1:100) showed much weaker results compared to equivalent live classes on campus [4], thus currently their main application area seems to be professional training for enterprises, but many universities still believe in the bright future of MOOC [60];

BYOD - allowing employees, business partners and other users to use personal devices - laptops, tablets, mobile phones - to execute enterprise applications and access data is inevitable trend, but brings along several dangers: the personal devices are usually connected with social networks, so " by 2017, 40 percent of enterprise contact information will have leaked into Facebook" and in average 20% of personal devices are compromised by malware ;

Big Data - used e.g. for analyzing MOOC performance issues

- **Sliding Into the Trough** :

Social Learning Platform for Education;
Mobile-Learning Smartphones;
Virtual Environments/Virtual Worlds;

- **Climbing the Slope:**

Open-Source Learning Repositories.

VIII. CONCLUSIONS

Universities have two contradictory roles in the world of learning: distributing knowledge, teaching and evaluating their own work with diplomas. Both of these roles - teaching and certificating - got their established forms long time ago, when world was small and closed; in our open, connected world both these functions have lot of pressure for modifications.

The university teaching should embrace all existing tendencies for non-formal, informal education and the new skills of students:

googling - ability to find quickly relevant information from the World biggest data resource - Internet;

hacking - understanding the information flows, the essential structure of a system - an enterprise, (local) government;

multitasking - simultaneous handling of several tasks;

being all the time connected etc

University teaching, distributing new knowledge should better use the rich and constantly renewing source of knowledge - Internet. Physical, i.e. publishing on paper is all the time becoming more expensive and the published materials are always first prepared in digital form - why to bother publishing on paper ? All materials are much easier to distribute and access on Internet. Scientific reports and journals, conference proceedings etc increasingly appear on-line.

People like to teach - creating a tutorial is the best way to understand a topic and new tutorials on newest IT application areas appear in Internet constantly. Most of them are free - the Wikipedia has been created for free. Wikipedia is referential, not designed for learning, but there are already many pedagogically well-designed tutorials covering many areas of IT.

Certification (giving diplomas) may disappear altogether. When diploma is not any more a guarantee for earthly 'summum bonum' and government ministers obtain their diplomas giving them 'a life of ease and pleasure' with copy-paste techniques [61], then - why to bother, spend 4-6 years under extreme strain and then when receiving the diploma discover that (at least) half of your hardly obtained knowledge is already obsolete [62], [63] ?

Microcredentials - digital “badges” [64] - are increasingly respected as a representation of a credential and accepted proof of knowledge, skills or achievements [65]. They digital, thus much more foolproof than classical paper diplomas. Contrary to diplomas, which reflect only the state of (academic) knowledge of the owner at the time of issuing, the microcredentials evolve with e.g. every proper answer in SourceForege or LinkedIn. They are open and public, everyone can support or disprove them.

Some creative minds, self-proclaimed “most experienced organization in professional licensure and certification services” [66] or "Institute for Credentialing Excellence" [67] already claim that they are the "only true source of this kind of information". It is difficult to believe, that this additional item of expenses for obtaining credentials can live long - diplomas are already too expensive. Everything is becoming open. The achievements obtained in open professional networks are easy to check and hard to forge - who needs additional (costly) 'certification services' ?

[1] Recovery. Job Growth and Education Requirements Through 2020. Georgetown University 2014, http://www9.georgetown.edu/grad/gppi/hpi/cew/pdfs/Recovery2020.ES.Web.pdf

[2] Top Technology Predictions for 2013 and Beyond. Gartner Webinars, http://my.gartner.com/portal/server.pt?open=512&objID=202&mode=2&PageID=5553&resId=2232827&ref=Webinar-Calendar

[3] http://www.mastersdegreeonline.org/

[4] Are MOOCs Really A Failure? http://www.forbes.com/sites/susanadams/2013/12/11/are-moocs-really-a-failure/

[5] M.I.T. Game-Changer: Free Online Education For All. http://www.forbes.com/sites/jamesmarshallcrotty/2011/12/21/m-i-t-game-changer-free-online-education-for-all/

[6] Galagan, P. A. (2000), The e-learning revolution. T+D, 54, December, 24–30

[7] M. Sloman. The E-Learning Revolution: How Technology is Driving a New Training Paradigm

[8] S Spencer. An eLearning Revolution, and It's Starting in Rural Africa. http://theinternationaleducator.com/2013/12/19/
[9] https://www.facebook.com/mLearningRevolution
[10] http://www.cloudtweaks.com/2013/09/the-education-revolution-cloud-in-the-classroom/
[11] http://www.timeshighereducation.co.uk/complete-the-digital-learning-revolution-take-to-the-cloud-say-david-roberts-and-blainegreteman/2006498.article
[12] http://www.ncomputing.com/classroom-in-a-box
[13] http://www.einstruction.com/solutions/federal-government-military-and-corporate/classroom-box
[14] http://classics.mit.edu/Aristophanes/clouds.html
[15] http://www.forbes.com/sites/halahtouryalai/2013/05/22/student-loan-problems-one-third-of-millennials-regret-going-to-college/
[16] http://www.ft.com/cms/s/2/ed70d986-5048-11e3-9f0d-00144feabdc0.html#axzz2qBVAV8kZ
[17] http://sandiegofreepress.org/2013/02/todays-college-graduates-in-debt-and-unable-to-find-a-job/
[18] D. M. Kurland, L. C. Kurland. Computer Applications in Education: A Historical Overview. Annual Review of Computer Science, Vol. 2: 317-358 (1987)
[19] J. Shue, J. McKendricks. History of Computers in Education EDUC 515 - Educational Technology & Learning San Diego Regional Center Master's Digital Teaching & Learning Professor: Summer 2009
[20] IBM 2010 Global CEO Study: Creativity Selected as Most Crucial Factor for Future Success. http://www-03.ibm.com/press/us/en/pressrelease/31670.wss
[21] Kyung Hee Kim. The creativity crisis: The decrease in creative thinking scores on the Torrance Tests of Creative Thinking. Creativity Research Journal, 23, (2011), 285-295
[22] http://c-learning.net/
[23] European Higher Education Area (EHEA). http://www.ehea.info/
[24] Patrick Werquin. Terms, Concepts And Models For Analysing The Value Of Recognition Programmes. OECD Directorate For Education, http://www.oecd.org/education/skills-beyond-school/41834711.pdf
[25] Patrick Werquin.. Recognising Non-Formal and Informal Learning: Outcomes, Policies and Practices.: OECD Publishing , 2010
[26] https://www.khanacademy.org/
[27] https://itunes.apple.com/us/app/khan-academy/id469863705?mt=8
[28] http://studentreporter.org/2014/02/higher-education-gets-another-online-challenger-as-wef-launches-its-own-web-courses/
[29] http://www.stanford.edu/dept/pres-provost/irds/ir/analytical_reports/stats_book/2.5_Staff_by_Unit.pdf
[30] United States Census Bureau. New Analyses of Census Bureau Data Examine Nation's 65 and Over Labor Force, Working Students and Changes in Self-Employment. http://www.census.gov/newsroom/releases/archives/american_community_survey_acs/cb13-15.html
[31] Beerkens, M., Mägi, E., Lill, L. University studies as a side job: causes and consequences of massive student employment in Estonia. High Educ (2011) 61, pp 679–692
[32] Pullmann, H.; Realo, A.; Kreegipuu, K.; Allik, J.; Mõttus, R. (2013). Fatigue does not predict academic performance among university students. In "E. Saar; R. Mõttus (Eds.). Higher education in the crossroad: The case of Estonia" Frankfurt am Main: Peter Lang Publishers, pp 92 - 117
[33] Ucionica. http://www.ucionica.hr/
[34] http://www.annaabi.ee/
[35] W.J. Schroer. Generations X,Y,Z and the Others. http://www.socialmarketing.org/newsletter/features/generation1.htm
[36] Ivanova A., Smrikarov A.. The New Generations of Students and the Future of e-Learning in Higher Education. Int. Conf. e-Learning and the Knowledge Society - e-Learning'09, http://www.elearningconf09.comhard.eu/Files/LEARNING_09.pdf
[37] http://www.indexmundi.com/facts/indicators/SE.XPD.TOTL.GB.ZS/compare?country=fi#country=hr:ee:fi
[38] Perna, L., Ruby, A., Boruch, R., Wang,N., Scull, J. Evans, C., Seher Ahmad The Life Cycle of a Million MOOC User. http://www.gse.upenn.edu/pdf/ahead/perna_ruby_boruch_moocs_dec2013.pdf
[39] Penn Researchers Calculate How Much the Eye Tells the Brain. http://www.uphs.upenn.edu/news/News_Releases/jul06/retinput.htm
[40] Esther P. Gardner, John H. Martin. Coding of Sensory Information. http://homepage.psy.utexas.edu/HomePage/Class/Psy394U/Hayhoe/IntroSensoryMotorSystems/week3/Kandel%20Ch%2021,%2022,%2023.pdf
[41] Justin C. W. Debuse, Andrew Hede, Meredith Lawle. Learning efficacy of simultaneous audio and on-screen text in online lectures. Australasian Journal of Educational Technology 2009, 25(5), 748-762
[42] Beck I.I., McKeown M.G, Worthy J..Giving a text voice can improve students understanding. Reading Research Quarterly 30:2, (1995)
[43] https://www.khanacademy.org/math/applied-math/informationtheory/moderninfotheory/v/symbol-rate--information-theory
[44] MCAT Video Competition Winners https://www.khanacademy.org/about/med-competition
[45] http://jsfiddle.net/
[46] List of oldest universities in continuous operation. http://en.wikipedia.org/wiki/List_of_oldest_universities_in_continuous_operation
[47] Felder R.M. Learning And Teaching Styles In Engineering Education. Engr. Education. 78 (7), 674–681 (1988)
[48] R. Scahnk. Every Curriculum Tells a Story. http://www.socraticarts.com/docs/sccwhitepaper.pdf
[49] https://www.imperialviolet.org/2014/02/22/applebug.html
[50] The Chronicle of Higher Education. Academic Calendars Enter a Season of Change. http://chronicle.com/article/article-content/141895/
[51] http://en.wikipedia.org/wiki/I_Am_Rich
[52] http://deepthought.ttu.ee/infoinst/conference/courswre/html5/WebGL/webgl_loader_collada_teemant.html
[53] http://www.youtube.com/watch?v=PLo3L8w9AUc&feature=related
[54] three.js. http://threejs.org/
[55] Gartner. Hype Cycles. http://www.gartner.com/technology/research/methodologies/hype-cycle.jsp
[56] Gartner. Hype Cycle for Education, 2013. https://www.gartner.com/doc/2559615
[57] Mozilla. OpenBadges. http://openbadges.org/
[58] Microcredentials Research Group. http://altcred.org/about/
[59] Lawrence Neal Garcia. The end of MOOCs and the future of education. http://ubyssey.ca/features/information-driveway-moocs-462/
[60] Peter High. For The Largest Not-For-Profit MOOC, edX, Experimentation Is The Path To Innovation. Forbes, http://www.forbes.com/sites/peterhigh/2014/01/20/for-the-largest-not-for-profit-mooc-edx-experimentation-is-the-path-to-innovation/
[61] D. Lukač, M. Čičin-Šain. Tracking the Roots of Plagiarism - Results of the Initial International Survey. Mipro 2013
[62] Philippe Kruchten. The Biological Half-Life of Software Engineering Ideas. IEEE Software, vol. 25, no. 5, pp. 10-11
[63] Robert N. Charette. An Engineering Career: Only a Young Person's Game? IEEE Spectrum, 4 Sep 2013, http://spectrum.ieee.org/riskfactor/computing/it/an-engineering-career-only-a-young-persons-game
[64] Mozilla. Open Badges. http://openbadges.org/
[65] Digital Badges in Higher Education White Paper. https://docs.google.com/document/d/1UqNeLzIu0i0EkiqdJEivIJrqVJ5Afikl7OSKKOzQgI8/edit?pli=1
[66] Professional Examination Service. http://www.proexam.org/
[67] http://www.credentialingexcellence.org/p/cm/ld/fid=1

The evocation of creativity and critical thinking: Analysis of information competencies and behaviour among students of electrical engineering and computing

Dina Vrkić* and Marko Pavlovski**
* University of Zagreb, Faculty of Electrical Engineering and Computing / Central Library,
Unska 3, 10000 Zagreb, Croatia
** University of Zagreb, Faculty of Humanities and Social Sciences / Department of Information and Communication Sciences,
Ivana Lučića 3, 10000 Zagreb, Croatia
dina.vrkic@gmail.com, mpavlovs1987@gmail.com

Abstract - One of the basic needs of contemporary engineering students is to, evoke their individual ability for creative and critical thinking in science, professional or research and development (R&D) field. This ability descends from an individual's level of information competency that, as the basis of lifelong learning, poses several distinctions compared to the level of one's skills in the fields of electrical engineering, computing and information and communication technology (ICT).

Previous researches show that the quality of those skills in the students' practices is markedly high, but the lack of information competencies is still notable.

The aim of this paper is to present the results of the information behavior and information competencies survey conducted among students of the Faculty of Electrical Engineering and Computing (FER) at all undergraduate and graduate levels. The results of the study will be processed to determine the skill level of information competencies and behavior of engineering students.

The outcome will be beneficial in the planning process for information providing strategies in academic libraries in the electrical engineering, computing, information, and communication technology (ICT) fields.

I. INTRODUCTION

Computer and information competence of students of technical disciplines is a component of information literacy. Computer competencies mean that students developed knowledge of information and computer and technical skills connected to their usage. But information competence means understanding necessity of actual and significant information, skills to find the information sources, using the most effective search strategy, estimate information critically and competently, to differ facts from opinions, to estimate the found information creatively, to use received habits on information search for personal purposes and professional activity, understanding information meaning in the development of democratic society, knowledge and following ethical standards in the sphere of information technologies, to operate in search and use of information and finally the skill to share results of one's activity [1].

II. LITERATURE REVIEW

Information behavior is the totality of human behavior in relation to sources and channels of information, including both active and passive information-seeking, and information use [2] or simply: how people need, seek, give and use information in different contexts [3].

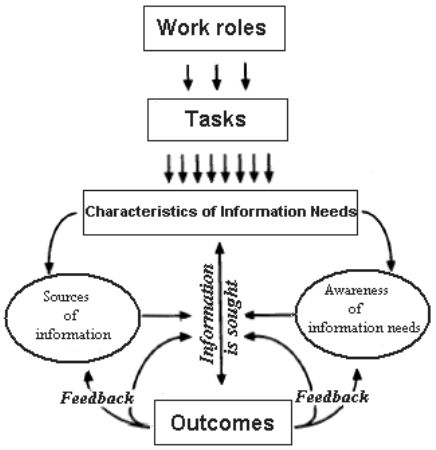
In other way, creativity is connected to engineers who possess a more thorough knowledge of information retrieval strategies and information sources, therefore they should be more efficient in educating themselves and should develop more creative solutions to problems, will practice more efficiently, and will be more competitive in the global economy [4].

Leckie and Fullerton study show that engineering students, as engineers in the academic area, tend to begin to develop serious information retrieval skills later in their program of study [5].

Previous research shows [6], [7] that students do not always examine the citation of the content found to evaluate information accuracy, which indicates that students don't validate the content through the citation given, as long as the information found answers the given academic question [7].

Kerin, Madden and Fulton examined the information-seeking patterns of final year undergraduate engineering students split evenly between two engineering institutes in Irish universities. It was reported that engineering students seem to have a preference for channels that require the least effort, such as the Internet. They explained that the result was due to the fact that student engineers viewed the Internet as a speedy, current information source which fed initial information needs quickly [8].

TABLE I. THE INFORMATION SEEKING OF PROFESSIONALS MODEL[9]



Awareness of information sources, including accessibility, quality, timeliness, trustworthiness, familiarity and previous success, has a direct impact on approaches taken in information seeking. For instance, engineers tend to rely heavily on informal sources of information, such as colleagues [10], [11], [12], and [13].

As Lasić-Lazić et. al. [6] said in a research that current educational systems thrive for student-centered, active, flexible and problem-based learning, which emphasizes the students' independent information seeking.

III. METHODOLOGY

The research method was an online survey (via GoogleForm/GoogleDrive) consisted of thirteen questions. The survey structure was inspired by a survey instrument of research carried out in Project information literacy (PIL) [14] but adjusted to the engineering students and oriented to the previous research conducted in the paper of Lasić-Lazić et. al. [6]. Questionnaire was published on the intranet pages of the News for students of the Faculty of Electrical Engineering and Computing (FER), University of Zagreb in the period from January 27, 2014 until February 7, 2014. The intranet page was selected because it is an instrument to connect information with students via push notification to their official e-mail with the content of the news for the 3516 students at the undergraduate and graduate levels.

The survey was approved by the Faculty Commission for quality management.

The survey was divided into four parts with the structure of the thirteen questions survey. The first, introductory part consisted of multiple-choice questions about types of research assignments they had for the courses they took last year. The second part was about information sources composed by several question types: two 5-array point choice, two close (with one open sub question), one 7-array point choice and one multiple-choice question. The third part was about research approaches with two 7-array point choice and one multiple-choice question. At last, the fourth part was about demographic information.

IV. RESULTS

The response rate includes sample of 75 voluntary students (2.13% students) of which 37 (1.84%) undergraduate, 31 (3.01%) graduate and 7 (2.17%) post-graduate students. The low response rate was due to the setup of the push notification (connected to e-mail) and lack of monitoring notifications on the News for students.

The first question asked concerned the insight on student activity in course-related research. The 46 (26%) of students already wrote seminar papers, 26 (15%) BSc thesis, 27 (15%) some kind of reports, 24 (13%) laboratory notebook, 12 (7%) essay and professional article 10 (6%) students. These results demonstrate the representation of the seminar and professional papers (except the BSc thesis at the undergraduate level). Seminar papers and essays are specific to (higher) education, but the students will continue to produce reports and laboratory notebooks in professional purposes after formal education.

The question related how often did they consult information resources during course-related research process, authors established that undergraduate and graduate students almost always use search engines as their primary source of information. Undergraduate and graduate students equally use encyclopedias, Wikipedia, personal collection of materials, or find their course-related source of information from colleagues. The usage of the library for graduate students is on fifth place and for undergraduate and postgraduate students at sixth.

In the previous question, authors did not include the use of databases because they wanted to evaluate the difference within the results of the previous research [6] conducted on the information competencies of FER students. Therefore, authors proposed it as a separate question. The total number of students who declare to use databases was 27 (36%) and the total number of students that declared they do not use databases was 48 (64%) students. From this large percentage of students at the undergraduate and graduate levels that do not use databases, authors can conclude that students at those levels are not aware of the usefulness of databases.

The problem can be found in the fact that a large number of students, especially at the beginning of their study, rely on textbooks in which they can find the information they need systematically written and have no need to look for sources that broaden the base knowledge for their courses. On the other hand, postgraduate students' results showed increased database use. The curriculum of undergraduate studies represents no courses on the topic of information competencies, while on the graduate level there are several courses that include instruction in the scientific information search, including a course designed from the perspective of library science: „Research, publications and responsibility in science". At the postgraduate level, professors at FER hold the course „Searching scientific information and publication of research results".

TABLE II. USE OF SCHOLARLY PEER-REVIEWED DATABASES

Students	Yes	No
Undergraduate	10 (27,03%)	27 (72,97%)
Graduate	12 (38,71%)	19 (61,29%)
PhD	5 (71,43%)	2 (28,57%)

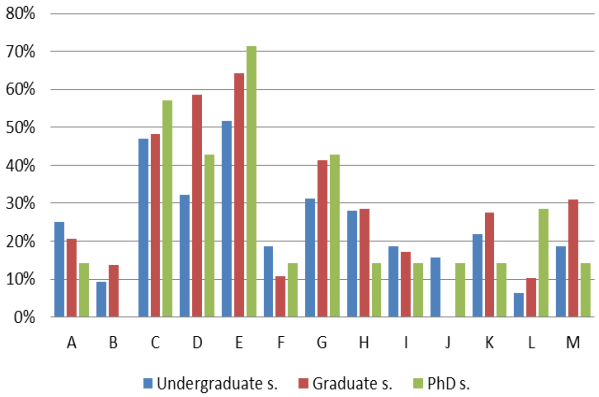
Furthermore, according to the research [6], authors set up a sub-question on a pre-positively answered question, requesting the students to list the options what they use while searching databases. They responded that they usually use advanced search options, search by uncontrolled expression (author-supplied keywords), Boolean operators, thesaurus, ISBN/ISSN, truncation symbols and wildcard characters and at last place DOI. The results form year 2007 and today are almost equal, except for usage of DOI.

In order to further evaluate the problem of students not using databases, authors asked students (of all study levels) if they had difficulties in finding the necessary information sources on the Web. The results indicate that 58 (77%) students do have difficulties, while 17 (23%) students responded that they do not have problems with information search. Students' stated in sub-question regarding the peer-review database search problems that their common problems are related with formatting information queries, determining the relevance and trustworthy of search results. Unfortunately they are not conscious, since the databases are relevant-trusted information source. They also mentioned that they are not familiar with the representation of databases at Faculty and the ways to reach articles (e.g. they do not know that FER has secured access to the IEEE/IET Electronic Library (IEL)).

The evaluation of library sources (books or articles from peer-reviewed databases) varies between the student levels. Undergraduate students almost always consider whether authors have vital information in the charts (e.g. not just attractive graphics), but rarely they consider whether the author/s give credit for someone else's ideas (e.g. footnotes). Graduate students almost always evaluate how current the library source is and also consider whether they have used the library source before. On the other hand, students rarely evaluate whether the library source has a bibliography and who is the publisher. Postgraduate students almost always consider the evaluation of library sources author's credential (e.g. where he/she is faculty or works) and whether the author gives credit for using someone else's ideas. They rarely consider using the library source mentioned by librarian, and whether they have ever heard of it or used it before.

With a notable lack in the use of information competencies, the issue arises: to whom students address for the evaluation of information sources? Undergraduate and graduate students most frequently consult among themselves, while postgraduate students equally ask help from professors and professionals (engineers). Undergraduate and postgraduate students rarely ask librarians for help in evaluating sources, while graduate students sometimes do ask librarians. In the last place, they all have in common demanding evaluation of information sources from friends and family.

TABLE III. COURSE-RELATED ASSIGMENTS (N=32 UNDERGRADUATE STUDENTS, N=29 GRADUATE STUDENTS, N=7 PHD STUDENTS)



A – Start over with new topic, after couple of searches w/out results

B – End research after finding number of citations required

C – Add in own perspective to assignment

D – Create thesis statement early on

E – Develop an outline (draft of paper)

F – Sit down and write, with no plan at all

G – One of the first things I do is to figure out search terms

H – Develop an overall research plan

I – Use system for organizing sources retrieved

J – Use Interlibrary Loan (ILL) for sources

K – Use same set of resources for different assignments

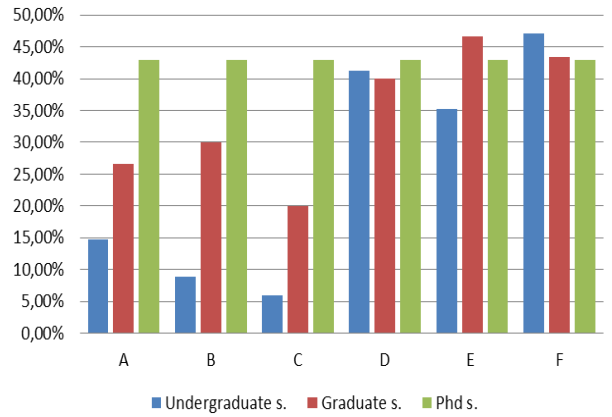
L – Use same topic for different assignments

M – Spend same amount of time on assignments

During course-related assignments, students show that most of them first develop an outline (draft of paper) and work their own perspective into the assignment, so the instructor/professor knows what they think. The results so far did not show significant growth of information competencies, so it is interesting that on the fourth place of the list of habits is that 'one of the first thing they do is to figure out search terms to use'. Undergraduate and graduate students tend to use same set of resources for different assignments, while postgraduate students tend to do it rarely. Students from all three levels of study are not inclined to give up, for example, if they do not find the necessary literature in one or two searches they will not begin with a brand new topic. They do not tend to write about the same topic from one assignment to the next. Undergraduate students rarely use interlibrary loan,

graduates do not use it, while postgraduate students use it sometimes.

TABLE IV. DIFFICULTY SEGMENTS IN COURSE-RELATED RESEARCH RESPONSES OF ″STRONGLY AGREE″ AND ″SOMEWHAT AGREE″ HAVE BEEN COMBINED INTO CATEGORY GENERAL IMPORTANCE. N=32 UNDERGRADUATE STUDENTS, N=29 GRADUATE STUDENTS, N=7 PHD STUDENTS)



A – Creating search terms

B – Finding articles in library databases

C – Finding Web sources

D – Determining credibility of sources

E –Figuring out where to find sources

F – Evaluating sources

As shown in table IV, in course-related research the most difficult tasks for all student levels are to evaluate information sources, finding up-to-date materials and determine credibility of sources. While, on the other hand, students of the undergraduate and graduate levels stated that they do not agree that is difficult to finding Web sources, finding articles in library databases and creating search terms. Postgraduate students are aware of the difficulty, better said, uncertainly of the overall information competencies because they are in a position to use above-mentioned skills on a daily basis.

TABLE V. USAGE OF P RODUCTIVITY TOOLS (N=65) (FROM MOST TO LESS USED)

Rank	Productivity tools
1	Document sharing Web application (e. g. Google Drive)
2	VoIP (e. g. Skype)
3	Highlighting feature on software program
	Digital “sticky notes”
	Forum
4	Wikis
5	Photo sharing sites (e. g. Flickr, Photobucket)
6	Virtual Research Environments
7	Reference manager software (e. g. RefWorks, EndNote, Mendeley)
	Web alerts (e. g. Google Alerts)
	Microblogs (e. g. Twitter)
	Online time management program (e. g. Google Notebook)
8	Social bookmarking (e. g. dig, Delicious)
	Blog (e. g. LiveJournal)
	Other

In the multiple-choice question about usage of productivity tools that students use in course-related research they marked that the most useful tools are the ones for document sharing Web applications (e. g. Google Drive) and VoIP (e. g. Skype), therefore, communication tools. Then, tools for helping them in the better processing of the research papers. Students rarely use tools regarding the reference management, probably, because they are not informed about the existence, offered opportunities and they're widespread of usage in professional and scientific (social) community.

Social Web component (Web 2.0) is the least used among students for course-related assignments.

V. CONCLUSION

Students at all levels of study show that they adjust their information behavior to contemporary media/research setting. For example, after they develop an overall research plan, they search for information sources out on the Web, not in the relevant information sources, e.g. peer-reviewed databases or in library sources. The results showed that the importance of information awareness (information competence) is developing exponentially with each additional level of education.

The library will have to be incorporated into contemporary concepts of supporting the learning process and supporting teachers through curriculum (interdisciplinary/cross-curricular segments). The study indicated that though behaviors can be changed through course design, values linked to learning tend to be much difficult to change.

Key sources for course related information are Google, Wikipedia, and personal collection of printed materials, rarely the library and almost never peer-reviewed databases. Authors can confirm the existence of disproportional usage of evaluating and consulting about information sources: almost never with librarians but usually with colleagues. To overlap the information gap, the library needs to pursue, as authors said above, the implementation of information literacy into the curriculum.

By providing timely information and customized education at each student level, they can gain necessarily required information skills that could extend the time for development of creativity either in information retrieval strategies and information sources either in a professional way.

REFERENCES

[1] A. Zeidmane and S. Cernajeva, “Interdisciplinary approach in engineering education,” in IEEE Engineering Education 2011: Learning Environments and Ecosystems in Engineering Education (EDUCON), Amman, 2011, pp. 1096–1101. Avaliable at http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=5773284 [2013-01-03].

[2] T. D. Wilson, "Models in information behaviour research," Journal of Documentation, 55(3), pp. 249–270, 1999. Available at http://www2.hawaii.edu/~donnab/lis610/TDWilson_Only_1999.pdf [2013-08-12].

[3] K. E. Pettigrew and E. F. L. McKechnie, "The use of theory in information science research," Journal of the American Society for Information Science and Technology, 52(1), pp. 62–73, 2001.

[4] M. England, "Information Literacy for Engineers: the Problem and Its Soulution," in 33rd Aerospace Sciences Meeting and Exhibit, Washington, D.C., American Institute of Aeronautics and Astronautics, 1995.

[5] J. Gloria Leckie and Anne Fullerton. "Information Literacy in Science and Engineering Undergraduate Education: Faculty Attitudes and Pedagogical Practices," College & Research Libraries, 60(1), pp. 9–29, 1996. Avalibale at http://crl.acrl.org/content/60/1/9.full.pdf+html [2013-01-20].

[6] J. Lasić-Lazić, S. Špiranec, M. Banek Zorica and Sanja Žunić, "Discipline–related infromation beahviour research: educational policy implications" in Proc. 30th Jubilee International Convention: Computers in Education = Zbornik radova 30. Jubilarnog međunarodnog skupa MIPRO 2007 : Računala u obrazovanju : savjetovanje, CE, Rijeka, 2007, pp. 21–26.

[7] H. Kahlal, "Information seeking behaviour: a comparison between business and engineering undergraduate students in tertiary environments – research project." Available at http://bp140-portfolio.wikispaces.com/file/view/Final_Research.docx [2013-12-20].

[8] G. Kerin, R. Madden and C. Fulton, "Information-seeking and the students studying for professional careers: The case of engineering and law students in Ireland," Information Research, 10(1), paper 208, 2004. Available at http://www.informationr.net/ir/10-1/paper208.html [2013-12-12].

[9] J. G. Leckie, E. K. Pettigrew and C. Sylvian, "Modeling the information seeking of professional: a general model derived from research on engineers, health care professional and lawyers," Library Quarterly, 66(2), pp. 161–193, 1996.

[10] M. Ward, "A survey of engineers in their information world," Journal of Librarianship and Information Science, 33(4), pp. 168–176, 2001.

[11] T. E. Pinelli, "The information-seeking habits and practices of engineers," In Information seeking and communicating behaviour of scientists and engineers, C. A. Steinke, (Ed.). New York, NY: The Haworth Press, Inc., 1991. pp. 5–25.

[12] E. J. Breton, "Why engineers don't use databases: indexing techniques fail to fit needs of the profession," Bulletin of the American Society for Information Science, 7(6), pp. 20–33, 1981.

[13] J. F. Young and L. C. Harriott, "The changing technical life of engineers," Mechanical Engineering, 10(1), pp. 20–24, 1979.

[14] Project information literacy. Available at http://projectinfolit.org/publications/ [2013-11-10].

Application of Modern Teaching Techniques in the Educational Process

M. Filipović Tretinjak*, A. Bednjanec* and M. Tretinjak*
*School of Electrical Engineering (Elektrotehnička škola), Zagreb, Croatia
martinafilipovic.tretinjak@gmail.com; andrea.bednjanec@gmail.com; mario.tretinjak@gmail.com

Abstract – Today students are using computers, cell phones and other electronic devices to communicate and to access information. New technologies fundamentally changed the way they interact with knowledge so the traditional approach to learning and teaching is no longer sufficient. In order to make the learning environment more interactive and enable students to learn easier and faster, modern teaching techniques need to be applied. There are many different teaching techniques that have emerged from the integration of technology in education such as: flipped teaching, whole brain teaching, gamification, social media, etc. Flipped teaching encourages students to study the course material before class by using readings, pre-recorded video lectures or research assignments. Class time is reserved for active learning strategies. Whole brain teaching helps stimulate both sides of the brain and emphasizes active learning in which students repeat core information and practice basic skills through humor and games. Gamification is a game-based learning technique that allows students learning through the use of games. Nowadays students are always connected to some social network. These students are easily motivated to engage with social media in the classroom. Experience from the School of Electrical Engineering in Zagreb shows that the use of different modern teaching techniques is making learning interesting and teaching more effective.

I. INTRODUCTION

Today students are using computers, cell phones and other electronic devices to communicate and obtain information. The use of technology in the classroom allows students to interact with knowledge in a new way and provided teachers with various online platforms for developing different learning materials and virtual classrooms.

Traditional teaching techniques were based mainly on a teacher explaining a topic and students taking notes. Teaching techniques in today's education should revolve around enhancing students' intrinsic motivation.

In our research we explored whether different modern teaching techniques can help students achieve better exam results and will students become more interested in course topics. Most mentioned methods require computers with access to internet in the classroom. Therefore we have based our research on courses that already have in-class computer-based activities such as: system software, technical documentation, computer networking and internet technologies.

II. FLIPPED TEACHING

Flipped teaching (or flipped classroom), also known as backwards classroom, reverse instruction, flipping the classroom and reverse teaching is a teaching technique in which students are encouraged to study course material online before class using readings, pre-recorded video lectures or research assignments. Class time is reserved for interactive activities, exercises, illustrating concepts, projects, discussions, etc.

Instead of lecturing, teachers offer students personalized guidance through learning materials and more interaction with course topics.

The main objective for implementing a flipped classroom is a change from a passive to an active learning model. Basic differences between traditional and flipped teaching are given in Table 1.

TABLE I. TRADITIONAL VERSUS FLIPPED TEACHING

	Traditional teaching		**Flipped teaching**	
	Teacher	Students	Teacher	Students
Before class	prepares lecture	mostly do no work	prepares reflective web-based lecture materials such as videos, content- rich websites, blogs, quizzes, etc.	complete interactive learning materials, prepare to participate in class activites, gain necessary knowledge or skills
During class	tries to get through all the lecture material, repeats what was in the lecture	listen to a lecture, follow guided instruction, take notes	guides the learning process, gives students support as needed	verify their understanding, extend their learning, practice skills
After class	grades past work	memorize notes, try to do homework	posts additional resources and explanations as necessary	continue applying their knowledge and skills, extend learning to more complex tasks

Four steps to successful flipped classroom implementation:

A. Identify what course content can be reversed

The flipped classroom can be used for flipping parts of each unit or an entire course. Our research has shown that reverse teaching is very useful for courses that already have in-class computer-based activities, such as system software, technical documentation, computer networking and internet technologies.

B. Create engaging class environment

To create an engaging class environment, use different instructional approaches such as peer instruction, team-based learning (TBL), case-based learning (CBL) and process-oriented guided inquiry learning (POGIL). Engage students in learning activities with an assessment. Evaluation tasks will provide feedback for students and teachers about mastered learning outcomes.

Peer instruction is a method by which students prepare for class by doing pre-class reading and answering questions. Answered questions give the teacher feedback about what students found confusing or difficult. For the next lesson, the teacher prepares questions based on the students' responses. During class the students discuss their answers, giving the teacher opportunity to evaluate their progress and to determine whether more explanation is needed before moving on to the next teaching unit.

Team-based learning (TBL) is an instructional approach in which students prepare before class by doing pre-class reading. At the beginning of the lesson students individually and as a team take a quiz or a test about the class topic they prepared for. The teacher estimates the students' comprehensions of the topic; gaps in understanding can be explained through mini-lectures (10 to 15 minutes long videos, presentation, etc.). Then teams engage in various assignments.

Case-based Learning (CBL) is an instructional design model that is quite similar to project-oriented learning. Students work together in small groups (3 or 4 students) analyzing a specific case study provided by the teacher. Case study is a complex, real-life scenario for the social sciences, natural sciences, computer sciences and other disciplines. The students through discussions analyze the given case study and evaluate the given course of action or determine an appropriate course of action on their own. The case study analysis develops the students' analytical thinking.

Process-oriented guided inquiry learning (POGIL) is a student-centered instructional technique in which students work together in small groups (3 or 4 students) on activities that are designed around a learning cycle (exploration, concept invention and application). Students are given a piece of data (graph, map, or image) or information and questions. Students use the questions to explore the given data and generate a conclusion, and then they apply these conclusions in new situations.

The POGIL activities develop the students' critical thinking, problem solving, information processing and team collaboration.

C. Developed and organized course materials

To create a dynamic environment within the flipped classroom, students need to come prepared for class. Define knowledge and skills that students need to acquire during class and then determine what they will need to read or view in advance of the lesson.

Students can prepare for class using traditional materials such as reading materials (textbook chapters or relevant articles) or using online video content (online lectures, simulations or demonstrations of specific procedures).

In order to manage the gathered and created course materials easily, organize them into three groups according to what students need to experience before, during and after the class.

Find the existing online content using websites such as TEDEd, YouTube, Vimeo, SchoolTube, KhanAcademy and OpenCulture. In our research, these websites served as an inspiration for creating our own course material. To create video lectures we used programs like Screencast-O-Matic, Photo Story 3 for Windows and Windows Movie Maker. Course material was shared with students through platforms such as Edmodo, Facebook and Moodle (CARNet).

D. Extend learning after class

In the flipped classroom, students apply course content during the class. In order to help students achieve the expected learning outcomes, teachers assign additional problem tasks (on the blackboard, course website or from the textbook). Also, the teacher can assign additional readings (from the textbook, particular article from a website or newspaper) to further expand concepts discussed in the class.

Our research has shown that flipped teaching offers many advantages and disadvantages like any other teaching technique. The advantages and disadvantages of the flipped classroom are given in Table 2 below.

TABLE II. THE ADVANTAGES AND DISADVANTAGES OF THE FLIPPED CLASSROOM

		Advantages		Disadvantages
Flipped classroom	**Teacher**	can devote more time to helping students, can easily identify students incomprehension	**Teacher**	needs to have access to the required technology, needs to motivate students for completing the pre-class assignment, needs time to collect and create the necessary learning materials
	Students	are able to access learning materials at home, learn and examine course material at their own speed, comprehend better, develop critical thinking	**Students**	need to have access to a computer and internet, spend more time on the computer

III. WHOLE BRAIN TEACHING

Students perceive and assimilate information in different ways. Whole brain teaching is a teaching technique that stimulates both sides of the brain and emphasizes active learning in which students repeat core information and practice basic skills through humor and games.

Seven methods on how to implement the whole brain teaching technique into the classroom:

A. Attention getter: Class-Yes

Class yes is an attention getter that the teacher uses to focus students on what they are going to say next. At the beginning of the class the teacher says "Class!" any way he/she likes and the students imitate the teacher's voice by responding "Yes!". Class-yes is also used during entire class to get students' attention to another short lesson.

B. Classroom Rules

Before teaching a specific topic, the teacher repeats the classroom rules with the entire class to ensure that students understand them.

There are five classroom rules supported by gestures [4]:

- Follow directions quickly! (Make your hand move forward like a fish.)
- Raise your hand for permission to speak. (Raise hand, make a talking motion).
- Raise your hand to leave your seat. (Raise hand; make a walking motion with fingers.)
- Make smart choices! (Tap one finger to your temple as you say each word.)
- Keep your dear teacher happy! (Hold up each thumb and index finger out like an "L" framing your face and smile really big!)

C. Whole brain developer: Teach-OK

At the beginning of each class, the teacher needs to divide the students into pairs. Then he/she presents a short lesson and the students turn to each other to teach the demonstrated short lesson. The teacher observes the students' comprehension and if necessary repeats the lesson. To begin this process the teacher says "Teach!" and the students respond with "OK!". If the students understood subject matter, the teacher moves on to the rule "Class-Yes" and begins another short lesson.

D. Scoreboard Game

In our research we used a scoreboard game based on the Teacher vs Student model.

Students receive a point when they perform the demonstrated short lesson well. The teacher wins a point if the students do not perform well. To avoid the students' disinterest, the difference in scores between the students and the teacher should not be greater than three points.

E. Focuser: Hands and Eyes

Hand and eyes is a focuser that can be used during the lesson. When the teacher wants students to pay special attention on what he/she is saying/doing, he/she needs to say "Hands and Eyes!". Students then fold their hands and closely monitor the teacher.

F. Mirror

In addition to using the focuser Hands and Eyes, the teacher can gain control of the classroom and get students to imitate his/her words and motions using another method called the mirror. When the teacher says "Teach", students respond "OK" and then mimic the teacher's' speech and motions.

G. The whole class involver: Switch!

Switch needs to be used with the teach-ok method. When the teacher says "Switch!", students respond "Switch!" and then change places in teaching each other. This technique ensures that every student in the class is equally engaged in speaking and listening.

Our research has shown that students of all ages wanted to learn and follow the listed techniques. The main problem was that we did not have enough time for implementing the learned methods. We mainly used the whole brain teaching for repeating the subject's topics.

IV. GAMIFICATION

Gamification is a game-based learning technique that allows learning through the use of games. Games can be used along with other teaching techniques and have a great potential in the educational process.

Incorporating game elements into non-game settings makes learning more interactive, increases motivation through engagement, transfers knowledge into skills, allows teachers to track the students' progress easily and to create a more dynamic lesson structure.

Examples of gamification:

A. Add points to assignments and tests

Students collect points through each assignment (homework, project) and test. Here is an example on how to gamify assignments and tests:

- One point for turning in the homework assignment.
- Ten points for turning in the project assignment before the given time limit.
- Two points for every passing grade on the test.
- Five points for every excellent grade on the test.
- Twenty points for every test that was solved without mistakes.

Students' grades depend on the amount of points they have accumulated at the end of the course.

B. *Award students with badges*

Award students with badges for each completed assignment (homework, project) and class attendance.

Example of assigning badges:

- Badge "Perfect attendance" for students that are always present in the class.
- Badge "Participant" for students that actively participate in the class activities.
- Badge "Hard worker" for students that have completed every assignment.
- Badge "Star performer" for students that have earned three excellent grades in a row or solved at least three tests without mistakes.

To assign various badges, websites such as ClassBadges, OpenBadges and Edmodo can be used. To show the best students in the class, teachers can create a leaderboard using a website like Leaderboarded.

Students have quickly accepted this evaluation method. As our research has shown they liked it because they could monitor their learning progress in a visual way.

C. *Organize a class tournament*

In our research we have found that class tournaments stimulate students to learn the course material and practice skills. Students gladly demonstrated their mastered skills; they were excited about seeing themselves on a leaderboard. TopHat and PurposeGames platforms can be used to organize a class tournament.

D. *Integrate educational games in the classroom*

Students can practice learned skills through the use of games. There are a lot of different websites that offer tools for creating various games that can be used in the classroom, such as interactive crossword puzzles, word searches, word scrambles, quizzes and other educational games.

Platforms that can be used as games creation tools are: ProProfs Brain games, PurposeGames, HotPotatoes and SuperTeacherTools.

V. SOCIAL MEDIA

Social media include different number of online technology tools that enable users to create and share information and resources or to communicate easily. Social media can include text, audio, video, images and other multimedia contents.

Here are the most popular social media platforms used in the educational process: Facebook, Twitter, LinkedIn, Google Plus+, Instagram, Pinterest, Wikispaces Classroom, Edmodo, Edublogs, YouTube, TeacherTube, SchoolTube, etc.

There are lots of ways to use social media as a teaching technique considering there are hundreds of social networks available.

Several ways to use social media platforms that have been applied in our research are given below:

A. *Create a virtual classroom*

In our research we have created a virtual classroom on Edmodo. Edmodo is a free website designed for teachers, students and parents. Through Edmodo teachers can share content, educational applications and important information. Students can connect to the classroom, access homework, learning materials and submit assignments even after the school day is over.

In order to observe their child's progress or to discuss with his/her teachers, parents can access the Edmodo website either under their child's username or they may create their own account. Parents can also receive alerts from teachers about school events, missed assignments and other important messages.

Edmodo is an easy to use learning management system similar to Facebook. After creating an Edmodo account, classes for students can be created.

Currently we are using Edmodo for class discussions, placing and viewing students' assignments, uploading digital learning materials and tracking progress of more than two hundred students from 13 to 18 years of age.

Edmodo is very suitable for sharing files and links. For example, the teacher uploads homework or an assignment and sends it to the class and all the students who are in that class automatically have access to the tasks. Students do not have the option to exchange files with each other. Students can send a file directly to the teacher or anyone in the class (including the teacher).

This platform is extremely useful in subjects that teach students to work in specific programs such as Microsoft Office, CADdy++ Electrical, Dev-C++, Photo Story 3 for Windows because teacher has the option to upload videos that explain program features.

Also, Edmodo can be used in all other subjects, so that students from home can review submitted learning materials. Our research has shown that students like to use Edmodo at home and in the classroom.

B. *Create a class account*

Allow students to choose a social media platform such as Facebook, Twitter, Myspace or Instagram and encourage them to create a class account for communication and sharing learning materials.

For example, students can easily upload homework assignments and other tasks, exchange learning materials such as links to videos, blogs, websites and other resources for a particular subject. Absent students can quickly obtain the necessary information. Students learn to find and manage information, create an online community, collaboration and develop skills that prepare them for the use of social media in the workplace.

C. *Provide access to resources*

As our research has shown, students like to have online access to different learning materials. Course materials were available to them through several social

media platforms like Facebook, Edmodo and Moodle (CARNet). Students were accessing materials even after classes, posted their comments and links and explored lessons that they have not learned yet.

By using YouTube and other similar platforms such as TeacherTube and SchoolTube, students can watch videos, discuss content or share created videos with others. If a video demonstrates a procedure that the student should learn (for example create and format a table in Microsoft Word) the student can review it countless number of times until they learn the required skill. These social media sites offer a lot of free resources that can be used for lesson ideas as well as to supplement a lecture during the class.

D. Create your personal learning network (PLN)

Each teacher should create their own personal learning network (PLN) using for example LinkedIn social networking website. Connections with other like-minded teachers can help find required resources for creating course content.

E. Gather content with Pinterest

Pinterest is a social media platform that offers collecting and organizing gathered information through pin boards. With Pinterest you can plan class projects, collect information on a specific class topic, organize collected learning materials and share ideas for the classroom with other teachers.

VI. SEARCH RESULTS

Our research was conducted in the last three years on two hundred randomly selected students from 13 to 18 years of age. Participants were students who attended system software, technical documentation, computer networking and internet technologies classes.

In teaching students we used both traditional and modern teaching techniques. Figure 1 shows the research results. Most of the interviewed students (76%) enjoyed learning in a new way. Students claimed that modern teaching techniques made learning fun and interesting, helped them achieve better results on exams, increased student-teacher interaction, etc.

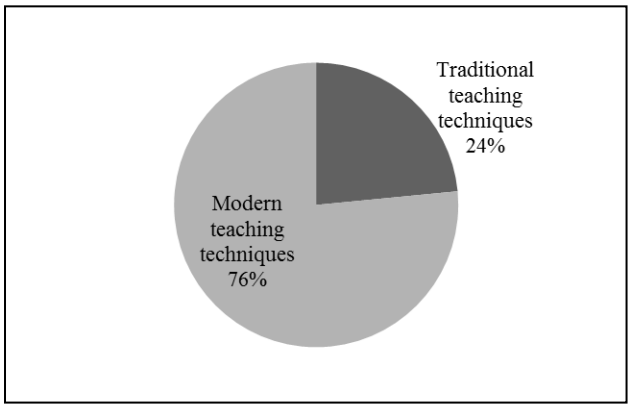


Figure 1: Research on traditional versus modern teaching techniques

Only 24% of the interviewed students claimed that they prefer traditional teaching techniques. When asked, they explain that is because they need to devote more time to completing various assignments during and after class, spend more time in front of the computer, etc. However, they agreed that doing different tasks helped them in acquiring the learning outcomes more quickly.

VII. CONCLUSION

Experience from the School of Electrical Engineering in Zagreb shows that the use of different modern teaching techniques such as flipped teaching, whole brain teaching, gamification and social media are making learning interesting, improve the students' comprehension, enhance the student-teacher interaction and improve critical thinking.

REFERENCES

[1] http://www.knewton.com/flipped-classroom/
[2] http://ctl.utexas.edu/teaching/flipping_a_class
[3] http://www.uq.edu.au/tediteach/flipped-classroom/what-is-fc.html
[4] http://www.wholebrainteaching.com/
[5] http://ashleytipton.weebly.com/whole-brain-teaching.html
[6] http://gamification.org/education
[7] http://www.gamifyingeducation.org/
[8] http://blog.tophat.com/4-ways-to-gamify-learning-in-your-classroom/
[9] http://www.teq.com/blog/2013/10/11-things-learned-social-media-classroom/#.UvZ-92JdUSt
[10] http://mashable.com/2013/08/18/social-media-teachers/

Open access initiative and its impact on use of digital information resources for research and teaching in higher education institutions in Croatia

Radovan Vrana
Faculty of Humanities and Social Sciences,
Department for Information and Communication Sciences,
University of Zagreb, Croatia
rvrana@ffzg.hr

Abstract – Digital information resources have become popular in academic community around the world as their quantity available on the internet increases. This is particularly evident in the case of open access information resources. The idea of open access to scientific information resources promotes use of mostly free of charge scientific information resources thus substituting use of fee based access to scientific information resources. Open access has both advantages and disadvantages. Main advantages would be greater visibility of published research, greater accessibility of scientific content in general, higher rates of citations etc. and disadvantages would be questionable quality of peer review, absence of peer review in some cases, charging authors for publishing their articles, lack of long term preservation policy etc. This paper also presents results of a study of use of digital information resources by students enrolled in two university courses at the Faculty of Humanities and Social Sciences in Zagreb over period of three years. The results showed that students prefer digital over printed information resources in researched two courses and that they are oriented more towards use of different types of free information resources for writing their term papers instead use of fee based information resources.

I. INTRODUCTION

The ongoing technological and information revolution presents us new challenges on a number of fronts on a daily basis. Two such fronts are educational process and research at higher education institutions. It is generally expected that higher education institutions benefit from the implementation of information and communication technology (ICT) in all their activities, especially in teaching and research and that ICT will help us all build a new information society. To achieve this goal, higher education institutions and academic community in general need support in form of quality information resources accessible without limitations i.e. open access information resources. The concept of open access to scientific information resources is related to modern science and its public, collective character as well as its commitment to cooperative inquiry and free sharing of knowledge [1]. Since part of the modern science is publicly funded, the results of the publicly funded research should also be freely available. As universities around the world are currently facing financial problems, they are changing their orientation from use of fee (subscription) based information resources towards open access information resources to provide the desired levels of availability and access of information resources to its users [2] thus helping the idea of open access to scientific information resources to become widely accepted. Without open access, new research would become more difficult to begin and, at the same time, it would introduce inequality in education between students who can afford paying the access to the educational material and those who can't. In the first part, this paper will discuss the idea of open access to scientific information resources and benefits and downsides of their use. In its second part, this paper will present results of the study of use of different categories of information resources in students' term papers in two university courses (one undergraduate course and one graduate course) at the Department of information and communication sciences at the Faculty of Humanities and Social Sciences in Zagreb. The next section of the paper focuses on the development of scholarly communication system which, eventually, led to the advent of open access.

II. DEVELOPMENT OF SCHOLARLY COMMUNICATION SYSTEM

Communication is integral part of science and it is present at all stages of the research process [3]. For centuries, scientific community and universities around the world have been depending on scientific journals for publishing the results of their research i.e. "(...) the diffusion of research results and the public documentation of copyright" [2]. In his article on digital distribution of academic journals and its impact on scholarly communication, Solomon [4] described briefly the history of development of the system of publishing of scientific journals. In the beginning of their existence, scientific journals belonged to scientific societies; they had small circulations and were funded by subsidies and page charges to authors (similarly to current efforts of some

commercial publishers which offer authors to pay a substantial fee to make their articles available in open access). Over time, commercial publishers started taking over scientific journals from scientific societies and increased journal subscription fees to maximise their profit (average journal annual fee has increased at a rate of more than 10% annually [5]) while the development of journals slowed down. During the same period, in the second part of the 20th century governments invested substantial money in research which resulted in growth of number of research articles and scientific journals as well as in increase of their subscription costs which led to scientific publishing crisis of the late 20th century when universities couldn't pay the increased journal subscription fees anymore. Consequently, the diminishing budgets of universities and university libraries initiated the cut backs in number of subscriptions to scientific journals [3] and, at approximately same time, universities initiated search for a solution of the this problem in form of a new scientific publishing paradigm. The next step in history of scientific publishing system development was invention of electronic and later on online journals. With the fast growth of the internet in 1990s the idea of open access as a scientific content publishing and distribution paradigm was born.

III. OPEN ACCESS INFORMATION RESOURCES AND BENEFITS OF THEIR USE

The most cited definition of open access is the one created by the Budapest Open Access Initiative: "By "open access" to this literature, we mean its free availability on the public internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of these articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. The only constraint on reproduction and distribution, and the only role for copyright in this domain, should be to give authors control over the integrity of their work and the right to be properly acknowledged and cited" [6]. In last two decades of their existence, open access publications have gained popularity [7] because they are "(...) free to read or use via the Internet and is supported by new technologies and databases" [8]. For Antelman [4] open access information resources could be potentially cited more by other scientists in comparison to fee based information resources. However, some authors like Anderson [10] claims that in case of journal articles "closed articles will be cited more often – simply because they are assumed to be of higher worth, given the older and more prestigious publications in which they appear". Although the use of open access journals has increased in numbers and in quality [10],[11] they are still not universally accepted. Liesegang [5] points out that open access will be finally accepted in scientific community when it becomes stable, when it produces sustainable and fair online business model which will preserve the scientific record and permits ease of exploration.

The goal of open access is to make the content of scientific journals freely available and that the cost of publication is funded via means other than subscription fees [4]. According to Mittell [12] open access should replace exclusive, limited-access, high-priced resources necessary for research and teaching. Gasparyan, Ayvazyan and Kitas [13] suggest that "for academics, broad access to periodicals, theses, books, and presentations is an indispensable tool for upgrading knowledge base and translating it into the skills of their trainees". For Goetsch [14] "open access helps achieve the fundamental mission of the academy - to disseminate knowledge. Users of that research can apply and build on it, and the public can reap the benefits of critical research more quickly".

Funders of academic institutions and research are especially interested in advancement of open access. They "(…) invest in research in order to accelerate the pace of scientific discovery, encourage innovation, enrich education, and stimulate the economy – to improve the public good"[15]. Open access brings academic institutions greater visibility and clearer management information [16]. The result of acceptance of open access could be not only greater accessibility of research but it could also help in attracting new research, new students and help kick-start new collaborations [17].

In education in general, open access brings the following advantages: contribution to core mission of advancing knowledge; democratization of access across all institutions – regardless of size or budget; provision of access to crucial educational materials; increase of competitiveness of academic institutions; enrichment of the quality of students' education; ensuring access to all that students need to know, rather what they (or their school) can afford and contributes to a better-educated workforce [15].

Students can directly benefit from open access because it facilitates unrestricted access to educational material and enables enriched education [16]. Furthermore, open access can ensure students getting the best possible education and these students are not artificially limited by the selection of information resources (they get access to the latest research findings throughout the world [18]). Open access is used by students for writing their term papers and final theses and they can access open access information resources regardless of their social status. Finally, after student finishes his / hers studying open access information resources will remain available [19] unlike fee based information resources. Open access is important for professors who must access to the latest research, so students' education isn't outdated.

Scientists can benefit from open access in a number of ways. First of all, open access can induce "(...) immediate visibility for research output and thus increased visibility and usage of their results" [16]. In addition to sharing the results of their work in form of published journal articles or books, they can share the research data to „(...) make genuine data accessible and preserve it in the long term" [20]. The research data are then often used in new research as well as in education after publication. Some universities in USA such as Harvard, MIT and Stanford recommended or even required deposit of researchers' works in campus digital repositories [14] making this type

of digital information resources more attractive to scientists for publication of their work and for its long term preservation. By helping scientists to disseminate and preserve scientific knowledge, open access information resources accelerate the rate of scientific progress [21].

Science (in general) and the general public can benefit from open access in the following ways [22]: all articles will become freely and universally accessible online - so an author's work can be read by anyone at no cost; the authors hold copyright for their work and grant anyone the right to reproduce and disseminate the article, provided that it is correctly cited; a copy of the full text of each open access article is permanently archived in digital repositories; authors are assured that their work is disseminated to the widest possible audience; the information available to researchers will not be limited by their library's budget, and the widespread availability of articles will enhance literature searching; the results of publicly funded research will be accessible to all interested readers and not just those with access to a library with a subscription; a country's economy will not influence its scientists' ability to access articles because resource-poor countries (and institutions) will be able to read the same material as wealthier ones, although creating access to the internet is another matter. Open access helps in enhancing and accelerating research cycle [16].

Open access has its downsides too. Some publishers have started to take advantage of open access and have become so called predatory publishers. These are publishers which ask hefty fee from authors to publish their paper after the paper is accepted and published [23]. Warlock [24] also discussed criticism of open access for charging authors or their institutions article processing charges which, then "(...) could create conflicts of interest and have a negative impact on the perceived neutrality of peer review, as there would be a financial incentive for journals to publish more articles". Additionally, some open access journals have a low article acceptance threshold so they accept as many papers as they can to collect money (from authors); they lie about peer review process thus ruining the credibility and reputation of the journal once author discover that there was no real peer review; such journals don't care much about digital preservation [23] and they even accept fake scientific papers [25] which can also destroy confidence in open access journals in general.

Despite these downsides, open access presents an important initiative and so far global academic community has benefited from it significantly mostly because works of countless scientists have become more accessible to wider readership. As these information resources are important to students in writing their term papers and final theses, a study was initiated to find out whether or not group of students at the Department of information and communication sciences at the Faculty of Humanities and Social Sciences in Zagreb (FHSS) are using different types of information resources including open access information resources in two courses they attend (one as a part of the undergraduate study and one as a part of the graduate study of information sciences).

IV. USE OF DIGITAL INFORMATION RESOURCES AT THE FACULTY OF HUMANITIES AND SOCIAL SCIENCES IN ZAGREB

The FHSS is one of the largest academic institutions in Croatia so its policies towards use of digital information resources and its position on open access initiative could accelerate the acceptance of this important initiative in other academic institutions in Croatia. The FHSS has one of the few operational digital repositories in Croatia which stores works written by researchers and students at the FHSS [26]. To help development of this and other open access digital repositories, special attention should be given to students' attitudes towards use of digital information resources in general and open access information resources in particular. The basis for the study in this papers are two researchers done in 2012 and 2013 [27][28] at the FHSS about attitudes of students at towards use of digital information resources and modes of use and preservation of digital information. The first research done in 2012 indicated that the students at this faculty are still oriented towards use of printed information resources in the Faculty's library. However, the same research also showed that students at the FHSS do search for digital versions of scientific information resources both in the FHSS library and on the internet [27] as an addition to use of printed information resources. Another research about use of digital information resources done in 2013 showed that students at the FHSS seek unlimited accessibility and availability of digital content that could be used on a number of different electronic devices they own and use. The same research confirmed that students would love to use digital content without any obstacles but they need someone else (a library) to take care of the long term preservation of the digital content they create during their years of studying [28]. The results of these two researches indicated a slow but definitive shift from use of printed towards digital information resources. Wide implementation of open access in Croatia would certainly help students in their studying as well as development of strategies for support to research in higher education institutions in Croatia. This is especially important in the present time of economic crisis when universities have less money for library acquisitions.

V. STUDY OF USE OF DIGITAL INFORMATION RESOURCES IN TWO UNIVERSITY COURSES

This part of the paper is dedicated to the study of use of digital information resources by students of information sciences at the Department of information and communication sciences at the FHSS in Zagreb enrolled in two courses from 2011 to 2013. The purpose of this study is to establish how many different types of digital information resources are used by students free of charge or for fee (full text databases paid by the Croatian government for use by the members the whole Croatian academic community). Such studies help university teachers to determine trends in use of digital information resources by students in a particular course or a program

of study in order to modify and update reading lists for their university courses. Content analysis was used as the method for analysis of students' term papers. The analyzed content was divided into several categories: journal articles (printed), journal articles (available in full text databases), journal articles (open access journals and freely available journals on the Web (which didn't declare explicitly themselves as open access), content on Web pages of academic institutions, content on Web pages of commercial companies, content on Web pages of newspapers and popular magazines, content on Web pages of Web portals, content on Web pages of miscellaneous organizations, content on private Web pages of the internet users, content on Web pages of private organizations, content on Web pages of government organizations, content on Web pages of encyclopaedias, books (printed), books (online), conference proceedings, university lecture materials and MSc and PhD theses. The study was carried out by analysis of the content of lists of references at the end of students' term papers written in courses "Digital collections" and "Internet culture" from 2011 to 2013 (during three consecutive academic years) and each information resource in the each individual list was categorized. The following number of term papers was analyzed in the course "Digital collections": 18 term papers in 2011/2012, 21 term papers in 2012/2013 and 26 term papers in 2012./2013. The following number of term papers was analyzed in the course "Internet culture": 29 term papers in 2011/2012, 23 term papers in 2012/2013 and 23 term papers in 2013/2014. The next part of the paper presents results of this analysis.

Students in both courses were encouraged to use fee based full text databases which offer peer reviewed scientific and professional content in digital format (access to these full text databases is paid by the Croatian government). Access is possible on computers in the library of the FHSS and its Web page.

The first two tables present the total number of term papers per year in the courses "Digital collections" and "Internet culture"; total number of references in one academic year in all term papers in these two courses and average number of references per term paper in each academic year in these two courses.

TABLE I. NUMBER OF REFERENCES IN TERM PAPERS IN COURSE "DIGITAL COLLECTIONS"

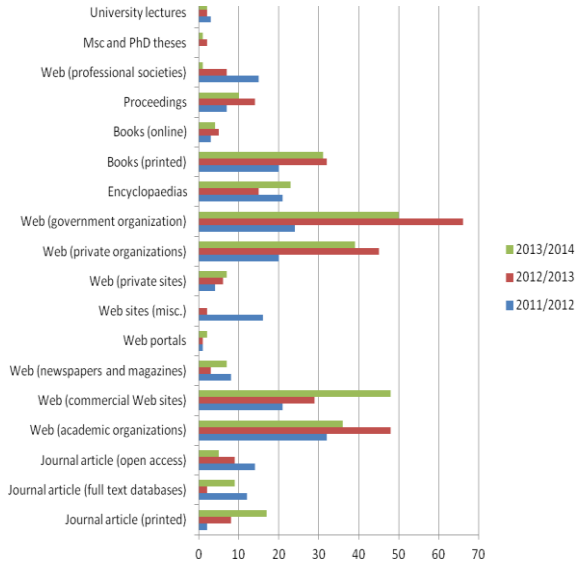
	2011./2012.	2012./2013.	2013./2014.
Total number of term papers per year	18	21	26
Total number of references in all term papers per year	223	296	292
Average number of references per term paper per year	12,38	14,09	11,23

TABLE II. NUMBER OF REFERENCES IN TERM PAPERS IN COURSE "INTERNET CULTURE"

	2011./2012.	2012./2013.	2013./2014.
Total number of term papers per year	29	23	23
Total number of references in all term papers per year	296	205	211
Average number of references per term paper per year	10,20	8,91	9,17

The following two figures present results of analysis of types of information resources listed at the end of each term paper.

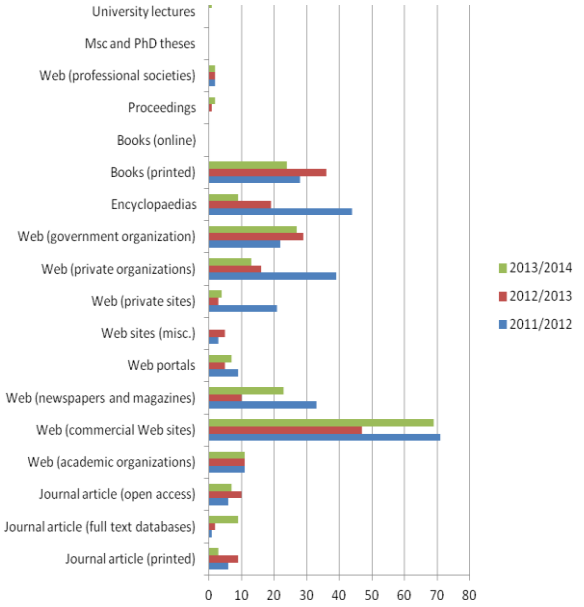
FIGURE I. TYPES OF INFORMATION RESOURCES USED IN COURSE "DIGITAL COLLECTIONS" FROM 2011/2012 TO 2013/2014.



The first figure presents results of use of different information resources in the course "Digital collections". It must be pointed out that the very nature of the course is to teach students different aspects of development, use and management of digital collections and that a significant quantity of content today is published only in digital format. Students are taught to pay attention to the digital nature of the (recently) created human knowledge and its preservation and significant portion of the information resources used in teaching in this course are also in digital format. In this course the following types of information resources were used most frequently by students for writing term papers: Web sites created by different governments (not just the Croatian government), academic institutions, private organizations and commercial Web sites followed by other types. Unlike expected infrequent use of printed journal articles, students in this course use printed books for writing their term papers (books listed in the course reading lists are available in the FHSS library). Additionally, students in this course use online encyclopaedias for writing their term papers which are popular among students [28], however, they are not encouraged to use encyclopaedias

too extensively as they contain abbreviated information and they should read and analyze primary information resources instead.

FIGURE II. TYPES OF INFORMATION RESOURCES USED IN COURSE "INTERNET CULTURE" FROM 2011/2012 TO 2013/2014.



The second figure presents results of use of different types of information resources in the course "Internet culture". In this course, the commercial Web sites are most frequently used as sources of information for writing term papers. Other popular information resources in this course are printed books which are available in the FHSS library, Other popular information resources are Web sites of government organization, followed by encyclopaedias and Web sites of private organizations. Other information resources including printed scientific journal articles and journal article available in full text databases are used less. Journal articles are apostrophized because of their importance in research and teaching and their subscription price. These two types of information resources are published more slowly than articles in online newspapers, popular magazine, on private Web pages etc. and are less attractive for students in this university course. Furthermore, not all of these information resources are scientific; some of them articles in popular magazines and articles newspapers which are simply freely available on the internet.

The analysis showed that students of two university courses has used predominantly digital information in the process of writing term papers while printed information resources were used significantly less during the analyzed period of three consecutive years. The analysis also showed clearly infrequent use of fee based information resources such as journal articles in full text databases. This paper didn't research specific aspects of use of digital informatin resources by students of study of information sciences at the FHSS (like motives for use of specific Web resources) as this will be part of the future reasearch. Nevertheless, it is rather disappointing to see such a low use of this category of digital information resources by students although these resources are highly recommended to them for use in their learning.

VI. CONCLUSION

Today, when digital information resources are essential for research and teaching, one of the biggest problems is access and their unlimited use because a large quantity of these information resources are accessible only on a basis of a paid fee. The fee for a single scientific journal article could be very high if a person paid from his or hers own pocket, so scientific journals are usually made accessible to scientists and students by university libraries. Since universities around the world started to cut back on funds for journal subscription, libraries started to look for other non fee based information resources they could offer to researchers and students. The solution to this problem came in form of open access to information resources, mainly of scientific nature. Academic community was given a number of information resources for use in teaching and research free of charge. However, vast number of journals and books are still unavailable freely and the number of open access information resources has still not grown big enough to replace the fee based resources. As a result, academic community uses both categories of information resources as needed. To see whether students of two university courses (one undergraduate and one graduate course) use open access information resources and to what extent, a study was initiated. The results showed low use of fee based information resources and extensive use of digital information resources available on the internet for free. Some of these information resources are marked as open access and majority is not but still fall under broad definition of open access. The results of the study clearly indicate orientation of students in two analyzed courses towards use of digital information resources and towards free information resources. Additional research would be needed to find out to what extent students at the FHSS are using open access information resources, with what motives and to find out their good and bad experiences in access to free of charge information resources on the internet. Such results would help teaching staff in creating new reading lists for their courses and the FHSS library to select information resources on the internet for use by students. Positive experiences at the FHSS could be shared by other faculties in Croatia in order to enrich teaching and research in their institutions.

REFERENCES

[1] P. David, The historical origins of open science. Retrieved January 4, 2014. from www-siepr.stanford.edu/workp/swp06008.pdf

[2] F. Mann, B. Von Walter, T. Hess, R.T. Wigand, "Open Access Publishing In Science", Communications of the ACM, vol. 52 no. 3, pp. 135-139, 2009.

[3] J.M. Russell, "Scientific communication at the beginning of the twenty-first century", International Social Science Journal, vol. 53 no. 168, pp. 271–282, 2001.

[4] D.J. Solomon, “Digital Distribution of Academic Journals and its Impact on Scholarly Communication: Looking Back After 20 Years”, The Journal of Academic Librarianship, vol. 39 no. 1, pp. 23-28, 2013.

[5] T.J. Liesegang, “The Continued Movement for Open Access to Peer-Reviewed Literature”, American Journal of Ophthalmology, vol. 156 no. 3, pp. 423-432, 2013.

[6] Budapest Open Access Initiative (BOAI) Retrieved December 27, 2013. from http://www.budapestopenaccessinitiative.org/read

[7] Happy 2012 Open Access Movement! Dramatic Growth of Open Access. Retrieved January 25, 2014. from http://poeticeconomics.blogspot.com/2011/12/happy-2012-open-access-movement.html

[8] F. Gu, G. Wide´n-Wulff, “Scholarly communication and possible changes in the context of social media: A Finnish case study”, The Electronic Library, vol. 29 no. 6, pp. 762-776, 2011.

[9] K. Antelman, “Do Open-Access Articles Have a Greater Research Impact?”, College & Research Libraries, vol. 65 no. 5, pp. 372-382, 2004.

[10] T. Anderson, “Open Access Scholarly Publications as OER”, The International Review of Research in Open and Distance Learning, vol. 14 no. 2, pp. 81-95, 2013.

[11] S. Dallmeier-Tiessen, et. al. “Open access journals – what publishers offer, what researchers want”, Information Services & Use, vol. 31, pp. 85-91, 2011.

[12] J. Mittell, “The Real Digital Change Agent”, The Chronicle of Higher Education, Retrieved January 2, 2014. from http://chronicle.com/article/The-Real-Digital-Change-Agent/137589/

[13] A.Y. Gasparyan, L. Ayvazyan, G.D. Kitas, “Open access: changing global science publishing”, Croatian Medical Journal, vol. 54 no. 4, pp. 403–406, 2013.

[14] L.A. Goetsch, “Open access starts with you”, Change: The Magazine of Higher Learning, vol. 42 no. 2, p. 50, 2010.

[15] Why Open Access? Retrieved January 22, 2014. From http://www.sparc.arl.org/resources/open-access/why-oa

[16] Benefits of Open Access. Retrieved 30, 2014. From http://sparceurope.org/open-access/benefits-of-open-access/

[17] The benefits of Open Access. Retrievede 30, 2014. From http://libguides.liv.ac.uk/benefitsofoa

[18] The Case for Open Access. Retrieved January 30, 2014. From http://www.plos.org/open-access/

[19] Why Open Access? Retrieved January 22, 2014. From http://www.righttoresearch.org/learn/whyOA/index.shtml#Students

[20] C. Scraba, “Open access at Romanian universities: benefits and challenges”, Bulletin of the Transilvania University of Brasov, vol. 52 no. 3, pp. 285-292, 2010.

[21] H. Doyle, A. Gass, R. Kennison, “Open access and scientific societies”, PLoS Biology, vol. 2 no. 5, pp. 549-550, 2004.

[22] P.J. Chiao, C. Schmidt, “Open Access gains attention in scholar communication”, Molecular Cancer, vol. 3, p. 23, 2004.

[23] J. Beall, “Predatory publishers are corrupting open access”, Nature Retrieved January 4, 2014. from http://www.nature.com/news/predatory-publishers-are-corrupting-open-access-1.11385

[24] K. Warlock, The Pros and Cons of Open Access, Retrieved January 4, 2014. from

http://www.nature.com/nature/focus/access debate/34.htm

[25] C. Shaw, “Hundreds of open access journals accept fake science paper”, The Guardian, Retrieved January 4, 2014. from http://www.theguardian.com/higher-education-network/2013/oct/04/open-access-journals-fake-paper

[26] OA in Croatia. Retrieved January 4, 2014. From https://www.openaire.eu/es/open-access/country-information/croatia

[27] R. Vrana, “Enriching the learning experience: use of scientific information resources and participation in scientific research in higher education: case of Faculty of humanities and social sciences Zagreb” in Jubilee international convention on information and communication technology, electronics and microelectronics : proceedings, P. Biljanović, Ed. Rijeka : Croatian Society for Information and Communication Technology, Electronics and Microelectronics - MIPRO, 2012, pp. 1539-1545.

[28] R. Vrana, “Modes of use and preservation of digital information resources as a support to the learning process at the Faculty of humanities and social sciences in Zagreb” in Proceedings of the 36th International Convention MIPRO 2013, P. Biljanović, Ed. Rijeka : MIPRO, 2013. pp. 685-690.

The organization of classes and assessment system using Moodle

T. Alajbeg, T. Horvat, T. Novak
Polytechnic of Zagreb, Department of Electrical Engineering, Zagreb, Croatia
trpimir.alajbeg@tvz.hr; thorvat3@tvz.hr; tnovak@tvz.hr

Abstract - The organization of classes and assessment system on the Personal Computer Applications (PCA) course, in the professional study of electrical engineering at the Polytechnic of Zagreb, is based on the Moodle system. Moodle is widely used open source application designed for e-learning. This system enables students to prepare for lectures and exercises, while emphasis is placed on their independent and individual work. Students have access to all the necessary teaching materials, applications, tasks and direct communication with the teacher even outside the class schedule. The aim of the PCA course is to point out the wide range of possible applications of personal computers, but still having in mind specific needs for a study of electrical engineering. This paper presents the organization scheme and content of the course lab exercises with the system of assessment, all based on the mentioned objectives and the use of e- learning system, Moodle.

I. INTRODUCTION

E-learning systems are in many aspects different from the traditional way of teaching both – for the teacher, learning – for the students and assessment –for both the teacher and students. They vary in the availability and transfer of learning (teaching) materials and information, method and time of interaction - communication with the teacher. The manner and speed of assessment and the impact of subjective parameters in the evaluation are also significantly different.

However, through the e-learning system, it is not possible to fully detect the sphere of interest of the students and their motivation. It is very difficult to distinguish and identify the sources of problems encountered in learning, use of instruments and computer applications. Missing component of empathy in communication makes inspiring and motivating students virtually impossible. In these segments direct contact between students and teachers is irreplaceable.

Combining the advantages of both methods of teaching / learning (e-learning and person to person), both modes of interaction and communication between students and teachers in the teaching process, it is possible to achieve a better effect for the mutual benefit.

The system of the teaching and assessment for the Personal Computer Applications (PCA) course, in the professional study of electrical engineering at the Polytechnic of Zagreb, is based on that ideas. Moodle, licensed and free open source application, is used as a e-learning system on the PCA course. Since the aim of the PCA course is to point out the wide range of possible applications of personal computers, the use of Moodle itself, since it is implemented in the course, may be considered as a part of the teaching course.

Laboratory (and construction) exercises are the practical and individual work of each student and make a total of half hourly rate of the course. As the course is at the beginning of the study, the laboratory exercises include the set of general computer applications a future engineer might need. Certainly, laboratory exercises are a complex form of teaching because, in addition to the readiness of students for the subject of a single exercise, their independent work and understanding of the results are expected. Students preparation for each topic is required.

However, availability of computer applications for work / exercise at home has to be considered. This may be achieved by allowing free applications to students, open source applications or demonstration versions of commercial software packages. Thus, wherever possible, the option of choosing an alternative, free tools is given. PCA laboratory exercises are designed with respect to the above requirements and needs. It requires a certain level of teachers' productivity within a predetermined time, therefore automated evaluation is preferred.

II. MANAGING ATTENDANCE RECORDS AND STUDENT'S ACHIEVEMENTS OVERVIEW

Administration tasks related to teaching are a necessity, although time consuming. Also, there is a need for reducing possible errors if not to the zero level, then certainly to a minimum. Administration tasks imply keeping record of each student's attendance during exercises as well as keeping track of the necessary conditions that need to be fulfilled so that a student could enter for an exam. The same refers to students' assessment. Since about 230 students enroll the course each year, this automated management administration shows its importance and advantages. Total time required to execute these types of tasks is reduced and the possibility of oversights, omissions and errors of the teachers are minimized. The ultimate goal of this kind of administration is to have an overview of the achievements, whether cumulative or individual for each student.

The particularity of attendance recording is that some students do not join their group. They attend exercises in one of the other groups, usually with another teacher. With the traditional „pen and paper" method of keeping

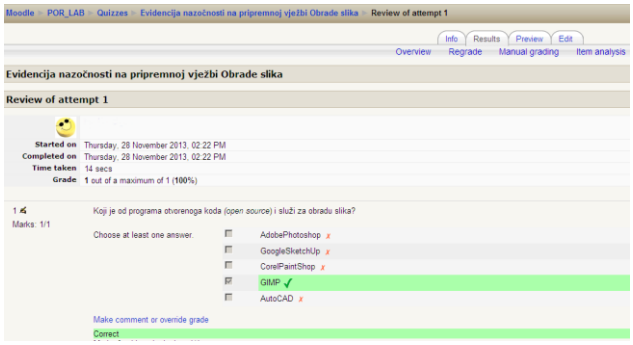


Figure 1 Example of attendance quiz question

records, that would represent additional work and thus an additional waste of time increasing the possibility of error due to a larger number of intermediate steps. Using Moodle capabilities with respect to the particularities mentioned above, records in the laboratory are made in the form of a quiz. All students present during the lab exercise receive a one-time password from the teacher in order to access the attendance quiz. The quiz is a multiple choice test with just one correct answer. Students have up to 2 minutes to submit their answers. When the test is finished, the password is changed to prevent possible distributing passwords and cheating. The only purpose of the quiz is keeping track of students' attendance, without any influence on the assessment issue. The question is simple so the students can easily and quickly choose the correct answer. If the student chooses the right answer, they get 1 point for attendance. Figure 1 shows an example of such a question.

Students answers are stored in the database. At the end of the exercise cycle, the review score clearly shows the attendance records (Figure 2). Of course, there is the possibility that a student has chosen a wrong answer. In that case the test shows 'zero' points, as shown in Figure 2., the teacher can delete that student's attempt from the database and give the student another chance. However, this is not the best solution. A better way for the teacher is to correct the answer manually. If the 'zero' grade is detected later, by the time the test was taken it can be checked i f the student had actually attended. Moodle offers several ways of verification when attendance is concerned, but this method proved to be the most effective.

Given that the administration of students' achievements in laboratory requires a plain and unique review in one place, all the relevant parameters are classified as 'Grades'. Selecting 'Grade' for each student displays the interface shown in Figure 3. The way of assessment of student's work during laboratory exercises will be presented in another chapter. It is possible to export this review in Excel or OpenOffice spreadsheet.

Prisustvo na vježbama			
Evidencija prisutnosti na vježbi "Osnovni rad s operacijskim sustavima"	1.00	0.00–1.00	100.00 %
Evidencija prisutnosti na vježbi "Obrada teksta"	1.00	0.00–1.00	100.00 %
Evidencija prisutnosti na vježbi "Proračunske tablice"	1.00	0.00–1.00	100.00 %
Evidencija prisutnosti na kolokviju vježbi Obrada teksta i Proračunske tablice	1.00	0.00–1.00	100.00 %
Evidencija prisutnosti na vježbi "Dijagram toka"	-	0.00–1.00	-
Evidencija prisutnosti na kolokviju vježbe dijagrama toka	1.00	0.00–1.00	100.00 %
Evidencija prisutnosti na vježbi "Crtanje električne sheme"	1.00	0.00–1.00	100.00 %
Evidencija nazočnosti na pripremnoj vježbi Obrade slika	1.00	0.00–1.00	100.00 %
Evidencija prisutnosti na kolokviju vježbe Obrada slika i Crtanje sheme	1.00	0.00–1.00	100.00 %
Ukupno prisutnost na vježbama	*8*	*0–9*	*89 %*

Figure 2 Automated attendance records for each student

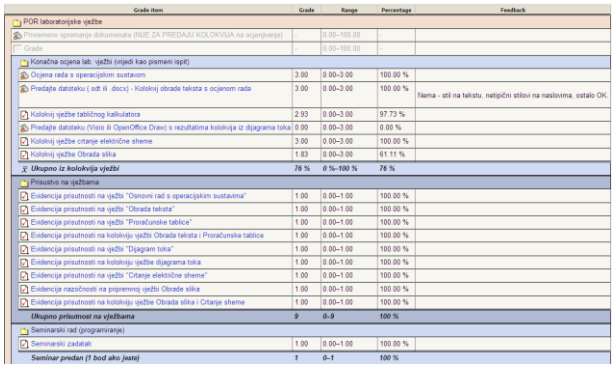


Figure 3 Unique (cumulative) overview of the achievements for each student

III. TOPICS AND PREPARATION

According to the Bologna process, each ECTS credit which a student receives when they pass, the exam includes a certain number of hours of work at home. Students need to master the use of general computer applications at the beginning of the study because they will probably use them in their private and professional lives, regardless of specific applications important in electrical engineering and will be addressed in special courses throughout the study. Taking these facts in account provides students access to those applications on their computers and at home. They can use them to attend lab exercises, practice after the exercises for exams and later in study or at work, preferably without additional investment in buying the commercial software packages. That is enabled by using open source applications or demonstration versions of commercial software packages.

Moodle allows opening a specific topic in advance, making it visible to students. All topics do not have to be visible in advance. Usually visibility is set in a way that students can see earlier themes, current topic and the first next topic. Information concerning the purpose of the exercise with activities and skills that should be mastered, as well as materials are included. Material required for the preparation of exercises can be found in 'Resources'. Materials usually consist of an Internet link to the literature and basic notes on the application software that will be used in the exercises. Informations on how and where a student can download it legally and free of charge, are also included. An example of the above-mentioned for two topics is shown in Figure 4.

Combining the above-mentioned principles and features, laboratory exercises enable students to work with applications that cover the following areas, some of them in a couple of exercises:

A. Basic use of operating system

The operating system used is Microsoft Windows 7. It is free and legally available for students through MS DreamSpark program for members of academic institutions. This exercise places an emphasis on working with files and folders and their organization and retrieval, installation and starting user and system applications, uninstalling programs and Windows components and managing peripherals.

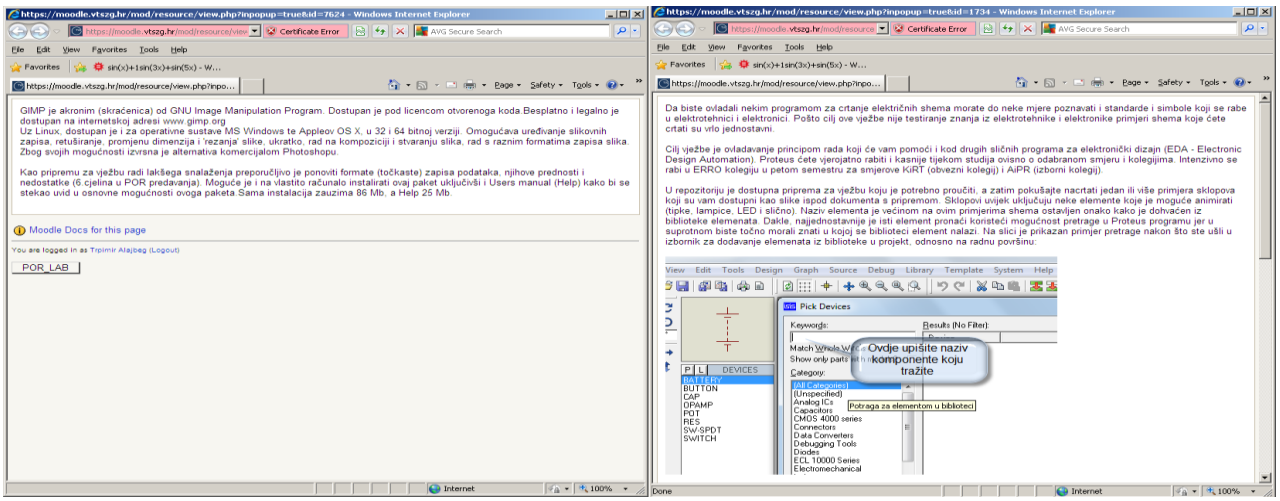


Figure 4 Access to the application and materials for the preparation

B. *Office Suite*

Word and Excel, from MS Office 2013 package, are used. Since this is not a free software package available to students, as an alternative possibility is offered work in the open source Apache OpenOffice package, which also includes word processing and work with spreadsheets. The aim of the text processing exercise is to master basic procedures which includes working with the application and adapting it to the user, creating documents, text entry and search, text formatting and working with objects and preparation for printing. In addition to the basic operation and the adaptation to the user, the goal of working with spreadsheets is to master cells and worksheets, formatting content, working with formulas, functions and graphs.

C. *Flowcharts*

This exercise includes working with Visio, form MS Office 2013 package. It is free and legally available for students through MS DreamSpark program for members of academic institutions. The aim of this exercise is to understand what flowcharts are and how to approach to a task or to solve a problem by using flowcharts.

D. *EDA (Electronic Design Automation) program*

Proteus ISIS package is used. Students can free and legally download demonstration version of the program. The main restriction, in comparison to a full version, is that the clip can not be saved. Of course, there are several more restrictions, but in this course, they are irrelevant. The aim of the exercise is to master the operating principle that can help with later use of other programs for electronic design or with simulation of circuits work.

E. *Image processing*

GIMP application, available under open source licenses, is used. The aim is to become familiar with the basic operations with the computer processing of images (photos). The exercise covers setting the working environment, retrieving, creating and saving documents, sizing images, work with different formats of images, their features and capabilities as well as the most suitable application. The purpose and the opportunity to work with layers and masks, is also included.

IV. ASSESSMENT – TESTS OF KNOWLEDGE AND SKILLS

In this segment are valid starting points listed in Section II. (reducing the number of intermediate steps and the amount of time spent, reducing the possibilities of teachers omissions to a minimum...). An additional requirement is reducing the potential teachers subjectivity in assessing students' works. In near future this influence should be reduced to zero. One possible way is the use of fully automated assessment. Of course, automated assessment is not in all cases the most appropriate and does not appropriately reflect student's level of skill working with a PC application. No less important in the verification of knowledge is to reduce and ultimately prevent the eventual inappropriate actions like copying. Proceeding from the above, the following options are applied on laboratory exercises:

A. *Demonstration*

The traditional method of verification is the students demonstration of work on a computer. This method is applied in Basic operating system exercise. Students present their skills of operating a computer to the teacher. The teacher manually enters the grade in Moodle.

B. *Evaluating submitted works*

For some exercises it's the most appropriate for a student to submit his work to Moodle and the teacher later reviews and evaluates it. This method of assessment is used in the text processing and flowcharts exercises. A module is opened for file submission in Moodle. Students, upon completion of work, submit their documents. The teacher, besides entering the grade (assessment), can enter a comment that will be visible to the student with the grade. Usually, as a comment are entered some notes and observations. This, along with criteria known in advance, makes the grading process more transparent. The student sees what was done well or omitted and can immediately respond in case of teachers oversight.

In this mode of assessment is not possible to use the Safe Exam Browser (details in the following paragraphs) because other (e.g. Word) applications should be opened along with the test. However, inappropriate actions can be made difficult by requiring the understanding of given informations rather than merely stating the facts. Another way is creating a sufficiently large base of questions that are randomly selected, changing the sequence of questions and responses (shuffle within questions, shuffle within answers).

C. Automated assessment

In addition to the significantly reducing the time spent evaluating and the possibility of erroneous results , another benefit is the reduction of the possibility of inappropriate actions during the test. This is made possible by using the Safe Exam Browser from which students access the tests. The window of that web browser opens over the entire screen without any navigational elements. Only permitted actions are solving the test, using calculator and logging off from the computer. Use of this browser disables, among other things, switching to another application, use of copy / paste, use of e-mail client and visiting other Internet sites, such as sites and objects with references to relevant materials.

In this mode of assessment Moodle offers more ways of answering the questions, one of which is in between the other tests used : Multiple Choice (single or multiple answer); Short Answer, Numerical answer (students write in response word, symbols or numbers); Matching (pairing of terms and their meaning or use). After the submission of the test , or when the time expires, students immediately see their score. The score is stored in the database and displayed in an overview as shown in Figure 3.

In this mode of assessment it is possible to analyze students' scores on tests in a manner that identifies which questions were too difficult or too easy. Questions can then be either reformulated or simply removed - deleted from the database without re- creating an existing test. Figure 5, in the lower right part, displays the analysis of the success in answering each question within a test for the current group of students (students of the academic year 2013/2014). There is also a display of successful answering a question that involves students and other groups (previous years).



Figure 5 Average number of points for a particular question

V. CONCLUSION

Basic use of computers is not easy to present to the participants who come from different backgrounds. They had different availability of information technology and the Internet, thus varying degrees of knowledge. Their previous interest in the subject can not be ignored as well. The use of a personal computer can not be reduced only to the use of electronic mail, social networking and sharing photos and video content. These are the reasons why the laboratory exercises of this course were designed in the manner described in the paper. ,

Moodle in being used in PCA lab exercises during the last six years, and the student's feedbacks are very positive. General impression is that using Moodle in combination with direct teacher - student contact is a good method and it is easy to identify where further focus is needed, on what topics and which students. This approach to the organization of teaching requires a good and thorough preparation materials in Moodle, but it constantly improves exercises, making them better for each succeeding generation.

REFERENCES

[1] Cole, Jason; Foster, Helen: Using Moodle, 2nd edition, O'Reilly, 2008

[2] https://moodle.org/

[3] Alajbeg, Trpimir; Ćika, Dražen; Ražov, Tomislav: Using moodle to automatically grade laboratory exercises, MIPRO, 2011

[4] Krznaric, Marija; Buinac, Rade: Can Online Learning Help Croatian ESL Students Achieve Better Fluency, MIPRO 2010

Perceived Quality of Verificator in Teaching Programming

T. Orehovački, D. Radošević and M. Konecki
University of Zagreb, Faculty of Organization and Informatics, Varaždin, Croatia
{tihomir.orehovacki, danijel.radosevic, mladen.konecki}@foi.hr

Abstract – Verificator is an educational tool meant for learning syntax and semantics of the C++ programming language at the introductory programming courses. With an objective to explore quality of Verificator from students' perspective, a pilot study was carried out. Data were collected by means of an online post-use questionnaire. The analysis of collected data uncovered pros and cons of Verificator perceived by novice programmers. Empirical findings are presented and discussed.

I. INTRODUCTION

Drawing on our teaching experiences at different programming courses, an initial version of an educational tool meant for writing code in C++ programming language named as Verificator was developed [1]. The idea behind the development of Verificator emerged from the observations of teachers. Students often tend to adopt some bad programming habits in their attempt to pass the exam. For instance, some students have learned even big fragments of program code by heart, almost without any understanding. There were also cases when students have written large amounts of program code without any syntax or logical testing. The set forth has resulted in a huge number of errors, discouraging students from learning programming. The use of Verificator prevents such bad practice and motivates students to adopt good programming habits.

The evaluation of the quality is an integral part of the life cycle of every software product. According to ISO/IEC 25000 [4], quality is a "degree to which a software product satisfies stated and implied needs when used under specified conditions". In the Software Product Quality Requirements and Evaluation (SQuaRE) series of the international standards [5][6] the quality evaluation framework consists of three models: the product quality model, the quality in use model, and the data quality model. Each of the aforementioned models offers a set of quality characteristics relevant to various groups of stakeholders.

This paper deals with the assessment of the perceived quality of Verificator from students' perspective. The remainder of the paper is structured as follows. Next section provides a brief overview of theories and models that form the background to our research. Features of Verificator are described in the third section. Findings of the pilot study are presented in the fourth section. Conclusions and future research directions are contained in the last section.

II. LITERATURE REVIEW

A. Theoretical background

The constructs that were used and examined in our research are based on several models: Technology Acceptance Model (TAM), Information System Success Model (ISSM), Unified Theory of Acceptance and Use of Technology (UTAUT), Expectation Confirmation Theory (ECT) and Theory of Flow (TOF).

TAM is a theoretical framework which provides insight into the determinants of the adoption and use of the information technology. From its latest version [7], we adopted four different constructs. Perceived ease of use is the degree of ease associated with the use of Verificator. Perceived usefulness represents the extent to which a student believes that using the Verificator will enhance his or hers performance in assignment performance. Self-efficacy denotes the degree to which a student believes that he or she has the ability to perform specific assignment by means of Verificator. Relevance refers to student's perception regarding the degree to which Verificator is suitable for the task at hand.

ISSM was developed as a framework and model for measuring the complex-dependant variable in IS research. For our study, we adapted one construct from its ten-year update [8]. System quality is the extent to which a student perceives the quality of specific aspects of interaction with Verificator.

UTAUT represents a synthesis of eight technology acceptance studies. From its recent extension [9] we adapted one construct. Performance expectancy (in our study added value) is the degree to which the use of Verificator will provide benefits to students.

ECT [10] examines cognitive beliefs and affects influencing one's intention to continue using information systems. Two constructs were adopted from this theory. Confirmation is the degree of fulfilled expectation when student was using the Verificator. Satisfaction refers to a pleasurable or positive emotional state resulting from the use of Verificator.

TOF [14] is a model which examines mental state, involvement and enjoyment of a person in the process of some activity. One construct was adopted from this model. Flow denotes the extent to which a student is fully immersed and enjoys interacting with Verificator.

Apart from the aforementioned, we added three additional constructs that are specific for our research. Helpfulness is the degree to which Verificator helps in better understanding of programming concepts. Good programming habits refer to the extent to which a student can utilize better his or hers programming skills and develop good programming habits when using the Verificator. Feedback represents the degree to which Verificator notifies users about the progress or status of the task at hand through appropriate messages.

B. *Related work*

There are many tools that were developed to help students' education process of computer programming. A brief overview of these tools is offered in [3]. Lane and VanLehn [18] found that tools developed for learning programming can positively influence on students' perception and motivation.

TAM is the most popular model for technology acceptance in the online learning domain [19]. There are a few studies that deal with constructs of TAM in the field of computer programming but they are not specific to a certain learning tool. Orehovački et al. [2] uncovered that perceived usefulness and perceived ease of use have influential role weather students will accept the learning tool and future behaviours towards its use. According to Ramalingam et al. [13], self-efficacy and mental models significantly affect the process of learning programming. Webster et al. [20] emphasize that flow plays a significant role in human-computer interactions. Results of the study conducted by Potosky [17] showed that there is a link between flow and post-training programming efficacy. Previous programming experience can also have significant role in programming education and acceptance of supporting learning tools [12][16]. Finally, findings of several studies indicate that computer anxiety correlates with the students' prediction of their final grade and with perception of their own programming skills and habits [11][15].

The lack of research on the acceptance of tools aimed for teaching programming together with the set forth findings motivated us to initiate a study on the assessment of selected aspects of the perceived quality in the context of the educational software named as Verificator.

III. VERIFICATOR

Verificator[1] is an educational tool developed with the aim to improve the process of teaching programming at the university beginners' level thus helping students to acquire programming skills more easily [1]. The user interface of Verificator is displayed in Figure 1.

After introducing some basic functionalities like preventing copy/paste of program code and obligatory syntax checks within given intervals, over years, Verificator has been enriched with several novel features such as program testing mode, tutor, time control, program structure analysis, etc. The aforementioned functionalities will be explained in more detail in the following sub-sections.

A. *Personalization of programs*

Students are asked to fill out the welcome form with some general information about themselves and the assignment (e.g. name, id, program name and description). These data are included into the program code in the form of comments together with a checksum by means of which it can be examined whether the program code was completely written in Verificator or not.

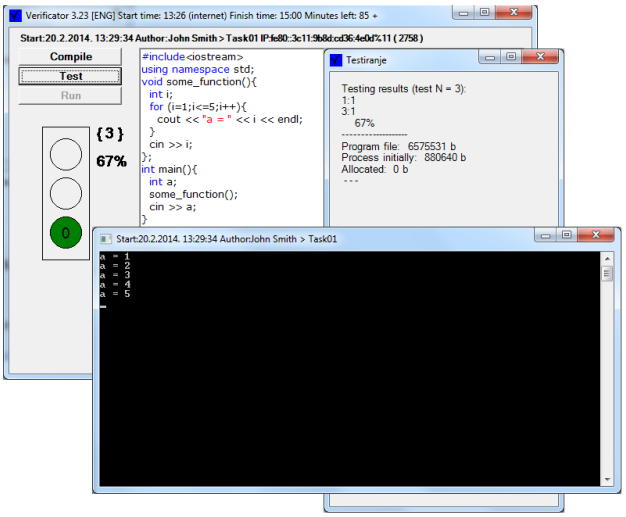


Figure 1. User interface of Verificator

B. *Time limitation for solving the assignment*

Students should solve the assignment within a predefined time interval. To avoid the problem with various system time at different computers, the time is controlled with the use of time servers on the Internet. After the time runs out, a student can still run the program, but he or she cannot modify it.

C. *Prevention of copying programs from colleagues*

Verificator prevents copying program code from external sources. In that respect, the overall code, apart from the given library, must be entered by the student. This feature is connected with obligatory intervals of syntax and logical program testing. Therefore, it is not possible to retype the program code. Instead, it has to be verified within the specified intervals.

D. *Adoption of good programming habits*

Findings of prior studies [1][3] indicate that students who employ standard programming environments to some extent tend to adopt bad programming habits such as learning programming code by heart, programming without syntax and logical checks, etc. Verificator requires from students to check the syntax of their programs after every 10 new lines of code which is graphically illustrated with traffic lights (e.g. red light indicates that is no longer possible to compile the program). Furthermore, program has to be checked logically within the given intervals (expressed in the number of program blocks) in a way that all program blocks have to be reached during the execution.

[1] http://gpml.foi.hr/teaching_tools.html

E. Help for easier detection of syntax and logic errors

Apart from the standard errors and warnings provided by the compiler, Verificator offers further explanation of possible causes of these errors (e.g. unmatched curly brackets or parentheses, incorrect use of certain operators, etc.). In addition, Verificator also contains a tool for the analysis of the program structure as well as the debugging tool.

IV. Pilot Study

A. Procedure and Apparatus

During the winter semester of the academic year 2013/2014 the study participants employed Verificator to solve fifteen diverse lab-based programming assignments. Students also had the opportunity to use Verificator in order to prepare themselves for the lab-based exercises which took place once a week. At the end of the semester, students were asked to complete the post-use questionnaire consisted of 7 items related to their demographic characteristics as well as 50 items meant for evaluating twelve different facets of the perceived quality of Verificator in the context of teaching programming. The responses to the items referring to specific aspects of the perceived quality were modulated on a five-point Likert scale (1 – strongly agree, 5 – strongly disagree).

B. Participants

A total of 121 respondents ranging in age from 19 to 28 years (M = 20.89, SD = 1.315) took part in the study. The sample was composed of 83.47% male and 16.53% female second-year undergraduate students enrolled in the Information Systems study programme. The majority of students (43.80%) perceived their programming skills as good. Similarly, most of the respondents (56.20%) had good knowledge of C++ programming language syntax. Considering the frequency of using the Verificator, 83.47% of students employed it between once and twice a week while 63.64% of them used the aforementioned tool between one and three hours per week. Finally, the majority of participants (84.30%) had up to four years of programming experience.

C. Findings

Responses to the post-use questionnaire items related to the perceived quality of Verificator in teaching programming are presented in Table 1 (see Appendix). Items marked with asterisk are reverse coded. Results of data analysis indicate that majority of pilot study subjects (61.49%) believe that the employment of Verificator contributes to the development of good programming habits. For instance, 63.64% of students reported that Verificator motivates them to be more disciplined while solving the problem-based assignments (M = 2.40, SD = 1.158). In addition, 59.50% of participants stated that Verificator encourages them to write their program solutions more clearly (M = 2.62, SD = 1.349).

The collected data also imply that 54.88% of students perceive Verificator as a helpful tool in the context of teaching programming. While 63.64% of respondents reported that Verificator is advantageous in the interpretation of syntax errors (M = 2.46, SD = 1.133), merely 44.63% of them believe that Verificator explains messages provided by the interpreter in a more intuitive manner (M = 2.74, SD = 1.061).

As can be observed from the responses to questionnaire items, 64.30% of study participants think that messages provided by Verificator are of high quality. Namely, majority of them (73.55%, 62.81%, 62.81%, 62.81%, and 59.80%, respectively) agree that messages displayed by Verificator are useful (M = 2.17, SD = 0.898), understandable (M = 2.42, SD = 0.955), accurate (M = 2.30, SD = 0.823), trustworthy (M = 2.28, SD = 0.839), and clear (M = 2.45, SD = 0.939).

The data gathered from study participants also indicate that 46.28% of students believe that Verificator has sufficient level of system quality. Namely, 69.42% of the respondents are pleased with the responsiveness of Verificator (M = 2.32, SD = 0.896). On the other hand, only 29.75% of them agree that Verificator is stable and reliable in its operation (M = 3.25. SD = 1.113).

Furthermore, the findings of the pilot study imply that Verificator made an impression on only one third (33.33%) of students. More specifically, 40.50% of participants reported that interaction with Verificator was a frustrating experience for them (M = 2.93, SD = 1.209). On contrary, 31.40% of study subjects were satisfied with using the Verificator (M = 3.13, SD = 1.072).

According to the results of data analysis, only 41.05% of students agree that the interaction with Verificator was as they expected. While 48.76% of participants reported that majority of their expectations related to interaction with Verificator were confirmed (M = 2.68, SD = 1.035), 34.71% of them stated that the employment of Verificator has not met their expectations (M = 3.03, SD = 1.016).

As can be observed from the responses to questionnaire items, 30.74% of study participants believe that frequent use of Verificator improves their programming skills. For instance, 45.45% of students think that there is a positive correlation between frequency of using the Verificator and number of points obtained during lab-based sessions (M = 2.93, SD = 1.219). On contrary, 36.36% of participants reported that frequency of using the Verificator has nothing with their efficiency in learning programming concepts (M = 2.78, SD = 1.092).

Considering the results of data analysis, Verificator is a relevant tool for teaching programming. Namely, 46.28% of students think that Verificator is convenient tool for learning programming concepts (M = 2.85, SD = 1.188).

Data displayed in Table 1 imply that 44.30% of study participants perceived Verificator as useful tool for teaching programming. More specifically, 50.41% of students agree that employment of Verificator improves their effectiveness in understanding both syntax and semantics of the C++ programming language (M = 2.78, SD = 1.165). Moreover, 38.02% of respondents believe that they can learn programming concepts more efficiently when they use Verificator than when they employ teaching materials (e.g. textbooks, educational artefacts deployed on LMS, etc.) for the same purpose (M = 2.93, SD = 1.146).

Results of the data analysis also indicate that majority of respondents (66.94%) believe that using the Verificator

does not require a lot of effort. Namely, 76.03% of students reported that is easy to learn to operate the Verificator (M = 2.03, SD = 0.921). Similarly, 60.33% of them stated that is easy to become skilful at using the Verificator (M = 2.33, SD = 1.028).

As can be observed from students' responses, 42.81% of them think that Verificator successfully focuses and holds their attention while making the solution of the problem-based assignment. More specifically, 53.72% of study participants reported that they are completely focused on solving the assignment while they are using the Verificator (M = 2.71, SD = 1.294). Furthermore, 30.58% of them stated that nothing can distract them while they are writing the program code in the Verificator (M = 3.16, SD = 1.162).

Finally, data presented in Table 1 indicate that majority of respondents perceived themselves as self-efficient in the context of using the Verificator. For instance, 8.26% of students believe that they could use Verificator only when someone would instruct them what to do (M = 4.07, SD = 0.941). In addition, 47.93% of study participants think that they could use Verificator even if they had not employed similar tools before (M = 2.73, SD = 1.162).

V. Discussion and Conlusion

The objective of this paper was to explore diverse facets of the perceived quality in use of the educational tool Verificator. For that purpose a pilot study was carried out during which the post-use questionnaire was developed and administered among information systems students. An analysis of collected data revealed strengths and weaknesses of Verificator. The high quality of the displayed messages, the effortlessness in use and the effect on the development of good programming habits are the main Verificator's advantages. For instance, 64.30% of students believe that messages displayed by Verificator are beneficial and constructive. In addition, 61.71% of respondents find Verificator easy to use. Finally, 61.49% of study participants think that Verificator significantly contributes to the development of good programming habits. On the other hand, the low level of the perceived satisfaction, relevance in the context of teaching programming, and confirmation of expectation are major disadvantages of Verificator. More specifically, only 33.33% of students finds the interaction with Verificator pleasant. Besides, 35.26% of study participants think that Verificator is appropriate tool for learning programming. Finally, 36.09% of respondents believe that the use of Verificator has met their expectations.

Considering all the aforementioned, the introduced questionnaire can be employed for measuring the quality of educational tools meant for teaching programming. Moreover, the proposed quality attributes can be used as a background for future research advances in the field. Given that the work presented in this paper is a part of an ongoing research, our future research efforts will be focused on modelling interrelations among determined quality attributes.

References

[1] D. Radošević, T. Orehovački and A. Lovrenčić, "Verificator: Educational Tool for Learning Programming", Informatics in Education, vol. 8, no. 2, pp. 261-280, October 2009.

[2] T. Orehovački, D. Radošević and M. Konecki, "Acceptance of Verificator by Information Science Students", Proceedings of the 34th International Conference on Information Technology Interfaces, V. Lužar-Stiffler, I. Jarec and Z. Bekić, Eds. Cavtat: IEEE, 2012, pp. 223-230.

[3] D. Radošević and T. Orehovački, "An Analysis of Novice Compilation Behavior using Verificator", Proceedings of the 33rd International Conference on Information Technology Interfaces, V. Lužar-Stiffler, I. Jarec and Z. Bekić, Eds. Cavtat: IEEE, 2011, pp. 325-330.

[4] ISO/IEC 25000, "Software Engineering -- Software product Quality Requirements and Evaluation (SQuaRE) -- Guide to SQuaRE", 2005.

[5] ISO/IEC 25010, "Systems and software engineering -- Systems and software Quality Requirements and Evaluation (SQuaRE) -- System and software quality models", 2011.

[6] ISO/IEC 25012, "Software engineering -- Software product Quality Requirements and Evaluation (SQuaRE) -- Data quality model", 2008.

[7] V. Venkatesh and H. Bala, "Technology Acceptance Model 3 and a Research Agenda on Interventions", Decision Sciences, vol. 39, no. 2, pp. 273-315, May 2008.

[8] W. H. DeLone and E. R. McLean, "The DeLone and McLean model of information systems success: a ten-year update", Journal of management information systems, vol. 19, no. 4, pp. 9-30, 2003.

[9] V. Venkatesh, J.Y.L. Thong, X. Xu, "Consumer acceptance and use of information technology: extending the unified theory of acceptance and use of technology", MIS Quarterly, vol. 36, no. 1, pp. 157-178, March 2012.

[10] A Bhattacherjee, "Understanding information systems continuance: an expectation-confirmation model", MIS Q. vol. 25, no. 3, pp. 351-370, 2001.

[11] E. C. Sheeson, "Computer anxiety and perception of task complexity in learning programming-related skills", Computers in Human Behavior, Vol. 21, no. 5, pp. 713-728, 2005.

[12] A. A. Anderson, "Predictors of computer anxiety and performance in information systems", Computers in Human Behavior, vol. 12, no. 1, pp. 61-77, 1996.

[13] V. Ramalingam, D. LaBelle and S. Wiedenbeck, "Self-efficacy and mental models in learning to program", SIGCSE Bull, vol. 36, no. 3, pp. 171-175, 2004.

[14] M. Csikszentmihalyi, M., "Beyond Boredom and Anxiety", San Francisco: Jossey-Bass, 1975.

[15] J. J. Beckers, R. M.J.P. Rikers, H. G. Schmidt, "The influence of computer anxiety on experienced computer users while performing complex computer tasks", Computers in Human Behavior, vol. 22, no. 3, pp. 456-466, 2006.

[16] B. C. Wilson and S. Shrock, "Contributing to success in an introductory computer science course: a study of twelve factors." SIGCSE Bull, vol. 33, no. 1, pp. 184-188, 2001.

[17] D. Potosky, "A field study of computer efficacy beliefs as an outcome of training: the role of computer playfulness, computer knowledge, and performance during training", Computers in Human Behavior, vol. 18, no. 3, pp. 241-255, 2002.

[18] H. C. Lane and K. VanLehn, "Teaching the tacit knowledge of programming to novices with natural language tutoring", Computer Science Education, vol. 15, no. 3, pp. 183-201, 2005.

[19] S. H. Chang, C. H. Chou and J. M. Yang, "The Literature Review of Technology Acceptance Model: A Study of the Bibliometric Distributions", PACIS Proceedings, 2010.

[20] J. Webster, L. K. Trevino, L. Ryan, "The dimensionality and correlates of flow in human-computer interactions", Computers in Human Behavior, vol. 9, no. 4, pp. 411-426, 1993.

APPENDIX

TABLE I. RESPONSES OF STUDENTS TO QUESTIONNAIRE ITEMS RELATED TO THE PERCEIVED QUALITY OF THE VERIFICATOR

Items	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
Good Programming Habits					
Verificator encourages me to prepare myself better for the lab-based exercises.	18.18%	43.80%	19.83%	9.09%	9.09%
Verificator stimulates me to write a program code more clearly.	20.66%	38.84%	14.05%	10.74%	15.70%
Verificator encourages me to be more disciplined while writing a program code.	22.31%	41.32%	17.36%	12.40%	6.61%
Verificator stimulates me to better manage my time during the lab-based exercises.	23.14%	37.19%	21.49%	8.26%	9.92%
Verificator encourages me to think more while solving the problem-based assignment.	15.70%	46.28%	19.83%	10.74%	7.44%
Helpfulness					
Verificator is helping me in the interpretation of syntax errors that occur in the program.	16.53%	47.11%	18.18%	9.92%	8.26%
Verificator is helping me find logical errors that occur in the program.	7.44%	51.24%	16.53%	16.53%	8.26%
Verificator enables me a better insight into the structure of the program.	13.22%	40.50%	23.14%	13.22%	9.92%
Verificator clarifies messages of the interpreter in a more intuitive way.	9.09%	35.54%	36.36%	9.92%	9.09%
Verificator is helping me in debugging the program.	11.57%	42.15%	23.97%	14.88%	7.44%
Feedback					
The messages displayed by Verificator are clear.	12.40%	47.11%	26.45%	11.57%	2.48%
The messages displayed by Verificator are understandable.	12.40%	50.41%	23.14%	10.74%	3.31%
The messages displayed by Verificator are useful.	20.66%	52.89%	17.36%	7.44%	1.65%
The messages displayed by Verificator are accurate.	14.88%	47.93%	30.58%	5.79%	0.83%
The messages displayed by Verificator are trustworthy.	16.53%	46.28%	30.58%	5.79%	0.83%
System Quality					
Verificator has all the necessary functionalities.	8.26%	34.71%	25.62%	22.31%	9.09%
Verificator is stable and reliable in operation.	4.13%	25.62%	25.62%	30.58%	14.05%
Verificator contains various forms of user help.	4.13%	31.40%	35.54%	22.31%	6.61%
The response time of Verificator is acceptable.	12.40%	57.02%	19.01%	9.09%	2.48%
Verificator contains mechanisms that prevent users to make mistakes during the use.	16.53%	37.19%	23.97%	15.70%	6.61%
Satisfaction					
I am pleased with using the Verificator.	3.31%	28.10%	33.88%	21.49%	13.22%
Using the Verificator was a frustrating experience for me.*	13.22%	27.27%	21.49%	28.93%	9.09%
I think that Verificator is an excellent tool for teaching programming.	5.79%	24.79%	34.71%	19.01%	15.70%
Confirmation					
Most of my expectations related to the use of Verificator were confirmed.	9.92%	38.84%	30.58%	14.88%	5.79%
The interaction with Verificator has not met my expectations.*	8.26%	19.83%	37.19%	29.75%	4.96%
I expected the benefits from using the Verificator would be higher than they actually were.*	9.09%	30.58%	35.54%	21.49%	3.31%

* reverse coded items

TABLE I. CONTINUED

Items	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
Added Value					
The students who employ the Verificator more often get higher scores than those who use it infrequently.	8.26%	37.19%	23.97%	14.88%	15.70%
The students who employ the Verificator on a regular basis pass the course earlier than those who use it less frequently.	7.44%	27.27%	35.54%	16.53%	13.22%
The frequency of using the Verificator does not have significant effect on students' efficiency in learning programming concepts.	15.70%	20.66%	38.84%	19.83%	4.96%
The frequency of using the Verificator has significant effect on students' effectiveness in gaining programming competence.	3.31%	23.14%	42.98%	18.18%	12.40%
The frequency of using the Verificator does not have significant effect on students' productivity in solving problem-based assignments.	14.05%	26.45%	37.19%	18.18%	4.13%
Relevance					
Verificator plays an important role in learning programming.	6.61%	18.18%	42.98%	19.83%	12.40%
Verificator is appropriate for solving various problem-based assignments.	5.79%	28.93%	28.93%	23.14%	13.22%
Verificator is suitable for learning programming concepts.	9.92%	36.36%	24.79%	16.53%	12.40%
Perceived Usefulness					
Using the Verificator improves students' efficiency in understanding the programming concepts.	6.61%	38.02%	28.10%	17.36%	9.92%
Using the Verificator enhances students' effectiveness in understanding the syntax and semantics of the programming language.	9.92%	40.50%	22.31%	16.53%	10.74%
Using the Verificator, students can learn programming concepts much faster than when employing teaching materials for the same purpose.	8.26%	29.75%	25.62%	22.31%	14.05%
Using the Verificator, students can understand the syntax and semantics of the programming language much easier than during lectures.	7.44%	34.71%	26.45%	19.83%	11.57%
Verificator is useful for learning programming.	9.92%	36.36%	23.14%	19.01%	11.57%
Perceived Ease of Use					
Learning to operate the Verificator is easy.	29.75%	46.28%	16.53%	5.79%	1.65%
It is easy to get the Verificator does what the user wants.	13.22%	35.54%	21.49%	19.83%	9.92%
It is easy to become skilful at using the Verificator.	22.31%	38.02%	27.27%	9.09%	3.31%
Flow					
Using the Verificator successfully holds my attention for a longer time.	6.61%	36.36%	21.49%	19.83%	15.70%
While using the Verificator, I am completely focused on solving the assignment.	18.18%	35.54%	14.88%	19.83%	11.57%
While using the Verificator, I am not aware of what is happening around me.	12.40%	26.45%	28.10%	19.01%	1405%
While using the Verificator, I do not realize the time elapsed.	20.66%	27.27%	22.31%	19.01%	10.74%
While using the Verificator, nothing can distract me.	8.26%	22.31%	28.10%	28.10%	13.22%
Self-efficacy					
I could employ the Verificator even if I had not used similar educational tool before.	13.22%	34.71%	28.10%	14.05%	9.92%
I could use the Verificator only when there would be someone around to tell me what to do.*	0.83%	7.44%	13.22%	40.50%	38.02%
Verificator cannot be used without consulting the help contents.*	2.48%	15.70%	24.79%	37.19%	19.83%

* reverse coded items

E-learning materials to study properties of liquid crystals at university level and in schools

R. Repnik*, A. Šabeder* and M. Krašna **
* Faculty of Natural Sciences and Mathematics, University of Maribor, Maribor, Slovenia
** Faculty of Arts, University of Maribor, Maribor, Slovenia
robert.repnik@uni-mb.si, aljaz.sabeder@student.um.si, marjan.krasna@uni-mb.si

Abstract – Physicists research numerous interesting physics problems, among them the behaviour of liquid crystals. These problems can be very useful for illustrating general principles of physics, and so we worked to include them in our natural science program. In the frame of a project entitled "Introduction of current research interdisciplinary topic into education - liquid crystals" we prepared static and dynamic components for e-learning materials to highlight some of the key properties of typical liquid crystalline phases and some structures that appear in different geometrical boundaries or in bulk systems. The materials produced, and an analysis of the appropriateness of different types of e-learning materials for different levels of education (including primary education, secondary education, and physics courses for educational and engineering studies) are described in this article.

I. INTRODUCTION

Physics in primary and secondary school is far from being a popular subject. Several reasons influence its popularity but at least one of them is an insufficient number of examples of the applicability of physics. In Slovenia, the physics major topics in the current curriculum are at least a hundred years old and they are not interesting from the point of view of many students. Yet, they need to be a part of curriculum, because without them, contemporary Physics cannot be understood. The integration of modern topics into the physics curriculum in primary and secondary schools can increase students' interest. One of the fundamental motivational factors is the perception of students that the topics they study are contemporary and relevant, and in physics we should pay more attention to these factors. It would be unwise, however, to put topics beyond the comprehension level of the students on the curriculum; therefore, contemporary topics should be carefully chosen.

Liquid crystals are a modern topic that is very suitable for the teaching of physics [1-4]. Students are exposed to numerous applications of liquid crystals in every-day life. When teaching the principles of liquid crystals, it is essential to explain the liquid crystals' phases. Liquid crystals can be demonstrated using mechanical models or with e-learning materials. We can create many different static and dynamic components for e-learning materials about liquid crystals. For this purpose, e-learning materials were constructed to help students visualize the main liquid crystal phases (nematic, smectic-A, smectic-C) and phase transitions. By using suitable arrangements of the dynamic components of e-learning materials, we succeeded in simulating the phase transitions of liquid crystals.

In this article, our E-learning materials dealing with different static and dynamic components are presented. The efficiency of different types of components is discussed, based on our observations. The intended users of our e-learning materials on liquid crystals [5-7] are primary and secondary school students. A large part of the materials presented were developed in the frame of a project entitled Development of Natural Science Competences [8].

II. DIDACTIC ASPECT OF E-LEARNING MATERIALS

For our purpose, we divided e-learning materials concerning liquid crystals into static and dynamic categories, based on the type of components used in the e-learning materials. Static e-learning material contains mostly text in two languages: around 3300 characters of text in the Slovenian language and 4300 characters of text in English. Static e-learning materials also contain graphic elements, for example, figures and equations [5]. The static e-learning material is composed of 5 screens. The first screen gives a short summary of liquid crystals. The second to fourth screens contain theory and mathematics equations pertaining to liquid crystals. The last one is reserved for conclusions and references.

We prepared two types of dynamic e-learning materials. The first type [6] contains around 2000 characters of text, which is around half the amount of text in the static e-learning materials. It also contains graphic materials and audio-visual elements, including video and animations. On the main page we can choose between (1) the theory of liquid crystals, which includes text or animations and (2), video of the main phases of liquid crystals. The second type [7] contains only VRML (Virtual Reality Modelling Language) simulations with short descriptions of the main liquid crystal phases (isotropic, nematic, smectic-A, and smectic-C) and liquid crystal defects.

III. KEY PROPERTIES OF TYPICAL LIQUID CRYSTALS

Liquid crystals (LC) are special a state of matter that has some properties of conventional liquids and others of solid crystals. For instance, a liquid crystal may flow like a liquid, but its molecules may be oriented in a crystal-like way.

LC phases are often referred to as a fourth state of

matter. LC phases are an intermediate state between ordinary liquids and crystals. At low temperatures, a typical LC material forms a solid state in which the centers of mass of the molecules lie on a lattice, and the molecular axes point in certain fixed directions. This solid state is said to have long-range positional and orientational order. At high enough temperatures, the material enters a liquid (isotropic) phase in which molecules have no order. There is some short-range clustering of molecules, but the molecular centers of mass diffuse freely and the molecular axes point in random directions. At intermediate temperatures, several different LC phases can become stable, characterized by orientational, and possibly some partial, degree of positional long-range order. These LC phases can be broadly classified as nematic, cholesteric and smectic (Fig 1).[9]

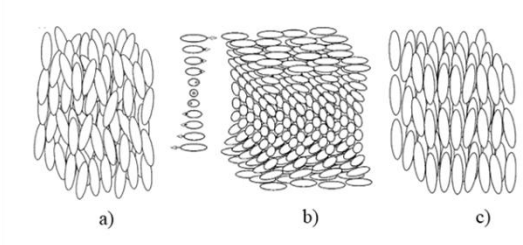


Figure 1: (a) Nematic, (b) cholesteric and (c) "smectic-A" liquid crystalline phases. In the nematic phase molecules tend to be aligned in parallel along a symmetry breaking direction. In the cholesteric phase, the planes with nematic-like ordering (the local nematic director field is marked by arrows on the left side of the figure Fig 1(b)) are twisted with respect to each other and form a helical structure. In smectic phases, molecules are arranged in layers. The various smectic phases differ according to the molecular arrangement within the layers. In the smectic A phase, the molecules tend to be aligned perpendicular to the layer [9].

In the nematic phase, LC molecules tend to be oriented locally in one direction, but they fluctuate around this direction. The local average orientation of molecules is traditionally described by a unit vector field $\vec{n}(\vec{r})$, called the director field. A molecule at the site $\vec{r}$ has its long axis in the average along $\vec{n}(\vec{r})$. Note that in the nematic phase the molecules display what is called "head to tail" invariance. A rotation of an angle π about the axis perpendicular to $\vec{n}(\vec{r})$ (i.e. the transformation $\vec{n}(\vec{r}) \rightarrow -\vec{n}(\vec{r})$) does not influence the physics of the problem. How intensive molecules fluctuate around the average direction is described with the nematic order parameter S:

$$S(\vec{r}) = \left\langle \frac{1}{2}(3(\vec{a}(\vec{r},t)\cdot\vec{n}(\vec{r}))^2 - 1) \right\rangle. \qquad (1)$$

Here $\langle ... \rangle$ indicates the spatial average over a small volume assigned to a "point" $\vec{r}$ in the continuum description and the time average over the "fast time-scales" of the system. An average temporal orientation of a molecule`s axis within the volume is described by the unit vector $\vec{a}(\vec{r},t)$. For randomly fluctuating molecules on fast time-scales, one gets $S(\vec{r}) = 0$. Conversely, $S(\vec{r}) = 1$ reflects perfectly aligned molecules along $\vec{n}(\vec{r})$ [9].

The classification of defects is based on group-theoretical theorems. In fact, group found its first impressive performance in physics in the area of defects. The essential group-theoretical quantity assigned to a defect is the integer topological charge q. This is defined in terms of the molecular orientational ordering enclosing the defect. To predict the potential cancellation of interacting defects, one introduces the winding number, noted as M. This reveals the rotation of $\vec{n}$ on encircling the defect counter-clockwise. Point defects are characterized by integer M, i.e. $M = \pm 1, \pm 2, \pm 3....$ For line defects half-integer values are also allowed $(M = \pm 1/2, \pm 3/2...)$ because of the head to tail invariance. Defects of opposite winding number may annihilate each other. One conventionally refers to the singularities with plus/minus winding signs as defect/anti-defect. [9].

IV. E-LEARNING MATERIALS FOR TEACHING THE PHYSICS OF LIQUID CRYSTALS

An example of static e-learning materials on the topic of liquid crystals is shown in Fig. 2. This e-learning lesson contains a large quantity of text about the theory of liquid crystals and a few pictures of the main liquid crystal phases and liquid crystal defects. It is suitable to give a student with a high level of knowledge of physics a broad perspective. It does not show the typical choice of color combinations because it is written in black and white.

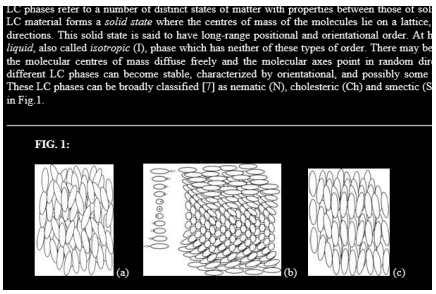


Figure 2: An example of static e-learning material on liquid crystals.

An example of dynamic e-learning material on the topic of liquid crystals, which contains a text about liquid crystals and video records and animations, is shown in Fig. 3. This e-learning lesson contains only a limited amount of text about the theory of liquid crystals and several video recordings and animations about liquid crystal defects. It is suitable for a pupil in secondary school. In these video records and animations, we can choose between different presentations of liquid crystal defects.

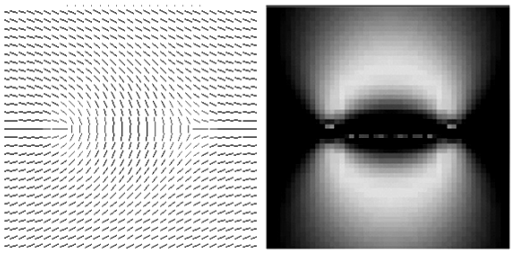


Figure 3: An example of dynamic e-learning material of liquid crystals which contains a video records and animations.

An example of a dynamic e-learning lesson on the topic of liquid crystals that contains VRML simulations of liquid crystals is shown in Fig. 4. This e-learning material contains some text about the theory of liquid crystals and many simulations of phases and defects of liquid crystals. These simulations are suitable for all levels of learning because they allow the student to better imagine the phases and defects of liquid crystals it in three dimensions (3D).

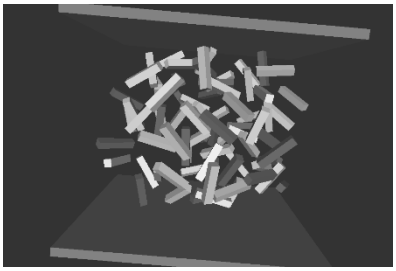


Figure 4: An example of dynamic e-learning material of liquid crystals which contains VRML simulations.

V. Analysis of the Appropriateness of E-Learning Liquid Crystals Materials

Our comparison of static and dynamic components in e-learning materials on the topic of liquid crystals was made with a particular focus on the learning effectiveness of two dimensional (2D) and 3D presentations [10]. The main finding was that a 2D representation (using different cross-sections) of a physical 3D system is preferable for experts in the field since the proper 2D visualization gives more concise information than the 3D view. On the other hand, a 3D active representation (in particular using VRML files) of the system is more appropriate for beginners. Another investigation [11] showed that the combination of frontal work with Information Communication Technology (where the use of VRML files and other visualization tools described in this paper is included) is especially useful for introducing some contemporary physics topics, such as the use of LC, in primary schools.

VI. Conclusion

To ensure the achievement of qualitative, applicative and sustained knowledge, it is important to find efficient ways of teaching, including research into the effectiveness of different teaching tools. Understanding physical phenomena often requires a comprehensive imagination of the course [11]. Visualization software uses two main approaches for presenting three-dimensional (3D) physical systems in two dimensions (2D): 3D virtual simulation (3D view - dynamic E-materials) or 2D cross-sections (2D view - static or dynamic E-materials). We focused on analyzing the appropriateness of E-learning materials for presenting topics related to liquid crystals at different levels of education: for primary and secondary schools and for students of physics in educational and engineering study programs. We found in our comparative study of comprehension of both approaches for two groups of students, those unfamiliar and familiar with LCs, that students with experience in this field prefer the use of 2D views, while other students are significantly worse in interpreting 2D views [10]. In our opinion, there is neither an advantage or disadvantage to using 2D static (web pages with a large amount of information: mostly text, images, formulae etc.) or dynamic E-materials (small amounts of text, more cross-polarizer-simulation videos, cross-section time development videos, etc.). Our experience shows that teaching the topic of liquid crystals at different educational levels is appropriate, modern, and motivational. We have to be careful, however; for students with experience in the field of liquid crystals, the use of 2D static presentations is more appropriate. Dynamic presentations, such as VRML simulations, are more appropriate for students with less or no such experience.

Acknowledgment

Our work was supported by the ARRS research project "Introduction of current research interdisciplinary topic into education – liquid crystals" J5-4002.



References

[1] J. Pavlin, N. Vaupotič, and M. Čepič, Eur. J. Phys. **34**, 745-761 (2013).

[2] K. Susman, J. Pavlin, S. Ziherl, and M. Čepič, Mol. Cryst. Liq. Cryst. **547**, 233-240 (2011).

[3] S. Ziherl *et al*., Mol. Cryst. Liq. Cryst. **547**, 241-248 (2011).

[4] J. Pavlin *et al*., Mol. Cryst. Liq. Cryst. **547**, 255-261 (2011).

[5] Available on: http://lizika.pfmb.uni-mb.si/projekti/kako_deluje_narava/ftk0.htm

[6] Available on: http://www.repnik.com/complex/computer_presentation_halle/index.html

[7] Available on: http://www.repnik.com/complex/tk_vrml/index.htm

[8] National project: Development of natural science competences, No. 3311-08-286011 (2008), Faculty of Natural Sciences and Mathematics, University of Maribor, Supporters: Ministry for education and sports and European Social Fund, available on: http://kompetence.uni-mb.si

[9] R. Repnik, L.Mathelitsch, M. Svetec and S. Kralj, “Physics of defects in nematic liquid crystals,” Eur. J. Phys. **24**, 481-492 (2003).

[10] R. Repnik, M. Cvetko, and I. Gerlič, “Development of Some Natural Science Competences in Undergraduate Study by Training Visualization Skills on Subject Liquid Crystal, Phases and Structures,” Molecular Crystals and Liquid Crystals, 547:1, 249/[1939]-254/[1944] (2011).

[11] R. Repnik, “Successfulness of traditional teaching methods in introduction of contemporary scientific discoveries within physics lessons in primary school,” Faculty of Natural Sciences and Mathematics, University of Maribor, PhD Thesis (2012).

Distance Learning: Challenges, New Solution

Pejman Panahi*, Farnoush Borna**
* Ankara University/Department of Computer Engineering, Ankara, Turkey
** Hacettepe University/Department of IT, Ankara, Turkey
panahi.pejman@gmail.com, borna.farnoosh@gmail.com

Abstract - **Distance learning using innovative use of technologies is becoming a good way in the last ten years to obtain knowledge separate from the traditional attending in the classrooms. The goal of this type of education is to provide maximum possible outcome for students. For developing societies and people who are in specific conditions, this method is a fastest growing form of education.**
The aim of this paper is to talk about recent remote education methods, their challenges and propose a new method to use offline distance learning. Nowadays internet is being widely used all over the world but for some territories the speed of internet is very slow or people don't have access to internet. Satellite communications are the main ways of interconnecting people all around the world. Distance education in terms of satellite communication can be done in two online and offline modes. Online method requires installing and buying special devices and services in which regarding different factors they are not cheap. In this paper we propose of using DVB-S devices to use offline remote learning. Using the proposed method is very useful especially for the regions that there is no good infrastructure for internet.

I. INTRODUCTION

Distance learning or distance education is not a new concept. They are different definitions for distance education. All of them have one thing in common: facilitating learning.

Based on Nipper [1][2] they are 3 generations of distance learning. Print technology is the first generation where the technology of printing press along with postal services is widely used for hundreds of years and known as correspondence learning.

In the second generation, broadcasting power of radio and television is playing an important role. Of course they are some bottlenecks in this way. For example the interaction between instructors and students remains minimal.

Computer and network communications are main elements of third generation. Rapid growth of computer networks and increasing the speed of communications among people especially in terms of internet and other digital technologies shows reduced response times to individual requests for information.

Typically, an activity can be considered as distance learning regarding the following features [2][3]:

- Distance: There is physical distance between the student and the instructor.
- Content: The content of program is provided and managed by an organization.
- Requirements: All requirements of system are well defined.
- Metrics: Measuring of learning output for improving its performance.

Media is an important factor for teaching in distance learning.

They are some basic comprehensions that the instructors collaborating in distance learning should aware of them. This includes the knowledge of using new instructional and communication technologies [10].

Distance learning makes it possible to learn without attending in classrooms. In this way individuals have opportunity to learn without considering boundaries include geography, time and other factors that finally will result in enthusiasm, hope and avoiding skepticism to continue education.

They are variety types of distance learning include [3][4][5]:

- Synchronous: training at the same time in different places. For example: Online chat, forums.
- Asynchronous: transferring information at different times for instance sending an email and replying.
- One-way: broadcasting information to other points, like TV and radio programs and.
- Two-way: Two sides can communicate together.
- Multi-point: many places can receive information at the same time. For example: video conferencing among different points.
- Multi-cast: delivery of information to a group of destination learners in a single transmission from the source at the same time [6].

Good aspects of distance learning are including [7][8]:

- It's a fast and effective way for global training.
- It provides the chance to use the best instructors and high-quality education.

- Information used for training is always up to date.
- Trainees are in charge of their personal success.
- Investment productivity will be increased. Is our society collectively getting the best bang for the bucks we are spending on education? If we can increase productivity we can educate more students to a higher level for the same investment.
- Reducing travel costs instead having more time for learning.
- Scheduling and managing distance learning sessions are easier.
- Adding new learners and instructors can be done easily.
- Learning resources and materials are easy access.

Some negative considerations of distance learning are [4][9]:

- Initial costs for implementation of distance learning courses are high.
- Generally, administrative services are more complex in contrast to live courses.
- Socialization may be reduced due to the reduction of interaction among students and instructors.
- All trainees learning the same, maybe there is only one educational package, lower chance for making mutual dialogue between student and instructor.
- Attending the maximum number of students may provoke numerous abandon because lack of good concomitance of the learning process.

II. REQUIREMENTS

Distance learning can be done by state or private organizations but generally we need the following considerations especially for web based training inside the system:

Administration: for example: unique identification numbers for students, managing databases, tracking trainee's participation.

Course design: using communication tools to share materials, work collaboratively among qualified instructors.

Development of materials: using multimedia along with sharing classic learning methods like e-books can increase the interactivity of materials.

Course delivery: Internet and other communication technologies are good way to distribute materials.

Student support: sending email, designing negotiation forums and availability of online chat to address learner's particular issues.

Assessment: online tests.

Evaluation: interactive communication tools like questionnaire and online feedback [8][10].

III. BASIC IDEA

The main idea behind this plan is to use satellite communications using Digital Video Broadcasting-satellite, second generation or DVB-S2. Satellite communications can be used to broadcast large amounts of data directly to user groups. With the latest advances of high-powered direct broadcast satellites and the changing traffic and usage patterns in the Internet, satellite communications are new ways of delivering data directly to the end user in which become viable alternatives especially to the territories that suffer from lacking of modern technologies like Africa or the speed of exchanging data is low due to different reasons. Figure 1 shows an instance of such networks for sending and receiving purposes.

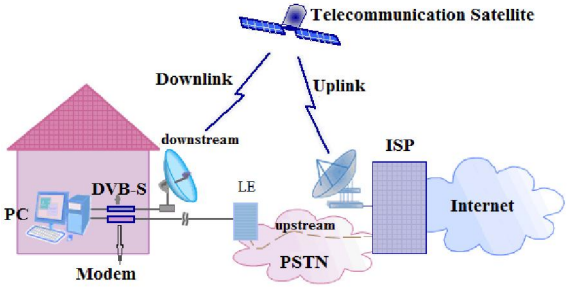


Figure 1. General schema of using satellite internet using DVB-S card (send & receive)

The goal was to deploy this satellite-based caching service for measurements and investigations into its benefits and disadvantages and to develop recommendations for the optimum usage of this technology. Protocols and connection issues were studied to develop an assessment as to the usefulness of incorporating caching into service delivery networks for to enhance their access to information resources available on the Internet.

The primary operational mode of the network augments the school's existing Internet access by providing a mechanism to update all caches within the satellite network with "popular" web objects. The main function of this mode is to inform the gateway cache of every request for web site content coming from any of the schools connected to the test network.

During normal school hours any request from any web browser was first be directed to the cache. The cache will immediately respond by sending the requested objects to the requesting computer. Requests for objects

not found in the cache are forwarded to the main server over the Internet connection, and then responses are stored in the cache and also sent to the requesting computer. These activities are logged in the cache.

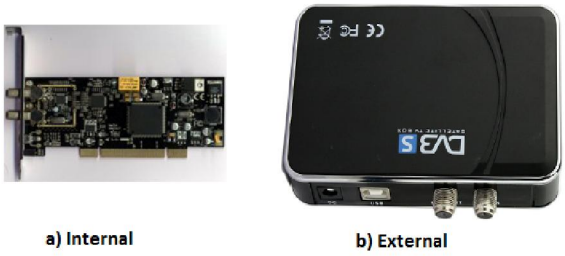


Figure 2. Internal DVB-S vs (b) External DVB-S

All log files are transmitted at regular intervals by phone line or Internet connection to the gateway cache and stored in a "master log" file for analysis. An algorithm called the "intelligence engine" evaluates all files in the master log file, and filters out the "popular objects." The gateway cache then retrieves these objects from the Internet, and adds them to the "popular objects" file. All objects appearing in this file are broadcast at regular intervals over chose satellite to all the receivers and caches in the network [11]. The secondary mode of operation for this network permits teachers to request whole web sites. Each cache has a graphical user interface, which is accessible from any computer in the system with a standard web browser. This graphical user interface (GUI) allows teachers and administrators to directly interface with the cache, to request satellite delivery of in a specific web site pages. The instructor simply types in a list of desired sites. Once a day, this list of sites is sent back to the gateway along with the regular cache log files. The web pages from these sites are treated as high priority items by the gateway intelligence engine, and they bypass the regular filtering process. All these items are simply retrieved from the Internet and added to the list of objects to be transmitted to all that night. This mechanism permits teachers to plan ahead and share information with other teachers in the network [12].

IV. PROPOSED PLAN

As mentioned before using DVB-S2 devices are fundamental of proposed plan. DVB-S2 devices are in two forms of internal and external as shown in Figure 2. Internal cards can be installed on computer main board. External DVB-S devices can connect to computer using usb ports. Generally internal DVB-S cards are cheaper and popular. No especial devices are required to use this method, one dish and LNB, one DVB-S2 and a computer are enough to gain remote education.

The main goal of using DVB-S2 is to watch television and listen to Radio channels, store favorite programs in computer. Along with watching television programs, the user is able to receive offline data from the satellite. They are different software types on the market that installing them will allow user receive offline satellite data like Skynet [13]. The mechanism of receiving data has been explained in prior part and based on the data cached in satellite.

The most important aspect of this method is to managing received data. The user is able to filter types of information that he or she is receiving by adjusting some sample settings in software installed in the computer. As mentioned before using this method there is no need to pay any tuition to gain satellite services since the mentioned services are expensive regarding some factors like bandwidth and etc.

They are several education channels or programs in each communication satellite. Administration is the most important criteria in this method. Several scenarios have been represented in Figure 3.

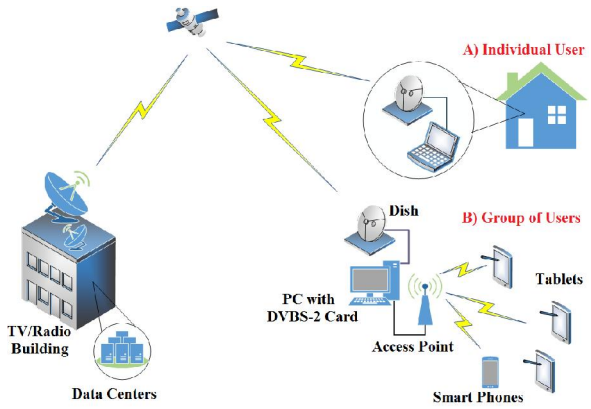


Figure 3. General schema of using DVB-S for distance education

As you can see in Figure 3, accessing to TV/Radio and offline date can be as individual or group of users. Typically in single scenarios, the trainee is able to watch training programs and record them inside computer. If he/she loses the some programs then he/she is able to receive the mentioned programs as multimedia sent and organized by TV/Radio channel through satellite. Other materials can be sent in the same way. For instance pdf's, training PowerPoints and other resources will be sent through satellite link.

In the second scenario there is only one PC equipped with DVB-S device. As it has been shown in Figure 3, this scenario is especially good for campuses, student dorms or rural areas. In this way the pc can be act as server and other devices include tablets, cell phones or any other devices that can receive wireless Wi-Fi signals can gain online education through TV signals or accessing to the materials store on server to study further. If number of trainees is increased or there is delay for receiving data from server, there is probability to install more access points or repeaters to cover more users.

Figure 4, shows another application of DVB-S. This scenario specifically is good for schools or classrooms. In this way students can watch online training programs through big LCD or projector. At the other hand since the data is stored in school server, students are able to watch and review multimedia data's.

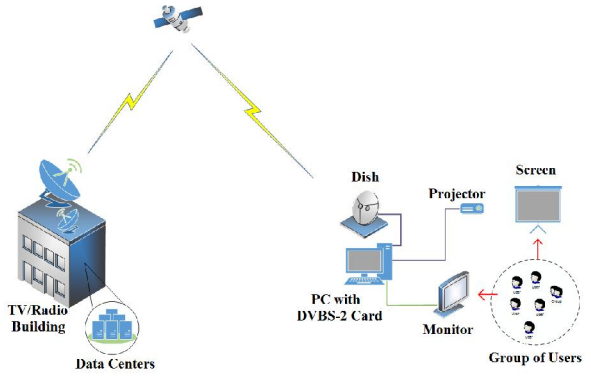


Figure 4. Using DVB-S for group of students

Using DVB-S has a lot of advantages like the act of installing these devices is simple, there is no need to pay for expensive satellite services, receiving dish and LNB's can be found everywhere, the user can chooses the required material but they are negative aspects using this device like communication is single way so the user can not send any request to satellite, large volume of data can be stored in computer therefore filtering and choosing file types is mandatory. However considering all cons and pros for these devices, DVB-S can be a good solution for remote education especially for territories like Africa or developing countries.

V. CONLUSION

The main goal of distance learning is to providing facilities to whom that can't attend in traditional classrooms for different reasons. For example because of financial matters the individual is working and cannot participate in typical educational environment or there are many rural places that are excluded from basic education equipment. Nowadays satellite communications are the most pervasive shape of exchanging data and information in the world. The plan represented here able trainees to get online training as watching TV and storing and receiving offline materials in order to study more using DVB-S devices.

REFERENCES

[1] Nipper, S., Third generation distance learning and computer conferencing. In R. Mason, & A. Kaye (eds), Mindweave:Communication, computers and distance education. 1989, Oxford: Pergamon Press, 63-73.

[2] Eugenia D. Conway, M.A, "Teaching Strategies for Distance Education: Implementing the Seven Principles for Good Practice in Online Education", 5th Annual Science, Engineering & Technology Education Conference, 3 January 2003.

[3] Michael Grahame Moore, William G. Anderson, HANDBOOK OF DISTANCE EDUCATION, Lawrence Erlbaum Associates, Inc., Publishers, 2003.

[4] Strategy for implementing a Distance Learning (DL) process in UNCTAD, for strengthening training capacities in international trade in developing countries, United Nations Conference on Trade and Development, Geneva, Revised January, 2004.

[5] Kim E. Dooley, James R. Lindner and Larry M, "ADVANCED METHODS IN DISTANCE EDUCATION: Applications and Practices for Educators, Administrators and Learners", Turkish Online Journal of Distance Education-TOJDE, July 2006 ISSN 1302-6488, Volume: 7 Number: 4.

[6] Merrill, D., "A Task-Centered Instructional Strategy", Journal of Research on Technology in Education. 2007, 40 (1), 5-22.

[7] Börje Holmberg, The Evolution, Principles and Practices of Distance Education, Volume 11, july 2008. Publisher: BIS-Verlag der , ISBN 978-3-8142-0933-3.

[8] Archambault, L.M., and K.J. Crippen. 2009. "K-12 Distance Educators at Work: Who's Teaching Online Across the United States." Journal of Research on Technology in Education, 41(4), pp. 363-391.

[9] Joi L. Moore, Camille Dickson-Deane, Krista Galyen, "e-Learning, online learning, and distance learning environments: Are they the same?", Internet and Higher Education, Elsevier, 2011.

[10] Simonson, Smaldino, Albright & Zvacek, Teaching and Learning at a Distance: Foundations of Distance Education, 5/e, 2012, Pearson.

[11] Fred Markhauser , "Satellite Web Caching for Schools", Telesat Canada [Online].

[12] A. Rousskov, D. Wessels, G. Chisholm, D. Newman, "Proxy Caches: Speeding up the Web", Data Communications, October 2005, pp 57-70.

[13] www-skynet-dvb-com

Online vs. Paper-Based Testing: A Comparison of Test Results

Sanja Čandrlić, Martina Ašenbrener Katić and Martina Holenko Dlab
Department of Informatics, University of Rijeka, Rijeka, Croatia
sanjac@inf.uniri.hr, masenbrener@inf.uniri.hr, mholenko@inf.uniri.hr

Abstract - This paper presents the results of a comparative research of paper-based testing versus online testing using MudRi. The research was conducted on three courses at the Department of Informatics, University of Rijeka, and included 1231 tests. In observed courses, online tests have largely replaced traditional paper-based tests in the last few years. Therefore, the aim of this research was to determine whether the online testing results differ from the results of traditional paper-based tests. The research results showed that, for the observed courses, the median values of the results achieved during MudRi online tests and traditional paper based tests are not significantly different when online tests include certain percentage of subjective test questions.

I. INTRODUCTION

An important aspect of the educational process is knowledge assessment. Knowledge assessment is required to determine the extent to which learning outcomes are achieved. The goal of knowledge assessment is not only to rate or rank students, but also to provide feedback. Therefore, in the educational process it is important to feature formative assessment as well as summative assessment. While summative assessment seeks to evaluate students' knowledge with a grade, the goal of formative assessment is to provide feedback to students about their learning in order to positively influence their learning motivation and help them to take control of their own learning [1].

Along with the development of information and communication technology (ICT), which is increasingly used in education, teachers today have computer-based testing (CBT) tools at their disposal. The popularity of such tools, especially in the sphere of higher education, stems from the fact that with the use of CBT it is possible to assess more students in less time because the computer program reviews and evaluates tests [2]. The advantages of using ICT in education have been recognized in a much broader context than just for the assessment process since the use of ICT makes possible to replace traditional classroom instructions with e-learning. Although courses could be realized entirely online by the use of web-based learning environments, a mixed or hybrid approach to e-learning is being used in practice, also called blended learning. This approach combines online learning with traditional face-to-face learning [3].

Learning Managements Systems (LMS) play a major role in the realization of e-courses. These systems enable teachers to deliver educational materials and implement collaborative activities online by the use of tools such as forums, chats and wikis. An essential part of every LMS is a subsystem that enables online testing. An online test is a test (a series of questions used to assess students' knowledge) that students solve using a computer rather than writing their responses on paper. Standard LMS systems provide various options that make the process of creating tests or quizzes very flexible. Test created in this way can be used for formative or summative assessment [4], [5]. By the use of LMS systems such as Moodle [6], Sakai [7], Dokeos [8] and Blackboard [9] it is possible to develop tests that feature different types of questions, randomly generate tests from question banks, allow students to solve the test several times, etc.

A big advantage of online testing is automated correction that makes the job easier for teachers. But, that is not the case with objective types of questions, the correction of which mainly requires teacher's involvement. In order to overcome this problem specialized modules have been developed. For example, the knowledge assessment subsystem AHyCo LMS [3], along with standard assessment provides automated correction of programming tasks. It is possible to evaluate responses automatically for questions the answer to which is code written in one of the following programming languages: C, C++, C# and SQL [10]. Tools that can be used for the evaluation of descriptive answers are being developed, e.g. Automatic Essay Assessor [11] and Intelligent Essay Assessor [12], but their use is rather complex and language-dependent.

Although assessing student's answers in case of subjective questions can be time consuming for teachers, this type of questions should not be neglected in the process of creating tests. By using subjective questions type like Essay it is possible to measure understanding of subject matter, but also students' skills in solving some types of problems and expressing ideas in written form [13].

This paper presents the analysis and comparison of results of traditional paper-based tests and online tests that students solve on a computer. The aim of the research was to determine if and what kind of online tests can replace traditional tests that students solve on paper. Based on the results, we propose a model for transition from paper-based to online testing. The paper is organized as follows. The second chapter lists the features of the system MudRi for online testing. The third chapter describes the research methodology, including the process of preparing questions

and creating tests that were used in the scope of this research. The fourth chapter presents the research results. The last, fifth chapter, presents conclusions and plans for future work.

II. TESTING WITH MUDRI

The academic community at the University of Rijeka (teachers, students and other University employees) uses the MudRi system [14]. MudRi is a customized solution for the realization of e-courses based on the open source product Moodle. The system is extensively used with the courses of the Department of Informatics. Since it offers many features in the context of online testing, MudRi has been chosen for this study. The following sections of this paper present the types of questions and features that are available while creating online tests.

A. Types of questions

MudRi offers a whole range of question types that can be present in online tests. For the purpose of this research, questions of objective type (*Multiple choices* questions, *Connection* and *True/false* questions) and questions of subjective type (*Essay* and *Short answers*) were used and are described in the remaining of this section.

The *Multiple choices* question type offers several answers of which one or more answers are accurate. It is possible to mix the order of offered answers and define the percentage of points that will be assigned for each correct answer. The *True/false* question type has only two possible answers – true and false – of which one is the correct answer and the other is the wrong answer. It is possible to define the percentage of points that each answer carries. In *Connection* questions students are supposed to connect questions or concepts on one side with answers or concepts on the other side. It is possible to mix the order of these concepts.

The *Essay* question type is used for questions that require a longer answer and that cannot be automatically evaluated through the system but the teacher must review and evaluate the response. In *Short answer* question type students enter a word or a phrase in the answer field, which is then compared with the correct answer (or more correct answers) that were entered by the teacher. It is also possible to subsequently award points for a correct answer.

When creating questions it is necessary to define a standard number of points awarded for a correct answer and a penalty factor for wrong answers. The penalty factor is an option for adaptive tests – tests that can be taken several times. The penalty factor enables the subtraction of a portion of points when the student is allowed to offer an answer to a question that was previously answered incorrectly. For each question, the teacher can enter the desired feedback that will be displayed to the student in case of correct or incorrect answers. For some question types, it is possible to define feedback for each possible answer and in this way, for example, provide detailed explanation for incorrect answers.

B. MudRi online tests

A test consists of selected questions that are stored in a question bank. Each question can be used in multiple tests. When choosing the questions that will be used in a test, the teacher can define a fixed set of questions, let the system randomly pick questions from a category or use a combined approach. In order to allow random selection of questions, it is convenient that questions are grouped by chapter, by question difficulty or by other criteria. The number of predefined points that a question carries can be modified in the process of creating a test. This means that the same questions can carry a different number of points in different tests.

When creating tests, it is possible to define a period of time during which the test is available, limit the time available for writing the test (in minutes), set the number of questions displayed on a page and set the access password. It is also possible to configure the system to shuffle the questions or answers thus reducing the possibility that a student can copy correct answers from other students during examination. The teacher can also decide whether to add a penalty for incorrect answers in order to discourage the students from guessing the correct answer. In order to prevent opening other web pages during the examination, it is possible to set up the test to be solved in a so-called 'safe window'.

Upon completion of each test, their own test results are available to students, while the teacher has access to the test results of all students. Information about each taken test is visible, such as test start and end time, time spent, number of points achieved and number of points per question. In addition, the total average of the test is calculated and the results are displayed graphically so that teachers can analyze the test as a whole and by specific questions.

III. RESEARCH METHODOLOGY

The aim of this research was to determine whether online tests can replace traditional paper-based tests and what types of questions should online tests contain. This includes an analysis of whether the online tests consisting exclusively of objective type questions can completely replace the paper-based ones, or the online test should include a certain percentage of subjective type that will be manually graded by teachers. The main hypothesis of the research is: *There is no significant difference in median values of the results achieved during MudRi online tests and traditional paper-based tests.*

The research analysed results for the tests taken on three courses at the Department of Informatics, University of Rijeka: *Data Modeling*, *Process Modeling* and *Information Systems*. These three courses offer fundamental knowledge in the field of information systems development and are suggested as important for the education of future practitioners in the field [15]. The transition to the online test was conducted in three phases performed by two teachers. The first phase included the adjustment of questions to an online test format and their insertion into the MudRi system (creating a question bank). The second phase included creation of online tests where a special attention was devoted to the composition

of the test (i.e. the ratio of subjective and objective questions type). The final, third phase refers to conducting the tests in a real environment.

A. *Creating the Question bank*

The study used questions from classic paper-based tests that were adapted to the MudRi testing environment.

Set of questions used in paper-based tests for observed courses had been created over a longer period of time. Questions used in paper-based tests were mostly *Essay* type which included description, model or picture description and listings. Questions that require a short answer or those that require to draw or describe a picture or a model were used less often. Questions of objective types were used rarely. The number of different questions used during paper-based testing for *Data Modeling*, *Process Modeling* and *Information Systems* course was 137, 62 and 149 respectively.

The transformation of questions from the previously used written format to available objective types included transformation of the question itself and the creation of possible answers. Question bank was also supplemented with certain number of subjective questions - existing *Essay* and *Short answer* questions were also translated into online form.

Table I shows the number of created questions in the online question bank, grouped by type and course. When compared to the number of questions used in paper-based tests, as stated earlier in this section, the number of different questions in online tests is significantly bigger. Besides that, the ratio between question types used in paper-based tests comparing to those created for online testing has significantly changed: in paper-based tests questions were mostly of *Essay* type while in online question bank *True/false* and *Multiple choice* questions prevail.

Questions in online tests are designed to be clear and unambiguous, not too easy and not too hard. For example, in *Multiple choice* questions it is necessary to write the question text and offer a few (one or more) correct and a few incorrect answers [13]. We tried to ensure that incorrect answers are not too easy i.e. that their identification and elimination as incorrect is not too obvious. Sometimes, it was not easy to do so. The same applies to *True/false* question type.

Compared to conventional paper-based tests, online tests are restricted when it comes to the use of questions that need to be answered by a drawing (e.g. drawing a diagram, graph or model). This type of questions is not supported by online tests. Therefore, apart from online tests, other assessment activities were used to estimate students' knowledge.

TABLE I. ONLINE TESTS QUESTION BANK

Category of questions	Data Modeling	Process Modeling	Information Systems
Essay	25	26	0
Short answer	3	0	0
Connection	31	9	50
Other: True /false, Multiple choice	159	184	540
Sum	218	219	590

Since inserting questions and answers into the MudRi system is a time-consuming task, we considered automatic import from a file offered by the system. MudRi offers the ability to import questions that were created in other solutions such as Blackboard and Hot Potatoes and import of files of the following formats [14]: Aiken, GIFT, Moodle XML, etc.

We prepared questions for each chapter in a textual document, and therefore chosen the Aiken format [16] to import them into the MudRi system. The format allows the creation of *True/false* and *Multiple choice* questions and these two types of questions were the most common in our question bank. The Aiken format is quite simple. The question is formulated in the first paragraph, and every subsequent paragraph provides a response. The proposed answers begin with a letter which is followed by a dot '.' or a bracket ')', and space ' '. The list of answers is followed by a paragraph that begins with the keyword "ANSWER:" and this is followed by a letter that indicates the correct answer. In case that a single file contains several questions, they have to be separated by an empty paragraph.

Essay, *Connection* and *Short answer* types were manually inserted.

During test grading, the points assigned to student's correct answers are summed up. For the process of grading *Multiple choice* questions that have multiple correct answers, we used the rule that an incorrect answer nullifies a correct answer in order to discourage students from guessing the correct answer. For example, if we have a task with five answers from which two are correct, each of the correct answers carries 50% of the question points (the sum is 100%), and each of the incorrect answers - 33.33% (the sum is 100%). This system ensures that in case that a student picks all answers, the sum will be 0% (i.e. no points will be awarded).

B. *Creating the tests*

Online tests were set up to be solved within a 'safe window' and in a specific timeframe. Each test had an access password defined which was given to students just before the test begins. Tests consist of 15 – 25 questions from different categories. For every online test a part of questions was predetermined and a part was randomly chosen by the system from specified categories in order to cover all course lessons.

Traditional paper-based tests included *Essay* type of questions, questions for which it was required to draw a sketch, and *Short answer* questions.

For the purpose of this study, tests for different courses were designed differently, in respect to the question types used. In the *Data Modeling* course and the *Process Modeling* course online tests one third of questions was subjective (*Essay* and *Short answer* types) and two thirds were objective (*Multiple choice*, *True/false*

TABLE II. TEST PROPERTIES

	Classic paper-based tests	Online tests	Online tests
	All three courses	Data Modeling, Process Modeling	Information Systems
Question type	Essay type of questions, short essays and listings, image drawing, short answer questions	Multiple choice, true/false questions, connection; essay and short answer	Multiple choice, true/false questions, connection;
Duration	60 minutes (paper-based)	20 minutes (computer-based)	20 minutes (computer-based)
Number of questions	15 (on average)	20 (on average)	25 (on average)
Number of tests	753	310	168

questions, *Connection* question types). Objective questions carried fewer points than *Essay* or *Short answer* type of questions. For the *Data Modeling* and the *Process Modeling* course points based on answers for objective type of questions was 50%, i.e. almost 70% of objective type questions carried 50% of points and 30% of subjective type questions carried 50% of points. Questions in online tests for the *Information Systems* course were only objective. Test properties are compared in Table II.

After students hand in the test for evaluation, they are presented with their test results, correct answers, feedback related to particular questions and overall feedback related to the whole test and scoring. *Essay* and *Short answer* questions are graded later manually by the teacher. After closing the test, points and overall feedback is visible to the student.

C. Participants and online testing properties

The research was held at the Department of Informatics, University of Rijeka, so the participants were students of Single major and Double major program of informatics at the Department (second and third year of undergraduate study).

We have started with online testing in 2009 in the *Data modeling* and the *Process modeling* course, using tests created as described in previous section. After two years of using online tests with combination of subjective and objective question types, our intention was to ease the correction of tests and therefore use only objective question types in the *Information systems* course. The motivation for such approach arises from the fact that in recent years the number of students, and consequently teachers' teaching load, has increased. This is especially the case when teachers hold several courses with two or three test planed.

Both, online and paper-based tests were solved in the classroom under controlled conditions. In order to solve the traditional test students had 60 minutes, and in this time, they had to solve about 15 questions. Although online tests had more questions than paper-based tests, less time was needed to solve them due to different questions types.

Grading traditional paper-based tests required significantly more time than grading online tests. Online tests were graded automatically by the system immediately after students submitted them. Teachers manually graded *Essay* type questions (if present in the test) and given points are automatically added to the total sum of points.

All the tests carried through the study were used as part of the summative assessment. Minimum of 50% test points was necessary for passing the test. Apart from online and paper-based tests, teachers used other activities in order to assess students' knowledge in mentioned courses.

IV. COMPARISON OF TEST RESULTS AND DISCUSSION

The goal of the research was to determine whether there was a significant difference between median values of the results achieved during MudRi online tests and traditional paper-based tests. The study included 753 paper-based and 478 online tests. In the *Data modeling* and in the *Process modeling* course used tests included both, subjective and objective types of question, while those for the *Information systems* course consisted only of objective ones. Descriptive statistical analysis for all three courses was performed. The results are presented in Table III.

High variance coefficient for paper-based test results shows that arithmetic mean cannot be considered as representative central tendency indicator. Rather low value of quartile deviation coefficient refers to median as better indicator of central tendency.

It is also important to notice that measures of variance are significantly lower for online tests. Based on the Levene's test for equality of variances, in all three cases the hypothesis of equality of variances can be rejected with extremely high significance ($p<0.0001$), i.e. we can conclude that variances for online tests and paper-based ones significantly differ.

The tests results for *Data Modeling, Process Modeling* and *Information Systems* courses are shown using the Box-Plot graph in Fig. 1, 2 and 3 respectively. In each figure the results for online test are shown in the left part of the graph, and for paper-based test in the right part of the graph.

TABLE III. DESCRIPTIVE STATISTICAL ANALYSIS

	Data Modeling		Process Modeling		Information Systems	
	Online tests	Paper-based tests	Online tests	Paper-based tests	Online tests	Paper-based tests
N	162	226	148	254	168	273
Arithmetic mean	0.6475401	0.6025221	0.5643795	0.5626378	0.6349981	0.5535165
Range	0.2415-0.9635	0-1	0.221-0.8915	0-1	0.33-0.93042	0-1
Standard deviation	0.140849	0.240482	0.135689	0.249787	0.136961	0.262828
Variation coefficient	0.2175147	0.3991249	0.2404207	0.4439565	0.2156873	0.4748339
Median	0.65875	0.63	0.56275	0.58	0.634095	0.57
Interquartile range	0.2125	0.27	0.1625	0.32	0.202145	0.36
Quartile deviation coefficient	0.164674	0.212598	0.146961	0.283843	0.160266	0.321586

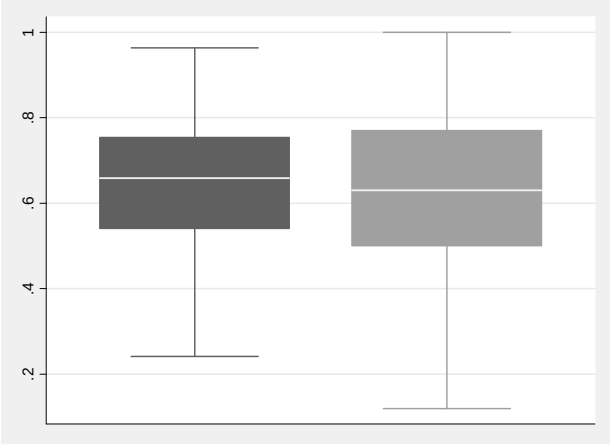


Figure 1. Box-plot diagrams for Data Modeling course (left online tests, right paper-based tests)

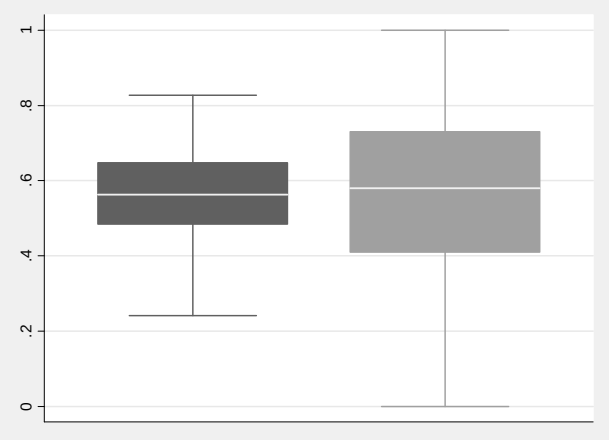


Figure 2. Box-plot diagrams for Process Modeling course (left online tests, right paper-based tests)



Figure 3. Box-plot diagrams for Information Systems course (left online tests, right paper-based tests)

Online tests, in which the majority of questions are of the *Multiple choice* or *True/false* type exclude the impact of subjective factors related to the teacher as the evaluator, which were potentially present while correcting *Essay* type of questions in paper-based tests.

Normality analysis (Table IV) indicated the use of nonparametric statistic methods for hypothesis testing. By using Mann-Whitney test it is stated that for the *Data Modeling* course and the *Process Modeling* course there is no significant statistic difference in median values (p=0.2401 and p= 0.3777, respectively) between online and paper-based tests. For the *Information Systems* course the difference is extremely significant (p=0.0066).

TABLE IV. NORMALITY ANAYSIS

	Data Modeling		Process Modeling		Information Systems	
	Online tests	Paper-based tests	Online tests	Paper-based tests	Online tests	Paper-based tests
Shapiro-Francia W'	0.98905	0.94746	0.99238	0.97006	0.99270	0.96782
P	0.20769	0.00001	0.53225	0.00009	0.47966	0.00002

Therefore, for the *Data Modeling* and the *Process Modeling* the hypothesis is confirmed – median values do not differ significantly. For the course *Information Systems* whose tests consist only of *Multiple choice*, *True/false* and *Connection question* types (i.e. objective questions), the difference in median values between online and paper-based test results is significant. Better results on online tests in the case of the *Information Systems* course can be explained by the fact that *True/false* and *Multiple-choice* questions can be answered even with lower levels of knowledge and based on recognition.

A large difference in the maximum number of points achieved in paper-based tests and in online tests can be noted. We conclude that the subjective factor of the examiner plays an important role in the assessment of descriptive answers, since the examiner has the possibility to assess the answer as partially correct, or to assess the "hidden knowledge" in the given answer. Differences can also be observed for the minimum number of achieved points in paper-based tests and online tests. On written tests, students will very often give up from finishing the test if they estimate that they will not be able to achieve the enough number of points. This is the reason why we recorded the minimum score of 0 in the analysis of paper-based tests.

Based on the result of the study, we conclude that online tests can replace traditional paper-based tests for student' knowledge assessment, but special attention should be paid to its composition. Therefore, we propose a following model for transition to the online testing.

In the phase of questions creating, existing question bank needs to be adapted to the online testing environment. The certain number of existing *Essay* type questions could be transformed to the questions of objective type (e.g. *Multiple-Choice*, *Connection*), while some of them could be just imported in online system as they are. Depending on the course domain, question types offered by the chosen system can or can not be adequate. In case there is no adequate question type (like drawing questions in our case), additional assessment activities can be performed [17]. When creating *Multiple-Choice* questions with multiple answers, we suggest to use the rule that an incorrect answer in certain percentage nullifies a correct answer. In addition, we highly recommend entering questions into the system using the Aiken or similar format because it is faster than entering questions

manually. Only minor modifications and manual input of questions that the format does not support are necessary.

Concerning the test creating phase, we consider that it is best to use a combination of online tests and other methods of knowledge assessment. In order to prove that they have adopted the course material, students have be able to explain what they learned, which is possible with the *Essay* type of questions. Therefore, instead of online tests in which the questions are posed in the *True/false* or *Multiple choice* types, we propose that online tests have around 30% of subjective questions (*Essay, Short Answer*). Although it is not possible to correct them automatically, this ratio does not represent great time workload for the teacher (it's about 2-3 questions per test).

V. CONCLUSION AND FUTURE WORK

This paper presents the results of a research that included 1231 paper-based and online tests. If the results are compared with the results of online tests, we can conclude that it is possible to replace paper-based tests with online tests in order to assess student knowledge. The biggest advantage of paper-based on the other hand is the short time in which it is possible to prepare the test and the possibility to ask questions that require a drawn response.

Based on students' comments that we received during the testing activities, we conclude that they prefer to use the keyboard rather than a pencil so they would rather take online tests than paper-based tests.

Due to above-mentioned, we plan to use them in the future in the scope of observed courses. Our future work will include efforts to further improve methods for online assessment. This includes revising the question bank. Students responses will be analyzed in order to identify to easy questions (the questions with 100% correct answers), as well as questions for which there are no correct answers (the questions that are potentially unclearly stated and therefore incomprehensible to the students). We also plan to update the question bank with sets of different types of questions for formative assessment. Students will be able to solve a self-assessment test after completing each course unit. This will enable them to test their knowledge, in the process of preparation for the tests that follow.

VI. ACKNOWLEDGEMENTS

The research has been conducted under the project "Extending the information system development methodology with artificial intelligence methods" (reference number 13.13.1.2.01) supported by University of Rijeka (Croatia).

REFERENCES

[1] D.J. Nicol, and D. Macfarlane-Dick, Formative assessment and self-regulated learning: a model and seven principles of good feedback practice, Studies in Higher Education, 2006; 31(2): pp. 199-218.

[2] S. M. Sidhu, Computer-Based Assessment. In Tomei, L. A. (Ed.), Encyclopedia of Information Technology Curriculum Integration, 2008; pp. 131-137.

[3] N. Hoic-Bozic, V. Mornar, and I. Boticki, Introducing adaptivity and collaborative support into a Web-based LMS, Computing and informatics. 2008; 27(4): pp. 639-669.

[4] J. P. Sewell, K. H. Frith, and M. M. Colvin, Online Assessment Strategies: A Primer, 2010; 6(1): pp. 297-305.

[5] J. Cole and H. Foster, UsingMoodle - Teaching with the Popular Open Source Course Management System, 2nd edition, O'Reilly Media Inc., 2008.

[6] Moodle.org: open-source community-based tools for learning, http://moodle.org (accessed on October 19, 2011)

[7] Sakai Project: collaboration and learning - for educators by educators, http://sakaiproject.org (accessed on October 19, 2011)

[8] Dokeos: Open Source E-learning, http://www.dokeos.com (accessed on October 19, 2011)

[9] Blackboard: Technology and Solutions Built for Education, www.blackboard.com (accessed on October 19, 2011)

[10] N. Hoic-Bozic, I. Budiscak and I. Boticki, Online assessment of programming assignments in a learning managment system AHyCo, Engineering Review. 2008; 28(1), pp. 51-60.

[11] T. Kakkonen and E. Sutinen, Evaluation Criteria for Automatic Essay Assessment Systems – There is much more to it than just the correlation, Proceedings of the 16th International Conference on Computers in Education, 2008.

[12] L. Streeter, J. Pstoka, D. Laham and D. MacCuish, The credible grading machine: Automated essay scoring in the dod. In Proceedings of Interservice/Industry, Simulation and Education Conference (I/ITSEC), 2003.

[13] A. W. Ward and M. Murray-Ward, Assessment in the Classroom. Belmont, CA: Wadsworth Publishing Co., 1999.

[14] MudRi: sustav za učenje Sveučilišta u Rijeci, http://mudri.uniri.hr (accessed on October 19, 2011)

[15] M. Pavlic, M. Marinovic and S. Candrlić, Model of Basic Training for Information Systems Developers, Proceedings of the 8th International Symposium on Operational Research in Slovenia SOR'05; Ljubljana: Slovenian Society Informatika - Section for Operational Research (2005); pp. 411-417.

[16] P. McElroy, D. Brabazon and E. McLoughlin, Analysis of use of multiple choice question (MCQ) examinations as an assessment tool for postgraduate engineering modules. Proceedings of the Irish Learning Technology Association Conference, IT Sligo. 2006.

[17] N.Hoic-Bozic, M. Holenko Dlab, and E. Kusen, A Blended Learning Model for "Multimedia Systems" Course. In Workshop on Learning Technology for Education in Cloud (LTEC'12), Springer Berlin Heidelberg, 2012, pp. 65-75.

Comparing classification models in the final exam performance prediction

Jasna Gamulin[a], Ozren Gamulin[b], Dragutin Kermek[c]
[a] University of Zagreb School of Medicine
Šalata 3, Zagreb, Croatia, Phone:+385 1 4590 268, E-mail: jasna.gamulin@gmail.com
[b]University of Zagreb School of Medicine, Department of Physics and Biophysics
Šalata 3, Zagreb, Croatia, Phone:+385 1 4566 797, E-mail: ozren@mef.hr
[c]University of Zagreb Faculty of Organization and Informatics, Department of Computing and Technology, Department of Theoretical and Applied Foundations of Information Sciences, Pavlinska 1, Varaždin, Croatia, Phone: +385 42 390864, E-mail: dkermek@foi.hr

Abstract – The use of Learning Management Systems (LMS) and the web-based formative and summative assessments during the traditional teaching in classroom provides the huge amount of data on students' behavior and results at the point of time when the course is still in progress. This data could be used for the final exam performance prediction so that the excellent as well as the students requiring help could be detected. The data on 302 students enrolled into the first year Physics course of a biomedical university study program in 2011/2012 have been used. The data were preprocessed by dividing Croatian grading system which comprises 5 grades (range 1-5; 1=fail, 5=excellent) into 2 categorical classes in one version of the experiment and into 3 categorical classes in the second version. Several up to date algorithms for classification were applied without and with attributes optimization by a genetic algorithm. Also, since students have 4 chances at 5 exam terms during an academic year, 3 different dependent variables were constructed. In order to evaluate the performance that each of the classification models had the following performance criteria were used: accuracy for all models and sensitivity and area under curve (AUC) for binary classifier systems.

I. Introduction

The hybrid learning as the form of learning environment in which the traditional teaching in classroom is mixed with the computer mediated interaction [1] is spreading rapidly in higher education in Croatia and worldwide. Students use some form of Learning Management System (LMS) on the regular basis during the course and also in some cases an additional web-assessment tool as well as the classical pen-and paper mid-term exams. This provides a vast number of data on students' behavior and results.

The goal of classification as a data mining type is to accurately predict the target class for each case in the data. During the process of a model building i.e. training process, a classification algorithm finds relationships between the values of the attributes and the values of the target class. Classification algorithms use different techniques for finding relationships. These relationships represent a model, which can then be applied to a different data set in which the class assignments are unknown. The historical data for a classification is divided into two data sets: one for building the model and the other for testing the model. Classification models are tested by comparing the predicted values to known target values in a set of test data.

The aim of this paper is comparing different classification models, which could not only differ by the classification algorithm but also by attribute selection (optimized or not by a genetic algorithm) and also by the selection and construction of the class label (definition of success classes in respect to grade and time of passing the exam).

The paper will be structured in the following way: the overview of the recent work in this field will be followed by the presentation of the used methodology and the obtained results with discussion. At the end, the conclusions with limitations and directions for the future research will be presented.

II. Literature Review

Educational data mining (EDM) is a relatively new research area which pursues the development of methods and computational approaches for exploring data originating from an educational context [2].

According to Baker and Yacef [3], during the period 1995-2005 the most prominent type of EDM research was the relationship mining and the prediction was the second most prominent research area. During 2008-2009, a shift occurred moving the prediction to the first place and discovery with models as a new type of research was moved to the second most frequently used type of EDM. The appearance of the discovery with models is explained by the integration of psychometrics and student modeling researches into the EDM community. In discovery with models, a model of a phenomenon is developed and then used as a component in another analysis, such as prediction or relationship mining.

A number of papers have been published elaborating the issue of predicting student performance using machine learning techniques and the majority refers to distance learning environment [4, 5, 6] or even intelligent tutoring

systems [7]. The papers referring to blended learning environment are not so frequent and not necessarily using DM [8, 9, 10, 11, 12]. Blended learning environment usually does not provide so many automatically available data as distance learning and especially intelligent tutoring systems but in Croatia blended learning represents the prevailing higher education environment also emphasized in Croatian universities mission statements.

Based on data from strictly e-learning environment Lykourentzou et al. [4] proposed a method for a dropout prediction. The authors used following machine learning techniques: feed-forward neural networks, support vector machines and probabilistic ensemble simplified ARTMAP. The decision schemes which combine the results of those three techniques were also proposed.

Jovanović et al. [5] described the application of a variety of classification algorithms as well as clustering students on grounds of their cognitive style. The data set was acquired from a completely distance learning environment.

Kotsianis et al. [6] deployed in distance learning environment several classification algorithms for predicting the students' performance and also assessed the relevance of the attributes involved (demographic data, face-to face meeting data, written assignments data). The data came from distance learning environment. The Naive Bayes was found to be the most appropriate for eventual construction of a software support tool.

De Nui at al. [10] used data from hybrid learning environment and investigated the relationship between LMS usage and student performance.

W. Dekkar et al. [9] used different modeling techniques on the data sets of students' past education as well as university grades and related data collected from the electrical engineering department to test their impact on students' performance.

Delgado et al. [8] used artificial neural networks for predicting students' marks from the logs to Moodle [13]. Thai-Nghe et al.[11] compared the accuracy of Decision Tree and Bayesian Network algorithms for predicting the academic performance of undergraduate and postgraduate students at two academic institutes.

Divjak and Oreški [12] used discriminante analysis for analyzing the effect of 30 attributes (the data collected by a questionnaire hybrid learning environment) upon the student success.

III. METHODOLOGY

A. Dataset preprocessing

Data on 302 students enrolled in the first year Physics course of a biomedical university study program in the year 2011/2012 have been used in this paper. Out of 302 students 121 (40,07%) are men and 181 (59,93%) are women.

As attributes we have used the data collected by Moodle [13] Learning Management System and a previously developed web-based formative assessment tool named *Pitalica* as well as 3 mid-term exam results. The application of *Pitalica* during seminars was extensively described in Gamulin et al. [14] and the application during laboratory teaching in [15]. The idea of *Pitalica* is to allow the teacher to assess a large student group during seminars and through their answers to detect the level of their former knowledge of the subject and understanding of the seminar topic and adjust the course of lecture accordingly. It is strictly formative assessment. In lab *Pitalica* allows the teacher to quickly assess students at the beginning of each laboratory exercise. The idea is to motivate students to prepare for their lab exercises and the result achieved at *Pitalica* has a mild summative component.

Using Moodle a student has the access to self-assessment test which could be taken only once, web-laboratory, simulations, video laboratory (short videos describing specific exercises), interactive exercises and PowerPoint presentations. After the course is over a student has 4 chances to take the final exam at 5 exam terms. Grading system in Croatia consists of 5 grades (range 1-5; 5=excellent, 4=very good, 3=good, 2=sufficient, 1=fail). This opens a variety of possibilities for constructing the class label (it is final oral exam grade, in further text FOEG) in a classification model. In this paper we investigated 12 different class labels. We used data discretization which is dividing the numerical data into categorical classes in order to increase interpretation and comprehensibility [16]. The 5 grades system is discretizated into 2-classes system in two different ways depending on whether we observe excellent and fair students or good and poor students in 2 classes (grades 5,4→class Excellent; grades 3,2,1→class Fair) or in 2 classes* (grades 5,4,3→class Good; grades 2,1→class Poor). We formed 3-classes system also in two different ways depending on whether we observe distinction students, very good and poor in 3 classes (grade 5→class Distinction; grades 4,3→class Very Good; grades 2,1→class Poor) or very good, standard and poor students in 3 classes* (grades 5,4→class Very Good; grade 3→class Standard; grades 2,1→class Poor).

Also, it is not all the same whether a student has passed an exam in his/hers first trial or fourth. Students have 4 chances at 5 exam terms during an academic year. Taking an exam repeatedly causes the exhaustion with students and also makes the heavy workload to teachers. In Table I there is an overview of all attributes and class labels.

TABLE I. DESCRIPTION OF THE ATTRIBUTES AND CLASS LABELS

Attribute	Description		Values
x_1	total scores of midterm exams (written).	integer	=[0 – 5]
x_2	scores of web-based formative assessment during seminars (*Pitalica* – seminars)	integer	=[11 – 42]
x_3	scores of web-based formative assessment during seminars and laboratory teaching (*Pitalica* – total)	integer	=[19 – 77]
x_4	scores of web-based formative assessment during laboratory teaching (*Pitalica* – laboratory)	integer	=[6 – 38]
x_5	number of logins to Moodle (learning material + general information about the course)	integer	=[15 – 869]
x_6	ability [17] calculated according scores of midterm exams	numeric	=[-4.184 – 405]
x_7	number of logins to Moodle learning material	integer	=[5 – 623]
x_8	number of login to Moodle general information about course	numeric	=[0 – 348]
x_9	ability [17] calculated according to scores of self-assessment test on Moodle	numeric	=[-2.673 – 4.102]
x_{10}	time used for self-assessment test on Moodle	numeric	=[0 – 5]
Class label			
y_1	FOEG, 1stT, 3 classes	polynominal	=[Distinction, Quite Good, Poor]
y_2	FOEG, 1stT, 3 classes*	polynominal	=[Very Good, Standard, Poor]
y_3	FOEG, 1stT, 2 classes	binominal	=[Excellent, Fair]
y_4	FOEG, 1stT, 2 classes*	binominal	=[Good, Poor]
y_5	FOEG, 3rdT, 3 classes	polynominal	=[Distinction, Quite Good, Poor]
y_6	FOEG, 3rdT, 3 classes*	polynominal	=[Very Good, Standard, Poor]
y_7	FOEG, 3rdT, 2 classes	binominal	=[Excellent, Fair]
y_8	FOEG, 3rdT, 2 classes*	binominal	=[Good, Poor]
y_9	FOEG, 5thT, 3 classes	polynominal	=[Distinction, Quite Good, Poor]
y_{10}	FOEG, 5thT, 3 classes*	polynominal	=[Very Good, Standard, Poor]
y_{11}	FOEG, 5thT, 2 classes	binominal	=[Excellent, Fair]
y_{12}	FOEG, 5thT, 2 classes*	binominal	=[Good, Poor]

Here we proposed three different ways of capturing the best image of final exam success when failures at some of the terms are possible:

1stT= the first term, final oral exam grade at the1st term calculated as the average of four questions' grades

3rdT=the third term, final oral exam grade calculated as the average of first three terms, failures are also registered

5thT= the fifth term, final oral exam grade calculated as the average of all five terms, failures are also registered

Tables with results accumulated by means of *Pitalica* during web-based formative assessment on seminars and laboratory teaching as well as data collected by Moodle (logs number, log date, type of accessed material, results from self-assessment test), and mid-term and final exam results were stored in a database. The data were exported to the format which can be processed by Rapid Miner 5.3 tool [18].

B. Modelling

In this paper we used 5 state of the art algorithms.

Logistic regression analysis (LogReg) which was also used by [5, 6, 7, 19] is a type of probabilistic statistical classification model used in research situations where predicting the outcome of a categorical class label is based on one or more predictor variables (attributes). The relationship between the classifier and attributes is a non linear function.

Neural Networks (NN) which was also applied by [5, 6, 8, 19, 20], are computational models inspired by neuropsychology where a nerve cell functioning principle has been used [21]. A multi-layer neural network consists of large number of units (neurons) joined together in a pattern of connections. Input units receive information to be processed, output units present the results of the processing and units in between are called hidden units.

Random Forests (RF), also employed in work of [5], operates by fitting a number of decision tree classifiers on various sub-samples of the dataset and use averaging to improve the predictive accuracy and control over-fitting.

The Support Vector Machines technique (SVM), also applied in paper by [6, 22], maps the data to a higher dimensional space in order to separate the positive from the negative instances in the training set.

K-nearest neighbor (kNN), also used by [6], operates on the principal that the instances within a data set will generally exist in close proximity with other instances that have similar properties. If the instances are tagged with a classification label, then the value of the label of an unclassified instance can be determined by observing the class of its nearest neighbors.

In Rapid Miner we generated 2 processes. In the first process, the split validation with stratified sampling was performed meaning that one set of data was used for model building (70%) and the other for validation of the model (30%). In the second process the proposed attributes were optimized by the genetic algorithm (GA) and then the model was built on grounds of the optimized selection of attributes and validated it as described in the first process.

The overall accuracy criterion is used to measure the number of positive and negative instances correctly classified in proportion to all the instances. In Table II it is shown that the accuracy is evidently higher in models where the attributes were optimized by the genetic algorithm and we achieved the same result for all class labels. That is why in further discussion we focused on the models with the GA optimization only.

TABLE II. ACCURACY FOR SELECTED CLASS LABELS

accur %	y_9	y_{10}	y_{11}	y_{12}
LogReg			72,09	72,09
LogRegGA			87,21	88,37
NN	52,14	62,79	77,91	80,23
NNGA	74,42	76,74	83,72	90,7
RF	52,33	51,16	75,58	67,44
RFGA	58,14	76,74	83,72	83,72
SVM			76,74	76,74
SVMGA			87,21	90,7
kNN	52,33	52,33	69,77	66,28
kNNGA	68,6	76,74	87,21	86,05

IV. RESULTS

In Table III there is an overview of accuracy indicators calculated by different algorithms.

Confusion matrix is the visualization of efficiency of an algorithm since it provides the overview of whether the actual and predicted classes are confused. Accuracy as a predictor could be misleading especially when the two classes are numerically unbalanced.

The precision criterion is used to determine the proportion of true positive instances versus the total number of true positives and false positive instances.

To measure the efficiency that each technique demonstrates in correctly identifying the positive instances, the sensitivity (recall, hit) criterion is used. This criterion measures the proportion of the number of positive instances correctly classified versus the total number of positive instances correctly classified and instances misclassified as negative.

Confusion matrices have been calculated for all class labels and out of that it could be concluded that discretisation into 3 classes (grade 5→class Distinction; grades 4,3→class Very Good; grades 2,1→class Poor) gives better performance than 3 classes* (grades 5,4→class Very Good; grade 3→class Standard; grades 2,1→class Poor), while 2 classes (grades 5,4→class Excellent; grades 3,2,1→class Fair) give better result than 2 classes* (grades 5,4,3→class Good; grades 2,1→class Poor) and in general, the best results are achieved by 2 classes (grades 5,4→class Excellent; grades 3,2,1→class Fair) when the stability of class recall and class precision for all the observed classes are considered. The only exception is noticed with NNGA where 2 class* gave better results. As an example, in Table IV and V, there are shown the confusion matrices for y_5 and y_{12} when RFGA and NNGA applied.

TABLE IV. CONFUSION MATRIX FOR Y_5 WHEN RFGA IS APPLIED

Accuracy 66,28	true Distinction	true Poor	true Quite Good	class precision
pred. Distinction	9	0	2	81,82%
pred. Poor	1	25	14	62,50%
pred. Quite Good	6	6	23	65,71%
class recall	56,25%	80,65%	58,97%	

TABLE V. CONFUSION MATRIX FOR Y_{12} WHEN NNGA IS APPLIED

Accuracy 90,7	true Good	true Poor	class precision
pred. Good	54	6	90,00%
pred. Poor	2	24	92,31%
class recall	96,43%	80,00%	

Area Under Curve (AUC) could be interpreted as the probability that a classifier will assign a higher score to a randomly chosen positive instance than to a randomly chosen negative instance under the assumption that positive ranks higher than negative [23] i.e. it is displaying the tradeoff between true positive rate and false positive rate.

The specificity indicator is used to determine the

TABLE III. ACCURACY FOR ALL CLASS LABELS, ATTRIBUTES OPTIMIZED BY THE GENETIC ALGORITHM

	1stT				3rdT				5thT			
accur %	**y_1** 3 classes	**y_2** 3 classes*	**y_3** 2 classes	**y_4** 2 classes*	**y_5** 3 classes	**y_6** 3 classes*	**y_7** 2 classes	**y_8** 2 classes*	**y_9** 3 classes	**y_{10}** 3 classes*	**y_{11}** 2 classes	**y_{12}** 2 classes*
LogRegGA			86,05	84,88			**87,21**	82,56			87,21	88,37
NNGA	68,6	75,58	80,23	87,21	70,93	80,23	83,72	81,4	**74,42**	76,74	83,72	**90,7**
RFGA	63,95	76,74	81,4	86,05	**66,28**	77,91	**83,72**	86,05	58,14	76,74	83,72	83,72
SVM GA			83,7	87,21			**87,21**	86,05			87,21	90,7
kNN GA	63,95	70,93	77,91	86,05	66,28	73,26	**87,21**	83,27	**68,6**	76,74	87,21	86,05

proportion of the number of negative instances correctly classified versus the total number of negative instances correctly classified and the number of negative instances misclassified as positive.

In Table VI there are presented AUC, sensitivity and specificity indicators for binary classifier systems.

When applying LogRegGA Class 2 comparing to Class 2* gives high accuracy, better confusion matrix and high AUC, sensitivity and specificity (y_7 and y_{11}). Employing NNGA with Class 2* gives better accuracy and confusion matrices for the first and fifth term data (y_4 and y_{12}). RFGA achieves better accuracy for Class 2* but confusion matrices are poor, so, it should be taken Class 2 as more reliable, especially since the accuracy is still high and the confusion matrices are better and AUC, sensitivity and specificity are high. SVMGA and kNNGA have the same picture as RFGA meaning that Class 2 is sorted out as more representative.

For 3 classes models we compared NNGA, RFGA and kNNGA. NNGA gave better accuracy for Class 3* model but the confusion matrices are better for Class 3 models. RFGA gave the same result as NNGA with remark that only confusion matrix for the Class 3 at the third term gave good score. kNN algorithm returned better accuracy for Class 3* but the confusion matrices implied that Class 3 is more reliable.

TABLE VI. AUC, SENSITIVITY AND SPECIFICITY INDICATORS FOR BINARY CLASSIFIER SYSTEMS

	1stT		3rdT		5thT	
AUC	y_3	y_4	y_7	y_8	y_{11}	y_{12}
LogRegGA	0,885	0,895	0,938	0,856	0,938	0,899
NNGA	0,87	0,901	0,893	0,837	0,893	0,938
RFGA	0,886	0,812	0,934	0,859	0,934	0,893
SVM GA	0,892	0,915	0,92	0,897	0,92	0,927
kNN GA	0,849	0,917	0,908	0,886	0,908	0,894
Sens in %	y_3	y_4	y_7	y_8	y_{11}	y_{12}
LogRegGA	89,47	81,67	89,47	80,95	89,47	91,30
NNGA	83,87	92,73	78,95	83,82	78,95	90,00
RFGA	89,19	84,00	88,00	85,14	88,00	81,82
SVM GA	83,72	88,14	88,57	87,93	88,57	91,43
kNN GA	75,00	89,06	86,11	88,52	86,11	86,36
Spec in %	y_3	y_4	y_7	y_8	y_{11}	y_{12}
LogRegGA	83,33	92,31	85,42	86,96	85,42	76,47
NNGA	78,18	77,42	87,50	72,22	87,50	92,31
RFGA	75,51	100,00	81,97	91,67	81,97	100,00
SVM GA	83,72	85,19	86,27	82,14	86,27	87,50
kNN GA	80,43	77,27	88,00	72,00	88,00	85,00

Starting from the observation that Class 2 gives better results than Class 2*, with exception of NNGA, we compared results for specific exam terms.

LogRegGA, RFGA, SVMGA and kNNGA produced the equally good results for the third and the fifth term and only NNGA produced the best results for the fifth term data. All this leads to the conclusion that the third term data is a good enough representative to build the model which has significant practical implications since the third term is still in summer and there is some time to remedy the possible issues before the academic year ends.

Similarly, if we take the starting point that Class 3 enables better results than Class 3*, we can see from Table III that NNGA and kNNGA give the highest accuracy (accuracy 74,42% and 68,60%) at the fifth term and RFGA at the third term (y_5, accuracy 66,28%).

When we take into account class construction, exam term, accuracy, confusion matrices, sensitivity, specificity and AUC, the best results were achieved for two classes by LogRegGA, SVMGA and kNNGA (y_7, accuracy 87,21%), and by NNGA (y_{12}, accuracy 90,70%) and RFGA (y_7, accuracy 83,72%). For three classes model the best results were achieved by RFGA (y_5, accuracy 66,28%) NNGA (y_9, accuracy 74,42%) and kNNGA (y_9, accuracy 68,60%).

V. CONCLUSION

Success of a prediction model depends not only on the attribute selection (genetic algorithm), algorithm selection, but also on class label construction and selection. In the specific case described in this paper we showed that the models with two classes have better prediction indicators but also have some flaws and a researcher has to decide about the trade-off between better prediction performance and other information that the three-class model brings (information concerning risk students and potentially good students).

For the future work the authors plan to take as attributes for the model also the LMS activity of students refined according to various LMS contents (simulations, videos, interactive exercises, PowerPoint presentations and self assessment tests).

Also, authors are planning to provide an algorithm that could follow the procedure of selecting the most appropriate combination of attributes, algorithms and class labels as described here, that would fit the desired option.

REFERENCES

[1] G. Bubas, D. Kermek, "The Prospects for Blended Learning in Croatian Academic Institutions", *CARNet Users Conference CUC* 2004.

[2] C. Romero and S. Ventura, "Educational data mining: a review of the state-of-the-art", IEEE Trans. Syst. Man Cybernet, C Appl. Rev., 40(6), pp. 601–618, 2010.

[3] R. Baker and K. Yacef, „The state of educational data mining in 2009: a review and future visions", Journal of Educational Data Mining, 1(1), pp. 3–17, 2009.

[4] I. Lykourentzou, I. Giannoukos, V. Nikolopoulos, G. Mpardis and V. Loumos, "Dropout prediction in e-learning courses through the combination of machine learning techniques", Computers and Education 53 (3), pp. 950–965, 2009.

[5] M. Jovanović, M. Vukičević, M. Milovanović and M. Minović, "Using data mining on student behavior and cognitive style data for improving e-learning systems: a case study", International Journal of Computational Intelligence Systems, 5:3, pp. 597-610, 2012.

[6] S. Kotsiantis, C. Pierrakeas, P. Pintelas, "Predicting students' performance in distance learning using machine learning techniques", Applied Artificial Intelligence, 18: pp. 411-426, 2004.

[7] N. Thai-Nghe, L. Drumond, A. Krohn-Grimberghe, L. Schmidt-Thieme, „Recommender System for Predicting Student Performance“, Procedia Computer Science, 1, pp. 2811–2819, 2010.

[8] M. Delgado, E. Gibaja, M.C. Pegalajar and O. Pérez, “Predicting Students' Marks from Moodle Logs using Neural Network Models“, in Proc. International Conference on Current Developments in Technology Assisted Education, (Sevilla, Spain), pp. 586-590. 2006.

[9] W.G. Dekkar, M. Pechenizky and M.J. Vleeshouwers, „Predicting Students Drop Out: A case study“. in Proc. 2nd Int. Conf. on Educational Data Mining, 2009.

[10] D. DeNeui, Daniel L.; T.L. Dodge, “Asynchronous Learning Networks and Student Outcomes: The Utility of Online Learning Components in Hybrid Courses“, Journal of Instructional Psycholog, Vol. 33 Issue 4, pp. 256-259. December, 2006.

[11] N. Thai-Nghe, P. Janecek, P. Haddawy, “A Comparative Analysis of Techniques for Predicting Academic Performance“, in: Proceeding of 37th IEEE Frontiers in Education Conference (FIE’07), Milwaukee, USA, IEEE Xplore, T2G7–T2G12, 2007.

[12] B. Divjak, D. Oreški, “Prediction of Academic Performance Using Discriminant Analysis“, in Proceedings of the 31st Int. Conf. on Information Technology Interfaces (ITI 2009), Cavtat, Croatia, 2009.

[13] Moodle, 2014, https://moodle.org/

[14] J. Gamulin, J. Gugić, O. Gamulin, “Improving classroom teaching in higher education environment using web-based formative assessment”, MIPRO 2010-33rd International Convention on Information and Communication Technology, Electronics and Microelectronics - Proceedings, art. no. 5533585, p. 1001-1006, 2010.

[15] J. Gamulin, O. Gamulin, “Enhancing laboratory teaching in higher education environment using web-based formative colloquiums”, MIPRO 2011-34th International Convention on Information and Communication Technology, Electronics and Microelectronics - Proceedings, art. no. 5967237, p. 1189-1194, 2011.

[16] C. Romero, S. Ventura and E. García, “Data mining in course management systems: moodle case study and tutorial“, Comput. Educ. 51(1), pp. 368–384, 2008.

[17] F.B. Baker, The Basics of Item Response Theory. University of Maryland, College Park: ERIC Clearinghouse on Assessment and Evaluation, 2nd edition, 2001.

[18] Rapid Miner, 2014, http://rapidminer.com/

[19] M. Jadric, Z. Garaca and M. Cukusic, “Student Dropout Analysis With Application Of Data Mining Methods”, Management, Vol.15, pp. 31-46, 2010.

[20] M. Paliwal, U.A. Kumar, “A study of academic performance of business school graduates using neural network and statistical techniques“, Expert Systems with Applications, Volume 36, Issue 4, pp. 7865–7872, 2009.

[21] G. Klepac, L. Mršić, Poslovna inteligencija kroz poslovne slučajeve, Lider press d.d., Tim press d.o.o., Zagreb, 2006.

[22] N. Thai-Nghe, A. Busche, L. Schmidt-Thieme, “Improving Academic Performance Prediction by Dealing with Class Imbalance“, in: Proceeding of 9th IEEE International Conference on Intelligent Systems Design and Applications (ISDA’09), Pisa, Italy, IEEE Computer Society, p. 878–883, 2009.

[23] T. Fawcett, “An introduction to ROC analysis“, Pattern Recognition Letters, Vol. 27, pp. 861–874, 2006.

The Role of Management in Career Development

Maja Lamza – Maronić, PhD*, Jerko Glavaš, PhD** and Ivan Uroda, PhD***
* Full Professor, Faculty of Economics in Osijek, Osijek, Croatia
** Assistant Professor, Faculty of Economics in Osijek, Osijek, Croatia
*** Senior Assistant, Faculty of Economics, Department of Computer Sciences, Rijeka, Croatia
maja@efos.hr, jglavas@efos.hr, ivan.uroda@gmail.com

Abstract - With a view to facilitate personal development and achievement of goals by means of career development planning, an organization facilitates the realization of personal goals aimed at achieving organizational goals. In this way, the career has both a subjective and an objective dimension. Subjective dimension refers to the attitudes, interests and motives of individuals, whereas the objective dimension refers to the organization's conditions, the structure of work posts, job promotion criteria, etc. It is essential to develop strategies for personal career development in order to eliminate the identified shortcomings and take advantages in achieving one's goals. The research results show the distribution of opinions on the role of an organization in an individual's career development.

I. Introduction

Nowadays there are four generations present in the labour market: traditionalists, big generation – Baby Boomers, Generation X and Generation Y. Generation Y replaces their parents Baby Boomers and they will almost certainly have great social and economic influence in the coming years. Different perspectives, motivation, attitudes, and needs of those four generations have changed the workforce dynamics. [4]

Generation Y, people born in 1980s are just arriving in the labour market. They spent their formative teenage years in a society in which post-war social security was not only questionable, but it was completely destroyed.

Internet and third generation of mobile telephony transform ways in which the ideas are exchanged and their creative contribution while social and economic innovations profit greatly from them.

However, in many countries small and medium companies work with obsolete technology, low-skilled human resources, weak management systems and entrepreneurial abilities, unavailability of appropriate and timely information, insufficient usage of information technology, low quality of products all of which contributes to low productivity. Hence, it is necessary to develop additional knowledge and skills on the individual level that will enable the individual to establish himself in today's labour market.

II. Career Development Phases

In different periods of their life people have specific relationship to career. During career development people constantly change, especially in view of attitudes towards values and life, goals, ambitions, possibilities and needs. First situation is at the beginning, second situation is in the middle and the third situation is at the end of a career. In simplest terms, people aged 15 and people aged 50 want different things from life as well as in career. Most authors have determined four key career development phases. Even though they are often connected with certain life phases and are determined on the basis of duration, their beginning and their length depend on the nature and the complexity of an occupation, length of the education process, length of the preparation for the occupation and the occupation itself. [11]

Popular way to analyze the career is to observe it as a sequence of phases. There are five-phase models that can be referred to the majority of people during their working life. Five-phase model includes:

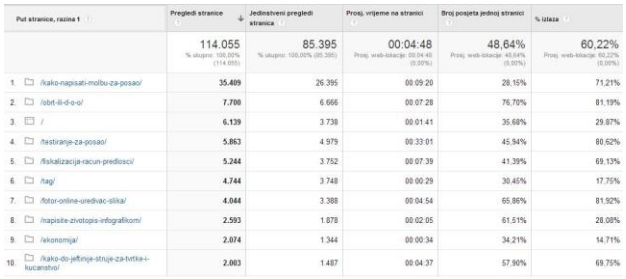
1. Investigation - when individuals analyze possible career options and when they make key choices
2. Establishment - when individuals start with job search and when they get their first job
3. Middle career - when individual is no longer perceived as a "pupil"
4. Late career - when individual can share his knowledge with the others within organisation
5. Withdrawal - when individual officially stops to work (retirement)

III. Crisis as the Chance

Ability to recognize the chances in the environment is a characteristic of top-line managers. Current economic crisis is a source of numerous chances for the development of personal knowledge and abilities in order to be more competitive in the labour market. On the basis of the analysis of the data available from web page http://www.ekonomskiportal.com [1] it is visible that time of crisis and high unemployment initiates the search for specific data and knowledge.

Because Figure 1. below is very small in size due to the two-column paper format and path of the sub-pages is written in Croatian the authors have prepared their translations in English as shown:

1. /how to-write job-application/
2. /trade-or-ltd./
3. /(root)
4. /job-tests/
5. /fiscalization-receipt-templates/
6. /tag/
7. /fotor-online-image-editor/
8. /write-resume-infographics/
9. /economics/
10. /how-to-cheaper-power-for-companies-and-household/



Put stranice, razina 1	Pregledi stranice	Jedinstveni pregledi stranica	Prosj. vrijeme na stranici	Broj posjeta jednoj stranici	% izlaza
	114.055 % ukupno: 100,00% (114.055)	85.395 % ukupno: 100,00% (85.395)	00:04:48 Prosj. web-lokacije: 00:04:48 (0,00%)	48,64% Prosj. web-lokacije: 48,64% (0,00%)	60,22% Prosj. web-lokacije: 60,22% (0,00%)
1. /kako-napisati-molbu-za-posao/	35.409	26.395	00:09:20	28,15%	71,21%
2. /obrt-ili-d-o-o/	7.700	6.666	00:07:28	76,70%	81,19%
3. /	6.139	3.738	00:01:41	35,68%	29,87%
4. /testiranje-za-posao/	5.863	4.979	00:33:01	45,94%	80,62%
5. /fiskalizacija-racun-predlosci/	5.244	3.752	00:07:39	41,39%	69,13%
6. /tag/	4.744	3.748	00:00:29	30,45%	17,75%
7. /fotor-online-uredivac-slika/	4.044	3.388	00:04:54	65,86%	81,92%
8. /napisite-zivotopis-infografikom/	2.593	1.878	00:02:05	61,51%	28,08%
9. /ekonomija/	2.074	1.344	00:00:34	34,21%	14,71%
10. /kako-do-jeftinije-struje-za-tvrtke+-kucanstvo/	2.003	1.487	00:04:37	57,90%	69,75%

Figure 1. Keywords according to the interest of webpage visitors
Rank of the visited sub-pages in comparison to total webpage visits

On the basis of the analysis of the available data authors conclude that the visitors of the web page search for two main types of knowledge/skills:

1. primarily necessary knowledge
2. other knowledge and skills

Consequently and on the basis of the conducted survey it is possible to identify individual's needs and his perception of necessary steps for career development.

IV. Career Management on a Business System Level

Career development includes career planning, treading of the possible career paths, development and training activities, formal staff policies, promotion criteria, estimating, counselling and promotion of mentoring. Besides career planning as a process in which there is a dominance of the individual, a business system is primarily associated with a term career management. That is a process of forming and implementing goals, plans and strategies that enables managers and human resources professionals to meet the needs for those resources, but it also enables individuals to reach personal career goals (Figure 2.).

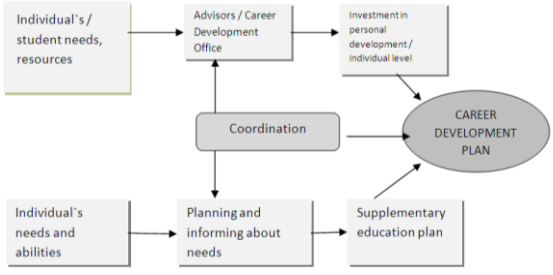


Figure 2. Career planning process [9]

Career management assumes great coordination and collaboration of managers, employees and human resources department. At the same time organizational unit human resources department should have the coordination role and the role of the one who provides the necessary professional help to managers and employees.

Individuals should be managing their careers like entrepreneurs manage their companies (Figure 3.). They must perceive themselves as self-employed even if they are employed in a big business system. For successful career management it is important to know yourself, to understand your potential and abilities as well as to correctly set yourself personal goals and to learn how to make right decisions.

TABLE I. Career Management in a Business System

CAREER MANAGEMENT		
INDIVIDUAL	MANAGER	ORGANIZATION
- determining personal preferences, interests and abilities - planning life and work - information on occupations - choosing organization - determining development and career options - work and professional preferences - introducing management with personal development preferences - activity planning for career development - self steering - developing abilities, knowledge and skills - tracing internal and external possibilities for career development	- catalyst and motivator for employee's development - estimating and tracing performance - determining abilities and talents - assessing the reality of goals and development needs - counselling for and supporting career development - providing information on organizational needs and possibilities for career development - steering - training - providing information on employee's development potential - matching individual and organizational needs	- estimating individual potentials and preferences - matching individual and work needs - performance grade - linking human resources planning with career development - establishing career ladder - professional steering - educating and developing employee - information system for human resources and needs - information on possibilities for careers development - counselling for career development - tracing, evaluating

Figure 3. Career management in a business system

Career management is a process that does not happen on a one-time basis, rather it is continuous and repetitive in a cycle of individual development. It is very hard for the individual to determine the ultimate career goal during the establishment of career goals. It is much more appropriate to establish career goals for a period of 10 to 15 years that can be even shorter because career goals are specific for every individual.

V. Case Study - Career Management On a Level of Higher Learning Institution: Faculty of Economics in Osijek

According to the curriculum in academic year 2013/2014, Faculty of Economics in Osijek at J.J. Strossmayer University of Osijek there are five courses of Graduate Studies: Financial Management, Marketing, Entrepreneurship, Management and Business Informatics.

Topic of empirical research in this paper is observation of the nature of relationship between career development and expectations about future job that was analyzed through career planning for the students of final year of graduate study (4th semester) from the courses Financial management and Marketing.

From a total number of enrolled students (106) there were 70 participants in the online survey.

Through general students' attitude towards career planning and career development options as well as thorough estimation of individual needs and expectations, the cognitive, behavioural and emotional aspects attributes of students have been observed. The majority of participants will, in a course of 2014 when they gain their Graduate study diplomas (mag.oec., eng. Master), find themselves in a labour market and/or in a business entity. In the course of online survey analysis the most frequent answers have been extracted presented.

. Consequently, with the realistic snapshot of participants' gender demographics it is confirmed once more that females are more dominant in the participants' sample and in the total student population that study economics. Above-mentioned matches the data about gender structure of Croatian students from Croatian Bureau of Statistics in which female students constitute 56% of total student body population and male students constitute 44%.



Figure 4. Gender structure of survey participants

Readiness for spatial and professional mobility is one of the very important factors to reduce unemployment. Promotion of workforce mobility within EU is treated as the one of new, important characteristics of the labour market. Creation of new jobs and rapid changes in labour demand generate the need for increased workforce mobility. [8]

Possibility of professional development is something that 79% of participants see as the most desirable form of advancement even though it was ranked as sixth in the ranking of desirable workplace characteristics. It can be assumed that the large portion of participants does not see possibility of professional development as an integrative part of career development, but as a form of reward for job well done and career advancement.

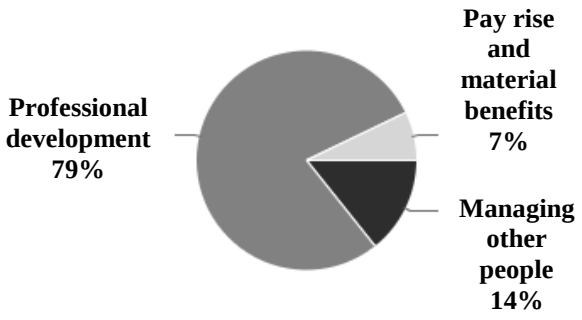


Figure 5. What career advancement means to survey participants

As opposed to male participants, female participants see professional development, in 63% of answers, as a desirable form of career advancement. From the presented results it can be concluded that there is a big difference between male and female participants in the terms of choosing "Professional development" as the desirable form of career advancement, while there is a small difference in their attitudes towards management positions and material satisfactions.

Members of Generation Y believe that technology will play the important role in their future work and they emphasize the need for organizations to keep up with technology. They wish to share knowledge on social networks through examples of good practice and collaboration with colleagues.

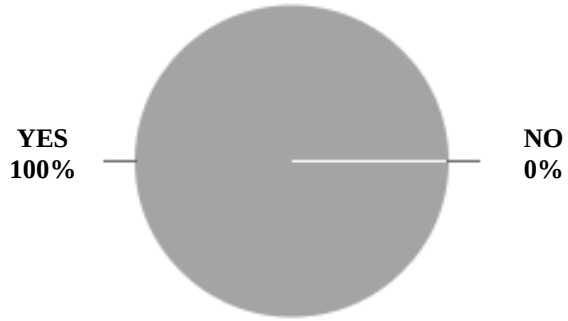


Figure 6. Social networks inclusion of survey participants

As per gender differences of participants there are no significant deviations and both genders are evenly represented in the utilization of social networks.

Which of the given possibilities do you use in order to gather data on potential jobs that you are interested in?

One of the reasons why presented results for personal contacts are so high is because of the inclusion in social networks.

TABLE II. POSSIBILITIES THAT ARE USED IN ORDER TO GATHER DATA ON POTENTIAL JOBS

Online job portals	18%
Employment agencies	5%
Job fairs	4%
Work experience in a company	7%
Career centre at the university	4%
Searching through companies' web pages, pages regarding employments/careers	14%
University mailing lists	0%
Job advertisements in press and media	4%
Company presentations at the university	7%
Professional collaboration with universities	0%
Employment services and offices	12%
Personal contacts	19%
Online news portals	7%

Youth's attitude towards lifelong learning and development is tested by various questions and grades on a scale from 1 (I do not value it) to 5 (I do value it). Data shows that the answer for which the most participants have awarded the highest grade is (with over 40%) "Mentor work". It is followed by "Formal education in scientific academic institutions with 35,9%. On the bottom of the list there is "E-learning, various short business courses" with only 23,2%.

TABLE III. THE MOST DESIRABLE FORM LIFELONG LEARNING – MENTOR WORK

1 (I do not value it)	7%
2	7%
3	7%
4	36%
5 (I do value it)	43%

In comparison, the answers that were given by the participants in a research on a world level (PwC, 3907 participants) have graded the E-learning much better (62%). [10] Youth from the other parts of the world, similar to the Croatian participants, have graded mentor work as the desired format. Such answers confirm the theories about Generation Y that state the importance of mentor's role in the young employees' career development, namely much more than the E-learning through technology that is very close to them.

VI. FINAL CONSIDERATIONS AND SUGGESTIONS FOR IMPROVEMENT

Research goal manifests in the attempt to, on the basis of research results analysis, provide appropriate guidelines and models that would act as a support with career planning and career development to young people during their education.

That also implies the importance of the increasing awareness of lifelong career development for every individual that will arrive in the labour market after education, but also the employer's awareness of the importance of employee's career and the realization of quality human resources in his business system.

Conducted research represents a kind of foundation for a pilot project whose target groups (users) would be:

- students in the final year of their graduate study
- young entrepreneurs without sufficient formal education who are wishing to improve their business by learning
- popularization of the entrepreneurship in target groups.

On the basis of presented research results analysis project activities can be defined even further with an aim to derive direct and indirect benefits from the project.

Direct benefits relate to competitiveness in labour market, additional practical knowledge as well as new approach to knowledge, technologies and information.

Indirect benefits imply decline of unemployment rate, increase of self-employment rate and recognition of the situation in the environment through the interaction with the attendees.

Throughout mentioned project activities and constant mentoring of the attendees the more quality promising professional future for the economists that are arriving in the labour market is foreseen.

REFERENCES

[1] Ekonomski portal (Google Analytics): http://www.ekonomskiportal.com, January 2014

[2] PWC, "Growth and value in a volatile world", http://www.pwc.com/en_GX/gx/ceo-survey/pdf/15th-global-ceo-survey.pdf, May 2012

[3] PWC, "How the downturn will change the future of work", http://www.pwc.com/gx/en/managing-tomorrows-people/future-of-work/companies.jhtml, March 2012

[4] V. Jovanovski, "Razvoj karijere pojedinca u okružju menadžmenta znanja", PhD thesis, Faculty of economics in Osijek, Osijek, 2012

[5] M. Lamza-Maronić, J. Glavaš, A. Vukašinović, "Primary school education system in Croatia – improvements through a strategic management model (case study)", International Journal of Knowledge and Learning (IJKL), Vol. 7, 2011, pp. 39-50

[6] PWC, "Managing tomorrow's people: The future of work to 2020", http://www.pwc.com/gx/en/managing-tomorrows-people/future-of-work/worlds.jhtml, March 2012

[7] PWC, "Millennials at work: Reshaping the workplace", http://www.pwc.com/gx/en/managing-tomorrows-people/future-of-work/millennials-survey.jhtml, March 2012

[8] F. Radin, V. Ilišin, B. Baranović, H. Štimac Radin, D. Marinović Jerolimov, "Mladi uoči trećeg milenija", Institute for social research in Zagreb, Zagreb, 2002

[9] P. Sikavica, F. Bahtijarević-Šiber, N. Pološki Vokić, "Temelji menadžmenta", Školska knjiga, Zagreb, 2008

[10] PWC, "The future of the global workface", http://www.pwc.com/gx/en/ceo-survey/key-findings/hr-talent-strategies.jhtml, May 2012

[11] D. Torringtonn, L. Hall, S. Taylor, "Menadžment ljudskih resursa", Prentice Hall Europe, Data status, Beograd, 2004

E-learning next step – learning materials for students

Marjan Krašna, Matjaž Duh, Tomaž Bratina
University of Maribor, Faculty of Arts, Department of Pedagogy, Maribor, Slovenia
marjan.krasna@um.si; matjaz.duh@uni-mb.si; tomaz.bratina@uni-mb.si

Abstract: E-learning materials were first introduced at the university level education in last years of previous century in Slovenia. From the initial steps and pilot projects we learned the trade and participate in many e-learning material developments for primary and secondary schools. But in the recent years we regretful discovered that the university's e-learning materials are not in the same league as lower level education e-learning materials. It is true that students' population is different from kids but "*digi-kids*" have grown up and become students and they are familiar with the e-learning materials. It is the fundamental difference between lower level education and university level education. Lower levels education e-learning materials were funded by the government but university does not have funds for e-learning materials production. Up until now lecturers used LCMS for communication and documents sharing. But it is become evident that this is not enough and upgrade of e-learning material is required. It is evident that this process will take time and should be efficient from all perspective (didactically suitable, cheap, and fast). In the last year e-learning materials was prepared for students to test what suits them most. A lecture for a whole course was prepared with the different composition of text, sound and video. In forthcoming semester we are going to test our e-learning materials and receive feedback from the students. Article presents the learning materials specification, expected didactical results, aesthetical considerations, individual preferences (visual, audio, kinaesthetic), and client's browser's limitations. It is still not clear if university level students require more elaborate learning materials than PDF which was their favourite type up to now.

Keywords: distance education, e-learning, learning materials, design, verification and validation

I. INTRODUCTION

Since 1998 when we first prepare web study materials for a whole course Multimedia for teachers [1] many years and many projects have passed. These initial steps were the first glimpse to the future of learning materials preparation. For a decade we have upgraded materials according to the advancement in the software technology and IT infrastructure. We were always one step ahead in the anticipation what are going to be the technologically feasible and acceptable by students. In all these years we have again and again witnessed the students' ICT paradox [2]. Older the students less they were familiar with the technology and more are conservative in their preferences. Even today we still have students that print their web materials and learn from paper.

From initial web textbooks we advanced to the more elaborate learning materials where we used animation (made of successive images) to shorten the learning process of certain skills. As the infrastructure advanced and computer processing power increased we started experimenting with video in education [3] [4] [5] [6]. In the beginning videos were post stamp size but it gives us valuable experiences with video editing. We experienced and studied the information overflow [7]. In those early years the aesthetic of e-learning materials was not an issue. The existence of e-learning materials was inherent aesthetic achievements.

In 2006 projects for better cooperation of educational institutions started [8]. In course of this project different e-learning materials were made (from distance learning didactics to support for teachers practical training in schools). These e-learning materials were the blueprints for later production of e-learning materials.

In 2008 projects for development of e-learning materials for primary and secondary schools were funded by European Social Fund and Ministry of Education and Sport of Slovenia [9]. What was started as an enthusiastic pilot project was now expand into the high quality production. The team was established and project started with tight schedule and different technological constraints. We have managed to prepare e-learning materials for primary and secondary schools and learn many pitfalls of distance education. Learning material was not enough for successful distance learning. Two years later (2010) the productions of e-learning materials were refined and more suitable for all participants in the educational processes – especially for teachers [10].

Whatever experiences we gained were for lower levels of education. In the same time we did not survey ours students. Information we get was just from the informal interview with the students. At that time we were happy to see more and more of lecturer using courses in Moodle than pay particular attention to the e-learning materials attributes. Moodle was mainly used for communication and file repository [11]. After lecturers have acquired basic knowledge they start to use Moodle for students' report submission too. Lecturers have acquired different styles for their materials structuring and some classes the structure was really opaque and annoying for students. Too many topics spread in too many divisions and

submission in different topics with practically the same names proves to be unsuccessful. Clear structure of the course is definitely successful approach and favoured by any users.

II. STUDENTS' PREFERENCES

Students' generation changes in different perspective [12]. Digital natives have come to the faculties. If it was a rare sight around five years to see students using laptops at the faculties then now it is a common sight. Students have laptops in our halls and corridors during their free time. They browse the web; read learning materials; or just watch YouTube videos. Current Wi-Fi architecture becomes congested and prevents users from accessing web on some occasions. Students generally occupy two connection one for laptop and the other for smart-phone. Inside the classroom we discovered that students generally do not use computer but they also write less notes as few years ago. Some of the students admit that they type faster than they write text by hand but our classroom equipment did not follow new needs. Especially electrical connections are insufficient and unavailable to the students therefore they preserve battery for other more needed tasks. Implicitly they only listen the lecturers and access learning materials on the web later. In our practice we have deep insight to the different students' groups. Observations of students from general pedagogy, art education and elementary education are presented in this topic.

In the winter semester of 2013/2014 we have a course of "*Didactics strategies information support*". This course enables us to work with postgraduate students of general pedagogy and acquire information about their preferences in studying and learning materials usage. Their future job will be understanding and organizing the educational activities and help students and teachers to perform better in the educational systems. Many of those graduates are later employed outside the educational system and works for different companies to organize in-house education or work on public relations or human resource management. In general they are specialist for learning. During their years of education they express that they acquire *spatially specific learning methods.*

1. In the classroom they listen to the lecturer and rarely actively participate. They could not explain their general passive behaviour in the discussion but from the lecturer perspective there are multiple reasons for their passiveness. Students in Slovenia are more or less shy in the classroom. They do not want to be regarded as silly if they ask something what others may already know. Unless they are explicitly instructed to do something they rather stay in the anonymity. The second reason is that students are focused on acquiring the data during the lecture and they do not process the data at the same time. This corresponds with the traditional structure of the courses where lectures are for knowledge acquisition, laboratory work or seminar is for discussion. Any deviation from "tradition" is new to the students. They are not accustomed to public improvisation and creative problem solving in the classroom therefore they are rather "*safely passive*".
2. During the laboratory work student prefer the detailed instructions what to do and how to do. Student often do not like open problems with multiple solutions. If we give students assignment to make a report of their research work about how can they use production of promotion materials in education, they generally ask what kind of promotion and how many pages should report have. Even if we give them option that they could use any type of promotion, they would generally all pick the same type and the creativity in their work is very limited. From students perspective they often express that they have a lot of work and they would like to know exactly: what to do; how to do; and it should not take them more than 2 hours of their precious time.
3. Their home or dormitory is for deep study; mental reasoning; and creative work. Most of them require their own study workspace (like protective cocoon where everything is at hand) and if they are not there they are often less effective and sometimes even incapable to do the required tasks. They often express that they need a certain time span to start the work. It they know they have less than three hours' time they rather not start the work at all.

Feedback from the student of art education is a bit different. They generally use laptops, tablets computers are still rare, but most of them use smart-phones regularly. Art students use ICT in multilevel approach:

1. On the **primary level** students use their ICT for communication and social networking. Both positive and negative impact can be observed. We do not approve their attempts to cheat on written exams and other types of knowledge assessments.
2. On the **second level** they use ICT in study processes. Some academic content in the professional and artistic courses require knowledge and creative use of computer tools for processing images and video materials. Freeware software is used for this purpose because they faculty does not provide licensed programs to improve the quality of their visual content creation on both technical and artistic level. Student acquire their basic knowledge to use these software tools on the web pages; specialised forums; and YouTube. The course Information and communication technology is scheduled in the 9th semester of study. This is too late and huge disadvantage for them because in the previous courses they have already acquired most of their professional and pedagogical content. Regarding the objectives of ICT course, which states that the student will achieve information literacy; understand and know the basics of computer science and digital technologies; and know how to use computer applications, we see that this course definitely scheduled too late. ICT skills are need at the beginning. Students should use them during their studies to see usefulness and effectiveness of ICT in education. To continue in the objectives of the ICT course curriculum we can read that the student will be able to electronically communicate with synchronous and asynchronous tools; known e-learning

technologies; be familiar with computer graphics; digital recording; media formats and possibilities of interactivity and multimedia. All these topics are required preconditions before they study them in details in other processional courses [13]. Despite the general ICT literacy of students they acquire in previous education, we see that they could perform better during their studies and professional and didactical assignments.

3. On the **tertiary level** use of ICT in didactical practical training (lectures in schools and pedagogical practice). ICT in education is mandatory and art education students should use them in full extend. Data sources for visualization of art assignments are available on the web. Contemporary information can be recorded on mobile phones or other recording devices. Images, video and web are building blocks for multimedia learning materials that are used in the classrooms. PowerPoint as a tool of choice can be used for static or dynamic presentations. In art education it is very important dynamic building of the visual presentation to highlight the graphics elements (central perspective lines, layers of colouristic perspective, contrasts presentations, colour lessons, etc). Animations in this occasion are not ornamental add-ons but fundamental necessity to teach the artistic expression. Students are aware of importance of multiple communication channels but beside images, video, animations and occasionally text they rarely use sound. Sound component of their presentation is included as their speech and explanation during presentation. It should be know to students that ICT is actually limited in the art education. It does not provide stimuli for all valuable modalities (sight, hear, touch, smell and taste) jet we know that students achieve better results in data processing and comprehensive information acquisition using multiple concurrent modalities. In their pedagogical practice the production and use of such materials prove to be highly motivational in the classroom; enhance the quality of content acquisition; and contribute to the deeper understanding of artistic phenomena. All this is the good basis for the practical art exercise solving and understanding the theoretical and history of art facts. Students prepare their own presentation for the pedagogical practice and no use of already prepared materials was observed.

Students of art pedagogy are sufficiently ICT literate for their own communicational and social. The less skilled are when they use ICT for education. No gender difference or social status difference was observed regarding the use of ICT. Only students' personal preferences make the differences.

Student for elementary school teachers study do not use tablets but they extensively use smart-phones. Observations show that less than half of them use laptops. Interestingly they do not use smart-phones for access to the LMS and they work exclusively with computers or laptops. Geographically they access the e-learning materials mainly from home or dormitories, and rarely between breaks on faculty.

During the laboratory work in computer room they do access the learning materials because they have to. The amount of access to the LMS increases before the exam and seminar work submission deadlines. Interestingly we have discovered from the log files on the LMS Moodle that students excessively visit the same topics again and again despite they should already acquired them during the lecture and lab-work.

Students use computers for word processing and communication. Social networking is ranked high in their computer activities where significant number of students started using smart-phones. Much less computers are used for preparation of multimedia elements for their study and pedagogical practice.

Life long students (part-time study) show different behaviour. They access the LMS even outside regular pedagogical activities. Their acquired knowledge is more effective used. Reasons are different and one is the limited ability to communicate with the lecturers face-to-face. Therefore electronic communication has higher value in their study process.

For formal communication with the lecturers e-mails are used. Informal communication between colleagues and other acquaintances is done in social networks and mobile communication (phone calls or SMS).

III. E-LEARNING MATERIALS PREPARATION

Our students are accustomed to the use of LMS Moodle and different types of learning materials. But those materials are generally all in separate files. We want to break this tradition and prepare learning materials differently. From the review of courses in our Moodle it is clear that courses mainly have learning materials in PDF, Word and PowerPoint files. Much less often additional types of learning materials are used (Excel, SPSS, images, and animations outside PowerPoint). Even rarely lecturers use video learning materials and video lectures. Video in education may be didactically controversial. A link to the YouTube video that last for an hour where important part of video for the students is only five minutes somewhere in between is not favourable between our students. Right composition and volume of data is crucial. In the past we discovered that we need to use different types of e-learning materials to boost the retention level but we need carefully orchestrated environment not to trigger information overflow [7].

Most of students start their study from the PowerPoint presentations and some students even stay there. PowerPoint should be used as a reminder and table of content of the lecture. The knowledge behind the slides is presented by the lecturer and/or can be found in the textbooks. But some lecturers prepare their PowerPoint presentation with the full text. Despite the bad taste of such presentations students actually like them. To pass the exam they often need no additional literature; web sources; journals; and textbooks. From aesthetic perspective such presentations are rubbish and bad didactical practice. From practical viewpoint it should be banned because some students actually think that is the way to produce presentations. This bad influence can be seen at the diploma theses presentations.

Our task was therefore not only to prepare learning material that is suitable for students but also present the good practice how to prepare e-learning materials and slides. We know who will use our learning materials therefore we could make a mental thinking (modelling?) about the intended users. From our previous researches we discovered that there is statistically significant difference between students who attend the lectures and those who study by themselves. Even 10% difference in grades can be observed. We have decided to prepare learning materials in blended learning paradigm. Blended learning materials is good for those who attended the lectures and wants additional explanations. In some occasion they can be used for self-study but little less successful without additional learning materials.

One would argue that with the better learning materials we make students even more passive they already are but the efficiency is the only thing that counts. If it is hard for students to switch from one learning material to another due to different reasons why not to ease this step. Regarding the research that shows increased efficiency in working with larger monitor or two monitors [14] due to more visible information on the screen; we therefore decide to prepare *integrated learning materials*.

In the preparation of integrated e-learning materials we wanted to address the different aspects: technical, practical, didactical and aesthetical.

Simple procedure and use of known tools were desired because we would need to support a lot of courses.

A. *Technical aspect*

In recent year there are shifts in the computer market. Number of sold PC declines and number of different tablets sale raises. We don't have enough funds to customize e-learning materials for different hardware therefore our product should be available for computers, tablets and smart phones. It is not going to be an application but a web page and it should look good and work in all browsers.

We already have experience with the production of e-learning materials for primary and secondary school that still works like a charm. It was produced using Adobe Flash technology that has many benefits over other technology but also some drawbacks. In general flash technology is not supported on new version of Android, iOS and Windows mobile. Therefore whatever we know and learned about the flash production it is out-dated. But in the recent years we have get almost working HTML5 browsers. We still have to figure out how to made resizable viewport for equal presentation of content on different screen sizes.

From technical viewpoint we need common encoding the diacritical characters present in the text which is solved with UTF-8 encoding, we need images in JPG or PNG encoding, we need support for MP3 audio encoding and MP4 video encoding. Text and image is supported in all browsers but the MP3 and MP4 are not. Though Firefox browser support MP4 from the version 23 upward (now version 26 is available) this is true in Windows and MacOS operating systems. In Linux this is not true by the default and additional settings that are a bit beyond the abilities of normal computer users are needed. All our requirements are covered with the Chrome browser that is available for practically every operating system.

B. *Practical aspect*

From the practical aspect we would like to reuse a large amount of already produced learning materials. Since we have no authoring tool available some things have to be done manually and we would like to minimize that. Authoring tools are generally limited but they enable authors to focus on content and not on the technology. In previous developments we were able to transform PowerPoint presentation to the FLASH. We have not tried commercial PPT to HTML5 conversion tools. Trial version and free conversion tools are not really what we want. Conversion is packed in the predefined layouts and these are not what we want. We did try the commercial product Hype authoring tool for HTML5 but it works on MacOS only and the produced HTML5 is hard to change.

Since we found from the feedback from the students that they are satisfied with the PDF files of presentations we do not need too much attention to the animated effects of the PowerPoint's presentations. If in general static image of presentation is enough then we can use free tools to do much of our work and only specific post processing are added later. This compromise enables us to reuse already made learning materials. Authors upgrade their existing PowerPoint presentations with the notes text. This PowerPoint can later be exported as HTML; include into the desired layout; and upgraded with HTML5 components. We managed to prepare the two hours blended learning materials in less than 2 hours if all multimedia components were available.

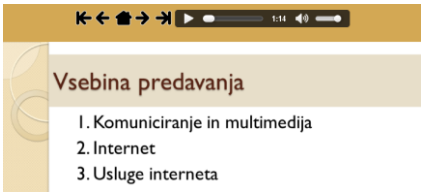


Figure 1: Sound narrated slides



Figure 2: Video narrated slides

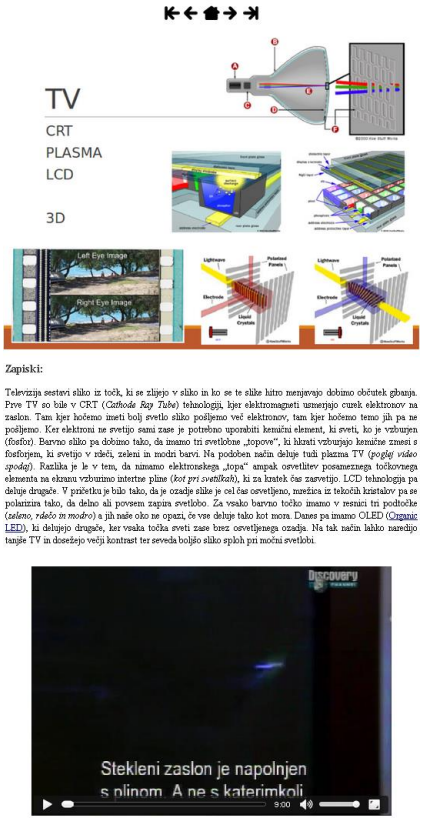


Figure 3: Integrated learning materials

C. *Didactical aspect*

Graduate level didactics is a bit different than those in primary and secondary schools. It is not even the part of general andragogy (learning of adults). We deal with the students who already have highly developed learning processes. We do not need to teach them how to study. In the preparation of e-learning materials we need to consider time constraint – topics should not require more than 15 min to process and we need to provide clear and unambiguous explanations. We can add outside sources and links to additional materials but for the time management it is still desired to include most important components into the package.

From the feedback of the students about the learning materials we discovered they do not care much about distractions which are necessary during the lectures. They are happy with PDF files but switch to additional materials is distracting. Either they skip the additional materials when reading the text and study it later or they do go visit other materials and have return problems. Even if they are using web daily they still have problems with browsing. By design the author can open link in the same window or in the new window. Browsers in general do not allow opening new window they open new tab instead. In the process of high concentration users too often close the window after reading the content of the tab and not only the current tab. In these occasions they cannot get back to the text they supposed to study further. Finding the e-learning materials takes time and additional effort is required to ascertain the position where to continue study. Such unwanted pauses lower the efficiency of study process.

Though we have not jet tested the new e-learning materials we have tested the components. We know that video is good for teaching skills. Sound is quite effective because it can be used where other types of learning materials are less appropriate (busses and walking). Text can be read on mobile phones. But on the other hand we discovered that even if we record the lecturer in the classroom and upload the video on the web students rarely watch such video. From the Moodle log file analysis we discovered that half of the students never even start playing the video. Around three quarter of students that start the video stops watching the video after 5 min. Only handful of students watch video for 15 min and no one watch entire video – except the lecturer. Therefore if we use video in the e-learning materials it should be short (no more than 5 minutes).

Data gathered from the faculty library shows that students rarely use audio books. Mainly students with special needs use audio books. Even on overcrowded buses or during their walks they rather listen to the music.

For the purpose of better communication between students we have bulletin board software. Analysing the posts we discovered that there is a huge free market for student's notes. In general we could not see the content of the notes but in some rare occasions someone have made the electronic version of the notes where we could analyse the content. Despite we do not know the content of the notes we can assume what is written from students' questions. From those sources we have highly confident insight that students sometimes makes a factual mistakes in the notes and those mistakes are not corrected and they live through multiple students' generations. Study literature and lecturer are available to the students' questions but students rarely use them.

D. *Aesthetical aspect*

In the e-learning materials each screen need to be designed according to the multiple parameters: technical, didactical, and aesthetical. Creatively designed e-learning materials must have following characteristics:

- *Applicable value* – assessed in pedagogical and didactical review;
- *Novelty* – new fresh idea, rarity, unusual and positive motivational impact
- *Appropriateness* – material must meet the problem based teaching [15]

High quality visualization is not important only for aesthetic pleasure. It gives the integrity of materials and learning for better processing of knowledge from e-learning materials. Contemporary upbringing and education build on the concept that activates auditory, haptic and visual perception. Those concepts therefore need to be fundamental part of e-learning materials. Through long observations we perceived preference shift toward visual perception. It is known that processing of same amount of information delivered through verbal messages takes more time than delivered through visual messages [16]. Therefore it is reasonable that authors extensively use visual perception activation concept in their e-learning materials design. The use of e-learning materials attracts visual capabilities. It requires an

organized observation, mental, emotional and other activities for processing the visual stimuli. The text takes much more time for processing and visuals are therefore more effective [17]. E-learning materials should be designed so that the introductory or intermediate sequence confronts the student with aesthetically ordered textual and visualized data to accept the content. "*The happiness regarding the transparent and organized screen with new content leads to the intrinsic motivation that positively affects the content acceptance*" [15]. Transparency and orderliness of screen can be achieved if elements (text, images, interactive elements, etc.) are positioned into the virtual orthogonal grid. Such perpendicular arrangement gives the required static and designed psychological simplicity and highlights isolation [18]. Unfortunately this arrangement also causes optical burden, inactivity and stiffness of composition. Important elements should therefore be positioned outside the grid. Such new composition achieves the correlation between content, importance, aesthetical attractiveness, and versatility.

Some authors [19] recommend that designers of e-learning materials should prepare the answers to the following questions before they set the design concept:

1. What is the overall structure of the e-learning materials?
2. Are the contents didactically appropriate and sufficiently reasonable?
3. How the information flows between screens from user perspective?
4. Can e-learning materials be used intuitively?
5. Which media elements are going to be used for visualization of individual content?

Additional multimedia elements have broad cognitive, educational, cultural, technical and social functions. With these we achieve a higher educational effectiveness in both educational directions wideness of knowledge and depth of knowledge [20]. It is therefore evident that for quality information transmission both aspects are important: learning content aspect and the presentation technique aspect. Didactically suitable structured e-learning materials have to be technically flawless and aesthetically pleased in order to achieve pedagogical objectives.

IV. RESEARCH

E-learning materials for course of 15 ECTS was made (15 hours lectures and 15 hours laboratory work). Apart of the lectures and laboratory work we also provide Moodle quizzes for self-assessment. Due to huge workload we did not provide the same topics in different composition but we prepare different topics in different composition (text, audio, video and integrated learning materials). Different compositions allow us to test the differences in human preferences on learning materials. It is known for a long time that people have different preferences in learning style (visual, audio or kinaesthetic) and some research shows that there are possible to get equal effectiveness in face to face as in online education [21].

Because we prepared learning material for blended learning we are going to allow access to the specific materials only after corresponding lecture. Students' performance will be tested with the quizzes and after that they will grade the materials according to different aspects.

A. *Presentation and narrative text*

Two types of presentation and narrative text are available to the students:

- Flash presentation where narrative text is displayed on demand, and
- PDF where each slide is accompanied with the narrative text.

In these two types of e-learning materials we would like to text students preference toward PDF files.



Figure 4: Presentation and narrative text – Flash technology



Čeprav danes veliko govorimo o prostorskem zvoku je zanimivo to, da lahko zelo dober prostorski učinek dosežemo že s pomočjo stereo zvoka. Seveda moramo v takšnem primeru snemati tako kot bi slišali – s pomočjo umetne glave. Posnetek je neverjetno dober, doživeti pa ga je mogoče le z uporabo slušalk.
V prostoru namenjenem za več ljudi je potrebno posvetiti pozornost ozvočenju. Odvisno od velikosti prostora in števila poslušalcev je potrebno razmišljati o geometriji postavitve zvočnikov in jakosti zvoka posameznih zvočnikov.

Figure 5: Presentation and narrative text – PDF file

B. *Presentation and audio*

Each slide in this presentation is accompanied with the audio that narrates the slide. To emphasise the value of audio we prepared slides with text only. This type of presentation has some advantage because everything is on one screen and no scrolling is needed. Students get feedback about the length of audio on the audio control slider.

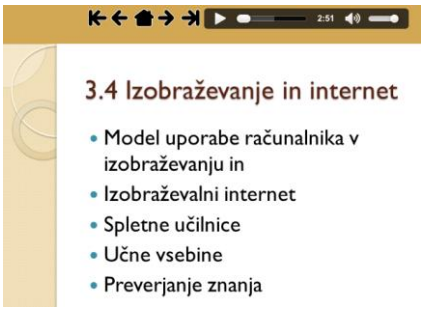


Figure 6: Presentation and audio narration

C. *Presentation and video*

Video was added to each slide to narrate slide. Since video occupy space and covers the part of the slide we add the control to hide the video. Hidden video still plays sound and enables students to hear the narrated voice of the slide. We have prepared this type of presentation according to observation of students' behaviour. They would never watch the whole video but in this occasion it is cut to the smaller chunks that covers only one slide. All video are in the time span of less than 5 minutes.



Figure 7: Presentation and video narration

D. *Presentation and text, images, links, video (integrated learning materials)*

This is probably the most favoured type of learning materials. It has drawbacks in the scrolling. Reviewers who have experience with the design and development or e-learning materials give higher score. Such materials are probably easier to implement since we do not need to record (sound or video) of lecturer. The only think we need is the narrated text and additional elements. But we need to check this finding with the students' preferences.



Figure 8: Integrated learning materials

E. *Research questions and intended users*

The research is going to go in two distinct directions: appropriateness and effectiveness. The first direction can be achieved easily in one generation but the second needs more time and search of trends through many generations of students.

Research questions:

- Number of perceived informations and layout.
- Ease of navigation.
- Personal preference to the tested type of e-learning materials.
- E-learning materials' building blocks preferences.
- How useful are controls on the e-learning materials (sound level control, hide/show video, pause, etc.)?
- How e-learning materials influence their understanding and knowledge acquisition?
- What devices were used for e-learning materials viewing?

In the summer semester of 2013/2014 we are going to test these materials for two distinct groups of students: natural science students and social sciences students.

V. CONCLUSION

University level student population is different than other students. They are highly skilled learners with established preferences in learning. Today's students are known as digital natives. This can be seen in different levels. They type better than write and they use ICT throughout whole day for different purposes. They are accustomed to different type of learning materials and only the lack of handy portable devices still bound them to the paper. During past years a lot of effort and funds were assigned to the production of e-learning materials for primary and secondary schools. University level education was not part of these projects and no funds were given to the e-learning materials production. E-learning materials were therefore limited to most common types of files (PDF, Word, and PowerPoint). Students therefore rarely make their notes and they study from PowerPoint slides. In the past we would fail such students but in recent years we see that this is not an option since we would need to fail almost all of them. Students therefore have huge width but at the same time very shallow knowledge. It becomes hard for them to search learning materials on different locations.

In an effort to enhance e-learning material and provide students with easier study we prepare different type of e-learning materials. To test their preferences we combine presentation and different narration types (text, audio and video) and prepare the integrated e-learning materials. According to their feedback we will provide support for e-learning materials production in other courses and gradually change all learning materials. We hope that this will positively influence the knowledge retention and provide good practice for further e-learning materials development.

VI. LITERATURE

[1] M. Krašna in I. Gerlič, „Multimedia for teachers," v *Proceedings of ED-MEDIA 1999*, Seattle, 1999.

[2] B. Kaučič, M. Ramšak in M. Krašna, „Rich media presentations in blended learning," *Informatologia*, zv. 44, št. 4, pp. 287-295, 2011.

[3] M. Krašna, T. Bratina in I. Gerlič, „Video supported learning material for distance education," v *14th International Conference on Information and Intelligent Systems*, Varaždin, 2003.

[4] M. Krašna in T. Bratina, „Video learning material in web education," v *Multimedia applications in education conference, MAPEC 2004*, Graz, 2004.

[5] M. Krašna in T. Bratina, „Educational value of instructional video," v *The Third International Conference on Informatics, Educational Technology and New Media in Education*, Sombor, 2006.

[6] T. Bratina in M. Krašna, „Analysis of instructional video efficiency for learning SPSS," v *The Fifth International Conference on Informatics, Educational Technology and New Media in Education*, Sombor, 2008.

[7] M. Krašna in I. Gerlič, „Multimedia - benefit or information overflow?," v *IS 2002, 13th International Conference on Information and Intelligent Systems*, Varaždin, 2002.

[8] University of Maribor, Faculty of Education, „Partnerstvo fakultet in šol 2," 7 November 2006. [Elektronski]. Available: http://distance.pfmb.uni-mb.si/course/category.php?id=4. [Poskus dostopa 9 January 2014].

[9] University of Maribor, Faculty of Arts and Faculty of Education, „E-spoznavanje okolja za OŠ," 17 August 2008. [Elektronski]. Available: http://distance.pfmb.uni-mb.si/course/category.php?id=7. [Poskus dostopa 9 January 2014].

[10] A. Vovk-Korže, „Okolje in trajnostni razvoj," 18 August 2010. [Elektronski]. Available: http://distance.pfmb.uni-mb.si/course/category.php?id=15. [Poskus dostopa 9 January 2014].

[11] B. Bedrač in M. Krašna, „Capacity testing/planing for successful implementation of LMS/CMS," v *Mipro 2009*, Opatija, 2009.

[12] T. Bratina, M. Duh in M. Krašna, „E-learning controversy in practical application," v *Mipro 2013*, Opatija, 2013.

[13] University of Maribor, Faculty of Education, „Študijski programi 2. stopnje," n.d.. [Elektronski]. Available: http://www.pef.um.si/content/Studij/2stopnja/LP_2stopnja_ucni_nacrt.pdf. [Poskus dostopa 9 January 2014].

[14] Core Communication, „4 Studies which Show that Using a Second Monitor Can Boost Productivity," 11 December 2010. [Elektronski]. Available: http://www.corecommunication.ca/4-studies-which-show-that-using-a-second-monitor-can-boost-productivity/. [Poskus dostopa 9 January 2014].

[15] M. Duh in M. Krašna, „Estetika e-gradiv," v *InfoKomTech*, Ljubljana, 2009.

[16] M. Schuster, Kunst-psychologie: Kreativität, Bildkommunikation, Schönheit, Hohengehren: Schneider-Verlag, 2000.

[17] J. Berger, Načini gledanja, Ljubljana: Emanat - Prehodi, 2008.

[18] A. Seyler, Wahrnehmen und Falschnehmen: Praxis der Gestaltpsychologie, Frankfurt am Main: Anabas-Verlag Günter Kämpf GmbH & Co KG, 2003.

[19] M. Duh in M. Krašna, „Aesthetic Principle in Design of Distance Learning Material," v *MIPRO 2009*, Opatija, 2009.

[20] M. Duh, „Mediji pri sodobni likovni vzgoji," *Pedagoška obzorja*, zv. 19, št. 3/4, pp. 104-110, 2004.

[21] C. Neuhauser, „Learning Style and Effectiveness of Online and Face-to-Face Instruction," *American Journal of Distance Education*, zv. 16, št. 2, pp. 99-113, 2010.

Configurable Image-Processing System for Education in Automotive Applications

C. Madritsch, T. Klinger
Carinthia University of Applied Sciences/Engineering and IT, Villach, Austria
c.madritsch@cuas.at, t.klinger@cuas.at

Abstract - This paper describes the overall image-processing system, which has been developed for the use in education and at various industrial projects. The added value for the students is the tight link between computer-based education in image processing and the real-world relevance. Complex and changing quality control demands of safety critical components for automotive applications require an innovative graphical systems design, which can be effortlessly reconfigured by the user. The quality control challenge claimed us to detect imperfections on the backside of brake shoes. These imperfections can cause safety related problems and therefore, the affected brake shoes need to be sorted out. From an engineering point of view, this was a technically challenging image-processing application. The surface of the brake shoe is highly reflective; the contrast between the imperfections and remaining areas is very low, the selection criteria are complex (depending on the size, number, and location of the imperfection). Due the fact that every car manufacturer and almost every car model needs different brake shoes, the most demanding requirement was, that the shape of the brake shoe needs to be freely configurable by the user.

I. Introduction

The Carinthia University of Applied Sciences (CUAS) offers R&D support to companies on a wide range of topics. Waldraff Technologies, a Germany based, well-known manufacturer of production equipment for the automotive industry, requested us to solve a quality control challenge for the inspection of brake shoes. Since other famous providers of image-processing solutions like KEYENCE have not come up with a suitable solution, we at CUAS were able to satisfy all needs and requirements of Waldraff Technologies.

The quality control challenge claimed us to detect imperfections on the backside of brake shoes. These imperfections can cause safety related problems and therefore, the affected brake shoes need to be sorted out. Figure 1 gives an example of such imperfections on the backside of a brake shoe. The glue, which was applied during the previous manufacturing step, is partly damaged. If during the next production step an aluminum plate is fixed to the backside, these imperfections will cause an uneven connection between those parts. The result can be that during the braking process of the car, the brake shoe will start vibrating. In the best case, only a high frequency audible noise will be present, the worst case is that the braking force can be influenced in a negative way.



Figure 1. Brake shoe with imperfections

II. Algorithmic Solution

From an engineering point of view, this was a technically challenging image-processing application. The surface of the brake shoe is highly reflective, the contrast between the imperfections and remaining areas is very low, the selection criteria are complex (depending on the size, number, and location of the imperfection). Due the fact that every car manufacturer and almost every car model needs different brake shoes, the most demanding requirement was, that the shape of the brake shoe needs to be freely configurable by the customer, Waldraff Technologies.

Figure 2 depicts the first part of the developed innovative image-processing algorithm. The acquisition of an image is triggered by a magnetic sensor of a light barrier, which is mounted to the conveyor belt of the manufacturing equipment. The first challenge is to detect the shape of the brake shoe. This is done using a contour

algorithm [1]. In parallel, a particle algorithm is used in order to detect reflective pixels, which represent the faulty areas on the brake shoe. Finally, both images are combined using a binary AND operation.

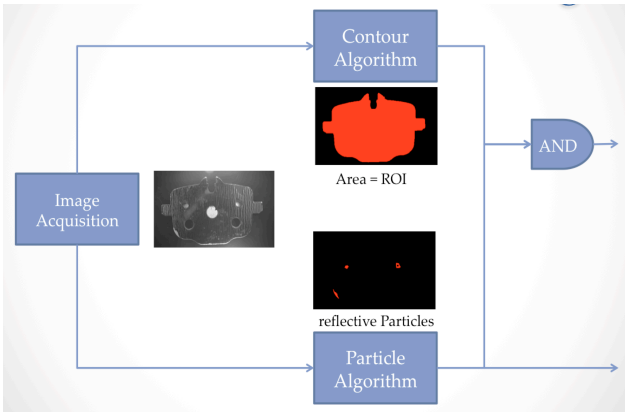


Figure 2. Image-processing algorithm Part 1

In Figure 3, the second part of the image-processing algorithm is shown. The resulting image from the contour and particle algorithm is the input into an Inverted Distance (Danielsson) Function. The Danielsson function [2] returns a distance map of the image. Here, the distance from the border of the object is computed and the result is a gradient, which allows for the classification of the faulty areas. The closer the faulty area is to the border of the object, the higher is its relevance in the quality assessment process, indicated by the grayscale value of the faulty area and shown as bright areas in Figure 3.

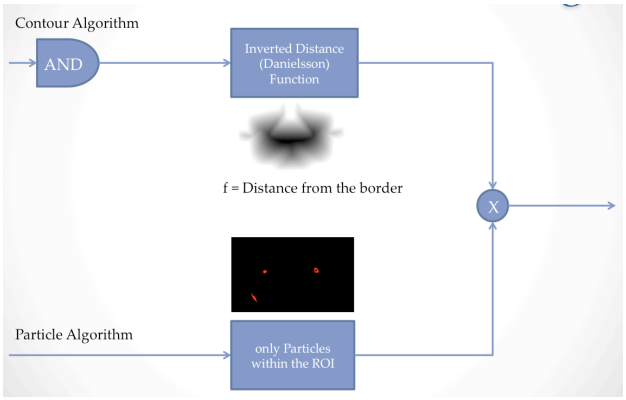


Figure 3. Image-processing algorithm Part 2

Figure 4 shows the last step of the image processing algorithm. The brightness of the resulting pixels is an inverse proportional function to the distance from the border of the object. Finally, using classification functions, the overall brake shoe is analyzed. The size, the position, and the number of faulty areas are used to compute a total impact number. If this number exceeds a given maximum value, the corresponding brake shoe is sorted out and removed from the conveyor belt. Otherwise, this brake shoe passes the quality inspection and proceeds to the next production step.

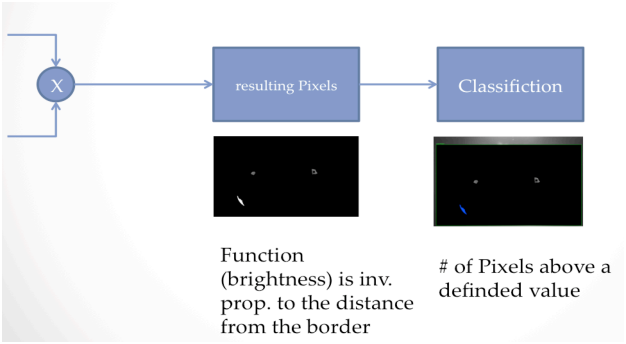


Figure 4. Image-processing algorithm Part 3

Some advantages of the developed image processing algorithm are that

- teaching of the contour of a brake shoe is not necessary, and
- assessment of the border areas of the brake shoe due to the Danielsson function is simple.

Although this algorithm is quite usable, there are some disadvantages, like

- high sensitivity against small deviations in the lighting conditions,
- the background needs to be homogeneous dark, and
- specific brake shoe elements, like holes, need to be taught manually.

An alternative algorithm to the previously depicted innovative solution is the use of a conventional image-processing mechanism. Both, the high sensitivity to changing lighting conditions and the homogeneous black background, become mostly obsolete. The teaching of the contour of brake shoes is now necessary. Since almost all brake shoes have some specific design characteristics (holes, pins, etc.), teaching might also be necessary using the innovative approach.

In Figure 5, this alternative algorithm is sketched. After the acquisition of the image, the first step is to filter ambient noise. This eliminates negligible small imperfections on the brake shoe´s image.

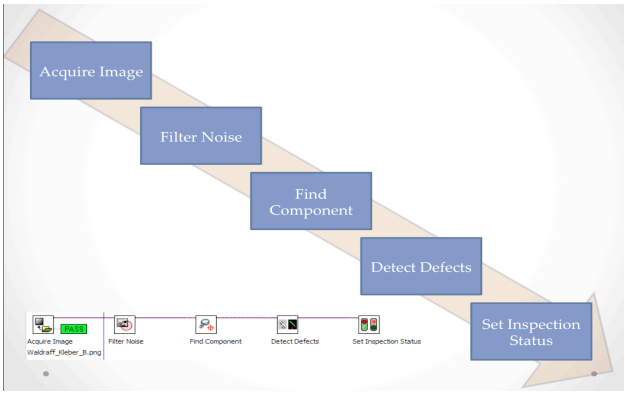


Figure 5. Classic Image-Processing Approach

After this step, a Find Component function is performed. The key to this function is Pattern Matching. Pattern Matching uses a previously thought image of the

brake shoe (see Figure 6). This function is insensitive to the rotation of the brake shoe on the conveyor belt; it can also introduce some minor perspective corrections [3].

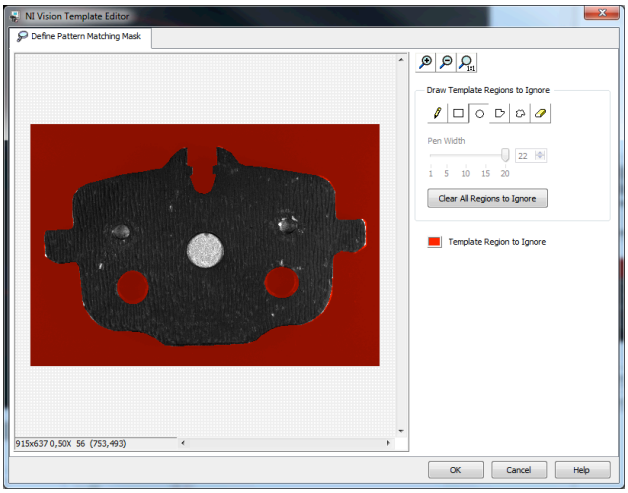


Figure 6. Pattern Matching Setup

The following step, Detect defects (Figure 7), highlights imperfections and faulty areas on the surface of the brake shoe. Here, a previously thought image of the brake shoe is used to eliminate holes and pins.

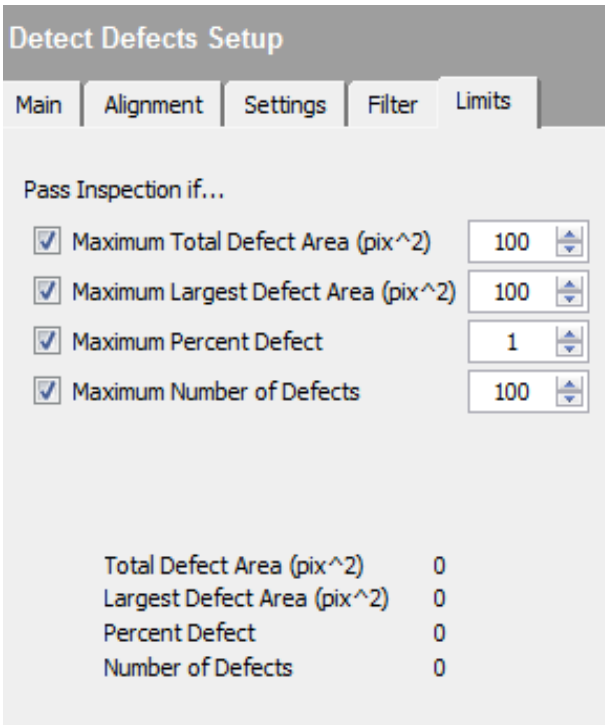


Figure 7. Classic Image-Processing Approach

The final step assesses the results from Detect defects and sets, if necessary, the inspection status of the current part to faulty which will result in the removal of this component from the conveyor belt.

The overall inspection time using an average PC is less than 50ms. This allows for a fast moving conveyor belt with an acceptable number of parts to be inspected per minute.

III. System Overview

The overall image-processing solution consists of hardware, software, and mechanical components. After the brake shoes leave the coating plant, they are transported to the sorting unit using a conveyor belt.

Due to difficult surface conditions of the brake shoes (their high reflectiveness and poor contrast between imperfections and remaining areas) the lighting unit is constituted of a big dimension light dome with four external high-power LED lamps [4] using a color temperature similar to daylight (see Figure 8).

A circular hole on the top of the light dome allows an industrial camera (Basler scA1390-17fm, 1392x1040 pixels resolutions) to capture images of the brake shoes, triggered by an external light barrier.

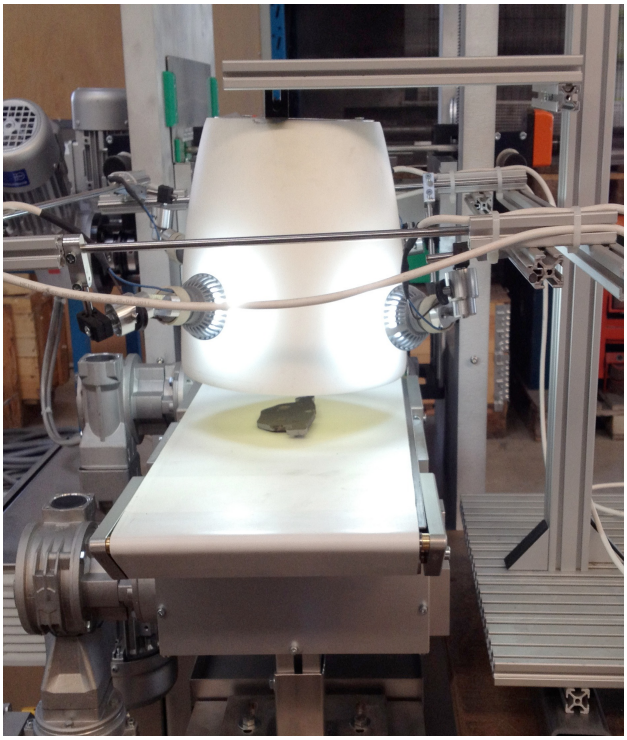


Figure 8. Light dome and brake shoe

A durable and powerful dual-core NI PXIe system, which can withstand the rough industrial environment, equipped with an NI PXI-8252 IEEE1394 interface, controls the triggering unit, the camera as well as the sorting mechanism using PROFIBUS/PA.

IV. Graphical System Design Approach

First, we developed a solution for the image-processing challenge together with students from the Systems Engineering (Bachelor level) and Systems Design (Master level) course in our labs. Then the prototype was tested with the customer's real manufacturing equipment. The most challenging requirement was, that the shape of the brake shoes varied. Our first approach was to develop a self-configuring solution, which would have automatically learned the shape of a new brake shoe. This approach leads to conflicting image-processing demands. Reliably detecting

the shape and inspecting the surface cannot be done with only one lighting setup.

The next and final approach was the one, where the customer would teach the image-processing system the shape and specialties (position of holes, cavities) of a new brake shoe with minimum effort.

CUAS has a long history of using LabVIEW for both educating our students and developing successful solutions for industrial partners. CUAS members, who are advanced LabVIEW Graphical Systems Designers on CLA, CLD, and CLAD level, attend trainings and certifications including classroom, online, or on-site custom training on a regular basis.

Without question, a Graphical Systems Design Approach was needed in order to satisfy all requirements. We have tested a variety of different configurations, for instance a LabVIEW standalone solution as well as a Vision Builder AI stand-alone solution. After intensive testing, a combination of both has been implemented. Now, the only tool the customer needs to use is NI Vision Builder AI [5]. There, we predefine the inspection process for quality control of brake shoes in general. More advanced test-steps, which cannot be accomplished using the build-in functionality of the Vision Builder AI, have been added using the NI Vision Builder AI Development Toolkit with LabVIEW.

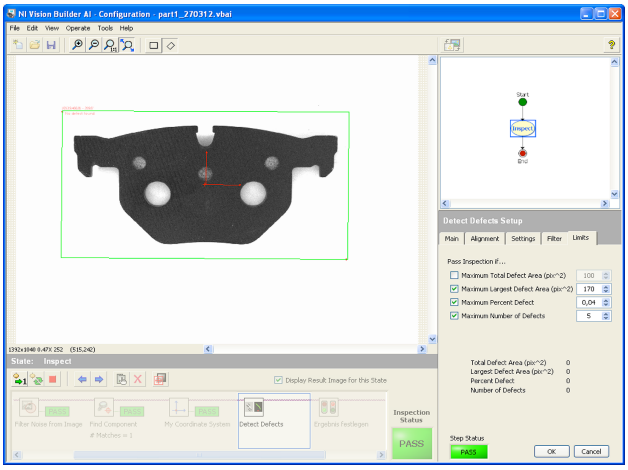


Figure 9. Custom Waldraff Technologies User Interface

The teaching of any new brake shoe requires the customer to configure two steps. The first step is the primary Find Component step, which needs to know about the shape of the part. The second step is the Detect Defects step, where holes or cavities need to be specified. The overall configuration time for a new component now is less than 20 minutes; the previous solution at the customer required a service call to the image-processing provider in order to redevelop the application; which costs a substantial amount of time and resources. During normal operation, a simplified custom user interface (see Figure 9) is used.

V. CONCLUSION

The developed image-processing solution shows that conflicting requirements can be combined into a successful application using NI software and hardware components. The chosen Graphical Systems Design approach can easily be configured by the customer; saving time and money. Furthermore, the quality control process for the production of brake shoes has been significantly improved.

Future requirements will be implemented using LabVIEW VIs which are incorporated into the NI Vision Builder AI testing sequence. Currently, Waldraff Technologies uses the prototype implementation as a demo plant for their customers. CUAS is currently developing extensions to the original challenge, implementing new types of quality control mechanisms for other types of production faults. These extensions can be seamlessly integrated into the existing solution without affecting the operating system.

Students from various courses of study like Systems Engineering (Bachelor course), Systems Design and Electrical Energy and Mobility Systems (both Master degree courses), are taking part together with lecturers and professionals from industry, in the process of specifying, implementing, testing, and installing manufacturing-ready solutions in the image-processing lab and bring them to the customer site. This unchallenged experience offers both, a great opportunity for direct contact to real-world projects and a relevant, useful and challenging image-processing platform for computer based education in automotive applications. Since this form of cooperation between industry and academia generates win-win situations on both side, it will of course be continued in future, possibly using other application domains.

REFERENCES

[1] K. R. Castleman, “Digital Image Processing,” Prentice Hall, 1996.

[2] R. Crane, “A simplifeid Approach to Image processing,” Prentice Hall, 1996.

[3] E. R. Davies, “Machine Vision; Theory, Algorithms, Practicalities,” Academic Press, 1997.

[4] T. Klinger, “Image Processing with LabVIEW and IMAQ Vision,” Prentice Hall PTR, 2003.

[5] J. C. Russ, “The Image Processing Handbook,” Third Edition, CRC, Springer, IEEE Press, 1999.

Using hybrid mobile applications for adaptive multimedia content delivery

Aleksandar Karadimce*, Dijana Capeska Bogatinoska*
*Faculty of Computer Science and Engineering
University of Information Science and Technology "St. Paul the Apostle" Ohrid Republic of Macedonia
aleksandar.karadimce@uist.edu.mk, dijana.c.bogatinoska@uist.edu.mk

Abstract - Increasingly more people are using smart mobile devices for delivering multimedia content such as text, video, images and audio in everyday communication. Therefore, there is an increased interest for developing mobile applications that will be supported by different mobile devices. The hybrid mobile applications are built using HTML5 and JavaScript native container that provide access to native platform features, such as built-in camera, accelerometers, GPS and gyroscopic sensors for navigation. This way application can extensively use the mobile device features in order to provide multimedia content that is adapted to learners in m-learning environment. The research focuses on methods and concepts for adaptive multimedia content delivery systems that are considering real-time multimedia delivery to hybrid mobile devices. The adaptive process requires considering the influence from interactive functionality and technical capabilities of mobile devices in m-learning environment. We will conclude the paper with a case study based on the hybrid mobile application used in mobile learning environment.

Keywords: hybrid mobile applications, adaptive multimedia content, m-learning

I. INTRODUCTION

In 2013, the number of mobile-cellular subscriptions had reached 6.8 billion that correspond to a global penetration of 96%, according to the latest report on global mobile-cellular subscriptions by ITU Statistics [1]. This increased revenue from smartphones and tablets sale and fast data connections provided by the mobile operators has created an explosion of mobile applications. Previously, using the web on mobile device with the limitations in screen size and processing capability was a fairly unsatisfactory and frustrating one. With the introduction of the *responsive web design* the website became flexible enough to work well at variating resolutions that can be controlled and fit to the screen, being viewed on phone or tablet or laptop. Considering today smartphones that are getting larger and faster with big displays and quad-core processors, they are also beating many of the desktop PCs used in education, business and home used ten years ago [2].

Besides all of this processing hardware the mobile devices, because of their mobility, have number of built-in sensors for orientation, location and capturing images or video. The sensor technology for smart mobile devices includes the accelerometers to detect the orientation of the device, Global Positioning System (GPS) capabilities and gyroscopic sensors for navigation. Embedding all of this technology into a single mobile device makes it perfect instrument for social networking, travelling, electronic payment, distance learning and etc. However, these built-in sensors can be used only with mobile applications provided by the smartphone vendor or developers that can create application for the specific mobile device as well as use them. If the owners of other type of mobile devices want to use the same application, which consumes the built-in sensors, the developers have to redevelop the mobile application using a different one. These applications can extensively use the mobile device features in order to provide multimedia content that is adapted to learners in m-learning environment. Therefore, today we can see increased interest for hybrid mobile applications, which are built using HTML5 and JavaScript native container that provide access to native platform features. We are proposing an effective adaptive multimedia content delivery for mobile devices by using the hybrid mobile application development platform. This research will provide proof of concept how it is easy to create hybrid mobile applications for different mobile OS vendors in the process of m-learning.

This paper is organized as following: Section II gives overview of the related work for the similar research. Section III discusses the difference between the native mobile applications and mobile web applications. Section IV presents the benefit of using the hybrid mobile applications for m-learning. Section V presents the results of considering the case study for mobile application development platform. Finally, Section VI concludes the paper and proposes future work.

II. RELATED WORK

Mobile applications can be generally classified into three categories: native mobile applications, mobile web-based applications and hybrid mobile applications [2]. Considering the differences in their underlying technology, each approach has characteristic benefits and disadvantages. The native mobile applications, as the name suggests, are built specifically for a particular mobile devices and its operating system. These applications can take advantage of the latest technology accessible on mobile device and can integrate with on-board core applications such as the calendar, contacts, and email or on-board hardware (such as GPS, camera, and

graphics) and software (such as email, calendar, contacts, picture/video gallery, file manager, and home screen widget areas) [3]. Native mobile applications are able to work across multiple devices, separate versions of the application required. On the other side, a mobile web application is normally downloaded from a central web server each time it is run, although applications built using HTML5 can also be run on the mobile device for offline use [3]. Major advantage of mobile web applications over native mobile apps is cross-platform compatibility, which allows reaching the widest audience for the least effort [3]. Unlike a web application that is accessed over the internet, a native application is downloaded from a web store and installed on the device [3]. The process of mobile web application development is relatively cheap, easy, and fast to build, while some device-specific customization is usually essential.

Another major paradigm of the hybrid mobile applications that they can save time by using web-based technologies to design nice user interface and the native shell to access mobile native functionality. These hybrid frameworks allow the use of web technologies for applications development with native access to device's resources [4]. These frameworks are based on open web-standards which mean that the application can be run and tested on common web browsers [2]. It is a cross-platform mobile application development framework which through the use of HTML5, CSS and JavaScript allows the development of applications for iOS, Android, Blackberry, Symbian, Windows, WebOS, Bada and Palm devices [2]. The hybrid mobile application, from the user interface, looks like browser based, with a native application wrapped around it providing access to device native functionality. The developers can feel an enormous difference, as they do not have to rewrite the entire application for each mobile platform, they simply write code in HTML, CSS, and JavaScript, and reuse it across different mobile devices [3].

The mobile learning (m-learning) has been defined as any sort of leaning that happens when the learner is not at a fixed, pre-determined location or learning that happens when the learner takes advantage of learning opportunities offered by mobile technologies [12]. Based on the paper review, given by the Nagella and Govindarajulu in [12], the process of adaptation is achieved in terms of the learner context, the learner's knowledge levels, by the content that is been presented, the learner's style which may vary from learner to learner and adaptation to the device's context. This process leads to personalized learning content adaptation, which becomes increasingly important to meet the diverse needs imposed by devices, users, usage contexts, and infrastructure. The proposed Personalized Learning Content Adaptation Mechanism (PLCAM) is using variety of data mining techniques, including clustering and decision tree approaches, to efficiently manage a large number of historical learners' requests in order that the adaptation engines provide the adapted mobile educational content for the learners [13]. The significance of the adaptive content delivery is a driver for new and competitive multimedia services that are based on hybrid mobile applications for end user learning services.

III. NATIVE MOBILE APPLICATIONS VS. MOBILE WEB APPLICATIONS

One of the main challenges for creating mobile solution is software development in a high-competitive and highly fragmented and rapidly evolving business market. Mobile applications developing require a fair amount of customization to run on various mobile devices and keep a continued stream of new hardware, OS versions and browsers. Therefore, due to differences in the underlying technology of native and web applications, each approach has inherent advantages and drawbacks.

A. Native mobile applications

A native mobile application (native app) is an application program that has been developed for use on a particular platform or device. Native applications are specific to a given mobile platform (iOS or Android) using the development tools and language that the respective platform supports (e.g., Xcode and Objective-C with iOS, Eclipse and Java with Android). It is "Native" because the code is essentially written for that platform and it is built for the native architecture of that device. The user can install the native applications through an application store (such as Google Play or Apple's App Store). Native applications have binary executable files that are downloaded directly to the device and stored locally. Once the application has been installed on the device, the user launches it like any other service the device offers. They provide fast performance and a high degree of reliability.

The native application is free to access all APIs that are made available by the OS vendor, they have the ability to use device-specific hardware and software, which means native applications, can take advantage of the latest technology available on mobile devices. According [5], there are two levels of APIs: low-level APIs and high-level APIs. Through *low-level* APIs calls the application can interact directly with the touch screen or keyboard, render graphics, connect to networks, process audio received from the microphone, play sounds through the speaker or headphones, or receive images and videos from the camera [5]. It can access the Global Positioning System (GPS), receive orientation information and, read and write files on the solid-state disk or access any other hardware element available today or in the future. *High-level* APIs include processes like browsing the web, managing the calendar, contacts, photo album and, the ability to make phone calls or send and receive text messages [5].

The portability of the code to other platforms always requires some level of rewriting which means that code written for one mobile platform cannot be used on another, making the development and maintenance of native applications very long and expensive process for multiple OSs. Native applications provide richer user experience and stronger engagement, but can cost significantly more because they've been developed only for certain OS vendor. The native approach is the best suited when you look for the best performance and maximum capability for a mobile application. However, while native mobile applications can be very powerful with rich feature offerings that have access to all device

capabilities and run seamlessly using the processing power of the devices, developing native applications can be a costly and resource-heavy approach due to each ecosystem using different native code development and separate deployment to each platform [6]. This means if you want as many users as possible to be able to use your mobile application then you have to build multiple versions to cover the variety of platforms.

B. *Mobile web applications*

A mobile web application is a web based application designed for smartphone and tablet use, and accessed through the mobile device's web browser. It combines the versatility of the web with the functionality of touch-enabled devices. The core technologies that are used to build mobile web applications are: HTML (for structuring content), CSS (for applying all visual styles) and JavaScript (for (almost) all interactive functionality). In addition, JavaScript toolkits have been created, such as jQuery Mobile or dojox.mobile or Wakanda, which generate user interfaces that are comparable in appearance to native applications. Mobile web applications are device agnostic and accessed at a standard URL through the device's web browser. The content is searchable (it is visible in search engine results), making it easy in delivering to a large mobile audience than which is available through the app store alone.

Since web applications are browser-based, they're intended to be platform and device independent, able to run on any web-enabled smartphone or tablet. They're relatively cheap to be hosted and deployed (we can host the app on our own web server - no approval processes and no added cost), fast to build - although some device-specific customization is usually required, and easy to support. Mobile web applications can reach the widest range of devices that are cross-platform compatible. Web applications generally cannot access the on-board hardware and software on a mobile device. Requirements such as camera control, direct GPS control (there is limited access to current location), or heavy/complex custom graphics like gaming cannot be supported.

The universal application that can run across any device, on any operating system, simply by using the web browser is the emergently thrilling promise behind HTML5. Inasmuch, HTML5 standard presents new semantic elements such as <nav>, <header> and <article> [6]. However, when it comes to mobile applications the most powerful thing it gives us is browser native media using the new elements tags for <audio> and <video> [6]. It provides possibility for local storage using the localStorage API in order to store their data locally on the user's mobile device. It provides location services by using the Geolocation API and enhances the user's interaction with your service by identifying exact position using the GPS sensors built into the mobile device. The Multi-threaded JavaScript provides the ability to use scripts that run in the background without interacting with users. The main power for developing mobile web-based applications comes with the combination of HTML5, with the presentation tricks delivered via CSS3 and behaviors and functionality through JavaScript [6].

IV. HYBRID APPLICATIONS

We define hybrid as a web mobile application that is wrapped inside a thin native container that provides access to native platform features. Hybrid applications are part native applications, part web applications. Like native applications, they live in an app store and can take advantage of the many device features available. Like web applications, they rely on HTML being rendered in a browser, with the caveat that the browser is embedded within the applications. To the user, a well-designed hybrid mobile application looks very similar to a native app: it is downloaded from an app store, stored on the mobile device, and launched just like a native mobile application.

The comparison of the mobile application is given in Figure 1, where we can clearly see that the position of the hybrid mobile application is between native and web based mobile application.

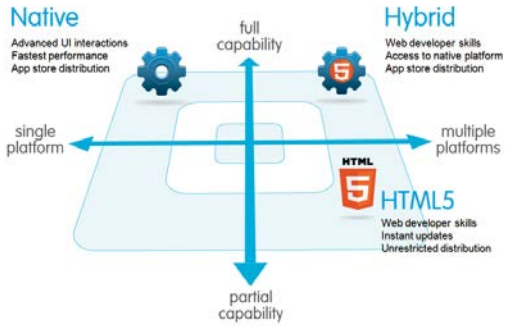


Figure 1. Native, web-based and hybrid mobile applications [7]

For the developers there is a huge difference, because they write at least some of the code in HTML, CSS and JavaScript, and reuse it across different devices rather than rewriting the entire mobile application for each mobile platform. Developers want to code less and accomplish more, to reuse and recycle across multiple platforms as much as possible, and they don't have to start from scratch for each OS. Hybrid applications combine the best of both the native and web worlds. They create a "bridge" between the browser and the device APIs, so the hybrid mobile application takes full advantage of all the features that smartphones and tablets have to offer.

There are few containers for creating hybrid mobile applications. Some of the most popular are PhoneGap [8], Appcelerator Titanium [9] or Intel XDK [10]. PhoneGap [8] is a widespread choice for developers of hybrid mobile applications, because it is a standards-based, open source development framework, free to download, with community-built development tools and plugins. Another popular solution, Titanium offers their SDK and basic mobile APIs for free, with add-on subscription packages for more advanced APIs and tools for Enterprise [9]. Benefit of hybrid mobile applications is that it can save time by using web technologies to develop the UI and the native shell to access lower-level functionality. For the purpose of this paper, we will introduce the tool Intel XDK. According its documentation, the Intel® XDK provides a comprehensive cross-platform development environment for building hybrid HTML5 applications for

mobile phone and tablet devices [10]. HTML5 applications are not limited to smart web pages viewed in a browser, we can also package our HTML5 code and deploy it directly on a mobile device as locally installed hybrid mobile applications. This development tool provides single development lifecycle to be used to create application compatible for multiple mobile devices. Using Intel XDK [10] we can code, debug, test and build mobile web applications and hybrid HTML5 applications for multiple target platforms. We can create an HTML5 project from scratch, or import existing HTML5 applications or modify any of the included sample applications. When a project is created or imported, the Intel XDK creates a special <project-name>.xdk file and locates it in the same directory as the project's index.html file. After the full project is finished, Intel XDK allows us to test our mobile application on a real device. Our mobile application is loaded into a container app on the real device that is representative of the native wrapper included with our mobile application when we build it.

Using this process we can evaluate how it behaves on a real device, and remotely debug it while it runs on the real device. The last step, in order to deploy our application as a hybrid mobile app, is to build it. The build process packs our app into a format that can be submitted to a mobile app store and, ultimately, downloaded and installed on customer's mobile device. All of the popular mobile platforms (Android, iOS, Windows 8, Amazon and Nook) are supported by the Intel XDK.

This great new tool is available for free and runs on Windows, OS X, and Linux. Abbreviated summary of features included in the Intel XDK[10]:

- Editor, Device Emulator and Debugger
- App for On-device Testing
- JavaScript UI library optimized for mobile
- APIs for game developers with accelerated canvas
- Prototype GUI quick-start wizard
- Intel cloud-based build system for most app stores
- No need to download Native Platform SDKs
- Tool to convert iOS applications to HTML5

With this research we were able to create context-aware and comfortable hybrid mobile application for discovering and learning the cave churches at the coast of the Ohrid Lake.

V. CASE STUDY

Cave churches are difficult to detect without careful examination of the environment, therefore we have created hybrid mobile application for m-learning. There are no written sources about occurrence of cave churches on the shore of the lake, presumably beginning emerged in the middle of the IX century with the spread of Christianity [11]. They were built in the recesses and cracks that have made nature at the foot of mountains or large rocks [11]. Cave churches are important because they continue to cherish the tradition of rewriting the Church Slavonic books and allow spreading education and culture. Bearing in mind that worldwide talk about common European cultural heritage, cave churches will be recognized on a wider level through this project. Creating m-learning application that can be used on a variety of mobile devices will mean fast widespread of the application between the learners. Giving the mobile application possibility to use the mobile device built-in sensors capabilities will provide adaptive learning capability. In order to achieve the following goals we propose to use the Intel XDK development platform to create hybrid mobile application. This application provides visual and textual presentation of each of cave churches, and their precise location.

The developed hybrid mobile application has been tested with HTC Desire X mobile device, which has display 4.0 inches (480 x 800 pixels), with camera 5 MP (2592 x 1944 pixels), Android OS (v4.0), CPU Dual-core 1 GHz and 768 MB RAM. The case study included testing of the hybrid mobile application for cave churches on site. At the location of the cave church of the Presentation this is located in the rocky slopes of the south side of the main temple in Caneo, near to the St. Johan Monastery Church in Ohrid. The first testing of the hybrid mobile application started without using Internet connection, which can be seen on screen capture of the application (see Figure 2).

The testing continued by getting closer to the cave church of the Presentation in Caneo, we were able to capture couple of images using the built-in camera. The screenshots of using the camera where made on the HTC mobile device (see Figure 3).

Using the camera from the mobile device we are able to capture the moment of visiting the cave church of the Presentation in Caneo. The image given in Figure 4 shows the cave church very near. Using this feature we were able to store the images in the image gallery for visited cave churches or the HTC phone gallery.

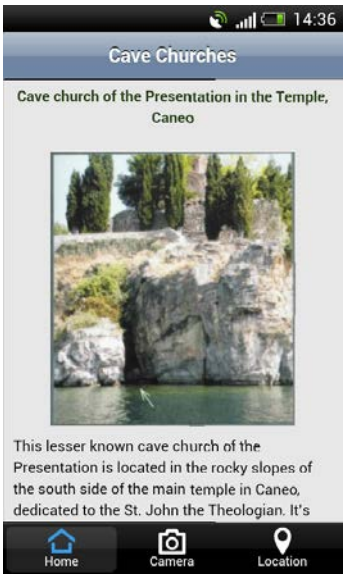


Figure 2. First look at the hybrid mobile application (no Internet).

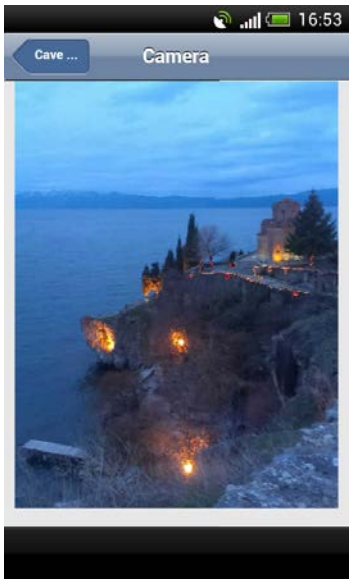


Figure 3. Using the camera from hybrid mobile application on site (no Internet connection).



Figure 4. Using of the camera from hybrid mobile application on site closer to the object (no Internet connection).

The location of the cave church of the Presentation in Caneo is determined using the geo-location plugin API. During our second visit to the site we were able to test the hybrid mobile application using the available 3G mobile data network connection. In order to increase the user interaction with the hybrid mobile application we have added location service. This service is provided by geo-location API, which uses simple JavaScript API and this provides the exact position using the GPS sensors built into mobile device. Therefore, we have created other two screen captures of the hybrid mobile application. The second testing of the hybrid mobile application was done using the HSDPA network, which allows us to have Internet connection on site. This can be seen on screen capture of the application, indicated with the symbol H (HSDPA) and two arrows (see Figure 5 and 6).



Figure 5. Using the location on hybrid mobile application on site.

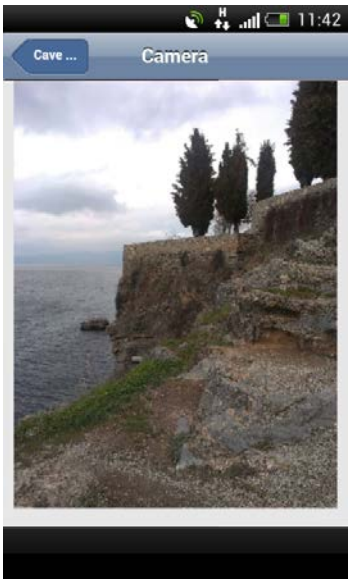


Figure 6. Using the camera from hybrid mobile application on site (with Internet connection).

The Figure 5 provides information for our current location on site, near the cave church of the Presentation in Caneo. Our current location is near the cave church of the Presentation in Caneo is marked with red circle. Whereas, in the Figure 6 is given screen capture of the use of built-in camera on site.

With this case study we have provided validation on use of the mobile device built-in sensors with the use of hybrid mobile application. This way the students using the benefit of hybrid mobile application development platform can easily create mobile app that can be distributed and used by variety of mobile devices. The creation of hybrid mobile applications in the same time this will increase the attention and provide better user experience in the process of m-learning.

VI. CONLUSION AND FUTURE WORK

This proposed model for adaptive multimedia content delivery is utilizing the benefit of hybrid mobile applications for m-learning. These mobile applications support the use of built-in functions on multiple OS vendors using the JavaScript wrap-up containers to enable the hybrid application to work on variety of devices. The main benefit of creating m-learning mobile application is the promotion of cave churches on the shore of the lake will include visual and textual presentation of each of cave churches, and their precise location. The continuation of the project will enable and contribute to the protection and promotion of cultural heritage of Ohrid-Struga region into an area not sufficiently made actual. In the future we plan to start covering the cave churches of the territory of the Republic of Macedonia, therefore we will ask for support from the Ministry of Culture of the Republic of Macedonia.

REFERENCES

[1] Global ICT developments, Report on Global mobile-cellular subscriptions by ITU Statistics, www.itu.int/ITU-D/ict/statistics/explorer/index.html <accessed: 05.01.2014>

[2] A. Nagesh, C. Caicedo, "Cross-platform Mobile Application Development," ITERA 2012 Conference, Indianapolis, IN 2012.

[3] Lionbridge, Mobile Web Apps vs Mobile Native Apps: How to Make the Right Choice, 2012. http://www.lionbridge.com/files/2012/11/Lionbridge-WP_MobileApps2.pdf <accessed 08.02.2014>

[4] K. Singh, "Practical Context-Aware Permission Control for Hybrid Mobile Application," Research in Attacks, Intrusions, and Defenses. pp. 307-327. ISBN: 978-3-642-41283-7.

[5] Native, Web or Hybrid Mobile-app Development, White Paper by IBM Software India on May 22, 2013. http://public.dhe.ibm.com/common/ssi/ecm/en/wsw14182usen/WSW14182USEN.PDF <accessed 09.02.2014>

[6] M. Power, "Delivering Web to Mobile," JISC CETIS Observatory. 2012. http://bitly.com/jiscobs-tw-2 <accessed 08.02.2014>

[7] M. Jern, "How to choose the right technology? Native, HTML5 and more," Published in Mobile Strategy Handbook Chapter 6. July 5, 2013. http://www.mobilemarketinguniverse.com/mobile-strategy-handbook-chapter-6-how-to-chose-the-right-technology-native-html5-and-more/ <accessed 06.02.2014>

[8] PhoneGap http://www.phonegap.com <accessed 03.02.2014>

[9] Appcelerator Titanium http://www.appcelerator.com <accessed 03.02.2014>

[10] Intel® XDK http://xdk-software.intel.com/ <accessed 03.02.2014>

[11] G. Angelicin-Zura, Cave churches: Ohrid coast, Matica macedonian- Skopje, 2013. <original cite: Г. Ангеличин – Жура, Пештерските цркви: охридско крајбрежје, Матица македонска – Скопје, 2013>

[12] U. B. Nagella and D. P. Govindarajulu, "Adaptive Approaches to Context Aware Mobile Learning Applications," International Journal of Computer Science and Security, Volume (2): Issue (2), 2009, ISSN:1985-1553.

J. M. Su, S. S. Tseng, H. Y. Lin, and C. H. Chen, "A personalized learning content adaptation mechanism to meet diverse user needs in mobile learning environments," User Modeling and User-Adapted Interaction 21, 1-2 (April 2011), pp. 5-49, DOI=10.1007/s11257-010-9094-0.

A Model of Adaptive Learning with Interactive Images - ADELE

Marjan Gusev, Sasko Ristov
University Ss Cyril and Methodius, FCSE,
Rugjer Boshkovic 16, Skopje, Macedonia
Email: {marjan.gushev, sashko.ristov}@finki.ukim.mk

Pano Gushev, Goran Velkoski
Innovation LLC,
Belasica 2, 1000 Skopje, Macedonia
Email: {pano.gushev, goran.velkoski }@innovation.com.mk

***Abstract*—Google has enabled the use of interactive image technologies for development of Google maps. The technology is based on presenting only a part of the image, which has sufficient details. This part is selected by appropriate zoom level or navigation through the image. The complete technology pre-processes the image in such a way to enable navigation and zooming and transferring of relatively small files over Internet, instead of transferring the complete high-resolution picture.**

In this paper we present several ideas how to use this technology in the learning process. We present how this technology can be used in analysis of medical images, and in realization of an adaptive e-Learning environment realized by e-Testing features. ADELE (ADaptive E-LEarning) is the name of the concept we propose in this paper to enable this integration. The realized applications are at least the following: 1) realization of a test with an interactive image, 2) opinion image, 3) collaborative learning with interactive images, and 4) voting image for collaborative learning. Each of these application areas will be described with functional description and analysis.

***Keywords*-e-Learning, Interactive Images, e-Testing, Cloud Computing, Medical Images**

I. Introduction

Learning refers to a cognitive process of acquiring knowledge and skills. Adaptive learning is a specific learning process adapted to the current learner's knowledge and skills. Adaptive learning can be efficiently implemented by e-Learning tools. According to Clark and Meyers [1] definition of e-Learning types, adaptive learning maps to *knowledge construction* method e-Learning type with perform-principle goals.

The realization of the adaptive learning as e-Learning may have various implementations, but we use our own approach by a system of online learning with interactive testing [2], based on a well known Socrates and Aristotle's methods, who educated students by continuously asking them questions. This process stimulates critical thinking, and illuminates ideas [3], [4] and finally leads towards knowledge construction with knowledge discovery technology. The students think that they have discovered the solutions, although it was the professor who led them towards careful analysis and deductive conclusions. Our approach uses the computer instead of a professor and acts as an adaptive software agent or a kind of intelligent agent, which will continuously ask the questions and react based on the answers provided.

The background to enable adaptive learning was the establishment of a sophisticated e-Testing system that was recently upgraded to a cloud based solution [5] and Assessment as a Service solution [6].

Recently we have realized an e-Testing systems that uses interactive images, called e-Testing with interactive images (eTii) [7]. The technology to realize interactive images is initially proposed by Google for their Google maps and its application in education [8]. Although, Google refers only to maps using this technology, we find that it can be efficiently used for various interactive images [9], such as medical images or images found in similar scientific areas.

In this paper we propose a new system for adaptive e-Learning based on e-testing with interactive images, called ADELE. The idea is to integrate the new concept of testing with interactive images into an adaptive e-Learning system.

The rest of the paper is organized as follows. The concept model of the ADELE system is presented with explanation of users and functions, interactions and functional description in Section II. Section III presents the main architecture model, along with intelligent software agent concepts and the adaptive e-Learning realization. Related work is presented in Section IV and discussion in Section V, along with presentation of obstacles, advantages and benefits of a such system. Finally, Section VI gives the conclusion and overview of future work.

II. Concept Model of ADELE

The main idea behind the ADELE concept is in realization of an intelligent software agent that acts as a professor in the learning process. The agent particularly has a professor role to evaluate the current student's knowledge and skills and lead him/her towards higher learning goals.

To present the ADELE concept model we will describe the actors in the learning process, and their interaction in the environment.

A. Users and functions

Stakeholders in the ADELE concept are instructors, students, external experts and interested parties, as presented in Figure 1.

Instructors are the users that are setting the e-Learning environment, by performing several functions: uploading samples, labeling characteristic points, lines, contours, identifying appearance of relevant knowledge items and learning objects, developing questions and learning materials. In addition to

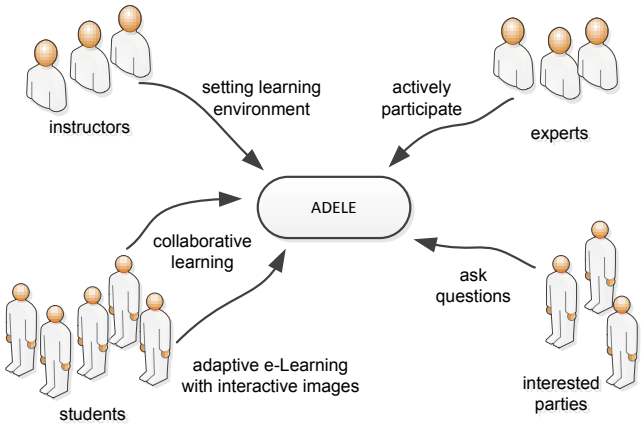


Fig. 1. Users and main ADELE functions.

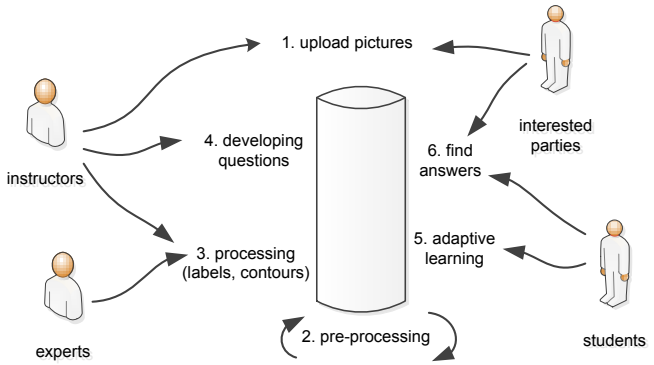


Fig. 2. Sequence of interactions of main ADELE functionalities.

this, they are capable to view and administer the progress of students, with functions that modern Learning Management Systems (LMSs) have.

Students use the system as an adaptive learning environment with interactive images. In this system, besides the learning materials with static samples, and basic textbook explanations, there are questions with interactive images, where they can analyze the question, and point or mark a contour where the knowledge item has to be identified on the interactive image. The adaptive system returns true or false responses for given answers, and the student is motivated to identify the appropriate region that is requested by the given question. In addition students may use collaborative learning approach and discuss with other colleagues about particular questions and answers.

External experts in the system can actively participate on a commercial basis, by giving their opinion for a specific uploaded sample. They can also contribute in the e-Learning system by setting new research results and discovered knowledge items.

Interested parties are those users that can use the system to analyze the latest research results. This is especially important for realization of the life-long learning concept. There is also another group of interested parties, a case when they would like to uploading images and ask for experts' opinion, and use the system with the conventional search tools.

B. Interactions

Interactions among users in the ADELE concept present how the users use the system in the adaptive e-Learning environment. A sequence of interactions that provide the main ADELE functionalities is shown in Figure 2. The following essential activities and functions are identified to be:

1) *upload pictures*, performed by instructors and interested parties;
2) *pre-processing*, realized by the automated system to enable realization of navigation and zooming of pictures;
3) *processing*, realized by experts and instructors, by setting labels and identifying contours for given knowledge items;
4) *developing questions*, realized by instructors, by setting questions using processed samples and enabling an interactive image for a given learning objective;
5) *adaptive learning*, realized by registered students for a given course, where the learning material with interactive images is accessed and used within the adaptive environment that adjusts to a given knowledge of a student; and
6) *find answers*, used by interested parties and students in the life-long learning concept, by enabling search functions and also opportunity to ask the expert for an opinion about a particular uploaded picture.

A more detailed functional description of these activities is presented in the next section.

C. Functional description

To realize the ADELE interactive system means to design and model a platform where text, images and other interactive multimedia elements can combine and integrate in a sophisticated learning environment. The idea is to include rich and intuitive interfaces for teachers and students and simulations and representations for teaching, learning and communicating about a topic.

Technology for interactive images includes upload of huge sized high resolution photos, their pre-processing to be used in web environment for different zoom levels and positioning, setting multimedia labels to certain points, establishing on-line quizzes based on point and click, or click and drag actions, use of multi touch pads etc.

Compact, mobile, primarily touch-based electronic devices could be used, but must be significantly improved for highly interactive medical e-learning. (Multimedia Messaging Service) MMS will research and develop novel user interfaces for e-learning on mobile devices, which may include touch, video or audio based interactions with medical training modules to maximize the user experience.

Interactive images and multimedia technology includes at least the following functionalities:

- *Small resolution pictures* offered via web to enable fast communication and response times, even if Internet speed is not fully broadband;
- *Human Computer Interface with input devices for pointing and sliding*, like point, click, drag, slide and other multi-touch pad interactions;
- *Picture navigation* technologies enabled by zooming and navigating though details in the image, even on multi-touch screen tablets;
- *Pointing and line (contour) drawing* functions realized by clicking, drag and drop or sliding actions, offering a possibility to point and mark picture regions;
- *Labeling and line identification* as technology to enable an environment with learning materials - descriptions of a certain point, region or picture object;
- *Multimedia interaction through sound and video* in addition to text annotations defined with corresponding protocols; and
- *Various picture processing components*, such as blurring, fading, etc.

D. Human Computer Interface

E-Learning human computer user interfaces for advanced image processing includes at least introduction of the following:

- *Touch screens*. The idea behind interactive images needs a pointing device that moves relatively fast over the screen surface. It is used for pointing, line drawing, and similar actions.
- *Multi-touch technology*. This technology is necessity to enable zooming, scaling and similar effects in image processing.
- *Portable cloud enabled devices*. Devices that are small enough and equipped with Internet access are desired to enable full outdoor access 24 hours a day.

Graphical user interface between users and the system can be realized by mouse, keyboard or touch pads. The system responds with presentation of textual and graphical objects. The expected input entered by the users in realization of tests is mostly click by a pointing device (mouse or touch pad) or entered value by a keyboard or selection object accessed by a pointing device (mouse or touch pad).

GUI model to realize the e-Testing with interactive images (eTii system) includes navigation and zooming [9]. For this purpose the system carefully administers the XY position and the size of the visible image part, and loads appropriate image parts stored on the server to present a figure with sufficient resolution. Interaction with image includes at least: object selection, region marking, setting a pin, commenting the pin, and selecting multiple objects.

E. Question Types for Interactive Images

Recently, we have introduced [7] three new question types intended for interactive images:

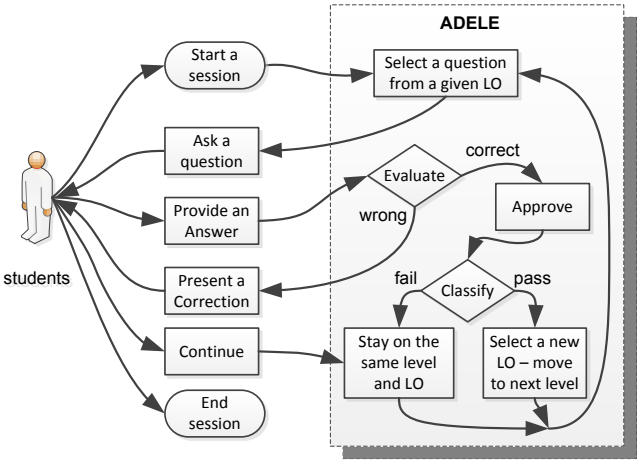


Fig. 3. Algorithms in ADELE adaptive e-Learning concept.

- *Single Graphic Choice (SGC)* question, which corresponds to a graphical variant of a classical multiple choice question with a single answer, but with enormous possibilities to include answer options, presented by dots in the image.
- *Opinion Map (OM)* question, which can be used to collect opinion about certain objects in the image. This is ideal for collaborative learning, where users can set a pin and write a comment on a specific graphical object.
- *Multiple Graphic Choice (MGC)* question, which is similar to the SGC, but the answer is constructed not on a single click on the image but on a selection of several objects.

The research realized afterwards showed possibility of new question types that are already built in our eTii prototype:

- *Voting Question (VQ)*, which corresponds to a graphical variant of a textual voting, where one can click on a specific object that is required by the question, in order to express the opinion with votes instead of comments.
- *Matching Question (MQ)* intended to enable matching among different graphical objects by assigning a line among these objects. Only pairs of objects can be matched.

III. ADELE Architecture Model

The architecture model will be presented by description of the realization of the adaptive e-Learning concept, the intelligent software agent approach and the ADELE main building modules.

A. Adaptive e-Learning

The concept behind establishment of an adaptive e-Learning environment is presented in Figure 3. A student starts the learning session and the ADELE system selects a question from a given Learning Objective (LO). Then the e-Testing engine is started and the question is presented to the student. The student provides an answer on the interactive image via

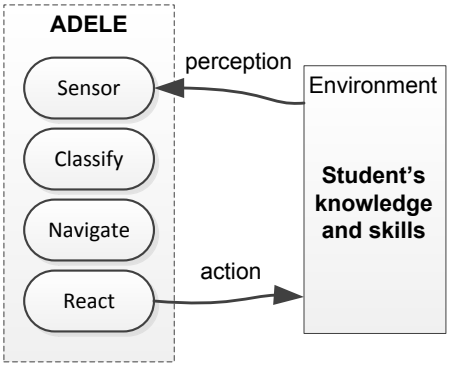


Fig. 4. ADELE adaptive e-Learning as an intelligent software agent.

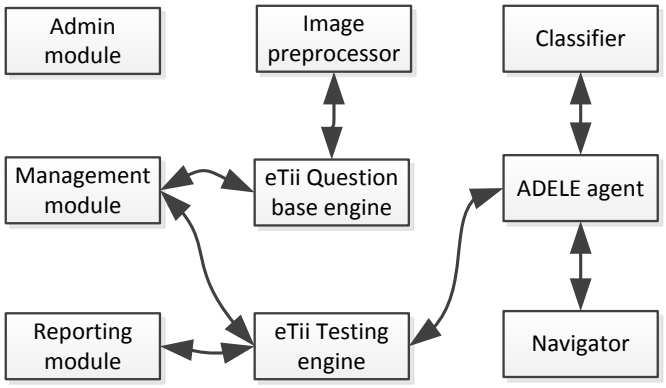


Fig. 5. Overall architecture of the ADELE model.

the graphical user interface, as presented in previous sections. The e-Testing engine is responsible to evaluate the answer and responds with wrong or correct options. In case of the wrong answer the e-Testing engine provides a corrective measure by displaying the correct answer and suppying explanation.

All previous functions and interactions were realized by the e-Testing engine, as described in the eTii prototype. It acts as a sensor part of the agent which makes the perception of the environment - presented by the evaluated level of provided answer.

In case of approval or when the student continues the session, ADELE starts with classification of the student's knowledge and skills. The classification algorithm responds with pass or fail as an estimation if the student has learned the corresponding LO.

The next part is realization of the navigation algorithm, which determines the level and LO adapted to the current student's level. It continues with the same level and LO, or in case of conclusion that the student has passed the corresponding LO, it moves to the next, usually higher level, with selection of a new LO.

B. Intelligent Software Agent

The intelligent software agent concept is presented in Figure 4. The environment is presented by the evaluated student's knowledge and skills. The sensor is realized by the eTii testing engine to realize the perception of current knowledge presented by the student answer. Later, another module realizes the classification of overall student's knowledge for the given LO and makes a decision. The next module realizes the navigation in the pool of LO presented by corresponding levels. This finishes with reaction, presented by asking the next question from a selected LO, and realized by the eTii engine.

A comprehensive explanation about classification strategy and navigation algorithm can be found in our recent publication [2]. We have analyzed several classification strategies to evaluate student's knowledge with pass or fail, by asking a subset of questions from a given pool of questions in the LO. The conclusion is that the strategy three correct answers in-a-row was the best alternative, when analyzing several criteria, such as avoiding boring repetition of the same question within a given LO, having sufficient probability of correctness if all questions were asked, analyzing user satisfaction, and analyzing results from provided experiments.

The navigation algorithm through levels was realized according to the already known tree traversal algorithm, particularly the preorder traversal algorithm for visiting all nodes in a given level before moving to a higher level (LRO preorder is visiting first the left beaches, then the right and at the end the node itself).

C. The ADELE architecture

The basic architecture is based on the eTii testing solution [7]. The overall architecture model is presented in Figure 5, presenting the 9 essential modules and some of the essential workflow communications among modules.

The architecture of the adaptive e-Learning concept of the ADELE system is using the following three modules:

- *ADELE agent*, which provides intelligent agent software functionality, communicating with sensor, classifier, navigator and finally reacts upon fulfillment of certain conditions;
- *Classifier*, which realizes the classification algorithm and makes decision about pass or fail; and
- *Navigator*, which realizes navigation in the knowledge levels and LOs, enabling staying in the same LO, or moving to the next levee and LO, and also selection of appropriate question from a given LO.

The overall architecture of background activities that support the ADELE system is using the following three modules:

- *eTii Testing engine*, which provides e-Testing functionality with interactive images and realizes the graphical user interface with the students, including functionalities, such as question asking, students answering, evaluation, presenting corrective measures, approval, etc;
- *eTii question base engine*, which provides functionality to fill the question base with interactive images, including functionalities, such as question developing, question processing, level organization, LO organization, etc;

- *Image pre-processor*, which processes images for exploitation as interactive images;
- *Reporting module*, which prepares and presents reports, statistics and analytics;
- *Management module*, which is responsible for management of classes, learning sessions, and overall system realization of a course; and
- *Admin module*, which realizes the administration of courses, users, authentication, authorization, and security.

IV. Related work

Two categories of adaptation are identified in learning environments [10]. The first is about *Adaptive Interaction* and refers to interface adaptations that facilitate the process of user's interaction. The second targets the *Adaptive Collaboration Support* intended to capture adaptive learning involving collaboration towards common objectives.

There are currently numerous organizations, like the Dublin Core Metadata Initiative, the IEEE, the IMS Global Learning Consortium, the Alliance of Remote Instructional Authoring and Distribution Networks for Europe, the Aviation Industry CBT Committee, the Advanced Distributed Learning Initiative, etc., that are dedicated to, or have committees and working groups active in the establishment of eLearning standards. It should be noted that the core standards are the various specifications of IMS1, ADL SCORM2, and the AICC specifications [11].

Having an intelligent (adaptive and interactive) online system is very important both for the students and teachers since the students gained knowledge and skills more easily when an interactive and adaptive environment for online learning is used [12]. Such a system should also recognize the students' motivation, emotion and mood condition since disengaged students usually gain less from learning than more motivated and engaged students. Several relevant attributes identified by Cocea and Weibelzahl [13] can be used to predict the students' disengagement. They even improved the prediction by analyzing the log-data of the system (the time of reading pages) [14]. Moridis and Economides [15] use the previously students' correctly answered questions to determine their mood during the assessment. They even improved their method by including the latest correct or incorrect answer [16]. An adaptive Learning Management System is proposed by Yaghmaie and Bahreininejad [17] by using link-level and content-level levels of adaptation to the learners' needs.

Several authors benefit of applying formative assessment in learning environments [18], [19]. Hwang and Chang [20] proposed an approach based on a formative assessment in order to improve the learning outcomes of students in a mobile learning environment.

However, although all these previous solutions, models and frameworks improve the learning in one or another way, all of them are using multimedia only as a support to the knowledge base. Revolutionary, in this paper we propose an adaptive e-testing system based on interactive images.

Browser technology to enable the required navigation and zooming depends on existing Google APIs [21]. Although the API was originally intended for maps only, we found that this technology is well suited for other scientific images, including astral images, medical images, etc. These are high resolution and huge size images and the only way to make a detailed analysis is by using the navigation and zooming features. Especially this is the case when the images are loaded from Internet. Current zooming enabled in the API allows up to 10 levels, which enables sufficient details. Storing the image in all zoom levels usually uses up to 10 times more storage capacity, than the original size, but the average load size of each image in a single session is at least 100 times less, as our experiments prove.

V. Discussion

The main obstacles in enabling a high resolution image presentation and manipulation of graphical objects is the size and enormous communication times to transfer the whole image from a server to the web browser that required it. Knowing that only a selected image part should be presented with sufficient details, the idea behind presenting the global figure with low resolution, enable navigation and zoom only the selected part saves a lot of time.

The application domain is not limited in only specific areas, such as, astral images or medical images. To realize the proof of concept for the eTii system, which is intended for e-Testing with interactive images, we have developed a set of 200 images for the course of Computer Networks [7]. The questions asked included typos of:

- Click where can you find an occurrence of a typical object or a concept on the figure!
- Mark a region with specific characteristics on a figure!
- Comment specific characteristics of each object!

For example, a figure contained a complex network with a lot of devices and the student had to identify which device is a router, or switch, or mark a subnet region. To realize this, the student has to navigate and zoom enough to read all details, including IP and MAC addresses assigned to each device. Another example is analysis of a sequence of signal levels and application of a specific data encoding scheme, such as biphase Manchester, NRZ or similar encoding techniques. There were several mistakes on purpose and the student has to click on all mistakes.

The motivation to include interactive images instead of classical multiple choice questions is based on several criteria, including extended domain of search options, extended answering possibilities, favoring creativity and solution discovery instead of memorizing, decreased cheating etc. We have already published a more detailed analysis of these advantages and analyzed in details in [9].

Combining the advantages of applying interactive images with the adaptive e-Learning concept tends to have even a more precise evaluation of current student's knowledge and skills, from one side and from the other side, to enable a better

learning platform, which initiates a higher degree of curiosity and establishment of a collaborative learning principles. Interactive images offered via an adaptive learning environment increase the overall understanding of learned material, avoiding memorization in multiple choice questions. Having a visual presentation instead of textual phrases is always a better concept to understand the corresponding concepts and analyze their relations. The system also provokes knowledge discovery instead of guessing in multiple choice questions. These concepts support the overall goal for knowledge construction method of the ADELE system.

VI. Conclusion and Future Work

In this paper we have presented the ADELE concept of an e-Learning adaptive environment that exploits interactive images. The work on building a prototype has already started and we expect that it will be realized in 2014.

The complete ADELE architecture consists of several modules, such as the eTii question base engine, image preprocessor, management module, reporting module, admin module and adaptive e-Learning module. Our solution for realization of the adaptive e-Learning module uses the already developed eTii engine for e-Testing with interactive images and on top of it, it builds the classifier and navigator modules. It acts as an intelligent software agent by sensing the student's knowledge and skills from the answer given, making a decision on how to proceed with corresponding classification and navigation algorithms and then reacting by asking a new question.

As future work we plan to test which strategy suites the best when nteractive images are applied, since we have used the three correct answers in-a-row strategy, mostly as a conclusion from realization of a conventional adaptive environment that uses multiple choice questions. The experiments will cover student's behavior and enable better estimation about classifying strategy to decide whether the student has learned an appropriate knowledge item or LO. We plan to realize experiments about the realization of the navigation algorithm, also. The user satisfaction and usability of the applied system are also foreseen to be realized as soon as the prototype is build and used. Development of a strategy to realize relevant learning content and development of a sustainability strategy are also planned.

We plan to continue with development of questions for other courses from computer science and create an extensive database of questions. Our experience shows that these type of questions are more difficult than the standard multiple choice questions, and that the students need a lot of concentration to solve the test. In addition, these questions actually present those analytical skills to understand the concepts and analyze complex figures, in comparison to the multiple choice questions which can correlate to memorizing capabilities of the student. We plan to report these findings in details after the finalization of the course and analysis of obtained results.

References

[1] R. Clark and R. Mayer, *E-learning and the science of instruction: Proven guidelines for consumers and designers of multimedia learning*. Pfeiffer, 2011.

[2] M. Gusev and G. Armenski, "e-Testing usage in e-Education," in *e-Learning Paradigms and Applications, Agent-based Approach*, M. Ivanovic and L. Jain, Eds. Springer Verlag, 2014, vol. Studies in Computational Intelligence, vol. 528, pp. 229–249.

[3] A. M. Collins and A. L. Stevens, *A Cognitive Theory of Interactive Teaching*. Bolt Beranek and Newman Inc Cambridge MA, 1981.

[4] J. Beck, M. Stern, and E. Haugsjaa, "Applications of ai in education," *Crossroads*, vol. 3, no. 1, pp. 11–15, 1996.

[5] S. Ristov, M. Gusev, G. Armenski, K. Bozinoski, and G. Velkoski, "Architecture and organization of e-assessment cloud solution," in *IEEE Global Engineering Education Conference (EDUCON), 2013 IEEE*. Berlin, Germany: IEEE, Mar. 2013, pp. 736–743, best Paper Award.

[6] M. Gusev, S. Ristov, and G. Armenski, "Interactive adaptivity in assessment as a service," in *Interactive Collaborative Learning (ICL), 2013 16th International Conference on*, Kazan, Russia, Sep. 2013, pp. 603–610.

[7] M. Gusev, S. Ristov, G. Armenski, P. Gushev, and G. Velkoski, "E-assessment with interactive images," in *Global Engineering Education Conference (EDUCON), 2014 IEEE*, Istanbul, Turkey, 2014, p. in press.

[8] B. Taylor. (2005, June) The world is your JavaScript-enabled oyster. [Online]. Available: http://googleblog.blogspot.com/2005/06/world-is-your-javascript-enabled_29.html

[9] M. Gusev, S. Ristov, and G. Armenski, "E-testing with interactive images - opportunities and challenges," in *Human Computer Interface Conference (HCI)*, Crete, Greece, 2014, p. in press.

[10] A. Paramythis and S. Loidl-Reisinger, "Adaptive learning environments and e-learning standards," in *Second European Conference on e-Learning*, 2003, pp. 369–379.

[11] P. Brusilovsky *et al.*, "Adaptive and intelligent technologies for web-based eduction," *KI*, vol. 13, no. 4, pp. 19–25, 1999.

[12] M. Gusev, S. Ristov, and G. Armenski, "Adaptive interactive online course on computer architecture and organization," in *Information Technology Based Higher Education and Training ITHET 2013*, 2013.

[13] M. Cocea and S. Weibelzahl, "Eliciting motivation knowledge from log files towards motivation diagnosis for adaptive systems," in *Proceedings of the 11th international conference on User Modeling*, ser. UM '07. Springer-Verlag, 2007, pp. 197–206.

[14] ——, "Disengagement detection in online learning: Validation studies and perspectives," *IEEE Trans. Learn. Technol.*, vol. 4, no. 2, pp. 114–124, Apr. 2011.

[15] C. Moridis and A. Economides, "Modeling students mood during an online self-assessment test," in *The Open Knowlege Society. A Computer Science and Information Systems Manifesto*, ser. Communications in Computer and Information Science, M. Lytras, J. Carroll, E. Damiani, R. Tennyson, D. Avison, G. Vossen, and P. Ordonez De Pablos, Eds. Springer Berlin Heidelberg, 2008, vol. 19, pp. 334–341.

[16] ——, "Mood recognition during online self-assessment tests," *Learning Technologies, IEEE Transactions on*, vol. 2, no. 1, pp. 50–61, 2009.

[17] M. Yaghmaie and A. Bahreininejad, "A context-aware adaptive learning system using agents," *Expert Syst. Appl.*, vol. 38, no. 4, pp. 3280–3286, Apr. 2011. [Online]. Available: http://dx.doi.org/10.1016/j.eswa.2010.08.113

[18] L. Gardner, D. Sheridan, and D. White, "A web-based learning and assessment system to support flexible education," *Journal of Computer Assisted Learning*, vol. 18, no. 2, pp. 125–136, June 2002.

[19] P. Orsmond, S. Merry, and A. Callaghan, "Implementation of a formative assessment model incorporating peer and selfassessment," *Innovations in Education and Teaching International*, vol. 41, no. 3, pp. 273–290, 2004.

[20] G.-J. Hwang and H.-F. Chang, "A formative assessment-based mobile learning approach to improving the learning attitudes and achievements of students," *Comput. Educ.*, vol. 56, no. 4, pp. 1023–1031, May 2011.

[21] F. Lardinois, "Google Launches Maps Engine API To Allow Enterprise Developers To More Easily Create, Share And Publish Custom Maps," Jun 2013. [Online]. Available: http://techcrunch.com/2013/06/05

User Satisfaction Evaluation of the Electronic Student Services System - iKnow

Ana Guseva
Innovation Dooel
1000 Skopje, Macedonia
ana.guseva@innovation.com.mk

Marjan Gusev, Sasko Ristov
Ss. Cyril and Methodius University
1000 Skopje, Macedonia
{marjan.gushev, sashko.ristov}@finki.ukim.mk

***Abstract*—In this paper, we report a survey that gathered the opinions on electronic student services by implementing the iKnow system. The methodology to evaluate user satisfaction was based on realization of questionnaires, and surveys with students, as well as interviews with the teaching and administration staff. We have realized several surveys and questionnaires, not just during the development cycle, but also during first two years of exploitation.**

The results show that most respondents have a positive attitude towards the iKnow software system. A very interesting fact is the one about swing of the attitude towards better recognition of the system, mostly due to the understanding of iKnow possibilities at all. Overall, our findings imply that iKnow has alleviated much of the basic and long-range problems that have challenged students in the past. Additionally, the study also reveals some shortcomings of iKnow, and lays ground for defining and building an even better user experience with the iKnow system.

***Index Terms*—*user interface design*; *software verification*; *software validation*.**

I. Introduction

iKnow [1] is an E-Student Services System developed as a delivery of an EU financed Tempus Project. It realizes the vision of a paperless university management and electronic workflow of student activities. Both students and staff use iKnow as a web-based application to automate the entire system of student programs, courses, the relationships among courses, courses with programs, mapping of teachers with courses, etc. The processes of studying, enrollment of students in semesters, enlisting in courses, exams passed, ECTS credits earned are also managed in this way. Members of faculty, classrooms, laboratories and all resources can be be administered with this module. Using iKnow, a total of 76 different workflow activities previously done by administration personnel are now automated and only a few clicks away.

Yet, the question whether iKnow will fully replace the outdated Student Services Desk largely depends on good user experience and acceptance by the users. iKnow is intended as an essential and indispensable component of daily university administration, collaboration and communication. In fact, users need to spend a significant amount of time daily interacting with iKnow. It encompasses a wide set of functionalities and presents a huge amount of data. For these reasons, it is evident that a seamless and pleasant user experience is of crucial importance and may have drastic consequences. Nielsen [2] identifies that usability has multiple components and is traditionally associated with five usability attributes: learnability, efficiency, memorability, errors and satisfaction. In this paper we are tackling only the user satisfaction. A user-friendly and user-centered design has become an increasingly important feature of all modern information technology systems.

User satisfaction refers to a domain, where the users do not have the relevant software development documentation that expresses the user requirements and software design, but can help in evaluation if the realized system corresponds to their expectations. In this paper we target the end users to find out how much they accept the system and to evaluate their attitude. We use a methodology with realization of questionnaires, surveys and interviews, realized immediately after deployment and also after two years of exploitation. This approach will show if there is a change in general attitude and if the software has been improved to meet the user expectations.

User experience (UX) shows a growing interest within the HCI (Human-Computer Interaction) community, both in academia and industry. Recently HCI researchers and practitioners have recognized that products are not merely means to achieve goals; they fulfill our need for stimulation [3], they evoke memories [4], and they provoke emotional reactions [5]. Products need not only to succeed in gaining initial acceptance, but also to become objects of increased emotional value over prolonged use.

In this regard, the traditional user satisfaction testing framework has its limitations. It focuses primarily on user cognition and user performance, In contrast, UX addresses human needs that go beyond the instrumental. In fact, UX highlights non-utilitarian aspects of human-computer interactions, shifting the focus to user affect, sensation, and emotion. As [6] observes, UX is associated with a broad range of fuzzy and dynamic concepts, including emotional, affective, experiential, hedonic, and aesthetic variables.

The rest of the paper is organized as follows. The iKnow system is briefly described in Section II. In Section III we present the used methodology and elaborate on implementation decisions. The results are presented and discussed in Section IV. Finally, relevant conclusions and future work is specified in Section V.

II. Description of the iKnow System

The eStudent Information System iKnow is designed to store and administer student's records and personal files, as well as related university data. It is developed using innovative approach and knowledge management techniques. The system provides exchange of electronic information among all stakeholders: students, professors, administration, university management and the Ministry of Education [1].

The system consists of two main components: the new students Enrollment Student Services (ESS) and the Core Student Services (CSS). Both components consist of modules and special software modules have been developed for data migration from legacy software applications into iKnow using Excel template files.

The ESS component consists of the candidate's enrollment wizard, the enrollment forms for manual entry of candidate's data, candidate's data processing, ranking module and the enrollment results publishing.

A total of 15 smaller modules define the CSS component, including the following:

- Module for administration (university and faculties), which encompasses users management, exams sessions, semesters and the integration with the learning management system Moodle;
- Module for study programs, which encompasses schedules, study programs management, courses management, equivalency of courses, connections between courses and programs, etc; and
- Module for academic records with goal to administer evaluation of academic activities that students perform while following the course, including laboratory assignments, homework, projects, partial and final exam results.

28 roles have been defined in the CSS component, including 6 different roles of technical, resource, IT and overall administrators on university, faculty or computing center level. Several roles have been assigned to university management administration, including HR department, student services, financial officer, ECTS coordinator, rector, deans, academic vice-rectors and vice-deans, and professors. The main intended role are the students, but the set is extended to tutors, evaluators, and other similar correlated roles.

76 workflow processes have been realized within the iKnow system, including enrollment in particular course subject, applying for exam, administration of academic records, repotting etc.

Overall the electronic student information system iKnow is a complex software solution used by more than 30.000 students and 3.000 staff at the Sts Cyril and Methodius University.

In the context of our study we tested both the CSS and ESS components.

III. Methodlogy

The goal of the methodology is to evaluate the user satisfaction with iKnow. According to Nielsen [2], from a usability perspective, questionnaires and interviews are indirect methods, since they do not study the user interface itself but only users' opinions about the user interface, and are direct methods when it comes to measuring user satisfaction. We have used questionnaire, survey, and interview methods to test the user experience and enable sufficient base for delving conclusions about user satisfaction evaluation.

A. Metrics

In the context of user satisfaction assessment, methods include focus groups, interviews, questionnaires, and surveys [7].

Interviews and focus groups are structured methods of inquiry, which are used to gather information about users' experiences and preferences. While surveys and questionnaires may also be used to gather such information, interviews and focus groups allow for more interaction with the users and for immediate answers to questions raised during the interview or focus group. In addition, focus groups consist of multiple users whose interaction may raise additional issues of interest [8]. Interviews and focus groups are generally conducted at early stages of product development, while surveys and questionnaires are generally used later in the product's life cycle. In an effort to incorporate the strengths and avoid the weakness of each method, we decided on a survey style with open-ended questions.

Regarding the survey design, it has been proven repeatedly that the way a survey question is asked can have influence on the answer. For instance, it is how political campaigns can generate opinion polls saying the exact opposite of what their opponent's poll revealed. It is evident, therefore, that the survey must be as neutral as possible, in order to yield an accurate, unbiased picture. Aside from the fact that the mere asking for their opinion will change the users' opinion of the product, a survey shouldn't in itself be considered a marketing tool. For this reason, the survey questions were asked in a plain, matter-of-fact way, and not in an advertising tone. In fact, much care was taken in designing the questions. For example:

- *NOT:* The iKnow team takes every attempt to please the students in regard to technical support. How would you rate our success in these efforts?
- *BUT:* How would you rate iKnow's technical support?

Another point requiring great care is designing the rating scale used. Lickert's scale [9] is used as a method of constructing an attitude with values from 1 to 5. The used system actually corresponds to the grading system used for the students in primary and and secondary schools, and the students are well used to it.

Specifically, giving an odd number of response options favors a 'neutral' response. In contrast, an even number of response options, without that comfortable middle ground will force a respondent to choose one or another of the extremes. To avoid similar traps, we decided to ask open-ended questions, and then evaluate the answers we received. Thus, the users' opinions were coded as:

- *(5) Strongly satisfied (strongly approve)*, flagged whenever a respondent expressed the most positive attitude,

the user is aware of, and strongly approves the statement;
- *(4) Satisfied (approve)*, flagged whenever a respondent expressed a positive attitude and approves the statement;
- *(3) Neutral (undecided)*, flagged whenever a respondent is neutral to express a positive or negative attitude, and had no particular comment or had no remarks;
- *(2) Not Satisfied (disapprove)*, flagged whenever a respondent had one or more remarks or stated problems encountered and disapproves the statement; and
- *(1) Strongly Not Satisfied (strongly disapprove)*, flagged whenever a respondent expressed the most negative attitude, the user is aware of, and had a lot of remarks or stated problems encountered and strongly disapproves the statement.

B. Designing the Surveys

User satisfaction questionnaires can be quite sophisticated [2]. For example, the QUIS (Questionnaire for User Interface Satisfaction) method reported by Chin et al. [10] rates 27 system attributes on 10-point scales, followed by a factor analysis. Furthermore [11], the provided factor analysis of a subset of subjective satisfaction questionnaire consisting of 18 questions shows importance of three underlying factors: system usefulness, information quality, and interface quality. It should be possible to assess these three factors with a reasonably small number of questions.

Another study provided by Bailey and Pearson [12] introduced 39 factors in the Computer User Satisfaction (CUS) questionnaire. Their conclusion based on rating of each factor, found that the users rated as most important: accuracy, reliability, timeliness, relevancy and confidence in the system. A lot of various studies also concluded that a simpler version of user satisfaction measurement instrument is necessary and sufficient to make relevant conclusions.

We developed a framework of questions (see Table I) during the project as a part of a collective team development of a corresponding control quality work package. The project participants including Technical University of Dortmund, Germany, London South Bank University, UK, University in Ljubljana, Slovenia, and University Ss Cyril and Methodius, Skopje, Macedonia, have actively participated in planing and developing the survey. A more detailed elaboration [13] shows the essential objectives of the survey, with target to evaluate the user satisfaction of the iKnow system. The focus was set on employees and students in development phase of the iKnow system, and after deployment on students as end users in the exploitation phase.

The final goal is to measure the user satisfaction with different aspects of iKnow. The used communication medium is digital (electronic) questionnaires, on line survey and personal interview. The techniques included closed questions defined as simple and clear statements using the already explained evaluation metrics Lickert's scale from 1 to 5. The sample size is 400 students in 2012 and 300 students in 2013. The questionnaire survey was conducted in a one week period during the semester as anonymous survey.

TABLE I: The content of the closed questioned survey

ID	Question
Q1	What is your general satisfaction from the iKnow system?
Q2	How do you evaluate the support that the system gives you as a student?
Q3	Do you support future usage of this system with enhanced functionalities?
Q4	The browser you are using displays all information in a correct way?
Q5	The loading time and the response time of the system correspond to the content which is displayed?
Q6	The system has a consistent look and feel?
Q7	The messages the system communicates are clear and descriptive?
Q8	There is an easy channel available to communicate with an administrator?
Q9	The content is understandable and up to date?

The planning process included analyses of several general user satisfaction criteria:

- Accuracy;
- Layout;
- Navigation;
- Exception and status handling; and
- User guidelines and online help.

The quality and information structure criteria included questions about content and efficiency.

Initially the draft survey consisted of more than 23 questions in the software development phase, when the survey was intended as customer response in the agile development. A careful analysis in the software exploitation phase after software deployment has recommended to reduce the number of questions to 9 essential questions. This was also necessary to make the survey as small as possible and applicable for greater number of students taking approximately 5-10 minutes on average for answering the survey. These nine questions showed to be the most essential for selected criteria in making final conclusions to realize the main objectives stated in this paper.

Finally, the realized survey consisted of those questions presented in Table I. The first four questions refer to the functionality criteria, the next three to the layout, and the last two to the navigation, user guidelines and online help criteria. We have reduced the questions that express the content and efficiency criteria, due to the fact that the software development phase is finished. This was also the reason to reduce the questions in the overall assessment criteria and also for exception and stats handling.

In addition to the closed question survey realized with relatively huge number of participants, we have realized also an open question survey for a smaller group of students, where they can express their comments for their user satisfaction. In addition we have added new criteria:

- New approaches; and
- Quality assessment.

The new approaches criteria were chosen mainly to include new sophisticated trends in realization of the software. It includes the user interface, which changes and improves

TABLE II: The content of the open questioned survey

New Approaches	
Q1	Could you share the application on Facebook using the iKnow system?
Q2	What do you think, what is that something that Facebook has, the iKnow system is missing as user interface?
Q3	What do you think, during the application, what information should be given, such as pictures, like on Facebook?
Q4	What do you think if your Facebook profile should be connected with your iKnow profile?
Q5	What do you think, if there should be an option for "Like" or "Comment" on the iKnow system, like there is on Facebook, to express yourself better?
Quality Assessment	
Q1	What do you think about the process of enrollment?
Q2	What remarks do you have about the steps of the enrollment application?
Q3	What do you think about the iKnow system?
Q4	What would you change in order to make the iKnow system better?

with new technologies to enable better user experience. Our intention was to include Facebook as the most popular social network and used application by the students (more than 95% students actively use Facebook).

The second criteria includes quality assessment of the both the ESS and CSS modules.

The survey presented in Table II included also total 9 questions, as the closed questioned survey. However, it was realized as interview and not as electronic version of a questionnaire, since most of the participants do not like to spend a lot of time on an electronic survey, while on interview will be able to provide more attention.

C. Conducting the Surveys

At the onset of the surveys, in September 2013, iKnow was well beyond the deployment stage, actually two years in exploitation and further improvement. The software has been officially accepted in October 2012, and in the last year several improvements have been realized (mostly in developing new reports, and improving some procedures). The process of re-design of workflow procedures for administration related to student services and academic records was finalized by September 2012 after alms a full year of meetings, presentations, and trainings. Probably the most difficult issue was persuading the university management to use the new system, although it had not yet been tested and rolled out. Luckily, all the processes were finalized and the system is officially used. The user satisfaction survey helps the university management to improve the system.

The interview method was conducted for a total of 32 participants randomly chosen among prospective students at the university who applied for enrollment. These users face the iKnow system for first time, in contrast to the students who provided answers for the closed questioned survey, who have been experienced in using the iKnow system. Statistically, this sample size yields still significant results intending to provide quality analysis of the used user interface technology, although there were more than 8000 prospective students applying for the enrollment at the university.

TABLE III: Statistics for the closed questioned survey

Question	avg. 2012	avg. 2013	st.dev 2012	st.dev 2013
Q1	3.10	3.68	13.81	12.65
Q2	3.18	3.52	16.07	11.27
Q3	3.72	3.90	12.22	14.72
Q4	3.18	4.32	17.82	23.93
Q5	3.18	3.56	17.82	13.52
Q6	3.35	3.75	21.71	13.23
Q7	3.32	3.70	16.99	13.17
Q8	3.16	2.77	21.45	07.79
Q9	2.92	3.33	13.86	09.80
Average	**3.23**	**3.61**	**16.86**	**13.35**

We have satisfied the selection criteria for socio-demographic characteristics, and also to cover various faculties the students were applying for. The sample included 16 males and 16 females and the origin town for half of them was the town where the University is situated.

Conducting the surveys was a relatively straight-forward task. We realized the survey by making interview, not in an electronic way with higher number of participants. This enabled to analyze open-ended answers and also ask questions in context to the discussions. Interpretation of this survey gave a set of suggested improvements and ideas.

To make the survey more attractive, an innovative method of organizing a lottery was included, where the interviewers draw a lucky question and were awarded with a small present (chocolate). Bearing in mind that the survey should take no more than 10 minutes, not all questions were asked of all respondents.

IV. Results and Discussion

The results include descriptive statistics including textual description and visual impressions, followed by discussion of trends and establishing conclusions about typical behavior. Finally the recommendations are given as a consequence of detailed analysis of the results, observed trends and typical behavior in the conclusions.

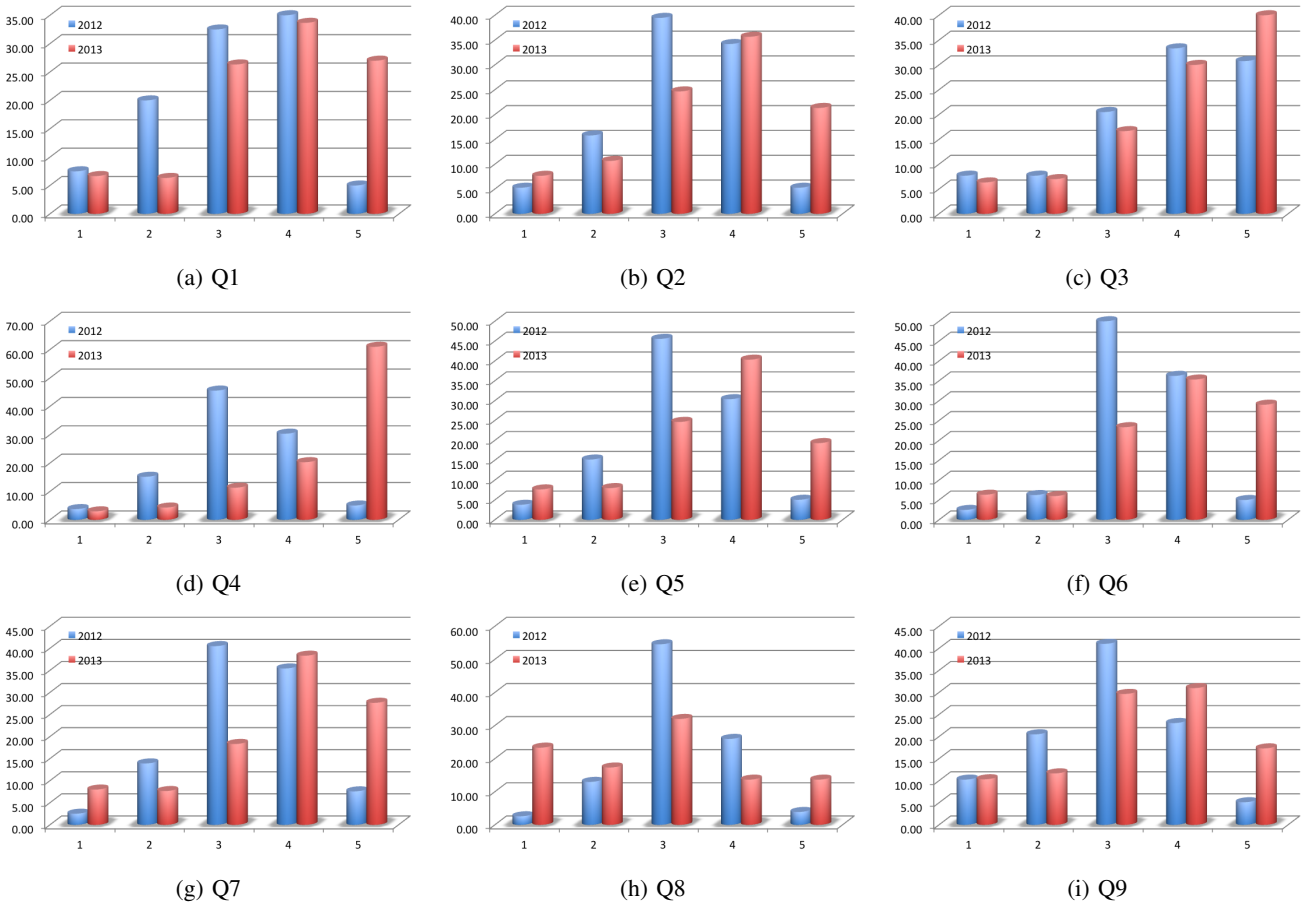
A. Discussion about results from closed questioned survey

A total of 400 students participated in the survey in 2012 and 300 students in 2013. Visual representation includes relative presentation of answers to avoid the difference of the number of students that answered a particular question and enable comparison of results in the two consecutive years.

We have compared the results from a survey conducted in 2012 [13] and a survey we have conducted in 2013. The relative responses are charted in Figure 1.

Statistical calculation of the average and standard deviation are presented in Table III. In the avg. columns we present the average grade calculated from provided discrete value answers, and in the st.dev we present the calculated standard deviation.

An interesting aspect of iKnow's UX evaluation is that of temporality, i.e. how the quality of users' experience develops over time. For this reason we compare our survey results conducted in 2013 with those conducted in 2012 [13].



(a) Q1 (b) Q2 (c) Q3

(d) Q4 (e) Q5 (f) Q6

(g) Q7 (h) Q8 (i) Q9

Fig. 1: Survey Results

One might immediately conclude that the results in 2013 are better than those in 2012, meaning that the small add-ons and improvements of the deployed software were successful and user satisfaction is increased from 3.23 to 3.61 on average.

Improvement is also concluded to the standard deviation as it gets smaller values, meaning that answers were convergent to the mean value and there were no huge discrepancies in provided answers. Relatively small values of the the standard deviation show that the curves follow the normal distribution, as it can also be seen in Figure 1.

A careful analysis shows a shift from *neutral* to *positive* attitude. This leads to a conclusion that in 2013 there are more satisfied students than those that were neutral in their answers. The charts undoubtedly confirm that the students have a positive attitude toward iKnow.

Normal distribution with average above neutral (middle) point is evident for most of the questions in both measurements, especially for Q1, Q2, Q5, Q6, Q7, and Q9. Raising trend of user satisfaction is obtained for Q3 for both measurements, meaning that there were more satisfied students than those unsatisfied. While Q4 had normal distribution in 2012, in 2013 it shows raising trend, meaning that great majority of students is completely satisfied.

Q8 shows discrepancy in answers for 2013 and normal distribution for 2012. We can conclude that there were no good channels for online help and direct communication with the administrators, i.e. the raised number of unsatisfied students showed that there were cases of students who were trying to reach the administrator, who, obviously had not responded. This is the only case of decreased average user satisfaction for value of 12%. However, this issue is not directly connected with the iKnow system, but with the administrators.

In all cases different from Q8, there is a positive shift of the user satisfaction of average 12% with minimum of 5% for Q3 up to 36% for Q4. The highest discrepancy in provided answers is recorded for Q4 in 2013 measurements and for Q6 and Q8 for 2012.

B. Discussion about results from open questioned survey

Most of the provided answers expressed satisfaction from the applying process and expressed what is missing by setting accent on paper forms with content same as the electronic form, some expected more graphics but were satisfied with the basic functionalities. A great majority expressed satisfaction that the system uses most of technology for user interface present at the FaceBook social network they are used to. Smaller system malfunctions were also reported to the software provider. Some provided answers are quoted as follows:

- "... All in all I'm satisfied with this tool, although personally I'd change few things. ..."
- "It's a very useful and effective website, which makes the university application process much easier. Furthermore it is a student-friendly tool and it's accessible ..."
- "... personally think that the aesthetic part of the system is overwhelming..."
- "... It's a bad idea to include pictures in the electronic application. The most impressive thing was the ranking of the student programs. ... "
- "I do not really have any remark about how to make the iKnow system any better. It is good as it is. And also I think it would be a great idea to connect it to Facebook since the youth, now, spends a lot of time on it. ..."
- "...I really like this system because I could fetch all the information I needed online, and there was no need for me to go to my university, which is great for the students who live far away. ..."
- "Personally I find the iKnow system useless, it just wastes my time when I am obligated to write the application on hand again. And also, the design on the site should be enriched. ..."

Judging by the realized analysis, we conclude that the students benefit by using this system, not only in the application for enrollment, but also in provision of overall electronic services for students.

V. Conclusion and Future Work

iKnow system is a delivery of an EU TEMPUS project to introduce electronic student services at universities. The project introduced change of university management workflow procedures, not by just changing the paper based activities with electronic, but changing the procedures themselves, to make usage of software more efficiently. Introduction of such a system included soft management skills to persuade the university management to believe that the new system will work more efficiently, even though the system was not tested in the moment of deployment.

Comparing this system with previous three attempts at the university for introducing automated student services information system, we conclude that iKnow has survived the initial roll out and is now in exploitation phase (second year), while the others were abandoned after its deployment due to software malfunctions, not trained staff to use the system and most importantly the high level of university management was not confirmed that it will successfully work. User satisfaction is an important part of a successful system, since it gives back a positive feedback on what and how to improve the system to satisfy the user needs. We are glad that iKnow resulted with a successful implementation, although, worldwide, there is a low successful rate of implementation of complex systems such as this one.

Nielsen [2] finds that the entire purpose of having a subjective satisfaction usability attribute is to assess whether users like the system, and therefore it seems highly appropriate to measure it by asking the users. Judging by realized empirical measurements of the user satisfaction, the students have positive or neutral attitudes about the system. Having in mind the critical approach by the students, especially since all of them were highly IT-literate and keen on using FaceBook social network at least 3 hours per day on average, we argue that the system is well adopted by the students and that they see the benefits in using the system further.

Thus, it can be concluded that iKnow will fully replace the outdated Student Service Desk. Not only the students, but also the general population will benefit from a change in the application method for university. We asked the students for their opinions, to which, most of them replied that they would like to apply for university via the internet. This will result in a massive reduction to the student queues during the applications process.

Moreover, more the students use, and new improved software versions are deployed, we witness the swing from *neutral* in 2012 to *satisfied* in 2013. All in all, electronic services build with the iKnow system has become an integral part of the university management and fulfilled the mission of the initial project setup.

We will continue to realize user satisfaction measurements and suggest software improvements of the iKnow system, and also realize a more comprehensive evaluation.

References

[1] I. Chorbev, M. Gusev, D. Gjorgjevikj, and A. Madevska-Bogdanova, "Architecture of an electronic student services system and its implementation," in *ICT Innovations 2012, (Editors Gusev and Markovski)*, 2012, pp. 381–400.

[2] J. Nielsen and J. T. Hackos, *Usability Engineering*. Academic press Boston, 1993.

[3] M. Hassenzahl, "The interplay of beauty, goodness, and usability in interactive products," *Human-Computer Interaction*, vol. 19, no. 4, pp. 319–349, 2004.

[4] D. A. Norman, *Emotional design: Why we love (or hate) everyday things*. Basic books, 2007.

[5] E. Karapanos, J. Zimmerman, J. Forlizzi, and J.-B. Martens, "User experience over time: an initial framework," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2009, pp. 729–738.

[6] M. Hassenzahl and N. Tractinsky, "User experience-a research agenda," *Behaviour & Information Technology*, vol. 25, no. 2, pp. 91–97, 2006.

[7] J. Hom. (1998) The usability methods toolbox handbook. [Online]. Available: http://www.idemployee.id.tue.nl/g.w.m.rauterberg/lecturenotes/UsabilityMethodsToolboxHandbook.pdf

[8] B. Battleson, A. Booth, and J. Weintrop, "Usability testing of an academic library web site: a case study," *The Journal of Academic Librarianship*, vol. 27, no. 3, pp. 188–198, 2001.

[9] R. Likert, "A method of constructing an attitude scale," *Scaling: A sourcebook for behavioral scientists*, pp. 21–43, 1974.

[10] J. P. Chin, V. A. Diehl, and K. L. Norman, "Development of an instrument measuring user satisfaction of the human-computer interface," in *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 1988, pp. 213–218.

[11] J. R. Lewis, "Psychometric evaluation of the post-study system usability questionnaire: The pssuq," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 36, no. 16. SAGE Publications, 1992, pp. 1259–1260.

[12] J. E. Bailey and S. W. Pearson, "Development of a tool for measuring and analyzing computer user satisfaction," *Management science*, vol. 29, no. 5, pp. 530–545, 1983.

[13] G. Armenski and L. Antovski, "E-student service system," in *The 9th Conference for Informatics and Information Technology (CIIT 2012)*, 2012.

Sign Language Tutor – Rebuilding and Optimizing

Nevena Ackovska* and Magdalena Kostoska*
* Ss. Cyril and Methodius University,
Faculty of Information Sciences and Computer Engineering,
1000 Skopje, Macedonia
{nevena.ackovska, magdalena.kostoska}@finki.ukim.mk

Keywords-Macedonian sign language; Education; Games; Deaf and hard of hearing

Abstract - In Macedonia there are around 6000 deaf people [4] and according to the National association of deaf and hard of hearing of Macedonia, this represents only 0.3% of the population. Because of this low percentage the community of deaf and hard of hearing feels very isolated and stigmatized. Only one television channel (the national channel) offers once a day a Macedonian Sign Language (MSL) interpreted news. Moreover, there are only two schools in Macedonia that offer education to deaf children. For that reason, when a new deaf child is born by non-deaf parents, the learning of the sign language represents a real challenge that affects the whole family. This is the reason why we created the sign tutor - to help this community and to introduce the whole family (the deaf children and the non-deaf parents) with the Macedonian sign language in an easy and entertaining way, with interesting animations and games.

This paper will give an overview of the MSL and the general recommendation for using IT tools in sign language learning. The project created as a result of the research – the Sign tutor will be resented, along with the results and comments received after promoting this project. The paper also elaborates on current upgrades and optimizations to this project, in order to make it more available.

I. INTRODUCTION

According to the World Health Organization statistics updated in 2013, there are around 360 million people worldwide that have disabling hearing loss (Disabling hearing loss refers to hearing loss greater than 40dB in the better hearing ear in adults and a hearing loss greater than 30dB in the better hearing ear in children) [1]. That represents over 5% of the world's population. According to the World Federation of Deaf there are about 70 million deaf people who use sign language as their first language or mother tongue [2]. However, they live isolated from the rest of the world. The reason for this is the inability for communication with the non-hard of hearing world. Regular world has little knowledge of the challenges that heard of hearing or deaf people experience in everyday activities. For a long time it has been generally accepted that the deaf and hard of hearing people are unintelligent, so that's why the offensive term "deaf and dumb" has been used. The terms deaf and hard of hearing are voted as an official designation in 1991 by the World Federation of the Deaf (WFD) [2] [3].

The language of the deaf and hard of hearing people is called Sign Language (SL). This language can be signed with usage not only of the hands and arms, but also the head, lips or the torso. This is not universal language and the languages differ from country to country, although there are universal features in sign languages. Also the signed languages are not based on the spoken language of the region.



Figure 1. Signs conserning human relations in the MSL dictionary [5]

In Macedonia there are around 6000 deaf people [4] and according to the National association of deaf and hard of hearing of Macedonia, this represents only 0.3% of the population. Because of this low percentage the community of deaf and hard of hearing feels very isolated and stigmatized. Macedonian Sign Language (MSL) dictionary exist only in printed form [5]. Fig. 1 depicts some signs about human relations included in the dictionary. Only one television channel (the national channel) offers once a day a MSL interpreted news, a weekly show called "The world of silence" and few MSL interpreted shows. All the other national programs and TV channels don't offer even subtitles to their contents in Macedonian language. Moreover, there are only two schools in Macedonia that offer education to deaf children: the primary school "Kocho Racin" in Bitola and the high school "Partenija Zografski" in Skopje. For that reason, when a new deaf child is born by non-deaf parents, the learning of the sign language represents a real challenge that affects the whole family. This is the reason why we created the sign tutor - to help this community and to introduce the whole family (the deaf children and the non-deaf parents) with the Macedonian

sign language in an easy and entertaining way, with interesting animations and games.

This paper will give an overview of the MSL and the general recommendation for using IT tools in sign language learning. The project created as a result of the research – the Sign tutor will be resented, along with the results and comments received after promoting this project. The paper also elaborates on current upgrades and optimizations to this project, in order to make it more available.

II. Related Work

Many have shown that the deaf children don't acquire sufficient lip-reading skills and that many deaf and hard of hearing individuals have difficulty with reading. In a large study conducted by Conrad [6] it has been discovered that deaf and hard of hearing children attending schools using an oral approach rarely acquire sufficient lip-reading skills. The children with hearing loss greater than 85 dB could only comprehend about 25% to 28% of the words through lip-reading that they could comprehend through reading. The children with hearing loss less than 65 dB could only comprehend about 36% of the words through lip-reading that they could comprehend through reading. Also it has been shown that the deaf are not truly acquainted to words and the writings are based on the spoken language [7]. Another study on deaf and hard-of-hearing children shows that there is a moderate association between the finger-spelling comprehension and reading comprehension [8].

A most efficient way to introduce the sign language to a deaf person is with the help of the visual perception. In that direction the IT tools that exploit the today's visual technologies are most efficient in the process of learning the sign language.

A lot of researches concentrate on the learning process of the sign languages. One research presents a HCI conceptual meta-environment framework to construct computational Intellectual Artifacts in SL to promote bilingualism (Libras-Brazilian SL/Portuguese) in order to increase family bonding activities and effective bilingualism for Deaf children and non-Deaf parents. This framework combines the SL with the cognitive theories (Sense Making, Common Sense, and Concept Maps). Its goal it is to give directions to design of bilingual artifacts in various genres like games, educational objects, storytelling environments etc. [9]. According to another of these studies a pedagogical (teaching and learning) architecture should have strong pedagogical conception, a methodological systematization and should provide a virtual learning environment [10]. Some results of researches that also exploit the visual technologies for learning the sign language are: the System for Sign Language Tutoring [11] which evaluates users' signing and gives multimodal feedback to help improve signing, the 3D Animation Editor and Display Sign Language System for Thai Sign Language [12], that exploits the XNA framework in order to train the general person and Deaf person with Thai Sign Language for ability to communicate to each other, and SignOn, which is a Model for Teaching Written Language to Deaf People [13].

Most popular commercial products on the market are the Vcom3D Sign Smith products which use avatar for signing the American SL. They offer the products like Illustrated Dictionary, Studio, ASL and Signing Science Dictionary. [14].

III. Macedonian Sign Language (MSL)

According to the National association of deaf and hard of hearing of Macedonia there are only 12 licensed interpreters of the Macedonian Sign Language (MSL) [15]. The signs of the alphabet along with the terms for mother, father and some basic colors were the only information about MSL that could be, until recently, found online. Since then the National association upgraded their web page and published this information online. Unlike the bigger countries that promote their national sign language, there are no e-books, videos or any other type of online content that can help learning the MSL.

The Macedonian sign language is based on hand gestures and body gestures, as all the other sign languages. The hands are the basic communication means. The signs are performed with predefined movement and location using one or both hands. Not every hand movement has meaning. There are additional elements that enable more efficient and understandable communication, like head movement, facial expression, mouth or body movement. MSL vocabulary is limited and it contains around 2800 terms (signs).

The Macedonian sign alphabet consists of 31 signs, the same number of letters in Macedonian alphabet. There are two versions of the alphabet: one version is signed using only one hand and the other is using both hands.

IV. Raising awerness among ICT students

ICT industry mat have big impact in providing a great help for persons with some kind of disability. Being ICT educator gives an opportunity to raise awareness about specific target groups, like the deaf and hard of hearing target group, to ICT students. In the University st. Cyril and Methodius, in the Faculty of Computer Science and Engineering, this is incorporated within the course Human-computer interaction (HCI).

In this course the students learn how to use the state-of-the-art technologies and create different kind of useful software products. It introduces the students to the basics of the complex interaction between the computer and the user. The students grasp the principles for building interfaces that compound the interests of the humans and the possibilities of the technology. During this course, the students understand that the humans are the main component of the computer systems and they learn how to develop the software and the hardware for utilizing different groups of users.

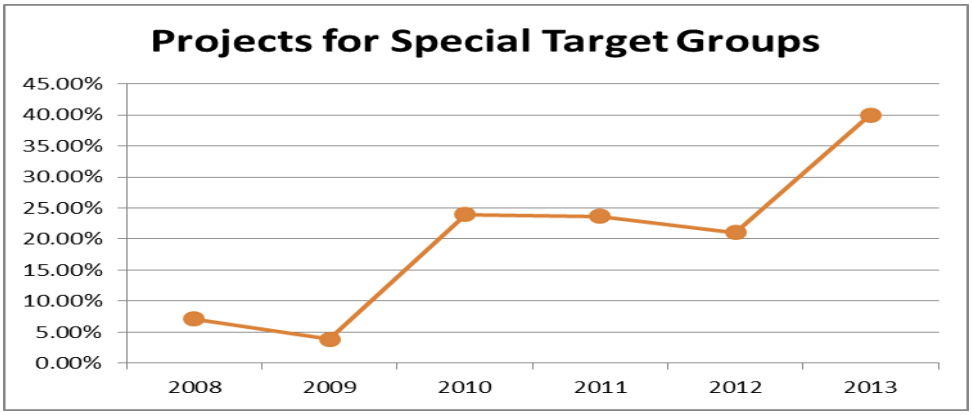


Figure 2. Percentage of projects intended for specific target groups per year

This course offers special experience and provides different perspective from other courses because it introduces knowledge about different kinds of target groups. It offers perspective of the challenges and opportunities for creating ICT projects for people with disabilities, when they are included as target groups. Since the software and hardware products for these clients come with special limitations, the students are thought recommendations how to build applications, working under strict constraints. This is challenging, but the experience shows that the students are very interested in creating such applications. As a result, out of 200 projects that were created in this course in the past 6 years, many were dedicated to special target groups. Some of the researches that students conducted for their projects were intense, since it's not common for computer engineers to gather knowledge that borders medicine, ergonomics, social sciences and psychology, especially not in the Computer Science curricula. It was shown that the students are interested to learn more, about the specific target group as well as about the technologies that enabled better application development. The interest for working with special target groups rises every year and the students show willingness to input extra effort, especially since these projects require additional knowledge in mathematics, optimization techniques, algorithms, special dedicated hardware etc. The trend of increased interest for working special target groups can be noticed from the following statistics: The first year (2008) only 7% of the student projects were intended for specific target group. The interest started to rise in 2010 and resulted with 24%, while in 2013 raised up to 40%. Fig. 2 depicts how the interest rises over the years. Furthermore, 33 projects were evolved to a diploma thesis, and are now fully functional applications for a specific target group or application domain.

Sign Language Tutor was such an example project. It started as a project in this HCI course, continued as a graduate work and at this moment a group of six students are working on the upgrade of this project.

V. SIGN LANGUAGE TUTOR

Sign Language Tutor is an interactive platform for MSL learning that is easily extendible. It represents a collection of games and modules that will ease the learning and increase the mental and memory skills of the deaf and hard of hearing children. The central part of this project consists of 3D animations of a girl that signs the chosen alphabet character or some objects. Fig. 3 shows our character performing signing.

The application consists of several modules, shortly described below.

3D simulation – this module provides signing of alphabet letters and words. The user chooses the letter or the object and the 3D animation illustrates the chosen sign. The user can rotate the animation in all directions and enables different views to better capture the signing. This is of outmost importance, since the signs and words are actually expressed in 3D, usually with more than just fingers and hands. The animation can be paused and started again.

Memory – this module represent a standard memory game where the goal is to merge an alphabet letter with the appropriate sign.

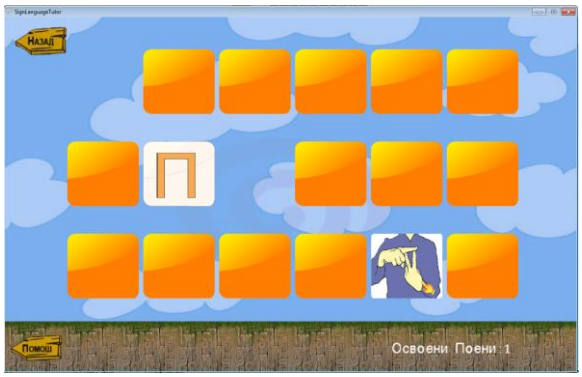
Explore – this module is a 2D game where our hero goes in an adventure against monsters and collects objects. Each of the collected objects is an award – an animated 3D sign of that object in MSL. The game doesn't require special skills and it is adapted for children.

Fig. 4 (a) shows the Memory module and (b) shows the first level of the Explore module.

The application enables the children to learn while having fun. However, the educational modules are interesting even for the other, possibly non-deaf members of the family, or the friends. The wider circle of people to a deaf or hard of hearing child can benefit from learning sign language in a fun and interesting way. This is especially true for the non-deaf children.



Figure 3. 3D animation of the girl signing



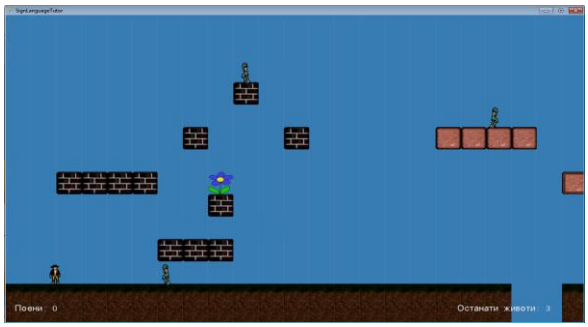


Figure 4. The Memory module (b) The first level of the Explore module

A. Sign Language Tutor Architecture

The Sign Language tutor is built from several software modules that are interconnected. This enables easier manipulation and expandability of the project. Fig. 5 depicts the architecture of the project.

The architecture is built of the following modules (and their descriptions):

- SignLanguageTutor - This module along with the core part is the central part of the software solution and depends of all the other modules. It represents a Windows Game type of project that work on XNA 4.0 framework. It contains the concrete implementations of the games, and respectively the entire user interfaces.
- SignLanguageTutorContent - This module is responsible for the management of the resources. It represents a ContentProject type of project and it is used for resources compilation, storage and preparation for use during the execution. It is divided in parts responsible for management of models, backgrounds, textures, maps etc…
- SignLanguageTutor.AnimationPipeline - This module is compiled to a dynamic link library. It contains the AnimationProcessor class that extends ModelProcessor. The processor contains two parts needed for texture processing: skeleton processing and animation processing part. This actions are called through the function Process that is inherited and replaced from the ModelProcessor.
- SignLanguageTutor.AnimationAux - The AnimationAux is an auxiliary module that includes animation processor con-nection. It contains entities for animation model storage, helpers for animation and models' binary formats loading directly in the source code.
- SignLanguageTutor.TileEngine, LevelEditor and LevelEditorContent - This TileEngine module is used in the 2D game Explore. It represents a framework that

defines all the basic 2D game functionalities like movement physics, game map and view. This is also popularly called game engine. The LevelEditor and LevelEditorContent modules are in charge for game maps creation and storage. The map editor is also XNA Windows Game application

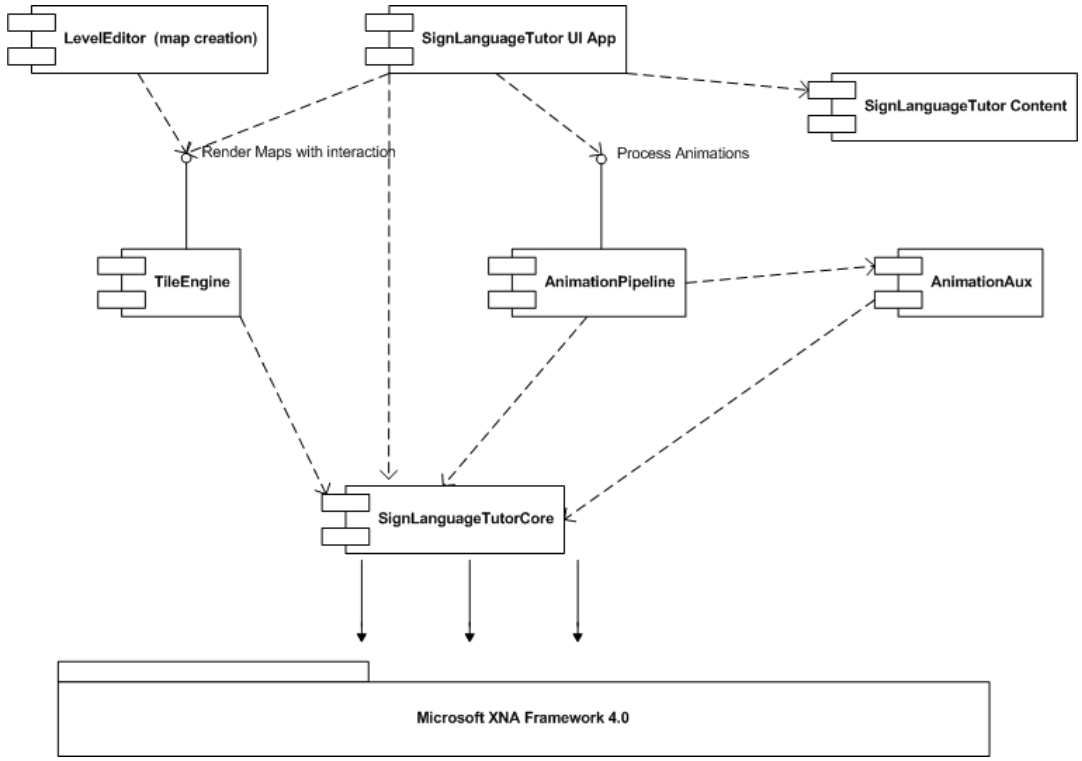


Figure 5. Sign Language Tutor architecture

VI. RESULTS AND CURRENT WORK

The first version of the Sign Tutor was presented to the final workshop of AEIF (Alumni Engagement Innovation Fund) Project - Equal Access through Service Learning for Persons with Disabilities, in front of persons with disabilities and especially people who work with deaf children. The project raised big interest and was approved and warmly welcomed by these people. They offered their help in the future extension of this project. Also, the interest resulted in filming an article about this project by the national television.

As expected, since the application has educational, but also entertaining components, the explore game module was found as one of the most interesting feature. At the moment of publishing the first version it consisted only of two levels. However, presently two additional levels are being created.

Since the initial project was a desktop application, it was very difficult to make appropriate upgrades. This was found to be limiting and less available to the user. For that purpose the project is rebuilt to a web platform. The new platform represents a type of content management system, where the content is easily manageable. It is expected that this new platform will allow adding new signs and terms in user-friendly manner, but will also allow leaving comments by the users. Another challenge that was faced during the migration was the size of the animations – they were large. For that reason a transformation process is currently taking place, and the animations are converted to gif files with good quality.

VII. CONCLUSION

Raising awareness about the needs of different target groups should be a goal to many ICT educators. ICT students welcome new knowledge about people they don't know much about, such as people struggling with disabilities and special conditions that limit their everyday activities. This was the case in building the Sign language tutor, an application for deaf and hard of hearing children, developed by ICT students in Macedonia.

The sign tutor offers an easy and entertaining way of learning of the Macedonian sign language, and it is the first application of this type in Macedonia. It is accepted by the deaf and hard of hearing community in Macedonia and it is precious tools for the deaf children and especially the non-deaf parents and friends, enabling them to learn through games and animations. This project also offers bigger inclusion and increases the awareness of this social group by helping the deaf and hearing people to learn MSL.

The project follows the recommendations given by human-computer interaction research concerning deaf children [16] [17] about deaf children psychology and abilities. It offers deaf and hard of hearing children, as well as their parents, an environment to learn through fun and games. The new, web based platform is expected to make this project even more accessible to the wider public.

REFERENCES

[1] World Health Organization. "Deafness and hearing loss", 2013, [retrieved: Noe., 2013]. [Online]. Available: http://www.who.int/mediacentre/factsheets/fs300/en/

[2] World Federation of the Deaf, 2013, [retrieved: Noe., 2013]. [Online]. Available: http://www.wfdeaf.org

[3] Creighton, N.: What is Wrong With the Use of The Terms: 'Deaf-mute', 'Deaf and dumb', or 'Hearingimpaired', [retrieved: Noe., 2013]. [Online]. Available: http://www.eamo.org/SNA/deaf%20PC%20terminology.pdf

[4] National Council of Disability Organizations of Macedonia, "Association of the deaf and hard of hearing of Macedonia",: 2013, [retrieved: Noe., 2013]. [Online]. Available: http://www.nsiom.org.mk/SGNM en.aspx.

[5] National Council of Disability Organizations of Macedonia, "Macedonian Sign Language Dictionary", National Council of Disability Organizations of Macedonia, 2000.

[6] R. Conrad, The deaf schoolchild, Harper & Row, 1979.

[7] A. Sears and J. A. Jacko, The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications, 2nd ed. CRC Press, 2007..

[8] C. Padden and C. Ramsey, "Reading ability in signing deaf children," Topics in Language Disorders, vol. 4, no. 4, pp. 30–46, August 1998..

[9] C. Guimaraes, D. Antunes, L. Garcia, A. Guedes, and S. Fernandes, "Conceptual meta-environment for deaf children literacy challenge: How to design effective artifacts for bilingualism construction," in Research Challenges in Information Science (RCIS), 2012 Sixth International Conference on, 2012, pp. 1–12.

[10] C. Guimaraes, D. Antunes, L. Garcia, L. Peres, and S. Fernandes, "Pedagogical architecture – internet artifacts for bilingualism of the deaf (sign language/portuguese)," in System Sciences (HICSS), 2013 46th Hawaii International Conference on, 2013, pp. 40–49.

[11] O. Aran, I. Ari, L. Akarun, B. Sankur, A. Benoit, A. Caplier, P. Campr, A. Carrillo, and F.-X. Fanard, "Signtutor: An interactive system for sign language tutoring," MultiMedia, IEEE, vol. 16, no. 1, pp. 81–93, 2009.

[12] P. Ittisarn and N. Toadithep, "3d animation editor and display sign language system case study: Thai sign language," in Computer Science and Information Technology (ICCSIT), 2010 3rd IEEE International Conference on, vol. 4, 2010, pp. 633–637.

[13] M. Hilzensauer and F. Dotter, "Signon", a model for teaching written language to deaf people," in IST-Africa Conference Proceedings, 2011, 2011, pp. 1–8.

[14] Vcom3D, "Sign smith," 2013, [retrieved: Noe., 2013]. [Online]. Available: http://www.vcom3d.com/signsmith.php

[15] National Association of Deaf and Hard of Hearing of Macedonia, "List of MSL interpreters" 2013, [retrieved: Noe., 2013]. [Online]. Available: http://www.deafmkd.org.mk/index.php?option=com content&view=article&id=55&Itemid=80.

[16] Z. Zafrulla, H. Brashear, P. Yin, P. Presti, T. Starner, and H. Hamilton, "American sign language phrase verification in an educational game for deaf children," in Pattern Recognition (ICPR), 2010 20th International Conference on, 2010, pp. 3846–3849.

[17] F. Alonso, A. De Antonio, J. Fuertes, and C. Montes, "Teaching communication skills to hearing-impaired children," MultiMedia, IEEE, vol. 2, no. 4, pp. 55–67, 1995.

Interoperability Challenges Among University Electronic Services Systems

Magdalena Kostoska, Marjan Gusev and Sasko Ristov
University Ss. Cyril and Methodius, Faculty of Computer Sciences and Engineering,
1000 Skopje, Macedonia
{magdalena.kostoska, marjan.gushev, sashko.ristov}@finki.ukim.mk

***Abstract*—This paper analyzes the interoperability issues of the information exchange among essential software applications that enable University electronic services. The analysis of the existing standards shows that there is no fully defined compatibility for these software solutions, which motivated us to conduct a research in order to detect and describe the interoperability challenges of creating a fully automated interaction.**

In this paper we present several case studies of challenges faced at our University, while integrating three existing different systems: iKnow as a custom developed university management system, a customized version of an open source Moodle learning management system and e-Testing as a custom developed e-assessment system. Although these systems operate independently, still there are many interactions among them; keeping the system consistent and tightly connected at the same time. Keeping the software development compliant with the elaborated standards and enabling a push or scheduled synchronization modules will eliminate a big maintenance effort to enabling interoperability with new software versions or when migrating to another type of service provider.

Keywords—Learning system platforms; eAssessment; University Management Systems; Inteoperability issues.

I. Introduction

A typical software solution to provide University electronic services uses at least three different software products for University Management Systems (UMS), Learning Management Systems (LMS) and eventually e-Assessment systems (EAS). These systems are usually developed by different vendors and are implemented separately, which means that they lack interoperability [1]. An additional challenge for the university is imposed by the new upgrades or changes of these softwares.

In this paper we give an overview of possible collaborations between LMS [2], EAS and UMS software solutions as well as existing standards that apply for them, in order to detect and describe the challenges of creating fully automated interactions between all of these systems in a manner of interoperability [3]. The main goal is to tackle the interoperability issues and to analyze if it is possible to go even further, to establish an interoperable environment, so the University can change the service type provider, or software provider, or both, without any loss or damage. Our idea is to explore the possibility of automated interaction and data exchange between these systems, as well as between general types of systems, in order to create a fully integrated environment.

Three systems used at our University will be analyzed in this paper: iKnow UMS (custom developed university management system) [4], Moodle LMS [5] and e-Testing EAS (custom developed e-assessment system upgraded as adaptive online learning [6]). All of these systems operate independently, limiting to only few interactions among them; while at the same time they are tightly connected. The reason why we use a dedicated EAS, rather than Moodle LMS assessment options, is the fact that it represents a safe assessment environment which imposes automated assessment, without access to any learning materials or any other software, randomized questions offering, prevents cheating, etc.

The rest of the paper is organized as follows. Related work about interoperability analysis of similar systems is presented in Section II. Section III gives a description of each of the system types and the idea of collaboration between these systems, where we define the possible data flows. These flows are used to identify the use cases for interactions between systems, which are presented in Section IV. Section V describes the possibility of using existing standards for each of these use cases, or the lack of standards, followed by identification and analysis of the interoperability challenges. Finally, in Section VI we evaluate the significance of each of the found interaction challenges and address the source of these challenges.

II. Related Work

Each of the analyzed systems needs a good specification and definition to provide integration of the University electronic services. Although most of the functionalities, specified data, authorization and authentication are standardized for each system, we can say that the LMS systems have been best defined and standardized [7], [8], [9].

The standardization of EAS is in process since the functionalities of these systems vary and initial standards are created [10]. Unlike LMS and EAS, UMS are neither well defined nor standardized, so integration of these systems is left to creating personal solutions. Only some parts of this system are included in IMS standards, like e-Portfolios [11].

Most of current research papers analyze the possibility of integration between two of the explored systems. For example, Maican targets the interoperability between LMS and Content Management Systems (CMS) like the open source tools Moodle Learning System and DotNetNuke framework [12]. Armenski explores the usage of SCORM, IMS Content Packaging and IMS Learning Design standards in order to create Service Oriented Architecture that will enable interoperability between LMS and AES [1].

Kudrass tackles the problem of integration of all three systems (UMS, LMS and EAS) as proposal of a high-level

architecture for coupling of existing systems [13]. However, this solution is not applicable to a variety of systems.

III. Overview of LMS, EAS and UMS

This section contains relevant definitions and gives an overview of the functionalities, the systems use to enable better understanding of the interoperability issues.

A. Learning Systems

LMS is a system that enables an environment where the teachers interact with students in the delivery of learning materials, assessments, homework assignments, and other relevant activities. It keeps track of all related activities for the learning process. It also enables discussion, forums, etc. The functionalities of these types of systems vary and may include many elements. One comprehensive example of the functionalities that these systems can have is the E-Learning Framework, shown in Fig. 1. This framework is developed through coordinated efforts of U.K Joint Information Services Committee (JISC) and Australia Department of Education, Science and Training (DEST) [14]. This Framework recognizes 59 distinct functionalities divided into three main groups: Sample User Agents, Learning Domain Services and Common Services.

One of the relevant standards in this area is developed by the Aviation Industry CBT (Computer-Based Training) Committee (AICC), which is an international association of technology-based training professionals. They have developed nine AICC Guidelines and Recommendations (AGR's) and in order to be "AICC Compliant" the training product should complies with one or more of these AGR's. They offer formal certification testing procedures for AGR-006 (File-based CMI Systems) or AGR-010 (Web-based CMI Systems) [15].

The most popular standard for LMS is the Sharable Content Object Reference Model (SCORM). It represents a set of technical standards and specifications for, as the name indicates, creating units of online training material that can be shared across systems [16]. It is developed by the Department of Defense's Advanced Distributed Learning (ADL), a research group sponsored by the United States Department of Defense. SCORM generally references existing standards and elements of the Institute of Electrical and Electronics Engineers (IEEE), AICC, and Information Management Standard (IMS) specifications.

At our University we use a customized Moodle installation. Moodle is a robust open-source learning platform and it includes most of the standard LMS functionalities and some additional features. It is compliant with most of the well-known international standards like IMS LTI, SCORM and IMS Question & Test Interoperability (QTI) [5].

B. E-Assessment Systems

E-Assessment systems are separated from the LMS, offering a possibility for question definition and administration, test creation, test delivery, test execution, evaluation, assessment and management of users and results [17].

A well-defined standard for EAS is the IMS QTI specification, which enables the exchange of item, test and results data between authoring tools, item banks, test constructional tools, learning systems and assessment delivery systems [18]. It is created by the IMS Global Consortium. EAS describes a data model for the representation of question and test data and their corresponding results reports. The specification is spread over a number of documents: Implementation Guide, Assessment Test, Section and Item Information Model, Metadata and Usage Data, Results Reporting, Integration Guide, XML Binding and Migration Guide.

Most of the LMS systems would like to upgrade their systems with all e-Assessment features, but the implementation are with limited perimeter and scope. Most popular EAS comprehensive systems are produced as stand-alone products, but still these products lack of interoperability with other LMS, or lack some other basic features, like adaptability according to learner characteristics and possibilities for content reuse. There is no evidence of sophisticated integrated LMS and EAS software [19] [20]. Although Moodle has a possibility for assessing the students' knowledge through different question types and grading scores, it builds only a small subset of the EAS functionalities.

A custom built EAS system is in use at our University since 1999 [6]. This solution has been upgraded with several newer versions. One of the latest upgrades of the system includes an adaptive testing realized via information and communication technologies, as well as on computer based assessment.

All versions support at least the following question types: multiple choice questions and single value input questions. Essay types of question or others are avoided since they require additional assessment by a human. The goal is to simplify the evaluation process and make it more efficient. In addition, the implementation of a single choice answer question is a subset of multiple choice answer questions. Single value input questions are those where the student is expected to provide a single answer that may be a number or a word. It has to be exactly the same as the expected answer, and different formats of writing the numbers or synonyms are not accepted by the system. This system also implements a negative grading schema. During the process of testing the instructor generates a strategy to create the test and sets a date when the test will be generated and valid for execution. The strategy is realized by selecting which knowledge items or learning objectives (LO) will form a test and the numbers of question of each selected knowledge item. If the number of available questions for the item is bigger than the number the instructor chooses, this system will generate different tests for each student. In addition beside the random selection of questions, this system also provides random positing of multiple choice answers. The test is available to the students that enrolled the course and that are present in the faculty laboratories.

Recent version of the e-Assessment system includes interactive images, where the users can navigate and zoom the image, and realize several actions on the image, such as answering by clicking on an appropriate graphical object, marking a region, annotating, setting an answer/comment on a given position, etc. A new set of question types has been developed, including Single Graphic Choice (SGC) which provides a single answer on a map sensitive figure; Opinion Map (OP) where a figure is presented to intended users and they can click on the map and enter their opinion; and Multiple

Sample User Agents						
Assignment marking tool	Authoring applications	Library System	Portal	Student Enrolment Portlet	Test	Timetabling
VLE / LMS						
Learning Domain Services						
Activity Management	Assessment	Competency	Course Validation	Curriculum	Grading	Learning Flow
Marking	Quality Assurance	Reporting	Resource List	Sequencing	Tracking	ePortfolio
Common Services						
AV conferencing	Alert	Archiving	Authentication	Authorisation	Calendaring	Chat
Context	DRM	E-mail management	Federated Search	Filing	Format Conversion	Forum
Group	Harvesting	Identifier	Logging	Mapping	Member	Messaging
Metadata Management	Metadata Schema Registry	Packaging	Person	Presence	Rating / Annotation	Resolver
Role	Rules	Scheduling	Search	Service Registry	Terminology	User Preferences
Whiteboard	Workflow					

Fig. 1. The E-Learning Framework[14]

Graphic Choice (MGC) which allows the user to enter multiple answers on a map sensitive figure [21].

C. University Management Systems

In this paper we refer to academic University management system as a subset of a complete UMS providing student services and academic records, meaning a system that keeps track of all student relevant information, including their records and files, and even financial and other administrative workflows involved among students and University.

Our University uses a custom developed solution as a delivery of an EU funded project Innovation and Knowledge Management towards eStudent Information System - iKnow. This system is designed to store and administer student's records and personal files, as well as related university data. This system is developed in order to enable conditions for realization of eStudent Information System for: Enrollment process; Identity management, with authorization and access control; Management of university services and resources usage (accounting); Administration of Europass CV, ECTS, diploma supplement and other certification issuing; Administration of student activities; Administration of registration and schedules of classes and assignments; and Administration of academic results, including course, exam and study results [4].

The system provides exchange of electronic information among all stakeholders: students, professors, administration, university management and the Ministry of Education. The system consists of two main components, each consisting of modules: the new students Enrollment Student Services (ESS) and the Core Student Services (CSS). The ESS component consists of the candidate's enrollment wizard, the enrollment forms for manual entry of candidate's data, candidate's data processing, ranking module and the enrollment results publishing. The CSS component includes: Module for administration on University and faculty levels, Module for study programs and schedules, Student activities module, Module for personal identification and access control, Module for electronic payment and use of resources and Module for migration of old data and interfaces to other systems [22]. The architecture of this system is shown in Fig. 2.

D. Interactions

The need for information exchange may vary, depending of the corresponding implementations and functionalities of the

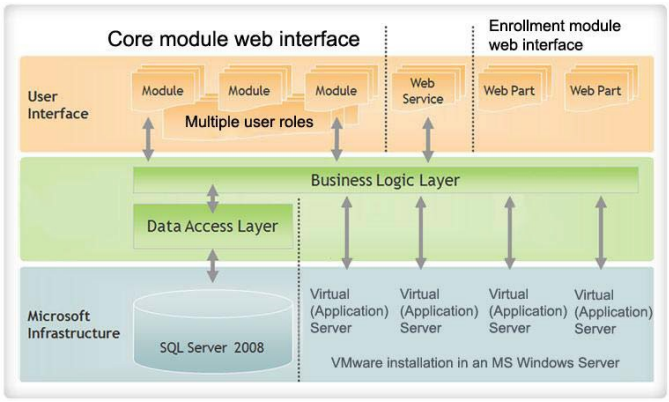


Fig. 2. iKnow Software Architecture [22]

described systems. Fig. 3 presents a possible interaction scenario for information and data exchange between the systems.

The interaction from LMS to EAS is to transfer the information about curricula, learning objectives, and students. In reverse direction, EAS sends assessment results of conducted exams and various reports used to improve the learning and teaching part of the course.

The LMS and UMS interaction is described next. The UMS sends information about relations of students and courses determined with the enrollment process, so the LMS can realize a proper authorization. In the opposite direction, the LMS sends academic records to the UMS about grades and obtained scores, along with various reports needed for the University Management.

There is no direct interaction between EAS and UMS, since the required interaction is performed via the LMS.

The central authentication system (single sign-on) is usually a part of the UMS. Its purpose is to authenticate the users and to inform the other analyzed systems for successful authentication, thus enabling the three systems with a possibility for authorization management.

IV. Interaction Use Cases

In this section we represent several use cases of interactions between the analyzed systems.

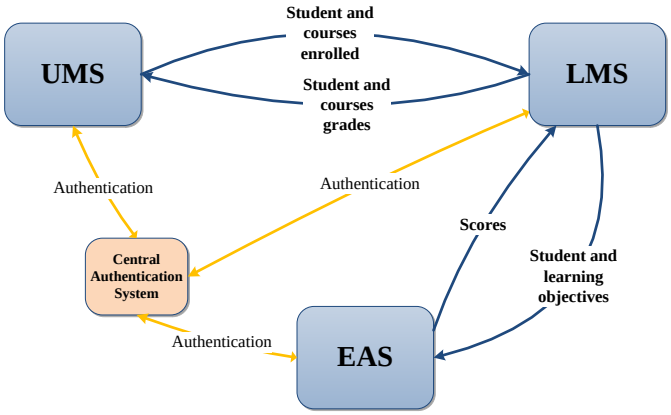


Fig. 3. Interaction among LMS, EAS and UMS systems

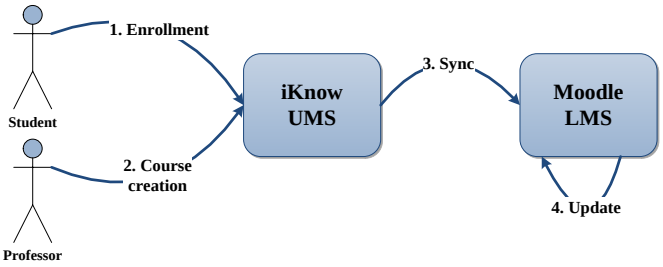


Fig. 4. UseCase: Transfer of student information for course enrolment

A. *Transfer of student information for course enrollment*

A special software module was developed and implemented at our University to enable transfer of student information. This module aims to enable a transfer functionality to the Moodle LMS filtering only those students that enrolled a particular course as registered by the iKnow UMS. This module relies on the Moodle enrollment functionality using an external database (in this case the iKnow database is the external database). This package uses synchronization to insert/update data in regular time periods. iKnow provides the relevant data for courses and students enrolled in these courses, as well as professors/instructors assigned to the course.

This interaction use case consists of at least the following steps, as presented in Fig. 4:

1) At the beginning of each semester, a student enrolls courses in the iKnow UMS;
2) When the enrollment period is over, every professor has an option (built in the iKnow UMS) to automatically create a Moodle course if he was assigned to or to create one Moodle course for multiple similar or same courses in iKnow (defined separately for different reasons);
3) A synchronization activity is started each hour on the Moodle LMS side; and
4) The Moodle course is automatically created and all the students that have enrolled the course are added, as well as the professor, with the appropriate permissions.

The current interaction is not standardized, so if the LMS is changed, it would require adaptations, which is not desirable. An improvement to this communication would be using IMS Enterprise standard [23]. The IMS Enterprise Information Model is designed to support interoperability of the following business process components: Personal Profile Data Maintenance, Group Management, Enrollment Management and Final Result Processing. Potential problem may occur if a new LMS system is introduced and that system does not offer all the functionalities that Moodle offers or doesn't support the IMS Enterprise standard.

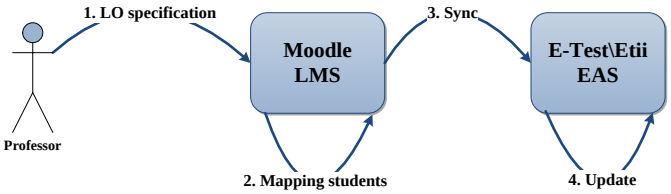


Fig. 5. UseCase: Transfer of information from LMS to EAS

Another possible improvement of this process would be creating SCORM learning packages for each course and import them in Moodle along with the creation of the course.

B. *Transfer of information from LMS to EAS*

Test creation process needs specification of learning objectives that have to be included in the test. The idea of this interaction use case scenario is to transfer the required specification from the LMS whenever an assessment activity is required. Created tests are delivered only to the students enrolled in the course to test their knowledge. The evaluation of results and assessment may be designed in a way that includes only some of the course topics or only students that satisfy some prerequisite to be able to take a test.

This interaction use case scenario for transfer consists of at least the following steps, shown in Fig. 5:

1) At the begging of each semester, the professor specifies the learning objectives for each course in the Moodle LMS;
2) The Moodle LMS keeps track of all relations between enrolled students and available courses;
3) A synchronization activity is started each hour on the Moodle LMS side to send the information about learning objectives and enrolled students with the EAS side; and
4) The EAS updates the learning objectives and list of enrolled students in the activated courses.

This interaction use case scenario is generally difficult to realize since it requires automated creation of questions. One possibility is to use semantics and ontology for each learning objective and to use them to create questions. But this would suggest that ontology is already created and available. Jovanov and Gusev [24] already explored this subject and created an initial model.

This use case has not been fully implemented in our systems yet and it is in development and testing phase, since it requires add-ons on the existing software.

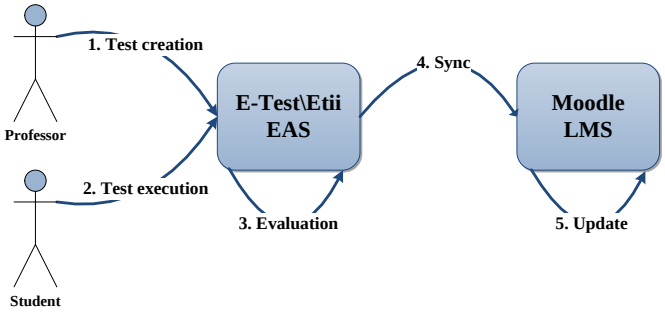


Fig. 6. UseCase: Transfer of assessment results from EAS to LMS

C. Transfer of assessment results from EAS to LMS

The goal of this use case is to enable automatic transfer of student scores from e-Test to Moodle for every assessment (test). This interaction use case scenario realizes the following steps, presented in Fig. 6:

1) The instructor creates the strategy and schedules test;
2) The student answers the test and EAS automatically evaluates and updates the results;
3) A synchronization activity is started as push activity on the EAS and corresponding results are sent to the Moodle LMS side,
4) A synchronization activity is started each hour each hour on the Moodle LMS to update the information received;

Initially this problem was solved by creating a dedicated Moodle activity for results import from the custom built EAS. This Moodle activity requires the testing id to enable the result import. This component had to be adapted with each Moodle upgrade since they appear frequently. This requires a permanent maintenance for the Moodle LMS and another solution is in effect at this moment. The alternative solution is based on exporting the results in a standard spreadsheet format, such as excel CSV or tab delimited data.

This interaction is not standardized, although it uses standard data representation. A better choice is to use the QTI standard. A problem may arise if the LMS system doesn't offer automatic results import or doesn't supports QTI standards. Another general problem of this interaction use case is defining how each test will influence the student's grade and how many points (in the course grading system) can be obtained through that test.

D. Transfer of student's academic records from LMS to UMS

The idea of this use case is to enable automatic transfer of grades upon course completion from Moodle LMS to the iKnow UMS. For this option to be possible, the professor has to define a criteria and relevant grading conditions in each course, as well as all the results (of any activity) to be available in Moodle. Another administrative issue is that the students in our University have to submit exam application for each course which will allow them to take the exam in given exams sessions and a grade can be set only for students that have submitted the application.

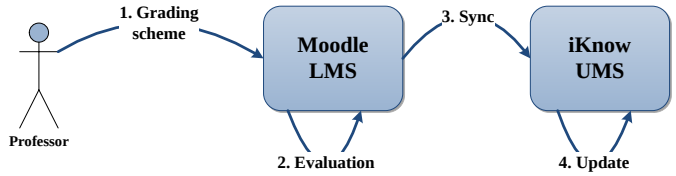


Fig. 7. UseCase: Transfer of student's academic records from LMS to UMS

This interaction use case scenario for communication from LMS to UMS consists of the following steps, shown in Fig. 7:

1) The professor defines the academic activities, grading criteria and passing conditions in the Moodle LMS;
2) The Moodle LMS has updated status of academic results for each student's activity and calculates the grade according to the defined criteria and conditions;
3) A synchronization activity is started each hour on the Moodle LMS to send the relevant information to the iKnow UMS; and
4) The synchronization module on the iKnow UMS updates the relevant academic records for the students.

Similarly to the previous use cases, due to the lack of sufficient maintenance for the LMS, this interaction is realized by using an alternative method. The results are exported in a standard spreadsheet format, such as excel CSV and later they are imported into iKnow.

This interaction can be improved and standardized by using the IMS Enterprise standard.

V. Results and Challenges

The presented interaction use cases are part of a possible scenario and do not include all the possible interaction details and data exchange information. For some of the use cases there are no applicable standards and some of the proposed standards may not be enough to complete the use case.

A formal specification of exchanged information between LMS and EAS includes various parameters, such as student information, course information, specification of learning objectives, test results, evaluation and assessment reports, etc. QTI standard specifies questions and test interoperability, but does not conform to challenges defined previously. SCORM is used as a standard to exchange learning materials, but does not cover all mentioned challenges.

The exchange between LMS and UMS includes at least the enrollment information, course identification, student information and academic results (grades). A possible usage of IMS Enterprise can be applied. Most of the efforts are done by developing specialized software solutions.

In order to enable partial interoperability among systems for the described interaction scenario, every system should support certain standards. Table I shows at least which standards the systems should support. It does not cover all relevant issues, but it covers the essential ones for exchange of learning related information.

Keeping compatibility with these standards will enable interoperability of solutions and new versions of the software, avoiding big maintenance efforts.

TABLE I. RECOMMENDED STANDARDS FOR SYSTEMS TO ENSURE PARTIAL INTEROPERABILITY FOR E-LEARNING RELATED CONTENT

System / Standard	SCORM	QTI	IMS Enterprise
UMS	X		X
LMS	X	X	X
EAS	X	X	

Another issue is the realization of a synchronization as a push-pull mechanism or as a scheduled activity. All software development should enable synchronized input and output of relevant information.

VI. CONCLUSION AND FUTURE WORK

The area of e-Learning offers a lot of well-established standards, while the area of assessment is still developing. In this paper we present our approach in defining interoperability challenges among LMS, EAS and UMS systems. As an example we comment on how we have faced and solved some of these challenges at our University. We also give recommendations on which standards may be used in order to provide partial interoperability. Yet all of the interactions among these systems can be delicated to orchestrate and very little number of available systems support all of the stated standards and standards are not always applicable.

A new challenge in this area which stresses the need for interoperability is service migration to the cloud. Problems arise when a university decides to migrate some features onto a cloud. Although the cloud offers good resources flexibility and scalability, storage, computational requirements and network access, and most important, lower cost [25], sometimes, this process is a nightmare, as the complex migration might cost time and information lost. We already started to migrate parts of the EAS system to the cloud [26] and faced a lot of interoperability obstacles.

Formal specification of these challenges will initiate recommendations for a possible standard and recommendations for interoperability framework and enable existence of independent cloud solutions, where information and services can be exchanged among various cloud providers and software solution providers without loss and damages.

ACKNOWLEDGMENT

This work was partially financed by the Faculty of Computer Science and Engineering at the "Ss. Cyril and Methodius University", Skopje, Macedonia through the project XaaSin-Cloud. The iKnow project has been developed as a part of the Tempus JPGR 511342 Project.

REFERENCES

[1] G. Armenski, "Designing Interoperable E-Assessment System," in *INTED2013 Proceedings*, ser. 7th International Technology, Education and Development Conference. IATED, 2013, pp. 5137–5144.

[2] M. Llamas-Nistal, M. Caeiro-Rodriguez, and M. Castro, "Use of E-Learning Functionalities and Standards: The Spanish Case," *Education, IEEE Transactions on*, vol. 54, no. 4, pp. 540–549, 2011.

[3] J. Fontenla, M. Caeiro, and M. Llamas, "Remote permission management in third-party tools and e-learning systems," in *Advanced Learning Technologies (ICALT), 10th IEEE Int. Conf.*, 2010, pp. 521–522.

[4] iKnow. [Online]. Available: http://iknow.ii.edu.mk/

[5] Moodle. (2013). [Online]. Available: http://moodle.org

[6] M. Gusev and G. Armenski, "E-Assessment Systems and Online Learning with Adaptive Testing," in *e-Learning Paradigms and Applications - Agent-based Approach*, M. Ivanovic and L. Jain, Eds. Springer Verlag, 2013, pp. 229–249.

[7] IEEE, "Standard for Learning Technology - Data Model for Content to Learning Management System Communication," *Std 1484.11.1-2004*, pp. 1–43, 2005.

[8] L. Zheng, Y. Liu, J. Wang, and F. Yang, "Multiple standards compatible learning resource management," in *Advanced Learning Technologies (ICALT), 8th IEEE Int. Conf.*, July 2008, pp. 657–661.

[9] A. del Blanco, A. Serrano, M. Freire, I. Martinez-Ortiz, and B. Fernandez-Manjon, "E-learning standards and learning analytics. can data collection be improved by using standard data models?" in *IEEE Global Engineering Education Conference (EDUCON)*, 2013, pp. 1255–1261.

[10] V. Gonzalez-Barbone and M. Llamas-Nistal, "eAssessment: Trends in content reuse and standardization," in *Frontiers In Education Conference - Global Engineering: Knowledge Without Borders, Opportunities Without Passports, 2007. FIE '07. 37th Annual*, 2007, pp. T1G–11–T1G–16.

[11] S. Soares and A. Figueira, "Creating interopearable e-portfolios for different educational levels," in *IEEE Global Engineering Education Conference (EDUCON)*, March 2013, pp. 1049–1054.

[12] C. Maican, "Integrated university management system based on open source tools," in *Internet and Web Applications and Services, (ICIW) 4th Int. Conf.*, May 2009, pp. 626–631.

[13] T. Kudrass, "Integrated university information systems," in *ICEIS*, Y. Manolopoulos, J. Filipe, P. Constantopoulos, and J. Cordeiro, Eds., 2006, pp. 208–214.

[14] The E-Learning Framework. [Online]. Available: http://www.elframework.org/

[15] AICC. [Online]. Available: http://www.aicc.org/

[16] SCORM. [Online]. Available: http://www.adlnet.org/scorm/

[17] JISC. (2007) Effective Practice with e-Assessment. [Online]. Available: http://www.jisc.ac.uk/media/documents/themes/elearning/effpraceassess.pdf

[18] IMS Global Learning Consortium. Ims question and test interoperability specification. [Online]. Available: http://www.imsglobal.org/question/

[19] W. M. Davies and H. C. Davis, "Designing assessment tools in a service oriented architecture," in *1st International ELeGI Conference on Advanced Technology for Enhanced Learning, Vico Equense-Naples, Italy. 15-16 March 2005*, 2005.

[20] N. Sclater, B. Low, and N. Barr, "Interoperability with CAA: does it work in practice?" in *Proceedings of the Sixth International Computer Assisted Assessment Conference, Loughborough University*, 2002.

[21] M. Gusev, S. Ristov, G. Armenski, P. Gushev, and G. Velkoski, "E-assessment with interactive images," in *IEEE Global Engineering Education Conference (EDUCON)*, Turkey, 2014, p. in press.

[22] I. Chorbev, M. Gusev, D. Gjorgjevikj, and A. Madevska-Bogdanova, "Architecture of an electronic student services system and its implementation," in *ICT Innovations 2012, Web proceedings*, 2012.

[23] IMS Global Learning Consortium. Ims enterprise specification. [Online]. Available: http://www.imsglobal.org/enterprise/

[24] M. Jovanov and M. Gusev, "Performance evaluation of a new approach for automatic question production," in *ICT Innovations 2009*, D. Davcev and J. Gmez, Eds. Springer Berlin Heidelberg, 2010, pp. 275–283.

[25] A. Fernndez, D. Peralta, F. Herrera, and J. Bentez, "An overview of e-learning in cloud computing," in *W. on Learning Techn. for Educ. in Cloud (LTEC'12)*, ser. Advances in Intelligent Sys. and Comp., L. Uden et al., Ed. Springer Berlin Heidelberg, 2012, vol. 173, pp. 35–46.

[26] S. Ristov, M. Gusev, G. Armenski, K. Bozinoski, and G. Velkoski, "Architecture and organization of e-assessment cloud solution," in *IEEE Global Engineering Education Conference (EDUCON)*. Germany: IEEE, Mar. 2013, pp. 736–743, best Paper Award.

Cloud Based Laboratory for Distance Education

Sasko Ristov, Marjan Gusev and Aleksandar Donevski
Ss. Cyril and Methodius University,
Faculty of Computer Science and Engineering,
Skopje, Macedonia
Email: {sashko.ristov, marjan.gushev}@finki.ukim.mk, aleksandar.donevski@outlook.com

***Abstract*—Today's dynamic life has changed the traditional teaching from the classroom and laboratory based classical learning to an approach using online and distance e-Learning. Universities usually follow this trend and enable their resources as online utilities for the students. There are several ways to offer these facilities online and cloud is the most promising one when analyzing several criteria. The cloud technology provides many benefits reducing the overall cost and improving the business continuity. In this paper we present an idea how to transfer the traditional computer laboratory into a cloud based utility using open source cloud frameworks. This transformation creates a dynamic, flexible, elastic, reliable and modern laboratory that can serve the students as a tool for distance learning for many courses simultaneously. It acts on a "Lab as a Service" principle.**

***Index Terms*—Education, Cloud Computing, Online laboratory**

I. Introduction

Universities are faced with a huge problem when they design their data centers and laboratories. The hardware equipment's life-cycle is usually three to five years and the managers must be smart enough to choose appropriate hardware, that is, to choose the most powerful equipment within a given limited budget. However, the experience shows that most of the time the resources will be either over-utilized (for example, during assignment deadlines, assessments or enrollments), or underutilized (during holidays or weekends).

Another problem, which maybe the worst, is to choose a sufficient and appropriate laboratory equipment. That is, different courses require equipment with specific technical characteristics. For example, multimedia-based courses require computing power and available storage space, parallel programming and algorithms need multi-core CPUs and a lot of RAM memory, while some courses demand only a high-bandwidth Internet connection, regardless of the equipment's power.

Apart of the hardware equipment, the various courses require different system and application software. A diversity of operating systems limit the use of learning tools, some might need Unix or Linux, while others Windows or Mac operating systems. Even more, 16-bit programs are not supported, and many 32 bit programs do not work properly on 64 bit operating systems since they run in an emulator [1]. Another problem arises when installing a program for one course damages already installed programs for other courses. There are file version conflicts, registry value conflicts, and their unnecessary duplication.

University laboratories are deployed to allow the students to learn the material more practically. However, there are many part-time students, which are part-time employed and cannot attend the hands-on laboratory in specific hours. They demand a distance learning environment available 24h per day. European Credit Transfer and Accumulation System (ECTS) on the other hand, gives a recommendation to the universities to evaluate the students' homework assignments and projects, which usually are developed at home, and the regular students also need some distance learning environment.

A possible solution for all these challenges is to use the emerging technologies, such as virtualization and cloud computing. Virtualization allows a creation of separated virtual machine images for each course and instantiating arbitrary number of active instances with customized virtualized hardware resources (CPU, RAM, I/O, HDD). By using the cloud computing paradigm, the universities can either build their own private clouds in their data centers, or rent resources from some commercial public cloud service provider, or both. A para-virtual laboratory can even improve the students' performance [2].

Our idea is to realize the laboratory based on a service principle offered by the cloud technology. In this paper, we present an idea how to develop a cloud based laboratory for distance learning by using the emerging techniques without or with a little additional cost. This online laboratory acting as a service is intended to be used by the arbitrary number of students of several courses simultaneously, 24 hours per day, 7 days in a week, instead of university working hours only.

The rest of the paper is organized as follows. Section II presents the related work in the area of distance education based on cloud. In Section III, we describe the architecture of our cloud based laboratory for distance learning. The usage of the laboratory is presented in Section IV. Section V discusses about further development of the laboratory for distance learning and self assessment, apart of classical online laboratory. Finally, we conclude our work and present our plans for future work in Section VI.

II. Related Work

This section presents the state of the art in the area of distance learning, especially the cloud based.

Many online courses exist that offer distance learning. Udacity offers free and chargeable online courses, which

can be started at arbitrary time [3]. Putting the Stanford's Introduction to Artificial Intelligence course online, Thrun got an enrollment of 160,000 students from all over the world [4]. StarHPC [5] (the Software Tools for Academics and Researchers) is one of the first projects where the cloud is used in distance learning, which "brings" arbitrary Amazon's EC2 (Elastic Computing Cloud) resources into the students' laboratory. StarCluster [6] is a similar project where the students can create a cluster computing environment in the cloud for parallel and distributed computing applications and systems. For more than 10 years, Cisco has a learning environment, which is migrated on the cloud [7].

Platform as a Service (PaaS) cloud service layer can be also useful in education, since it allows the students to focus in the tasks, rather than on spending a lot of time to set up the necessary software for course activities [8]. Tian's framework [9] manages PaaS in a virtual cloud laboratory and manages the users, resources and their access rights. Yin et al. [10] built a multimedia courseware sharing platform, which is based on Google App Engine. NVLab is another tool, which uses virtualization. It is a web-based laboratory for management of network designs [11].

Gusev et al. [12] developed an e-Learning and benchmarking platform for distance learning of the Parallel and Distributed Processing course. They proposed a cloud model for the platform, as well as [13]. Bastidas [14] proposed the implementation of the computer networking laboratory infrastructure to distance students in order to extend the learning.

III. The Online Laboratory Architecture

In this section we present our model of a distance learning lab working as a service on a cloud. The architecture consists of two open source cloud frameworks, which are dedicated with arbitrary number of physical cloud server nodes.

A. The Physical Equipment

The laboratory is based of the existing 20 workstations, one router and one switch, as depicted in Figure III-A. All workstations are grouped in several VLANs connected by a switch. A router is used to enable an access to the university internal network.

B. The Architecture

We have a lot of experience during the last several years with various open source cloud frameworks, such as OpenStack, Eucalyptus, OpenNebula and CloudStack. Realizing various cloud deployments we have faced several challenges, such as Java incompatibility with applications in the cloud, or interaction with the cloud interfaces. Therefore, our recommendation is to implement two cloud frameworks, i.e., OpenStack and Eucalyptus, as most appropriate to deploy our online laboratory. This provides a sufficient redundancy, as well as extension for research options.

Figure III-B depicts the architecture of this online laboratory offered as a service. Two workstations are used as Eucalyptus

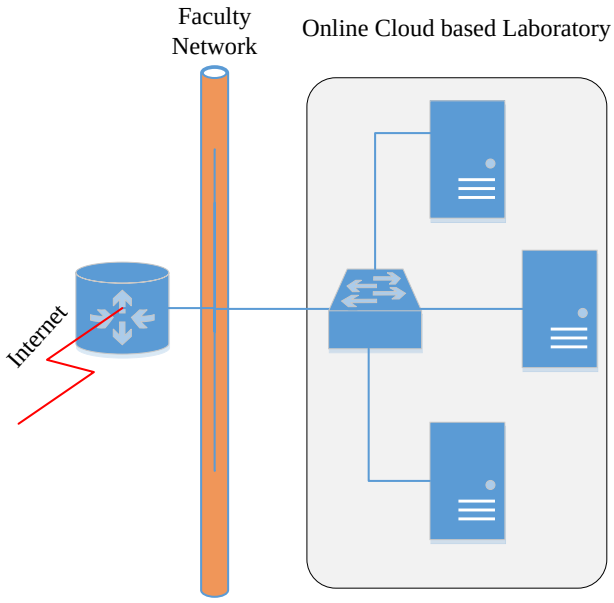


Fig. 1. The online laboratory equipment

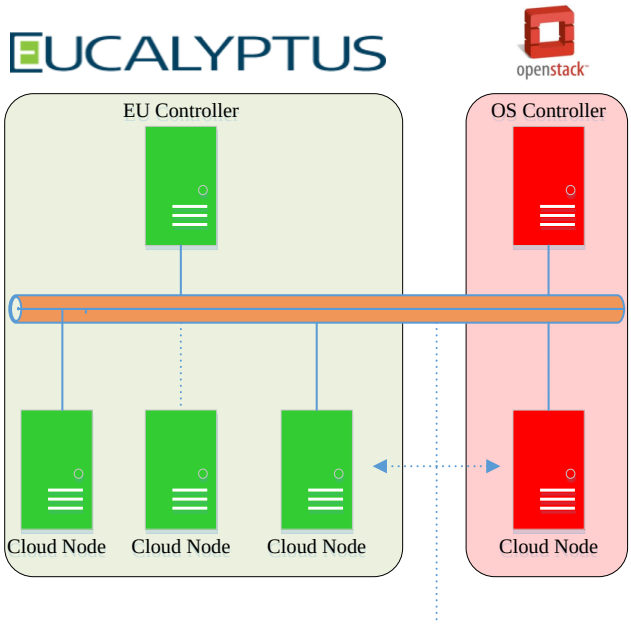


Fig. 2. The online laboratory architecture

cloud and cluster controllers, and one workstation for OpenStack. All others are used as cloud node controllers.

In normal mode, the Eucalyptus cloud uses all other workstation to instantiate virtual machines, while all OpenStack's services are disabled. When necessary, the OpenStack's services will be enabled, while the Eucalyptus' services will be disabled on several cloud nodes.

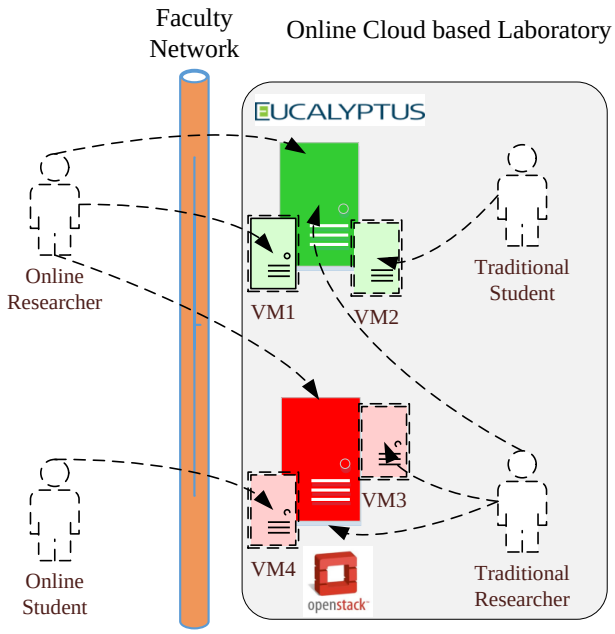


Fig. 3. The usage scenario

IV. Functional Description

This section describes the usage of our deployed cloud based online laboratory realized as a service. Figure IV depicts its three main purposes:

- *Traditional Physical Laboratory* opened for the students only in working hours;
- *Distance Learning Laboratory* available for the students 24/7/365; and
- *Research laboratory* [15], available both for the students and researchers.

A. Using as a Traditional Physical Laboratory

Since each workstation is installed with Linux with GUI (Graphical User Interface), the whole laboratory can be used as a traditional physical laboratory in the faculty's working hours. The students use the host operating system for those courses that require Linux operating system, while they can still run the virtual machine instance of the cloud for a particular course that requires Windows based applications.

B. Using as a Distance Learning Laboratory

Our currently prepared image with Windows operating system has 124 installed applications, which includes several IDEs (Integrated Development Environments) for various programming languages, simulators, network and database tools, etc. Each of the laboratory workstations is installed from this image. Although this is a huge list of installed applications, it is only a Windows platform, and something else is always missing, such as some plug-in or some configuration file is changed by other application.

Update of the image is conducted each semester, which limits the flexibility of the traditional laboratories. Therefore, we migrated one laboratory in the cloud to become a distance learning laboratory. We developed several images, each for a particular course. For example, the Computer Networks image is Windows based, which contains Cisco's learning materials, several network tools and simulators, etc, while the image for the High Performance Computing course is Linux based, which contains C++ compiler, OpenMP, MPI and BLAS libraries.

Since most of the regular students are at the faculty and work on the host operating systems, while the other busy students are at work, the laboratory is used mostly as a traditional physical laboratory in the working hours. Just in case, the administrator can limit the number of available virtualized resources through the cloud according to the requirements of the laboratory exercises of the current course. For example, if the current course requires two CPUs of four core cloud server nodes, the administrator can limit the cloud scheduler to instantiate virtual machine instances with maximum two CPU cores per physical server.

C. Using as a Research Laboratory

Additionally, apart of the already prepared customized virtual machine images for several courses, we prepared "blank" virtual machines for the master and PhD students to realize research, or for the undergraduate students to realize a team-oriented project. These virtual machines are allowed to interact with the cloud controllers in order to develop the elastic and scalable applications as a service on the top of the cloud, according to the model for e-Assessment cloud solutions [16].

V. Discussion

This section discusses details of current and further lab development to realize the overall goal as a distance education lab providing its resources as a service to the students, teachers and researchers. The benefits of migrating the traditional laboratories in the cloud are elaborated along with description of possible challenges.

A. The Benefits

The university can achieve many benefits if it migrates some or all of its laboratories in the cloud and offer it on a service based approach. In this section we present the most important ones.

1) Isolated Course Applications: The online laboratory is used for several courses: Computer Networks, High Performance Computing, Parallel Programming, Teamwork, Computer Network Design, Microprocessor Systems, etc. Each virtual machine instance is independent of others and works isolated when it is instantiated. This improves the performance of the guest operating system because it has less system files, rather than our image with 124 installed programs.

2) Balanced Load of the Equipment: Since different courses require various hardware resources and cloud based online laboratory allows the students to work in the arbitrary time, the utilization of the laboratory equipment will be balanced instead of either over-utilized or underutilized. Even more, the laboratory can serve much more online students rather than a maximum of 20 if it is used as traditional laboratory only. That is, usually less than a maximum of 20 students are enrolled in the exercises, and not all of them attend all exercises.

3) Reducing the Cost: Migrating the laboratory in the cloud can reduce the costs in several ways.

E-Learning should be developed towards open access to learning, open source software, open standards, and OER [18]. The additional software that our pilot cloud based laboratory uses is open source, that is, OpenStack and Eucalyptus. Although the expert's assistance is required to setup a fully operational, open source workbench [19], the cost for man hours for the initial efforts for installing and configuring the clouds is much smaller than the cost for formating and configuring each computer at the beginning of each semester, and the administering during the semester (viruses, no storage place, deleted files, etc).

The reduced number of students in the laboratory will also reduce the cost for laboratory maintenance. The cost for electricity can be reduced because all physical machines (except the cloud controllers) can be started or deactivated dynamically, according to the students requirements.

B. Further Challenges

Despite the benefits, there are additional challenges for the online cloud based laboratory, such as security or availability.

All cloud based environments are more exposed than being on-premise due to virtualization and multi-tenancy [20]. However, the student exercises' data are less sensitive.

Another challenge is the laboratory availability. That is, the laboratory room is neither dedicated for mission critical services, nor its room has sophisticated climate and redundant power supply. Still, our design with two clouds and many server nodes offer good availability.

Maybe the most important challenge is the usage of the online laboratory for hardware based courses, which requires working on real equipment, rather than on simulators. However, the remote engineering is a new trend to share the lab equipment, which will allow the students with more time for learning, instead of traveling to the University [21].

C. Further Extensions and Improvements

Our pilot project is a huge step forward, both for the university and the students. However, we are not going to stop at this level, instead we are going to further improve and enhance our cloud based online laboratory. Our intention is to develop the laboratory not only for distance learning, but more wider, that is, for distance education.

Applying formative assessment in learning environments increase the benefits [22], [23]. The learning outcomes of students can be improved in a mobile learning environment [24].

Recently, we have developed a scalable and elastic cloud solution for e-Assessment based on OpenStack cloud framework [17] and E-Assessment system with interactive images [25]. Both of them will be hosted in our cloud based online laboratory in order to allow the students to learn through e-Testing. The following sections briefly describe them.

1) Scalable and Elastic e-Assessment Cloud Solution: The e-Assessment cloud system is deployed on OpenStack cloud, which allows to dynamically instantiate and deactivate the virtual machine instances. Figure 4 presents its architecture. The implementation creates an instance of already prepared virtual machine for assessment with particular size according to the number of assessed students, which is deactivated after the assessment is finished.

2) E-Assessment System with Interactive Images: This system uses sophisticated technologies for interactive images, such as Google Maps Engine. Interactive image technology offers a web based environment where the images are sent to the students. After loading the image, the students can navigate and zoom to an appropriate level to analyze the image details. The students answer by setting up one or more marks on some regions of the image. This system enriches the e-Assessment environment, enables more correct evaluation and prevents several methods of cheating and unfair grading.

VI. Conclusion and Future Work

This paper presents the architecture and the usage of the multi purpose cloud based distance education laboratory. It is deployed by migrating the existing traditional laboratory in the cloud by using the open source cloud frameworks. It is offered as a service free of charge for enrolled students and University researchers that demand more hardware resources. It acts as an extended Platform as a Service, since it offers more than standard platforms the conventional cloud providers offer. It is not a fully Software as a Service offer for the students, since it offers a limited set of software tools and hardware in comparison to the full software the conventional cloud providers offer. Finally, the students are expected to develop the final application code or realize experiments with the available tools and hardware.

The laboratory can be used in three manners. The students still can work on-site when they have hands on exercises, such as hardware based courses, which require working on equipment. Apart of being the traditional on-site laboratory, it allows the students to work remotely from home whenever they can and want, for each course. Finally, it is also used by the young researchers, such as master or PhD students.

The usage of the laboratory can be also extended in distance education, that is, it can host faculty's online learning and online self assessment systems. We will migrate our interactive e-Assessment and e-Learning systems in the laboratory in order to make them scalable and elastic capable to handle arbitrary number of assessed students.

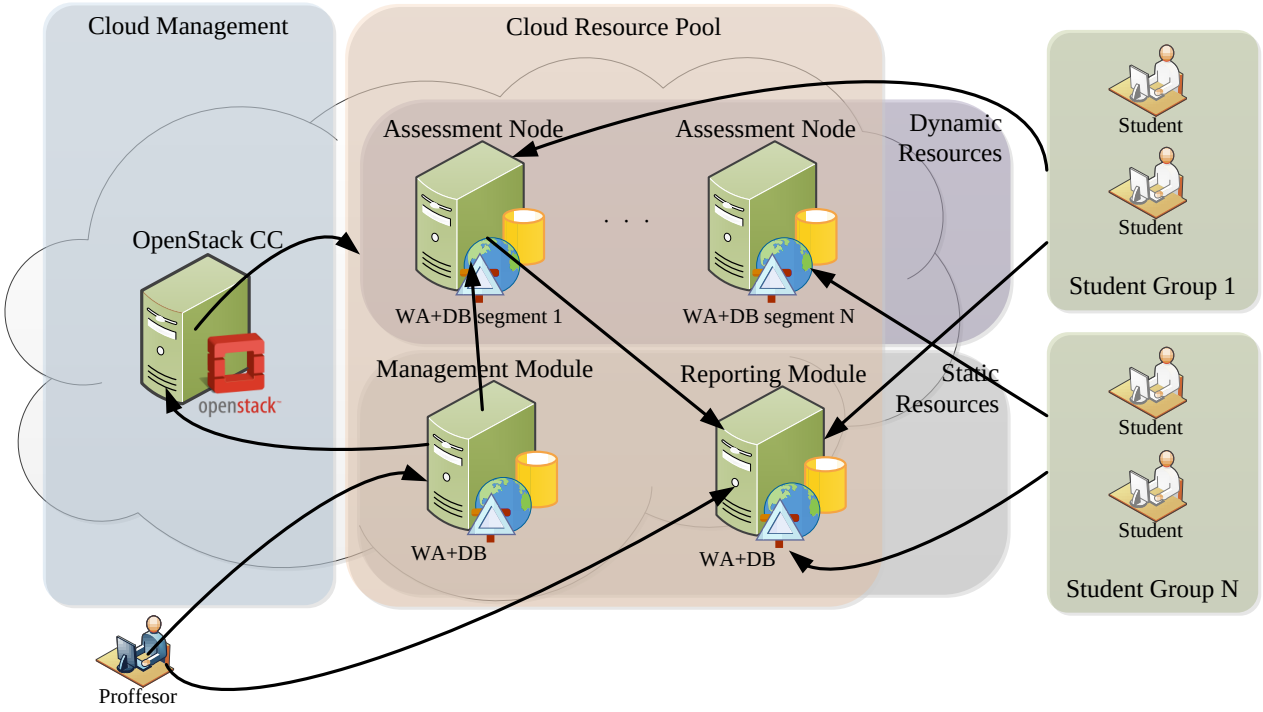


Fig. 4. e-Assessment cloud system [17]

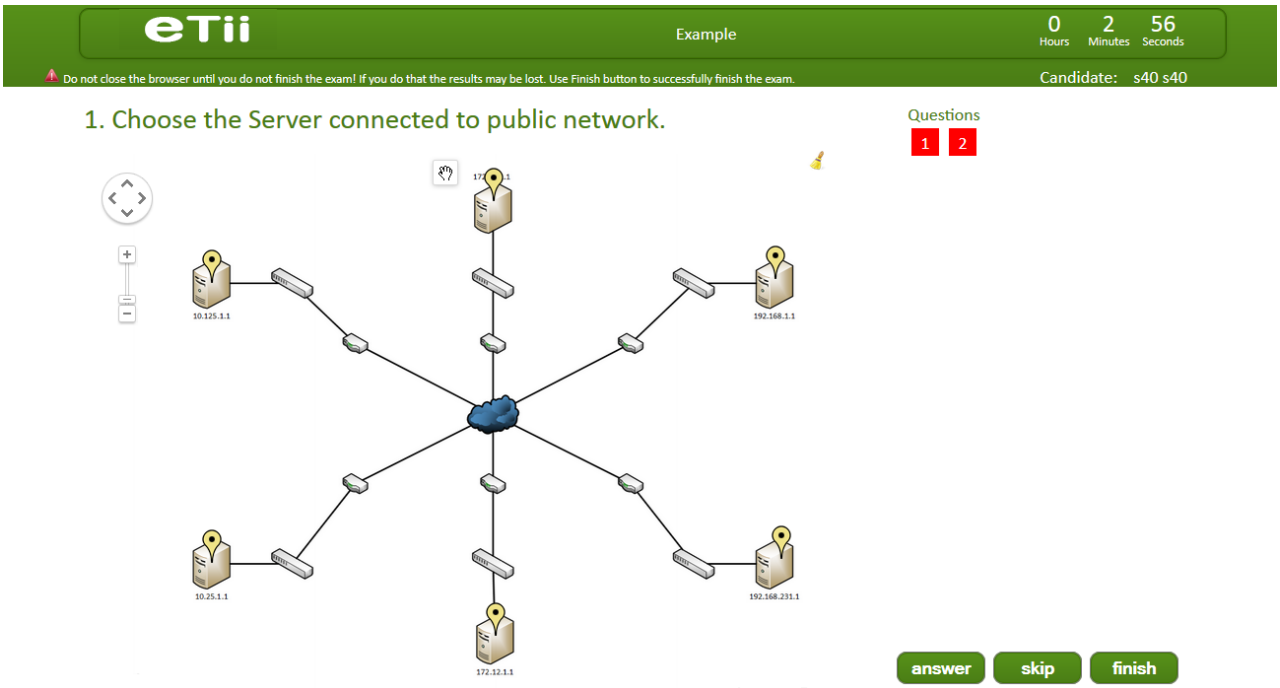


Fig. 5. An example of a MGC question for e-Assessment system with interactive images [25]

Our pilot laboratory will be extended and improved in several directions. We will develop and implement virtual machine images for more courses and we will test the performance of the increased number of the users. Also, we are challenged to integrate several distributed laboratories on several locations into one huge virtual online laboratory for distance education.

REFERENCES

[1] Microsoft, "Running 32-bit Applications," [retrieved: Jan, 2014]. [Online]. Available: http://msdn.microsoft.com/en-us/library/aa384249(v=vs.85).aspx

[2] M. Anisetti, V. Bellandi, A. Colombo, M. Cremonini, E. Damiani, F. Frati, J. Hounsou, and D. Rebeccani, "Learning computer networking on open paravirtual laboratories," *Education, IEEE Transactions on*, vol. 50, no. 4, pp. 302–311, 2007.

[3] Udacity, "Courses," [retrieved: Jan, 2014]. [Online]. Available: https://www.udacity.com/

[4] F. Salmon, "Udacity and the future of online universities," 23 Jan. 2012.

[5] C. Ivica, J. Riley, and C. Shubert, "StarHPC: Teaching parallel programming within elastic compute cloud," in *Information Technology Interfaces, 2009. ITI '09. Proceedings of the ITI 2009 31st International Conference on*, 2009, pp. 353–356.

[6] MIT, "StarCluster," [retrieved: Nov, 2013]. [Online]. Available: http://star.mit.edu/cluster/

[7] CISCO, "Cisco networking academy," [retrieved: Nov, 2013]. [Online]. Available: https://cisco.netacad.com/

[8] L. Vaquero, "Educloud: Paas versus iaas cloud usage for an advanced computer science course," *Education, IEEE Transactions on*, vol. 54, no. 4, pp. 590–598, 2011.

[9] W. Tian, S. Su, and G. Lu, "A framework for implementing and managing platform as a service in a virtual cloud computing lab," in *Education Technology and Computer Science (ETCS), 2010 Second International Workshop on*, vol. 2, 2010, pp. 273–276.

[10] H. Yin, J. Han, J. Liu, and X. Hongyun, "Development and research of multimedia courseware sharing platform based on gae," in *E-Health Networking, Digital Ecosystems and Technologies (EDT), 2010 International Conference on*, vol. 2, 2010, pp. 289–292.

[11] M. Wannous and H. Nakano, "Nvlab, a networking virtual web-based laboratory that implements virtualization and virtual network computing technologies," *Learning Technologies, IEEE Transactions on*, vol. 3, no. 2, pp. 129–138, 2010.

[12] M. Gusev, S. Ristov, G. Velkoski, and B. Ivanovska, "E-learning and benchmarking platform for parallel and distributed computing," *International Journal of Emerging Technologies in Learning (iJET)*, 2014, in press.

[13] S. Ristov, M. Gusev, and G. Velkoski, "Cloud e-learning and benchmarking platform for the parallel and distributed computing course," in *Global Engineering Education Conference (EDUCON), 2014 IEEE*, Istanbul, Turkey, 2014, p. in press.

[14] C. Bastidas, "Enabling remote access to computer networking laboratories for distance education," in *Frontiers in Education Conference (FIE), 2011*, 2011, pp. F3C–1–F3C–6.

[15] LiiT, "Laboratory for Internet and Innovative Technologies," [retrieved: Jan, 2014]. [Online]. Available: http://liit.finki.ukim.mk/

[16] S. Ristov, M. Gusev, G. Armenski, K. Bozinoski, and G. Velkoski, "Architecture and organization of e-assessment cloud solution," in *IEEE Global Engineering Education Conference (EDUCON), 2013 IEEE*. Berlin, Germany: IEEE, Mar. 2013, pp. 736–743, best Paper Award.

[17] S. Ristov, M. Gusev, G. Armenski, and G. Velkoski, "Scalable and elastic e-assessment cloud solution," in *Global Engineering Education Conference (EDUCON), 2014 IEEE*, Istanbul, Turkey, 2014, p. in press.

[18] D. Dinevski, "Open educational resources and lifelong learning," in *Information Technology Interfaces, 2008. ITI 2008. 30th International Conference on*, 2008, pp. 117–122.

[19] C. T. Delistavrou and K. G. Margaritis, "Towards an integrated teaching environment for parallel programming," in *Proceedings of the 2011 15th Panhellenic Conference on Informatics*, ser. PCI '11, 2011, pp. 3–7.

[20] S. Ristov, M. Gusev, and M. Kostoska, "A new methodology for security evaluation in cloud computing," in *MIPRO, 2012 Proceedings of the 35th International Convention*, 2012, pp. 1484–1489.

[21] D. Pop, D. Zutin, M. Auer, K. Henke, and H. D. Wuttke, "An online lab to support a master program in remote engineering," in *Frontiers in Education Conference (FIE), 2011*, 2011, pp. GOLC2–1–GOLC2–6.

[22] L. Gardner, D. Sheridan, and D. White, "A web-based learning and assessment system to support flexible education," *Journal of Computer Assisted Learning*, vol. 18, no. 2, pp. 125–136, June 2002.

[23] P. Orsmond, S. Merry, and A. Callaghan, "Implementation of a formative assessment model incorporating peer and selfassessment," *Innovations in Education and Teaching International*, vol. 41, no. 3, pp. 273–290, 2004.

[24] G.-J. Hwang and H.-F. Chang, "A formative assessment-based mobile learning approach to improving the learning attitudes and achievements of students," *Comput. Educ.*, vol. 56, no. 4, pp. 1023–1031, May 2011.

[25] M. Gusev, S. Ristov, G. Armenski, P. Gushev, and G. Velkoski, "E-assessment with interactive images," in *Global Engineering Education Conference (EDUCON), 2014 IEEE*, Istanbul, Turkey, 2014, p. in press.

Mediated Heritage Learning in Formal and Informal Contexts: Antiquity CyberEd

G. Zlodi*, Ž. Miklošević*
* Faculty of Humanities and Social Sciences, University of Zagreb/
Department of Information and Communication Science, Zagreb, Croatia
E-mail: gzlodi@ffzg.hr, zmiklose@ffzg.hr

This article explores the ways a physical educational resource is transformed by ICT into a virtual learning material on the example of Ancient Greek sculptures at Zagreb's Faculty of Humanities and Social Sciences. The sculptures, acquired as casts from the originals in the British Museum and Louvre, serve as a learning resource in teaching museum study courses, while the results of this formal education provide a new learning material for an online environment. The theoretical framework is based on Laurillard's classification of educational media that facilitate participative learning in formal education and consequently in informal online heritage education.

I. INTRODUCTION

Cultural institutions, including museums have in the past several decades been oriented to users and the ways users themselves construct meaning communicated through heritage objects and sites. This trend corresponds well with contemporary educational theories, which replace transmissive approach in which the body of knowledge is conveyed to learners, with the approach that gives the central importance in the learner and his ways of constructing knowledge. In other words, learner becomes more important than the body of knowledge.

A. Cultural Heritage in Contemporary Theories of Learning – new media-enhanced education

Priority is given primarily to individual experience and social context in which learning takes place. Self-exploration, participation, experience, social interaction, personalization are only some of the terms which comprise contemporary view on learning in museums and heritage sector and have been used in literature dealing with informal learning, in museums and in general [1], [2] [3], [4], [5]. Presented cultural heritage exerts influence on users and their cognitive and perceptual characteristics change only if they come from the familiar cultural and social contexts and have previously known experiences related to the topic [6].

Furthermore, museums and interpretation centres are expected to provide users with opportunity to establish personal links with heritage items, topics and experiences, to enable participation in activities which go beyond mere observation, and finally, offer possibility of sharing experiences and opinions of people belonging to certain groups.

As one of the currently most used concept in heritage sector which is heavily oriented to communication, heritage interpretation has developed in the last thirty years, and especially with the introduction of the internet, into functional framework. Online heritage interpretation started with web pages that resemble an electronic brochure, end continued with the pages whose content is more of a digital database of museum objects, which turned into complex sites offering information related to different heritage contexts and preferences of online users instead of exclusively to museum objects [7], [8].

B. Museum Object in the Virtual Environment

The influence of new media on heritage or museum objects can be seen, on the one hand, as a threat to professionally created knowledge, and on the other hand as an opportunity to broaden the impact heritage can have on people if it becomes more accessible, both in terms of space and time, which is somewhat the internet provides, and in terms of understanding which can be achieved by inclusion of different contexts and perspectives on heritage.

The principles of numerical representation, modularity, automation, variability and transcoding [9] enable limitless copying, appropriation and modification of digital heritage objects by a network of wide varieties and groups of users. Digital objects are qualitatively and materially different from their "physical" counterparts. While the physical objects are relatively stable (Figure 1 and 2) entities in space and time, digital ones are mobile

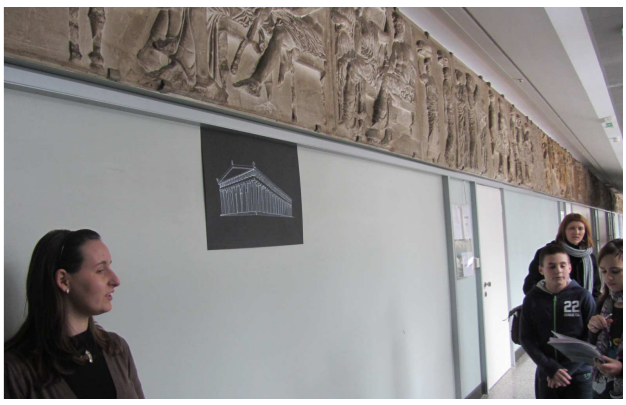


Figure 1. Casts of original sculptures from Parthenon - guided tour for 5th grade of elementary schoool

and volatile. Perry, however, observes changes in the comprehension of the "static" museum object in the virtual environment. The definition and value of the object became scaled and distanced from the issues of materiality and immediate physical presence, characteristic for the industrial period, and more in tune with the post-industrial perception of material culture in a wide sense – as units of human experience which witness the individual and collective epistemology and ontology [10]. In the institutional context, the use of digital integrated multimedia enables distancing from the limited information distribution and better representation of a wide range of contexts pertaining to certain information [11]. As far as users are concerned, access to digital objects forms continuous stimulation in creating new interpretation and knowledge.

Affordances offered by new technology design can easily be used as advantages for multiple interpretation and contextualisation of objects in virtual environment. Additionally, these online environments can help in overcoming certain limitations imposed by traditional methods of organisation and presentation of information and can contribute to the use of multiple levels, various aspects and quantity of information about particular topics.

Narratives which emerge around objects, created by various users, entice much more discussion than objects themselves since stories are the main thing that brings close users by encouraging knowledge acquisition [12], [13] and identification with heritage [14].

Still, materiality of heritage items and their auratic characteristics will always be important for experience. However, as opposed to the physical encounter with objects which is strongly supported by the desired visual, and more narrowly defined aesthetic experience, the physical world could (and should?) be extended by virtual spaces in terms of quality of heritage communication. The use of new media in heritage context should be taken as a positive direction towards digital practices in which elements of greater significance include connectedness, access, fortuity, network, collaboration, participation, freedom of expression, and cultural and social empowerment of users. These "key words" have served as guidelines for the project Antiquity CyberEd which comprises a virtual museum as an environment of informal education for a wide range of users and created by students within university courses in museum studies as part of their formal education.

II. ANTIQUITY CYBERED PROJECT

Antiquity CyberEd project is ongoing cultural heritage project launched at the Department of Information and Communication Sciences at the Faculty of Humanities and Social Sciences, Zagreb University. It envisages collaborative work of various faculty departments on the protection and communication of a collection of Ancient Greek and Roman sculptures. The Antiquity CyberEd forms only one part of the project and focuses on the educational process revolving around the collection, in terms of both formal learning within several courses in museum studies programme and informal learning in the form of virtual museum which results from the formal educational process.



Figure 2. Students of museology taking notes and discussing about damages on the sculptures

A. *Collection of casts of antiquity artworks located in Zagreb's Faculty of Humanities and Social Sciences*

The collection was formed in 1892 through acquisition of casts of ancient Greek sculptures housed in European museums. The biggest number of works belongs to the Parthenon frieze (Athens, Greece, ca 438-432 BC) and was cast in the British Museum in London. The man most responsible for the acquisition was Isidor Kršnjavi whose intention was to use the collection to bring awareness and aesthetic sensitivities to the general public [15].

Every few years, the collection moved from one institution to another, without the much needed care and museological interpretation and presentation. Recently, Zagreb's Faculty of Humanities and Social Sciences has made investments in conservation of several most damaged works. Still, objects are not equipped with labels and there is no theme panels providing contextual information to the collection as a whole. Furthermore, there is no navigation inside the faculty building which would direct people's attention to the collection.

This all has consequences on general visibility of the collection. Students know very little or nothing about it and a lot of individual artefacts have been damaged by carvings, graffiti and other kinds of devastation. After introducing students with the value and status of the collection, they showed high motivation and interest in the topic.

B. *Productive educational framework - from formal to informal learning contexts*

The theoretical framework of the project relies on the theory of experiential learning [16], [17] and embraces both the sense of institution-sponsored learning (i.e. university) in that it involves a direct encounter with the studied phenomena (in this case the physical collection), and the type of learning which occurs as a direct participation of individual in his or her life (i.e. informal learning or self-directed learning). The formal learning, emphasizing both process and product is made possible by the content management module which enables the

creation of digital "heritage content" for the virtual environment that facilitates informal learning.

The focus in both of these types of learning is placed on the technology-enhanced production and organization of knowledge which supports at the same time content and process intelligence and employs acquisition and contribution learning forms.

The framework presented in this paper (Table I.) comprises the following categories:

- learning resources (physical collection in the center of formal, and virtual collection as the core of informal learning)
- learning process supported by ICT tools which require students' participation and their individual creation, adaptation, structuring and organization of information related to the collection
- learning process results (Collection documentation and virtual catalogue and web site as an informal education resource)

TABLE I. PRODUCTIVE EDUCATIONAL FRAMEWORK

Productive educational framework - from formal to informal learning contexts		
Resources	***Learning process via ICT tools***	***Results***
Physical collection	Collection management and documentation	Collection documentation and virtual catalogue
Literature and related audio-visual material	Content management and adjustment	Web site (as an informal education resource)

User segmentation has been done in order to determine the tasks and goals of each group in the process of content creation and organisation, since they are all at different levels of knowledge about and relatedness to the collection in question. Museology students, as the primary group in formal education, have a task of creating and structuring a cultural heritage content, first from the perspective of museum research and documentation and then from the perspective of museum and heritage communication. Other students of the faculty who are envisaged to be a potential source of contribution as authorized online users with their expert knowledge about the sculptures, historical period, ancient languages, myths as well as personal stories from their everyday encounters with the collection at the faculty. In addition to these two groups of participants, there are also all other online users which will be able to visit the site and engage in informal learning about the collection in the virtual environment in their own pace, space and time.

C. Museum documentation database as a tools for productive education

At the beginning of the project museology students have to enter information about the collection from the already existing published museum catalogues, the information which can be obtained directly from the casts (measures, colour, damages etc.), and create links to audio-visual documentation which they also have to either create or include in the database. For this purpose they use M++ museum documentation database where each object is represented with multiple records about the origin, condition, related references etc.

This sort of information is highly formalized and structured by international standards and guidelines for data structure and data values. For example, students have used ICONCLASS classification system for vocabulary control for mythology and other subject information.

D. Content management module as a tool for productive education

After finishing the first main objective – to document the collection, students have to reuse digital content used in the collection documentation process and then adjust this content for communication purposes and publish it via content management system. Students use the content management module of M++ database (Figure 3) which enables developing of virtual collections and exhibitions.

This module enables following key features: development of thematic structure with related texts and visuals and linking of associated information about objects form the collection. It is important to stress at this point that catalogue items are integrated into already defined and structured topics which contain cultural, social and historical stories which are then supported, or better yet, evidenced by deliberately chosen and most suitable objects. The module is also characterized by a feature that enables flexibility of the created information (i.e. student-created knowledge). Since the main purpose is to develop a website as an informal learning resource the module facilitates modification and adjustment of the linked information in order to make the online communication easier, more comprehensive and effective.

Another very interesting functionality of content management module is the possibility to choose terms from controlled vocabularies that will be visible on the end user interface. That enables a fine adjustment of the complexity of terms and hierarchical structures on the

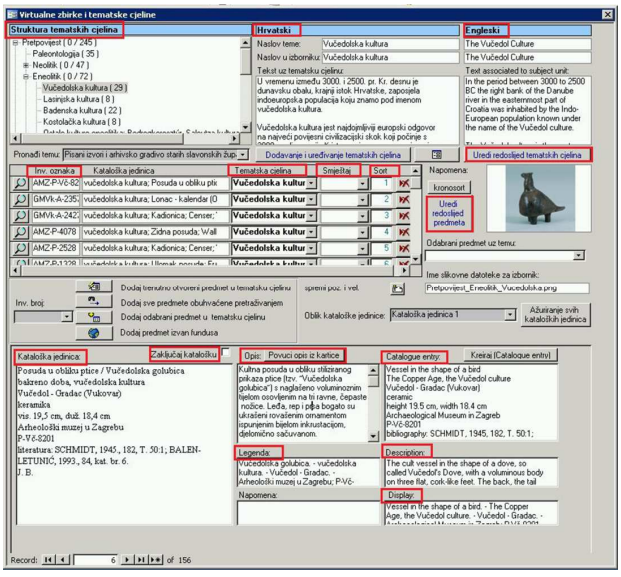


Figure 3. Content management module of M++ museum collection management database

website. It is also possible to map items (i.e. terms) from hierarchically structured vocabularies to visual controls such as timelines or ground plans.

E. Website as an informal education resource

In this ongoing project the website would be a continuously growing resource where final results of formal learning process would be presented for many generations of students. The projects' long term goal is to communicate continuous care for physical collection via web pages whose content will be constantly updated and new functionalities implemented.

Students who decide to continue their contribution to the project after their formal obligations in museology courses would be able to update web site segments such as stories and interesting facts related to the collection in informal context as website users.

A virtual catalogue, as the central part of the website, provides multiple points of access to information about the heritage objects and associated contextual information through timelines, interactive maps and thematic structures, and allows the inevitable search through the entire database. Interactive ground plans would be a very effective way to give directions from the virtual environment to a specific position in the physical environment [18]. Visitor experience in physical environment could also be enriched by interactive video projection mapping on sculptures. Simulation of colors in which marble originals at the Parthenon were painted has been projected, as you can see on Figure 4.

To ensure communication in the opposite way, from the physical to virtual environment, to find additional information about the objects and collection, QR codes would be included as an integral part of object labels.

Technology-mediated creativity and research done by museology students thus forms the basis of a further development of the heritage content by virtual users, which can further stimulate the development of new solutions in both technological and content-based aspects of the framework. This approach therefore makes students/users not only participants in the creation of heritage content but also co-creators of learning resources through the use ICT. This knowledge-making process reflects also on the learning outcomes for students in the formal aspects of education which is beneficial not only to the quality of their academic achievement but to their personal intellectual and social growth. Or, as Jarvis says: "Experiential learning is the process by which individuals, as whole persons, are consciously aware of a situation and make sense, or try to make sense of what they perceive, and then seek to reproduce or transform it and integrate the outcomes into their own biography." [17]

From the point of practical contribution of the project particularly, and museology courses in general, one of the most important though implied goal is to provide students with an insight into a part of the custodian work, which comprises a wide range of procedures, from collection care and documentation to different ways of communicating heritage – from exhibitions in the physical or virtual spaces, narrative development and the like.



Figure 4. Projection of simulation of colors in which marble originals was painted at the Parthenon

III. CONCLUSION

Experiences and results of the work on the presented project show that it is possible to narrow the gap between a documentary approach, as a set of traditionally rigid procedures of fixating identities of cultural heritage objects in controlled documentation media, and heritage communication in the virtual environment, which by definition involves a flexible approach based on interactivity, multimedia and open access.

This project, relying in its goals on the productive educational framework and the described content management module, promotes knowledge construction and organization through the use of physical and digital heritage artifacts. It also helps students and online users to capture the totality of learning experience which is based on participation and contribution of individually created content.

REFERENCES

[1] G. Hein, Learning in the Museum. London, New York: Routledge, 1998.

[2] J. H. Falk and L. D. Dierking. Museum Experience. Washington, D.C.: Whalesback Books, 1992.

[3] J. D. Brandsford, A. L. Brown, and R. R. Cocking, Eds., How People Learn: Brain, Mind, Experience, and School. Washington, D.C.: National Academy Press, 2000.

[4] M. Glassman, "Dewey and Vygotsky: society, experience, and inquiry in educational practice," Educational Res., vol. 30, no.4, pp. 3-14, 2001.

[5] E. Hooper-Greenhill, Museums and Education: Purpose, Pedagogy, Performance. London, New York: Routledge, 2007.

[6] E. von Glassersfeld, "A constructivist approach to teaching," in Constructivism Education, L. P. Steffe and J. E. Gale, Eds. London, New York: Routledge, 1995.

[7] W. Schweibenz, "The development of virtual museums," ICOM News, no.3, pp. 3, 2004.

[8] M. Sigala, "A learning assessment of online interpretation practices: from museum supply chains to experience ecologies, " in Information and Communication Technologies in Tourism - Proceedings of the International Conference in Innsbruck, A. J. Frew, Ed. Wienna, New York: Springer, 2005, pp. 67 – 78.

[9] L. Manovich, The Language of New Media. Cambridge MA: MIT Press, 2001.

[10] R. Parry, Recoding the Museum: Digital Heritage and the Technologies of Change. London: Routledge, 2007.

[11] G. Hoptman, “The virtual museum and related epistemological concerns,” in Sociomedia: Multimedia, Hypermedia, and the Social Construction of Knowledge, E. Barrett, Ed. Cambridge MA: MIT Press, 1992, pp. 141-160.

[12] S. G. Paris and M. J. Mercer, “Finding self in objects: identity exploration in museums,” in Learning Conversations in Museums, G. Leinhardt, K. Crowley and K. Knutson, Eds. Mahwah, N.J. : Lawrence Erlbaum, 2002, pp. 401 – 423.

[13] D. T. Schaller and S. Allison-Bunnell, “Practicing what we teach: how learning theory can guide development of online educational activities,” in Museums and the Web 2003, J. Trant and D. Bearman, Eds. Toronto: Archives & Museum Informatics, 2003. [Online]. Available: http://www.museumsandtheweb.com/mw2003/papers/schaller/schaller.html#ixzz1vX9VmNTK

[14] Ž. Miklošević, G. Zlodi, D. Babić, “Story sharing,” in The Future of Information Sciences INFuture 2011: Information Sciences and E-Society, C. Billeness, A. Hemera, V. Mateljan, M. Banek Zorica, H. Stančić, S. Seljan, Eds. Zagreb : Department of Information Sciences, Faculty of Humanities and Social Sciences, University of Zagreb, 2011, pp. 109-119

[15] M. Matijaško, “Gipsani odljevi antičkih umjetničkih djela smješteni u prostorijama filozofskog fakulteta u Zagrebu,” Muzeologija, no. 46, pp. 37-135, 2009.

[16] D. A. Kolb, Experiential Learning – Experience as the Source of Learning and Development, Englewood Cliffs, N.J.: Prentice Hall, 1984.

[17] P. Jarvis, Adult Education and Lifelong Learning: Theory and Practice. London, New York: Routledge Falmer, 2004.

[18] G. Zlodi and I. Maroević, “Od muzejske dokumentacije do tiskanoga i virtualnog kataloga: modul muzejskoga informacijskog sustava M++ za upravljanje i uređivanje sadržaja i tematskih cjelina virtualnih zbirki i kataloga,” in Što je muzejski katalog, S. Radovanlija Mileusnić, Ed. Zagreb: Muzejski dokumentacijski centar, 2013, pp. 76-95.

Development of an Online Multimedia Course

M. Troselj*, M. Matetic**, M. Tkalcic***, M. Brkic**, P. Pahljina*
* Bura Znanja, University of Croatia, Rijeka, Croatia
marin.troselj@gmail.com, ppahljina@gmail.com
** Department of Informatics, University of Rijeka, Rijeka, Croatia
majam@inf.uniri.hr, mbrkic@inf.uniri.hr
*** Faculty of Humanities and Social Sciences, University of Rijeka
mladenka.tkalcic@ffri.hr

Abstract – At the University of Rijeka the theoretical knowledge is mainly transferred by the traditional model consisting of professor explaining material to a group of students in front of a blackboard. We believe that this model dates from the period of the industrial revolution when it was necessary to educate large number of people to work on new machines, but there was no technology for more efficient knowledge transfer. Nowadays, such a model cannot keep up with the speed of generating new knowledge and technology. A larger number of participants cannot be educated due to time and space constraints, and the knowledge usually becomes obsolete by the end of studies. The result of this model are students with skills unadapted to the labor market requirements. We believe this is an extremely important issue for our university. One of the possible solutions might be digitizing theoretical parts of courses. Using web technologies, theoretical lectures can be held online and in that way more efficient methods of teaching on campus can be used. In this paper, the process of digitizing the university course Introduction to Psychology is presented, as well as the benefits that can be achieved by creating and using multimedia online courses.

I. PROBLEM DESCRIPTION

A. *Outdated model of knowledge transfer in Croatia*

At the University of Rijeka the theoretical knowledge is mainly transferred by the traditional model consisting of professor explaining material to a group of students in front of a blackboard or using power point slides. We believe this model is outmoded and dates from the time when there was no technology for more efficient knowledge transfer. Nowadays, such a model cannot keep up with the speed of generating new knowledge and technology. A larger number of participants cannot be educated due to time and space constraints, and the knowledge usually becomes obsolete by the end of studies. The result of this model are students with skills unadapted to the labor market requirements. We believe this is an extremely important issue for our university.

B. *Practical problems*

Students are faced with a number of problems in mastering course content. Inadequate, outdated and sometimes raw literature, the lack of literature, badly maintained course websites, no playback of lectures, and the lack of interaction during the class are some of these problems. Students who wish to choose and design their own curriculum, by attending only those courses that interest them, are forced to make compromises in their work week, due to the spatial and temporal limitations such as overlapping classes. Hence, the maximum capacity of their learning is not utilized [1]. Teachers, on the other hand, are overloaded by teaching and by bureaucratic obligations, which greatly inhibits their scientific productivity.

C. *Saving resources*

The University of Rijeka organizes a large number of courses each academic year - about half of the teaching workload comprises theoretical lectures. For example, the Department of Informatics offers 41 course at the undergraduate level, which is approximately 1125 hours of theoretical lectures [2]. When these theoretical parts of the courses could be mastered over the Internet by watching multimedia video lessons, the University could redirect significant resources in research and scientific projects. By using computer technology, universities can reduce the number of classrooms, staff and hard copies of teaching materials which are required if the number of students is to be increased [3]. Payne and Stergioulas argue that technology enhanced learning could be a significant driver for progressive change and economic growth, which can assist in the economic recovery, as well as in the democratization of education [4]. Teachers would not need to spend time in teaching and preparing the content year in year out, and students would be able to adapt the learning process to themselves so that they learn where and when they want and at a pace that suits them [3]. Moreover, such an approach would evidently solve the problem of overlapping courses.

II. TECHNOLOGY AND EDUCATION

Using technology in education can enhance the process of learning; therefore it is called Technology Enhanced Learning (TEL). It can make the process of learning more effective, efficient and flexible for students and it can improve teaching quality and add more value [3, 5, 6]. Technology evolved to the level that it can support individualized and collaborative learning [7, 8, 9]. There are even reports that students prefer TEL to the traditional face-to-face learning methods [1]. Some people in academia fear that TEL can never be as effective as face-to-face teaching, but one might argue

that this held true in the past, due to the negative experience with poorly designed content or inappropriate use of technology [4]. With education becoming more and more personalized, and enriched with multimedia and social media, that should not be the case anymore.

A. *Democratization of knowledge*

Until recently, access to the most advanced knowledge, created by leading experts, has been reserved for a small number of students who passed a rigorous and financially very demanding process of enrolling in one of the prestigious universities. All that changed with the open curriculum initiatives, which began to spread across prestigious universities. Since 2012, many online platforms have emerged with the goal of providing top quality education for free. Some of them are Coursera, edX, Udacity, Venture Lab, Udemy in USA or OpenupEd and Iversity in Europe. Since the courses are open to the general mass, they are called Massive open online courses or abbreviated MOOC. One might ask why institutions and educators would give their knowledge for free. Bonk argues that by doing it they promote their talent and enhance their reputation, they potentially help faculties with recruitment and retention, and they create connections with life long learners [10]. They can also build business models around this free content, like charging for knowledge certification or charging other institutions for material usage.

B. *MOOCs*

MOOCs are online courses with no formal entry requirements, and usually no participation limit. They can be enrolled free of charge and they do not earn credits. [11]. According to Siemens, there are 2 different categories of MOOCs [3, 12]. By adopting a connectivist philosophy, cMOOC, where c stands for connectivity, emphasizes creation, creativity, autonomy, and social networking learning. xMOOC is based on the traditional learning approach, and comprises of lecture recordings and short quizzes and tests. Lisa Lane suggests that MOOCs can be grouped into three types: network-, task-, and content-based [17].

C. *Big data in education*

More and more activities are carried out via the Internet. This results in an increasing amount of data to be analyzed in order to develop intelligent tools for improving the decision-making process. The same applies to the process of learning. With the development of MOOCs, the educational process began blending with the Internet in much higher rate. Users interact with LMS systems (Learning Management System) in different ways, and thus generate enormous amounts of data. These data should be analyzed to better understand the process of learning over the Internet and thus improve the LMS systems and the design of the courses themselves.

Learning analytics "is the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs" [13]. By building predictive models based on learning data, researchers can answer questions like: "What sequence of topics is most effective for a specific student? What student actions are associated with more learning (e.g., higher course grades)? What student actions indicate satisfaction, engagement, learning progress? What features of an online learning environment lead to better learning? What will predict student success [14]?" Long and Siemens argue that by using learning analytics we can innovate and transform university systems and pedagogical approaches, so its biggest value is in reforming activities in higher education and assisting educators in improving teaching [15].

III. COURSE DEVELOPMENT

The course Introduction to Psychology from the Faculty of Humanities and Social Sciences at the University of Rijeka is digitized for the purpose of this paper. The course is offered at the undergraduate level and intended as an elective course for the students of the entire university.

By enrolling in a course at the university, a student is expected to adopt certain theoretical concepts and develop specific skills. Since the purpose of this course is to provide a basic overview of the entire scientific field of psychology, most of the course loads are theoretical lectures and it is, therefore, not possible to go into the depths of matter or to develop certain skills. The course is digitized with the aim of making students learn theoretical concepts before coming to class, and thus a content-based xMOOC is created.

To create this online course, a web application for distributing the content needed to be developed, and the course content needed to be digitized.

A. *Web application development*

For the development of the web application, the "Lean Startup" philosophy is used [18]. Since this approach is based on fast and agile development, technology which enables developing a fast, dynamic and flexible application needed to be chosen. The back end part of the application is developed in Python programming language and Django web development framework driven by MySQL database. The front end is developed using HTML language and jQuery Javascript library. An important criterion for choosing these technologies was the fact that they are free and available to everyone.

B. *Course digitization*

The course is divided into 12 thematic sections, each divided into concepts that are shown in the form of video lessons with the average duration of two minutes. Recent research suggests that the format of short videos provides learning outcomes comparable to the traditional on-campus lectures [19]. To keep the users engaged, a fast exchange of short videos and short single choice questions with immediate feedback is employed. The

purpose of the questions is not evaluating users' knowledge, but rather pointing out key points and helping with memorizing them.

The course is available to the general public since one of the goals is the popularization of psychology. To make the course materials interesting and fun to watch, video lectures are made in the form of comics and drawings in the making. In order to improve users' memory, mnemonics is used, a strategy that helps in storing information in long term memory and recalling it when needed [20]. Fig. 1 shows an example of mnemonics which describes the role of glial cells inside our brain.

The whole digitization process, as shown in Fig 2, is conducted in 5 steps. The first step consists of preparing the course content in the form of narrative, and the questions that go with video lectures. Thereafter, two processes take place in parallel. While the illustrator creates graphics using the drawing tablet, the lecturer is recorded in a specially prepared studio. These steps are then followed by the assembly of video in a video editing software. The final step is the publication of the content on the web platform.

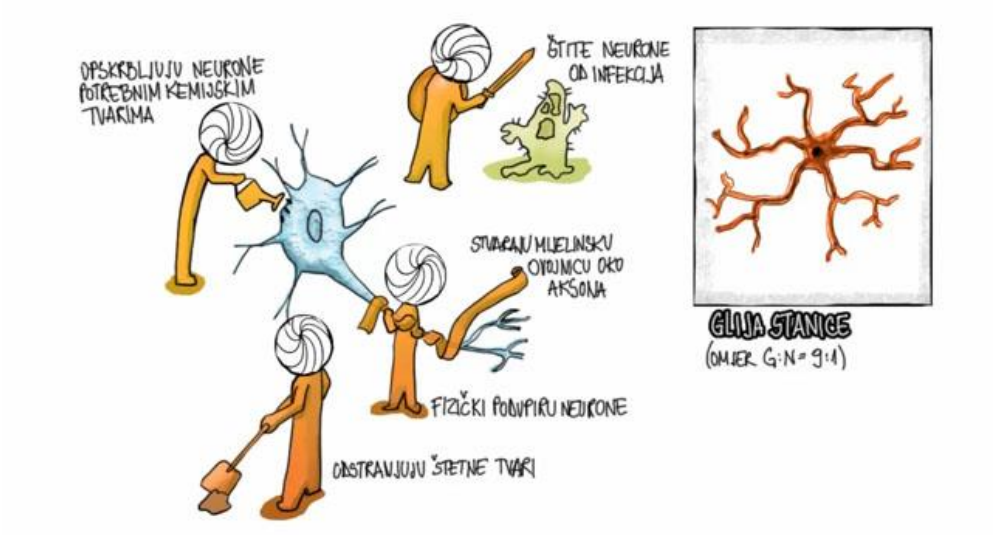


Figure 1. Using mnemonics in describing function of glial cells

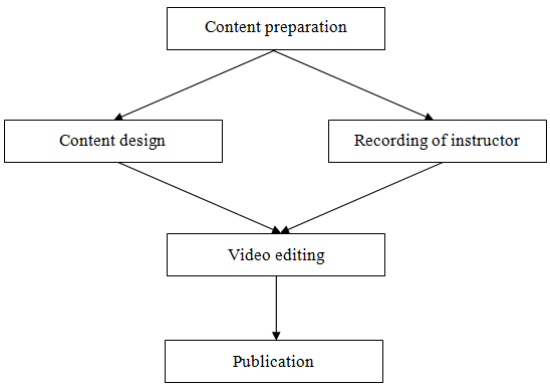


Figure 2. Course digitization process

IV. PRACTICAL RESULTS

Following the example of developing platforms for MOOCs mentioned in this paper, we opt for a similar approach. A standardized platform is made, enabling all courses to have the same structure. User interface currently consists of two main parts: a promotional page and a virtual classroom.

The main element of the promotional page is a motivational video, in which the course instructor is trying to motivate visitors to enroll in the course. This page also provides information about the course instructor, description, thematic segments and literature. Once participants enroll in the course, they get access to the classroom where they can watch video lectures and answer lecture questions, as shown in Fig 3 and Fig 4.

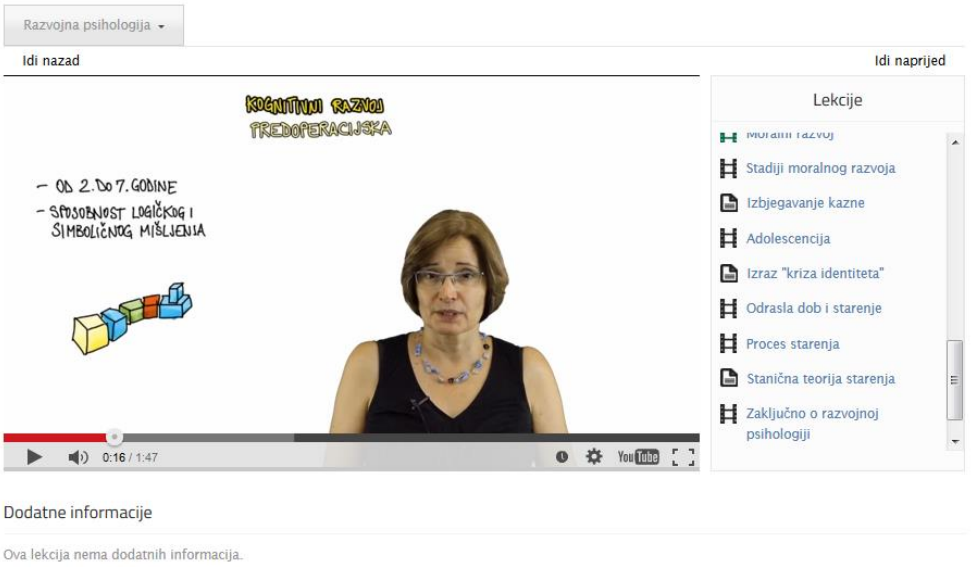


Figure 3. Watching video lessons inside virtual classroom

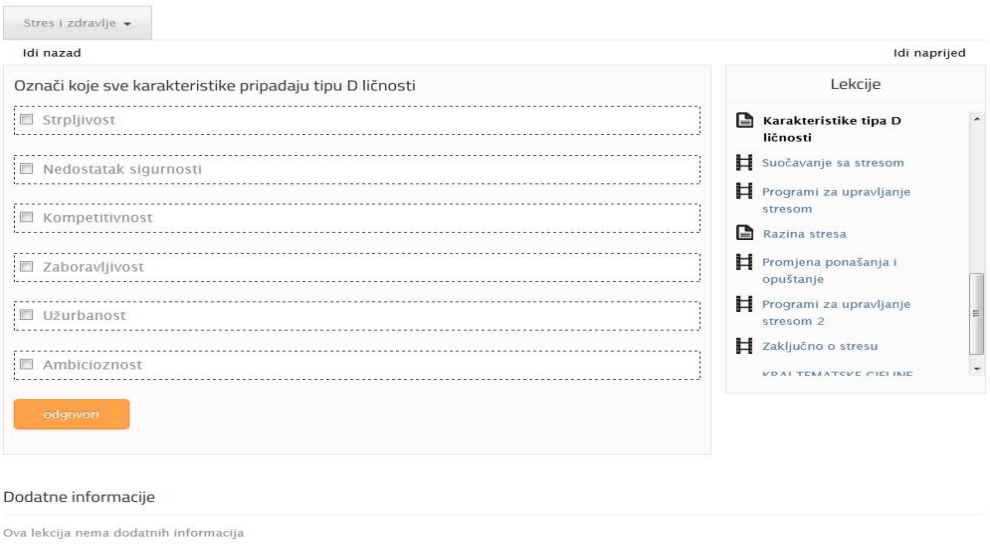


Figure 4. Answering questions inside virtual classroom

The course, which can be found on web address http://www.buraznanja.uniri.hr/courses/uvod-u-psihologiju, consists of a total of 12 thematic sections and 223 lessons, which makes 347 minutes of video material. By the 15th of January 2014, when the data were analyzed, the course was enrolled by 849 participants. Fig. 5 shows a huge drop in the number of video lessons watched after the first thematic section. The reason probably lies in the fact that a large number of people enrolled this course out of curiosity rather than interest in the material.

Fig. 6 shows times at which the participants were watching video lessons. This supports the idea that people need flexibility in organizing their learning, and if the flexibility is given, they use it.

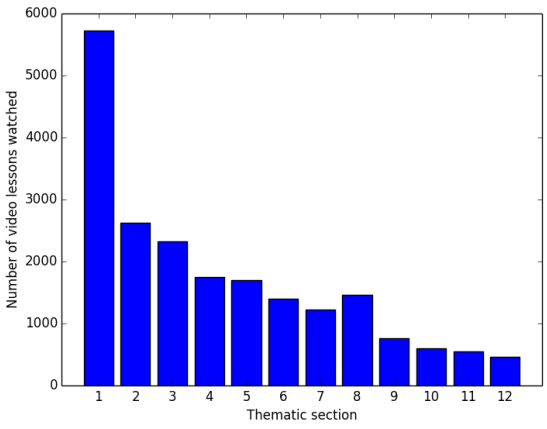


Figure 5. Video lectures watched inside each thematic section

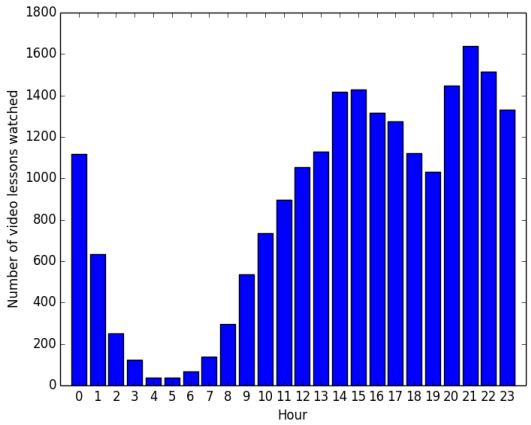


Figure 6. Video lectures watched in 24 hour day

V. CONCLUSION

The growth of popularity of MOOC courses in the world opens door to new research in various domains of science, from pedagogy and education to computer science, and based on the data created within a variety of learning systems. With new technologies evolving every day, teachers are faced with new challenges. For students, the biggest contribution of this type of courses is the possibility to watch lectures all over again whenever they want. The biggest benefit for teachers is that they can spend more time on practical work and research with their students. They can focus more on directing social communication, identifying students capable of mentoring, and identifying and selecting the most successful method of learning.

Using online multimedial learning and developing analytical tools, we are able to personalize and customize learning for each student individually. We can raise student motivation and student achievements by providing immediate feedback while learning (lecture questions) and by visualizing complex information. Our future work will focus on identifying the impact of combining multimedia courses with on campus teaching, and developing tools for estimating student motivation in an online environment.

ACKNOWLEDGMENT

We thank Marina Skocigoric for providing great illustrations for the course content.

REFERENCES

[1] H. Kelly, M. Ponton and A. Rovai, "A comparison of student evaluations of teaching between online and face-to-face courses", Journal of Internet and Higher Education, 10(2) pp. 89-101, 2007

[2] Department of Informatics, University of Rijeka. Retrieved from: http://www.inf.uniri.hr/index.php?option=com_classes&Itemid=40&task=displayStudy&studyid=109

[3] R. Papp, "Critical success factors for distance learning", Americas Conference on Information Systems, Long Beach, CA, USA., 2000

[4] A. Payne and L.K. Stergioulas, "Technology Enhanced Learning in the Economic Crisis: How Learning Technologies Assist in Bridging the Budget Gap and Overcoming the Crisis", Proceedings of IFIP Working Conference "Addressing education challenges: the role of ICT", Manchester Metropolitan University, Manchester UK, July 2012.

[5] R. Newton, "Staff attitudes to the development and delivery of e-learning", New Library World, 104 (1193), pp. 412-25, 2003.

[6] M.J. Broad, M. Matthews and K. Shephard, "Audit and control of the use of the internet for learning and teaching: issues for stakeholders in higher education", Managerial Auditing Journal, 18 (3), pp. 244-53, 2003.

[7] G. Suddaby and J. Milne, "Coordinated, collaborative and coherent; Developing and implementing e-learning guidelines within a national tertiaryeducation system", Campus-Wide Information Systems, 25(2) , pp.114-122, 2008.

[8] J. Bell,"Feature articles E-learning: your flexible development friend?", Development and learning in organizations, 21(6) pp. 7-9, 2007

[9] T. Volery and D. Lord "Critical success factors in online education', The International Journal of Educational Management, 14(5), pp.216-223, 2000.

[10] C. Bonk, "The World is Open for a Reason", 2009, Retrieved from: http://www.usu.edu/ust/pdf/2009/september/itn09010938.pdf

[11] M. Gabel, "MOOCs Massive Open Online Courses", EUA Occasional papers, January 2013.

[12] G. Siemens, "Connectivism: A learning theory for the digital age." 2004.

[13] 1st International Conference on Learning Analytics and Knowledge, Banff, Alberta, February 27–March 2011, Retrieved from: <https://tekri .athabascau.ca/analytics/>.

[14] M. Bienkowski, M. Feng and B. Means, "Enhancing Teaching and Learning Through Educational Data Mining and Learning Analytics", An Issue Brief. U.S. Department of Education, 2012.

[15] G. Siemens and P. Long, "Penetrating the Fog: Analytics in learning and education" EDUCAUSE Review, vol. 46, no. 4 (July/August 2011).

[16] S. Downes. "New technology supporting informal learning" Journal of Emerging Technologies in Web Intelligence, pp. 27-33, 2010

[17] L. Lane, "Three kinds of MOOCs", Retrieved from: http://lisahistory.net/wordpress/2012/08/three-kinds-of-moocs/

[18] E. Ries, "The lean startup", 5th ed., Crown Publishing Group, 2011.

[19] D.G. Glance, M. Forsey and M. Riley, "The pedagogical foundation of massive open online courses", First Monday, May 2013

[20] R.D. Balaji, B. Ramniklal, N.Balasupramanian and R.Malathi, „Mnemonics for Higher Education Using Contemporary Technologies", Second International Conference of the Omani Society for Educational Society, 2013.

[21] Online multimedial course Introduction to psychology, http://www.buraznanja.uniri.hr/courses/uvod-u-psihologiju

Are they a perfect match?
Analysis of usage of author suggested keywords, IEEE terms and social tags

Dina Vrkić
University of Zagreb, Faculty of Electrical Engineering and Computing /
Central Library,
Unska 3, 10000 Zagreb, Croatia
dina.vrkic@gmail.com

Abstract - The majority of the scientific papers include keywords besides the obligatory title and abstract. The use of keywords is not just for the description of content, but it is a viral part of the scientific paper which is later used for information retrieval function. The aim of this paper is to present comparison of author suggested keywords in scientific papers of Faculty of Electrical Engineering and Computing (FER), IEEE terms and user tags in academic social network Mendeley.

This paper will examine the scientific papers represented in IEEE Journals and Magazines indexed in the IEEE/IET Electronic Library reviewing the overlap between author keywords and IEEE terms from a controlled vocabulary (IEEE thesaurus) and user tags in Mendeley. The results showed no significant use of controlled vocabulary and social tags, which leads to the conclusion that the authors are not aware of the importance of controlled keywords.

I. INTRODUCTION

The documentary (or indexing) language is the conventional language used by the information unit to describe the contents of documents with a view to the storage and retrieval of information [1].

Their function is shortening and simplifying natural language which controlled in this way enables a quick and successful search. This kind of example is provided by a thesaurus.

Thesaurus is a controlled and dynamic documentary language of semantically and generically connected terms which cover a particular field of knowledge. They usually consist of descriptors (standardized keywords) and non-descriptors (prohibited terms, for example, synonyms which usually stand with reference to the preferred term). Thesauruses are usually defined by the following relationships between terms: equivalence (synonymy), hierarchy (generic, partitive etc.) and associativity. According to the number of languages, there are monolingual and multilingual thesauruses; according to the level of analysis, thesauruses are divided into macro thesauruses and micro thesauruses. In its function, thesaurus is a tool for controlling the terminology which is used in the translation of natural language of a document, level of contents or user into a more limited "system language" (documentary language) [2].

The main users of a thesaurus are primarily professionals (indexers), and then users. Searching only by the title, author or keywords (freely formed) is not enough, but it is also necessary to connect the field of research to what is being searched in a certain way.

This paper will present comparison of author suggested keywords in scientific papers of Faculty of Electrical Engineering and Computing (FER), IEEE terms and user tags in academic social network Mendeley.

II. IEEE THESAURUS

IEEE thesaurus is a controlled vocabulary with more than 9600 descriptors for engineering and scientific terminology, as well as IEEE-specific society terms. The thesaurus represents a terminology and oral use of terms drawn from IEEE publications (transactions, journal articles, conference papers and standards). The basis of language was a natural language, English language from the speaking area of the USA with cross-references of the British spelling varieties. The scope and structure of the Thesaurus reflect the engineering, that is, technical and scientific discipline which are consisted of the societies, councils and communities of IEEE, together with the technologies which IEEE serves. Generally, the IEEE Thesaurus is a very simple controlled vocabulary consisted of descriptors for favorable terms and technical terms.

Descriptors are used for indexing within the IEEE/IET Electronic Library (IEL) (popularly IEEE Xplore Digital Library), subject analysis, search and inclusion in the publishing business through keywords which are considered preferred terms for forming keywords [3]. Since IEEE is also a publisher, journals and conferences which are under its jurisdiction are generally required to include the controlled vocabulary, and in this way include descriptors instead of freely formed keywords [4].

Thesaurus construction is based on the ANSI/NISO Z39.19-2005 standard, Guidelines for the Construction,

Format, and Management of Monolingual Controlled Vocabulary.. Each descriptor included in the thesaurus represents a single concept or unit of thought. The descriptors are considered the preferred terms for use in describing IEEE content. The scope of descriptors is based on the material presented in IEEE journals, conference papers, standards, and/or IEEE organizational material. A controlled vocabulary is a specific terminology used in a consistent and controlled fashion that results in better information searching and retrieval [3].

Each descriptor included in the thesaurus represents a single concept or unit of thought. The descriptors are considered the preferred terms for use in describing IEEE content. The scope of descriptors is based on the material presented in IEEE journals, conference papers, standards, and/or IEEE organizational material. A controlled vocabulary is a specific terminology used in a consistent and controlled fashion that results in better information searching and retrieval [3].

Taxonomy is a strictly hierarchically structured list of terms in which all the conditions are correlated and all the cross-references are excluded. Each term in taxonomy is in one or more parent-child relations to other concepts in the taxonomy.

IEEE taxonomy contains the first three hierarchical levels below any broader concept (branch). In this document, conceptual relations are presented in alphabetical order and highlighted in bold. Every parent term does not have more than tri sub terms with levels marked by dots (in groups of four dots) of previous conditions to the next level. Terms can appear in more than one hierarchical branch in a parent-child relation. IEEE taxonomy is defined in such a way that it is always a subset of the IEEE Thesaurus [5].

Taxonomy is used when it is necessary to review the list of acceptable terms, but a thesaurus is more convenient when it needs to be decided which term to use. Professionals will use a thesaurus because an alphabetical list of terms is enriched with descriptor relations which enhance the detection of terms (broader and narrower terms, related terms and references). Furthermore, taxonomy will be used as a reference document when the authors will ask for a list of keywords with which the authors describe their works or any situations when they will need an IEEE list of words.

IEEE Explore uses the IEEE Thesaurus for indexing articles, but not only those of IEEE publishers, but all the papers which become included in the database itself. Today IEEE publishes more than a quarter of all the publications related to electrical engineering and computer science. Through its scientific and technical publications, conferences and IEEE norms it publishes over 30 percent of the world's literature in electrical engineering and computer science and organizes over 300 major conferences annually and has more than 800 existing norms with another 700 in preparation.

The most famous of the IEEE guidelines for helping editors and authors in publishing papers is "IEEE Author Digital Toolbox" [6]. It contains valuable information, guidance and assistance in the preparation of an article and its submission to the editor, as well as in reviews and ordering reprints. It is enriched with frequently asked questions related to the topic of scientific communication. The part about keywords is particularly highlighted. Authors are encouraged to use Instructions for Authors with the use of keywords from the list of IEEE taxonomy, stressing that if the author does not find the appropriate term, he/she should add his/her own within the paper. This is ultimately another method to enter new words into controlled vocabulary. It has been noted: Authors are encouraged to select required keywords from the list of "recommended keywords for authors." They contain the first three hierarchical levels below which there are terms in parent-child relations of the most frequently used terms from the IEEE Thesaurus. If you cannot find the appropriate terms, you can always add your own [6].

The list of keywords (or thesaurus) is available as auto retrievable ASCII file which is sent to e-mail keywords@ieee.org. In a reply to e-mail an alphabetical list of approved IEEE index terms is sent. The latest version supersedes the previous one and in this way changes all the previous versions, which makes only the last one up-to-date and usable. If an outdated list is used, it will be harder to search the works [7].

III. BACKGROUND OF THESAURUS AND KEYWORDS

Keywords are listed after the abstract, but in the language of the summary [8]. They are of particular importance for detecting topic and focus of the paper. They reflect the content and the method described in the paper. The content is mainly described by nouns, and the method and procedure by nouns and verbs. Basic keywords represent important concepts and thus completely describe what the paper is about. In general, the author can choose keywords himself, but in some cases there are needs for choosing keywords with the use of a glossary or a particular thesaurus [8]. In order to successfully communicate with the information database, the user needs to formulate the aspect concepts using symbols – words, phrases and codes, as in communication with other people. If natural language is used as a language for finding information, the existence of homonyms and synonyms will increase problems in online search [8].

Therefore, keyword or term lists are essentially stripped-down versions of subject heading lists with little or no synonym control and often lacking cross-references; they generally do more harm than good for indexing purposes [9]. Subject heading is a word or phrase from a controlled vocabulary which is used to describe the subject of a document or a class of documents. Subject headings may be represented in individual bibliographical records [9].

Articles which include descriptors help to identify and summarize the importance of the work. Also IEEE stated that: Electronic searches for articles are often more productive when authors prove ide keywords (or "indexing terms") in addition to the abstract to locate articles.

- Concepts inherent (but unstated) in an article should normally be indexed unless they are so rudimentary as to be presumed by any novice looking for informal ton. An author should use judgment, but it is always better to err on the side of inclusion rather than exclusion.
- IEEE policy - and good indexing practice: assigns the most specific index in keywords available to describe the major points or topics covered in the art ice.
- Choosing keywords from the current IEEE Indexing list promotes standardized indexes [7].

A thesaurus provides access to words which are used in bibliographic records which have derived from a controlled vocabulary. However, there are also disadvantages. In this way they cannot compensate for the complexity of natural language and accordingly, expression, while the context cannot be recovered.

Uncontrolled languages give the user more possibilities with using them, and do not require precision for data input for the search, but have a higher probability of errors in the information output [7].

Indexing, or in this case, user labeling, is a popular method for users to allocate their own freely chosen words (tags) regardless of the fact whether or not they are in the controlled vocabulary [7]. That is why we speak about freely formed words. On the other hand, labeling has its advantage because it is free of charge and allows the user to connect the term with some terms which are no longer in controlled vocabularies (for example, the term *grid* as an old term and the term *cloud computing* as a contemporary one). IEEE Explore does not allow readers to tag articles. Therefore, we have introduced a new component into our research which allows such action – Mendeley, software for managing references.

Mendeley is a free reference manager and academic social network that can help scientists to organize your research, collaborate with others online, and discover the latest research [10]. Why Mendeley? Because altmetrics is popular today, which has dual function: scientific communication and a new approach to scientific measurability, associated for instance in Scopus, but mostly because of search in Mendeley which is indexed by Google as well.

Connecting author supplied keywords to the tags, Kipp mentioned [11], [12] adapted Voorbij's[13] thesaural analysis method to compare tags, author keywords and descriptors assigned to LIS articles tagged on CiteULike. Tags were more likely to match author keywords exactly (33%) than descriptors (16%), but author keywords were as likely to match exactly or be related terms of descriptors (19%) [12].

Thomas et al. [14] adapted identical thesaural comparison for books tagged in LibraryThing with associated Library of Congress Subject Headings finding that 6% of tags matched LCSH exactly while 31% matched thesaural categories and 35% were related to LCSH but not subject headings.

According to a notable number of recent studies carried out in the area of medicine, social and humanities sciences it is noticed a significant lack of literature and examples in area of technical sciences.

IV. RESEARCH METHODOLOGY

The research question is: do the users of the Faculty of Electrical Engineering and Computing use a recommended controlled vocabulary, prescribed by IEEE journals, while forming keywords in an article?

IEEE Xplore Digital Library is one of the most important databases of scientific and technical articles in the world, which covers a third of the world's literature in electrical engineering, computing, information and communication technologies, as well as many other interdisciplinary areas. By accessing the database you have the access to complete texts (PDF) for more than 3 million documents (2 million authors, 170 journals, 900 conferences, 2000 IEEE standards and 25000 new scientific articles per month).

After searching the database and extracting data, the collected results will be subjected to the method of comparing the keywords with IEEE controlled vocabulary and Mendeley (freely formed) tags.

On January 3, 2014 the data were collected from IEEE Xplore Digital Library with advanced search strategy in which only metadata elements were selected, Author Affiliation: Zagreb, Publisher: IEEE, Type of publications: Journals & Magazines, Publication year spam: from 1996 to 2014. Since the first year from the literature was 1995, in which for the first time instructions for authors and controlled vocabulary were mentioned, the year 1996 was taken as the first complete year which could go through implementation and indexing.

The results were later narrowed by Author Affiliation facet and forms of the name for the Faculty of Electrical Engineering and Computing of the University of Zagreb, which was represented less than 34 different terms. The forms which contained the term University of Zagreb in 4 versions were not taken into consideration.

The problem of diverse forms for a particular institution causes a difficult search of institutional production if the right manner of stating the name of an institution in which the scientist works is not properly regulated. In 2012 the Faculty of Electrical Engineering and Computing published an internal document "Specifying the name of the University and the Faculty within the papers of the FER employees" in Croatian and English. Since this provision was introduced in May 2012, we had to take into consideration the probability that a very small number of papers would be listed in the following order. Therefore, the decision was made that all forms of the name which were used by then would be extracted. All journal articles were coded for an easier tabular display.

V. RESULTS

The result of the search was 76 articles ranked according to the age from the oldest to the newest. Only the ones which fulfilled the conditions that articles contain keywords listed by the authors and IEEE index terms in the database IEEE Explore were taken into consideration. The final number of articles for comparison was two thirds of the total, that is, 53 articles.

A problem with reaching authors' keywords was perceived because they were not listed in the bibliographic record. Therefore, any paper which did not have the elements had to be opened and imported into the table. It was found that 9 articles did not have keywords and 14 articles did not have assigned IEEE terms in the database. In addition, the presence of social marks (tags) was verified on the network/software for managing Mendeley references.

The scientists predominantly focused on keywords taken from the titles of their papers. They used terms which were either too important or not important enough. A low number of keywords predominated. According to the instructions, a recommended number of keywords are up to 10 words, in an alphabetical order, separated by commas.

During the period between 1996 and 2009 IEEE Xplore listed 10 IEEE terms (in 31 papers), but from 2009 the number of terms has significantly varied: from 5 to 14 terms. The alphabetical order of keywords was noticed in 39 works, that is, three quarters. In the analysis only those terms which are almost identical were taken into consideration (except for tilde). The terms containing abbreviations with the full name were eliminated, as well as abbreviations as such, together with the relationship singular-plural in terms (for example authors' words) flexible manufacturing systems (FMSs) -> flexible manufacturing systems (IEEE terms); e-learning -> electronic learning (IEEE terms). It was noted that there were 25 papers with the terminology which in a small part coincides with the suggested IEEE keywords.

TABLE I. THE NUMBER OF KEYWORDS USED IN ARTICLES

Number of keywords in article	Number of articles with keywords
3	4
4	9
5	22
6	9
7	3
8	3
9	3

Thus we see how the authors predominantly decided for 5 keywords to describe the paper, 9 papers use 4 or 6 keywords, while a smaller number of papers contains 4 or more than 6 keywords.

TABLE II. COINCIDENCE OF THE NUMBER OF TERMS IN THE PAPERS WITH IEEE TERMS IN THE DATABASE

N of matching terms	N of articles with identical terms
1	17
2	6
3	2

By the number of terms that coincide with the authors' keywords and allocated IEEE index terms, we can conclude that the authors did not use *Authors instructions*, that is, IEEE taxonomy.

TABLE III. RELATED TAGS FROM MENDELEY AND IEEE TERMS

N terms	N articles with related M tags & I terms
1	3
2	1

Correlation of (freely formed) tags in Mendeley with IEEE terms does not show correlation to a greater number of terms, due to the fact that we can compare it only in two cases; with one or two terms.

TABLE IV. DISPLAY OF CORRELATION BETWEEN TAGS FROM MENDELEY AND IEEE TERMS (E.G. ARTICLE UNDER CODE N.11)

Author suggested keywords	IEEE terms	Mendeley Tags
losed-loop control hybrid stepper motor microcomputer implementation of optimal control algorithms optimal control algorithms	Acceleration Choppers Microcomputers Micromotors Motor drives Open loop systems Optimal control Permanent magnet motors Rotors Torque control	acceleration average torque bipolar supply chopper amplifier choppers (circuits) closed-loop control closed-loop optimal control algorithms closed loop systems control design control performance

The tag *choppers (circuits)* is already marked with the term from a higher hierarchical level in order to eliminate the possibility that these are, for example, choppers as engines. The tags that appeared in the research are mostly tags which mean something only to the authors who created them, while they are meaningless to a wider audience, for example: he, phaseread00, folder 5, hjacs, inscope, tech, etc. The presence of tags was unexpectedly below the expected number. A greater presence was expected due to the distribution of the net users, but also because of the concept of open access which enhances authors' visibility. Therefore, a well denoted document provides easy search ability which provides good visibility.

TABLE V. CORRELATION BETWEEN AUTHORS' KEYWORDS (AKW) WITH THE TAGS FROM MENDELEY

N terms	N articles with related AKW & M tags
1	2

The correlation of authors keywords with freely formed tags is special because it is the equal level of indexing, thus we speak about freely formed words. The result is realistic because we need to have in mind that the

user denotes the text according to the choice of the keywords which he or she knows the best.

TABLE VI. DISPLAY OF CORRELATION OF AUTHORS' KEYWORDS TO TAGS FROM MENDELEY (E.G. ARTICLE UNDER CODE N.32)

Author suggested keywords	IEEE terms	Mendeley Tags
Constrained systems constrained time-optimal control (CTOC) control invariant set discrete-time (DT) dynamic programming electronic throttle friction limp-home (LH) nonlinearity piecewise affine (PWA) systems	Combustion Computational modeling Control systems Electric variables control Engines Friction Optimal control Sampling methods Table lookup Valves	electronic throttle model predictive control

TABLE VII. AUTHORS' KEYWORDS CONNECTED WITH THE SAME IN MENDELEY (M-ASK) WITH IEEE TERMS

N terms	N articles with related AKW, IEEE terms & M-ask
1	2
2	1

The correlation of author suggested keywords to the ones from Mendeley (M-ASK) with the IEEE terms in the database IEEE Xplore show that a small number of authors have created words using a controlled vocabulary. This relation is an example of how you should, when creating a particular paper, already have in mind the use of a controlled vocabulary because process of taking a bibliographic record (either from the database or manual entry) requires keywords. If they have already been well formed, the possibility of better visibility is also improved in Mendeley. Although the term permanent-magnet motors was specified, it was decided to ignore the tilde between terms, in order to show better correlation between different terms.

TABLE VIII. DISPLAY OF CORRELATION OF KEYWORDS AND THOSE FROM MENDELEY WITH IEEE TERMS (E.G. ARTICLE UNDER CODE N.30)

Author suggested keywords	IEEE terms	Mendeley Tags	M-ASK
Air-gap permeance magnetic fields permanent-magnet motors	Air gaps Finite element methods Forging Geometry Magnetic analysis Magnetic fields Magnetization Permanent magnet motors Surface waves Torque	Air-gap permeance air gaps conformal transformation electromotive forcefinite element analysis finite-element simulations flux density Folder - zarko magnetic field distribution magnetic field modelling	Air-gap permeance magnetic fields permanent-magnet motors

TABLE IX. CORRELATED AUTHOR SUGGESTED KEYWORDS, TAGS AND AUTHOR SUGGESTED KEYWORDS IN MENDELEY WITH IEEE TERMS

N terms	N articles with related AKW, M tags, M-ASK & IEEE terms
1	2

The correlation of author suggested keywords, tags and author suggested keywords in Mendeley is an example of a well denoted article. If the users perceive in M-ASK how a controlled term is specified they can, what they most frequently do, rewrite it or enter the tags.

TABLE X. DISPLAY OF CORRELATION OF AUTHOR SUGGESTED KEYWORDS, TAGS AND AUTHOR KEYWORDS IN MENDELEY (E.G. ARTICLE UNDER CODE N.30)

Author suggested keywords	IEEE terms	Mendeley Tags	M-ASK
Air-gap permeance magnetic fields permanent-magnet motors	Air gaps Finite element methods Forging Geometry Magnetic analysis Magnetic fields Magnetization Permanent magnet motors Surface waves Torque	Air-gap permeance air gaps conformal transformation electromotive forcefinite element analysis finite-element simulations flux density Folder - zarko magnetic field distribution magnetic field modelling	Air-gap permeance magnetic fields permanent-magnet motors

TABLE XI. DISPLAY OF CORRELATION OF AUTHOR SUGGESTED KEYWORDS, IEEE TERMS, TAGS AND AUTHOR KEYWORDS IN MENDELEY (E.G. ARTICLE UNDER CODE N.30)

Author suggested keywords	IEEE terms	Mendeley Tags	M-ASK
Air-gap permeance magnetic fields permanent-magnet motors	Air gaps Finite element methods Forging Geometry Magnetic analysis Magnetic fields Magnetization Permanent magnet motors Surface waves Torque	Air-gap permeance air gaps conformal transformation electromotive forcefinite element analysis finite-element simulations flux density Folder - zarko magnetic field distribution magnetic field modelling	Air-gap permeance magnetic fields permanent-magnet motors

By the connectivity of author keywords, IEEE terms, tags and author keywords in Mendeley we can conclude that there is a circle between the way of connecting them, which shows that there is equal presence of freely formed terms – tags and M-ASK. User tagging, with its lower apparent cost of production, could provide additional access points to traditional controlled vocabularies and provide users with the associative classifications necessary to tie documents and articles to time and task relationships as well as other associations which are new and novel [11].

The correlation between author keywords, IEEE terms, tags and Mendeley author keywords (M-ASK) is not present because there is not a single example of such connection. Kipp examined author keywords in comparison to tags and subject headings using a modification of Voorbij's categories. She found that author keywords were more similar to tags in terms of term choice, but that there was still a high degree of overlap between tags, author keywords and subject headings when related term (RT) matches were taken into account [15].

As a reason for a small number of this we need to take into consideration that: some papers perhaps did not have author instructions in a form as they are now, editorships

(obviously) do not pay attention to forming keywords. But above all, the scientists themselves may not be aware of the advantage of using a thesaurus, by the Association and the library, natural language in technical sciences is suitable for abbreviations which scientists have relied on, but which do not match a controlled vocabulary – the terms are being resolved.

Regarding the possibility of outdated terminology, it can be clearly found in the IEEE Thesaurus, but the question is to which extent the authors will use it as it has already been noted in the paper that thesaurus is used by informational professionals, while the authors use taxonomy.

VI. CONCLUSION

Since IEEE continuously follows technological trends and in this way wants to maintain a quality organization, that is, engineering sciences and profession, it is recommended to use a controlled vocabulary. Since the results showed that the scientists do not use the proposed controlled vocabulary, it will be necessary to establish a connection between informational professionals, in this case a higher education library, which will raise awareness of the benefits of using the vocabulary and of the scientists themselves. Raising awareness of the use of a controlled vocabulary and better indexing for better searching, as shown in Mendeley, certainly has its benefits, especially when it comes to science where terminology is continuously changing.

We have shown how scientists adhere to the concept of forming keywords, but do not use the IEEE controlled vocabulary. Therefore, we can say that there is a problem of awareness about the importance of implementing a controlled vocabulary in writing papers. Some concepts are quite narrow, while others are quite broad. We have also managed to find a few errors in the IEEE indexing (symmetric terms, typos). Mendeley failed to give the results we were hoping for, that is, that a larger number of freely formed tags would be used. There is a large number of papers in which by the use of metadata a great number of metadata was extracted, so there is not a great number of social tagging.

VII. REFERENCES

[1] C. Guinchat and M. Menou, General introduction to the techniques of information and documentation work. Paris: Union Nations Educational, Scientific and Cultural Organization, 1983.

[2] Rječnik odabranih pojmova. Available at http://dzs.ffzg.unizg.hr/text/Obavijest%20i%20znanje/oiz_rjecnik.html [2013-11-28].

[3] IEEE Thesaurus. IEEE. Available at http://www.ieee.org/publications_standards/publications/services/thesaurus2.html [2013-12-20].

[4] 2013 IEEE Thesaurus: Version 1.0. Available at http://www.ieee.org/ucm/groups/public/@ieee/@web/@org/@pubs/documents/file/20042208.pdf [2013-12-20].

[5] 2013 IEEE Taxonomy: Version 1.0. Available at http://www.ieee.org/documents/taxonomy_v101.pdf [2013-12-20].

[6] Author Digital Tool Box. IEEE. Available at http://www.ieee.org/publications_standards/publications/authors/authors_journals.html [2013-12-29].

[7] IEEE 1998 Keyword List. Available at http://code.ucsd.edu/keywords.html [2013-11-01].

[8] M. Gačić, Pisanje znanstvenih i stručnih radova. Zagreb: Sveučilište u Zagrebu, Učiteljski fakultet ; Školska knjiga, 2012.

[9] H. H. Wellisch, Indexing from A to Z, 2nd. Ed. New York: H. W. Wilson, 1995.

[10] Mendeley. Available at http://www.mendeley.com/ [2013-11-01]

[11] M. E. I. Kipp, "Complementary or Discrete Contexts in Online Indexing: A Comparison of User, Creator, and Intermediary Keywords," Canadian Journal of Information and Library Science, 29, 4, 419–436, 2005. Preprint available at http://eprints.rclis.org/8771/ [2013-12-01].

[12] M. E. I. Kipp, "User, Author and Professional Indexing in Context: An Exploration of Tagging Practices on CiteULike," Canadian Journal of Information and Library Science, 35(1), pp. 17–48, 2010. Avaliable at http://eprints.rclis.org/14867/1/kipp-2011-userauthorindexerlis-cjils351.pdf [2013-11-10].

[13] J. H. Voorbij, "Title keywords and subject descriptors: A comparison of subject search entries of books in the humanities and social sciences," Journal of Documentation, 54(4), pp. 466–76, 1998.

[14] M. Thomas, D. M. Caudle and C. M. Schmitz, "To tag or not to tag?," Library Hi Tech, 27(3), pp. 411–434, 2009.

[15] M. E. I. Kipp "Controlled vocabularies and tags: An analysis of research methods,". In Smiraglia, Richard P., ed. Proc. from North American Symposium on Knowledge Organization, Vol. 3. Toronto, Canada, 2001, pp. 23-32.

Lab assessments in undergraduate course in Compilers for students with no prior knowledge in assembly

Vesna Kirandziska, Mile Jovanov, Marija Mihova, Marjan Gusev
Faculty of Computer Science and Engineering
Ss. "Cyril and Methodious University" in Skopje
{vesna.kirandziska, mile.jovanov, marija.mihova, marjan.gushev }@finki.ukim.mk

***Abstract*—Compilers is an important course in the curricula of Computer Science Faculties. A problem that we have faced while organizing the course at our institution was lack of prior knowledge of some students in the area of assembly programming and microprocessors. Due to the fact that the number of classes for the course was limited, we were not able to teach the necessary material on assembly programming. In this paper we present our idea for the lab project to be done by students through the course, that should enable them to grasp all necessary concepts in Compilers without a need of assembly programming. The idea was tested a few of years in a row, and most of the students managed to finish the project, and achieved better results in the final exams.**

***Index Terms*—compilers, learning**

I. Introduction and related approaches

By definition, a compiler is a computer program by which a high-level programming language is converted into low-level programming language that can be acted upon by a computer. Hence, for someone to be able to make a compiler, she needs to know the high-level programming language, and also to have a great understanding of the low-level language.

The course Compilers or similar courses named Compiler construction, Compiler and interpreter, Compiler Practicum, Programming languages and Translators are all in the computer science program at many Universities. At our Faculty of Computer Science and Engineering in Skopje this course is elective course in the sixth semester, by the proposed program. It is a compulsory course for students enrolled in some of the modules like 'Computer science' and 'Software engineering'. Each year we have around 40-70 students enrolled in this course. It is of a significant importance how this course is taught to the students and whether they can achieve the course goals.

There are different experiences worldwide. At the University of Virginia [1] a big accent in the curriculum is given to code generation. The course project consists of two programming assignment. The first is building a compiler for COOL (the classroom object-oriented language which is a high-level programming language) and the second is optimizing it. COOL has the essential features of a realistic programming language, but is small and simple enough so that its compiler can be implemented in a few thousand lines of code. The course Compilers [2] held at Stanford University also uses the COOL language as an example for building parts of a compiler through five different programming assignments. Even more, an optional course project is to write a complete compiler for COOL in C++ or Java.

Other universities use part of Java programming language as the high level-language that has to be translated. For example, at the University of Texas[3] as a student assignment, each student writes a compiler for Pascal and code is generated for a real processor and run on hardware. Beside heavy programming skills, hardware knowledge is crucial for creating a program written in a low-level language. One of the four distinct assignments in the project at this university involves code generation. A major part of the course held at the University of Waterloo [4] consists of an implementation of a compiler for a simplified Java-like language called Joos1W. The course project represents 75% of the student grade.

At all mentioned universities the project assignment represents the great deal of the student evaluation. Building their own compiler enables students to get more deeply involved in the crucial parts of a compiler and it enables understanding it better through all challenges that came up while working on their programming assignment.

Other thing in common for all Compiler courses lectured at different universities is that the compiler built usually translates in some version of an assembly programming language. At Stanford [2] the MIPS (Microprocessor without Interlocked Pipeline Stages) assembly is used, while at the University of Waterloo[4] use the Netwide Assembler dialect that runs on he Intel x86 architecture. So, for building a translator into low-level language the knowledge of assembly language is essential.

All programs for Compilers at other universities discussed above require knowledge form Data Structures, Algorithms, Programming, Microprocessors and Formal Languages.

At the Faculty of Computer Science and Engineering in Skopje Structural and Object Oriented Programming are courses taught in the first student year (by the proposed schedule), Data structures, Algorithms and Formal Languages are in the second student year, but the course Microprocessors is an elective course in the third or forth student year. The

curriculum of Microprocessors includes assembly language and programming 8086 microprocessor. As a result, usually students enrolled in the Compiler course, had not been enrolled on the course Microprocessors previously. This was a problem that had to be considered.

In this paper we are going to present our solution on how to teach students Compiler with no previous microprocessor and more precisely assembly knowledge.

In the next sections our idea for organizing the Compilers classes as well as the rules for student evaluation are presented. Afterward, our approach for presenting code generation to our students is explained in more details. Next, the results taken from our four year experience in teaching Compilers are presented. At the end a conclusion is given.

II. Course curriculum

The topics of this course follows the book "Compiler Construction: A Practical Approach" [5] and its main objectives are students to get familiar with a compiler and also to learn how a compiler is build. Here the Inger language is defined and used as an example language for which the compiler is built.

'Compilers' weights 6 ECTS credits and is organized with 5 classes a week: 2 classes for lectures, 1 for auditorium exercises and 2 for laboratory exercises. The lectures continuously cover each part of a compiler construction. In the curriculum of the Compiler course all parts of a compiler are considered in details.

The language for which a compiler is made can be represented through syntax diagrams, Backus-Naur Form or Extended Backus-Naur form. These approaches for language representation are contained in the first lectures.

Following these lectures lexical analysis is introduced to students. The part of the compiler which perform lexical analysis consists of a lexer implemented as a finite automaton and a parser implemented as a push-down automaton. Beside checking if one program is lexically correctly written, a compiler could detect and report the type of error that is found in a program and even where the error is found. This is also included in the curriculum.

Next, the part of the compiler concerned with checking the semantics of the programed code is presented. Between many different kinds of semantic errors that can be found, most of them involve variables or symbol names. A symbol table is usually used as a tool in semantic analysis and is included in the course.

The last part of the lectures is about the code generation module of a compiler. This is actually what we are concerned about in this paper. The classes dedicated to this part of the compiler are given in two lectures. The dedicated time does not allow the professor to teach about this part of a compiler, and additionally to show the basics of an assembly language since our students usually do not have any background knowledge in microprocessors. Consequently, we used our approach presented in this paper to teach students about this part of a compiler.

While during the lectures we cover the topics and we use auditorium exercises for more practical examples, at the laboratory exercises students work on the practical lab assignments given. In the next subsection the student evaluation for this course is presented.

A. Students' grading

Students are graded by the midterm exam(20%), final exam (25%), homework(15%) and by the grade for the final project assignment (40%). Exams cover the theoretical knowledge in the Compiler construction. Homework is different in different years, and for 2 years we have included an original on-line collaborative activity named "Ontology" also developed at our insitution [6].

Here we present the work on the project which was constructing a compiler. The whole project consists of several separated tasks:

- **Task 1:** Make a syntax diagram for the language A
- **Task 2:** Program a lexical analyzer (lexer) for the programming language A
- **Task 3:** Program a parser for the programming language A
- **Task 4:** Make a complete syntax analyzer with error detection
- **Task 5:** Make a complete semantics analyzer with error detection
- **Task 6:** Generate code in language B from the program in language A
- **Task 7:** Test the compiler

Divided in teams with 3-5 students in each group, students should do the assigned tasks needed to do the final project. Their involvement in the project should be on weekly basis, by upgrading the compiler with the newly learned module introduced on the lectures. The laboratory exercises are consultation exercises in which students ask question about their project and advance to its construction on regular bases.

Working in teams allows some students to work more then others and be more involved in the project, while others can be not involved at all. This arises the problem of how to grade students. One solution was giving the students a chance to grade by them selves their activity in the project. After grading the project itself, the students should all get the appropriate part of the points based on their involvement. Another approach is the teacher to grade each student in the group from an oral presentation about his/hers involvement in the project. Both approaches were used in the Compiler course in different years.

Because different students have different possibilities, background knowledge and ambitions, several variations of the final project were given. The hardest project scored 100 points, while the easiest one 65 points. In this way the projects ware leveled to the student knowledge. Students were given a possibility to choose the project level they think that suits them. Afterward, teams were formed out of the survey results.

In the next section the project assignment is defined in more details.

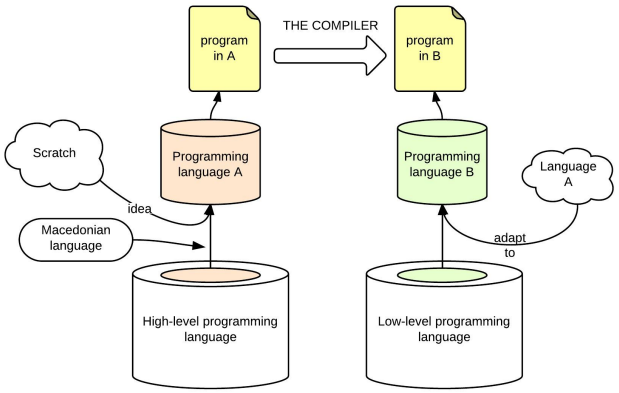


Fig. 1. A schematic view of languages A and B

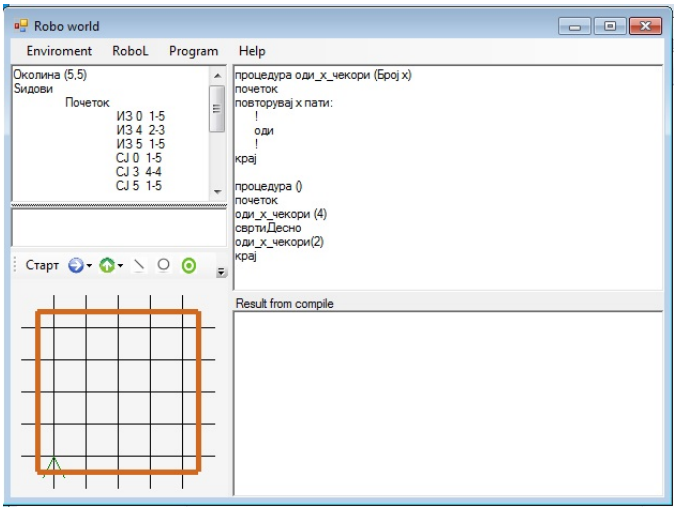


Fig. 2. A shell application for the compiler

B. Project Assignment

The project was stated as follows: Make a compiler for one given programming language A and translate it to another assembly-like programming language B.

The languages A an B are new languages that were introduced in our course. They are created just for this specific purpose. In all years the course was held, new (similar but different) languages A and B were introduced. On 1 a schematic view of these languages compared to existing programming languages is shown. As shown on the figure the programming language A is based on the procedural high-level programming languages like C, Basic or Pascal. A is not an object-oriented language. It includes the three most basic statements and other basic features. The key words in the language A are written in Macedonian language and so it differs form all other languages.

The inspiration for this language comes from turtle-based programming languages used for beginners in programming languages. Indeed, the elementary statements of the languages are actually commands for the movement of the specific turtle-object like a frog, robot, bird and so on. We have used all these objects in our course. The interesting content of the language and its statements made the project more interesting to students. As an example, a robot's program should move him in a specified environment and it must be careful not to hit the walls. Another example is a frog's program. This program should enable a frog to jump form one field to another considering water in the sense it should not step on it. Using these languages as introductory languages for young students that show interest in informatics, gives great results and we have included them in the strategy for education in informatics in our country [7].

As stated, these programs should be executed in a specific environment. A language for environment definition was used to define the environment. But, program execution is not in the scope of the course Compilers because here an interpreter of the language is needed.

The programming language B as explained on Fig. 1 is based on assembly, but has only the most basic statements needed for the translation to be possible. All hardware knowledge for the execution of a program written in this language is excluded.

In the rest of the paper we are going to explain in details only one of the project assignments (given in 2009). The language A defined was called RoboL (extension - rl) and was specialized for moving a robot in its environment. For the project assignment the students were given a shell application. It is shown on Fig. 2. This is just the interface of the compiler students should do. In this application a compiler for the language should be build in. The application is build in Microsoft Visual Studio. In the lower left corner a picture for the robot environment is given. Bold lines represent walls. On the right side a program written in RoboL is shown and on the left side a program for environment definition is given. The menu of this application should be implemented by students.

In the example project RoboL was translated into RIMAL (Robo Imaginary Meta Assembly Language). More about these languages will be explained in the next sections. The result from the compilation is the program translated in the RIMAL language (extension - rml) if the program is correct. But if the program is not correct, students should output messages for all errors found in the program code. These messages should be displayed in the lower right part of the application window.

III. The Languages for the Project

Next, the languages used for building a compiler are introduced.

A. The high-level language

RoboL poses the three simple, hierarchical program flow structures: sequence, selection, and repetition. There are 3 basic statements which are direct commands for the robot movement: go straight (one step), turn left or turn right.

```
процедура ()
почеток
променлива к,бр
к=4
бр=0
ако ознака
!
  бр=1
!
повторувај к пати:
!
    оди
    ако ознака
    !
      бр=бр+1
    !
!
                          example.rl
```

```
свртиДесно
повторувај до (бр==0):
!
    оди
    ако ознака
    !
      бр=бр-1
    !
!
крај
```

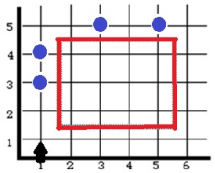


Fig. 3. An example program in RoboL and an environment in which it should be executed

Selection structure is realized trough the if-like statements. While-like repetition is not implemented, but a variant of do-while statements and for-statements is present in the language. There are five different variants of repetition (loop) statements.

In the language only integer and a kind of character types can be declared. The character variables are specific variables that can be considered as an enumeration. Indeed, these variables contain information about the direction of the robot position that can be South-S, North-N, West-W and East-E. As a result, arithmetic operators have no meaning with these character variables. Integer variables can be only positive. Arithmetic operators for addition, subtraction and multiplication can be used with integer variables and constant values. For logical expressions relational operators for comparison are used. These can be used as conditions in selection and repetition statements.

Other conditions used in these statements address the environment in which the robot is. The robot environment has walls (red lines) and marks (blue lines). In each moment a robot is in one point that is an intersection of one vertical and one horizontal line. Also the robot heads in one of the four directions (S,N,W,E). Considering this, other conditions implemented in the language are based on:

- **Wall**: If the robot is exactly in front of a wall the expression will be true
- **Mark**: If the robot is on a mark the expression will be true.
- **Some direction (S,N,W or E)**: If the robot is pointed in the stated direction the expression will be true

RoboL supports the concept for procedures. Functions and return statements are not supported.

The definition of the language was given to the students with a written tutorial. Example programs in this language were also available.

B. The low-level language

As stated in the Introduction, because our students in general have no knowledge in assembly as low-level language we made an artificial language RIMAL simulating a real assembly.

Robo Imaginary Meta Assembly Language (RIMAL) is a simulation of a simple assembly. The idea of this language is not to be used for program execution but to be used for educating students. The code written in this assembly-like language can not be executed by some hardware. This language has the most basic features of an assembly. Indeed, the language has similar instructions as the assembly language. But, some of the instructions are omitted from this language and some are added to it. The final product of student project beside giving error messages for the syntax and semantic errors, is the code generated in RIMAL. By making this code students learn and practice how code generation is implemented in a compiler.

As in assembly 8086, in RIMAL registers, variables and stack can be used for storing data. Because in the high-level language RoboL arrays are not implemented other memory are not used in RIMAL. 3 registers can be used:

- regN - for storing integer data
- regD - for storing character data
- regC - for constant values

As shown, all registers have a specific usage and should be used for the appropriate data. Note, that although logically there are registers in the language, there are no actual registers. Other important similarity to assembly language is the usage of labels to label code lines. In RIMAL also, any string different then the key words can be labels. After the label : (colon) should be written.

Instruction type	Instructions	Assembly
Move instructions	move	modified
Arithmetic instructions	inc add sub mul	✓
Function instructions	call, ret	✗
Comparator instructions	cmp	✓
Robot movement instructions	rl (go left) rr (go right) go (go)	✗
Jump instructions	je, jne jl, jle jm, jme	✓
Stack instructions	push, pop	modified

TABLE I
RIMAL INSTRUCTION SET

In Table I the instruction set of the newly defined language is represented. For each instruction type the instructions are given. Also, a comparison with the assembly language is roughly given. For example, move instructions which are used for coping data from one location to another is modified with respect to the mov instruction in assembly. Beside the instruction name, the operands can only be: variables and registers. Similarly, the stack instructions are modified so that the instruction 'push' pushes the value in the register regN on the stack.

Arithmetic, comparison and jump instructions are similar with the instructions in assembly language. There is a difference only in the operand values in these instructions.

```
odi_do_x_oznaki:
data x
mov regN,x
start:
go
cmp reC,$M
jne next
sub regN, 1
next:
cmp regN,0
jne
ret
main:
data k
data br
move k, 4
move br, 0
cmp reC,$M
jne next
move br, 1
next:
```

```
move regN,0
start:
go
cmp reC,$M
jne next
push
move regN, br
add regN, 1
move br, regN
pop
next:
cmp regN,k
jne
rr
call
odi_do_x_oznaki(br)
ret
```

Fig. 4. An example program translated in RIMAL

Call instructions are different then in assembly language. The call instruction has one operand which is the label where the function is written in the code. But in assembly the operand is the address where the first instruction in the function is written. When calling a function in assembly the IP (Instruction pointer) register is pushed on the stack and the address of the new instruction is put in the IP. But, here the knowledge for the instruction pointer and also the instruction stack is omitted. In in RIMAL the function definition ends with the key word ret. After the code in the function is executed the program continues after the function call. This approach allows does not allow recursion functions.

The RIMAL language is enriched with three robot movement instructions. Because in the RoboL language there are instructions for moving the robot straight ahead, left or right by one step, RIMAL also has some version of these instructions. These are the instructions go, left and right, accordingly.

As a conclusion RIMAL has a very similar syntax with assembly. The valuable feature of RIMAL is that students are not concerned with the semantic of its instructions and their execution. This help students to see this language just like another (non high-level) programming language. Computer hardware on which the low-level language should be executed is not known. This makes the problem of understanding the language easier, but also gives students a little sense of what an assembly language looks like.

In Fig. 4 an example code translated in RIMAL is given.

IV. Results from the lab project assignment in the Compiler course

With the laboratory project given, successful student project included more then 5000 lines of programming code for building a compiler. By doing the project students went through the whole process of making a compiler by which they practically learned what a compiler is and how does it work.

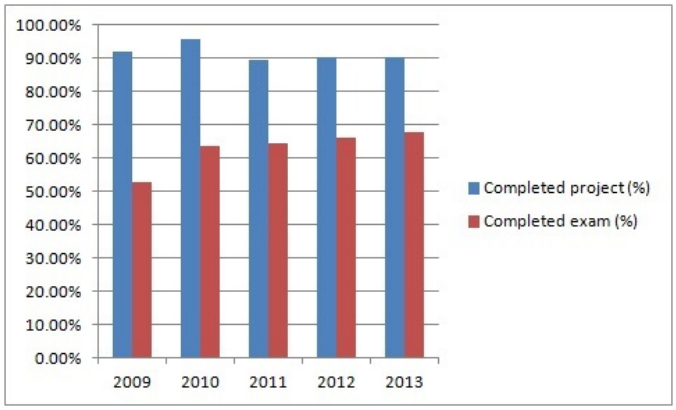


Fig. 5. The percentage of students with completed projects compared to the percentage of students with completed exam

Students worked in groups with 2-5 students in each group. In the laboratory exercises all team members could consult with their lab demonstrator or teacher about the project. In each exercise students were given a specific task towards creating the compiler which was the end objective in their project. As a result the compiler was built step-by-step on weekly bases. Each class students were given a chance to ask some questions about the task from the previous week. This approach gave students many opportunities to learn more about their project and to improve on their teamwork skills.

Also this approach yield great results in completion of the student projects. On Fig. 5 a graph for the percentage of students that completed the project is given. The results presented are since 2009 until today. Only around 10% of the students do not complete the project at the end every year. In 2010 only 4% of the students did not make a compiler enough for passing the exam. Even more in 2012 15 out of 17 given projects, which is more then 88%, were completed.

The percentage of completed projects is a promising result. There may be several reasons and some of them are given here.

First, working in groups is a challenging, but motivating task in which students make a great effort to do the assignments given. Groups may have different profiles of students and due to their difference they could not only get better results, but also learn from their teammates.

Another explanation is that students choose their project from more project variants with different difficulty levels. Because students are usually aware of their skills, knowledge and ambitions they choose a project with appropriate difficulty and in most cases manage to complete the project. As a result, suitable projects are available for each student. The most difficult project that can be chosen includes visualization module with which a program in RoboL is interpreted. This is the most interesting project for students, and each year more and more students are interested in taking the most difficult,

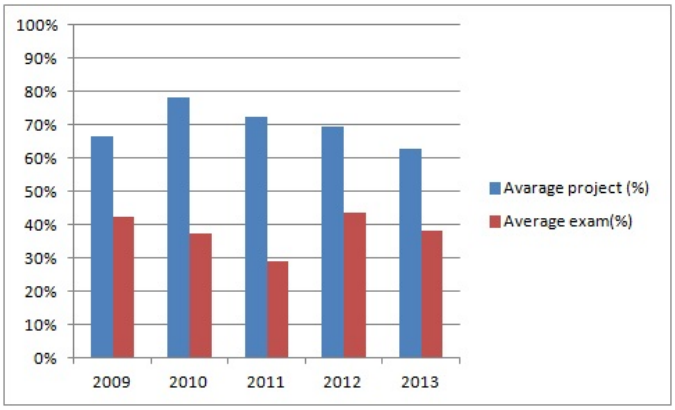


Fig. 6. Comparison of student project assignment and student exam percentage in 2009-2013

but most interesting project.

The last explanation for the high percentage of project completion is due to the organization of the laboratory exercise. Step-by-step tasks given to students enable students to complete most part of the projects during the classes. The help from demonstrators on the laboratory exercises also improves the project accomplishment result.

In Fig. 5 the percentage of students that completed their project is compared to the percentage of students that have completed or passed the exam. Because the project covers all the material from the lectures it is expected that the project would improve the percentage of students who completed the exam. Every year, more than 50%, and up to 70% of the students were able to pass the exam.

In all years from 2009 to 2013 the percentage of students that completed the project is higher then those who completed the exam. This is expected because the project assignment is much more attractive and interesting because of, among other previously stated features of the project, teamwork included.

The project in general helps students learn the course material and pass the exam. Indeed, through students activity in the project, students are stimulated to learn the theoretical part of the exam, even though it is hard and rather ”boring” for most students.

The final course grade depends on the project assignment grade as well as from the final exam. Fig. 6 shows that students got higher results on the project then on the exam. The exam validates student's theoretical knowledge for compilers, while the project checks the practical understandings of compilers and its implementation in some programming language. Students could choose from Java and C++ as programming languages for implementation.

The average project grade given in percentage in the last 2013 was something above 60%. Best results are noticed in 2010 where the average student score was around 80%. The average exam grade in 2010 was little below 40%.

The difference between the the project and exam grade (%) is notable decreasing from 2011 until today. In the future more close average grades for the project and exam are expected. Students should be motivated to work more on the project. One possibility is to assign roles for each student. Also, a group manager could be assigned. The manager should give an evaluation for the effort and activities of each student in his/hers group. The fact stated that working on the project helps students learn the theoretical material is our assurance for future improvement.

The last accomplishment of our approach presented here is that most part of the students managed to learn the RIMAL language without previous knowledge of microprocessors. Truly, completing a project means that students also did the code generator part of the compiler where RIMAL language was used for translation. Even though they did not have any knowledge in assembly, they could understand this language and use it in the project.

V. Conclusion

In this paper we discussed the organization of the course Compilers at our Faculty of Computer Science and Engineering. We presented our original method of teaching compiler production trough a practical project in the conditions where students does not necessary have prior knowledge of an assembly language and in the field of microprocessors.

The presented results from 5 years teaching the course confirm that this approach is efficient, and it can be used by others when similar conditions occur. This was the main intention of the authors when writing the paper. We plan to continue our work in the following years, making adjustments with every following generation of students.

Acknowledgement

The research presented in this paper is partly supported by the Faculty of Computer Science and Engineering in Skopje.

References

[1] “The university of virginia engineering: Compilers practicum,” 2014. [Online]. Available: http://www.cs.virginia.edu/

[2] A. Aiken, “Compilers,” 2014. [Online]. Available: https://www.coursera.org/course/compilers

[3] G. S. NovakJr., “Compiler construction,” 2014. [Online]. Available: http://www.cs.utexas.edu/~novak/cs375.html

[4] “University of waterloo: Compilers,” 2014. [Online]. Available: http://uwaterloo.ca/

[5] F. Benders, J.-W. Haaringm, T. Janssen, D. Meert, and A. van Oostenrijk, *Compiler Construction: A Practical Approach*, 2003.

[6] M. Jovanov, M. Gusev, and D. Martinovikj, “A new model of on-line collaborative activity for building ontology in e-learning,” in *Proc. of the IEEE Region 8 Conference EUROCON 2013)*, 2013, pp. 25–31.

[7] M. Jovanov, B. Kostadinov, E. Stankov, M. Mihova, and M. Gusev, “State competitions in informatics and the supporting online learning and contest management system with collaboration and personalization features mendo.” *Olympiads in Informatics*, vol. 7, 2013.

Analysis of elective courses selection in post-Bologna programmes

I. Rožanc and B. Slivnik
Faculty of Computer and Information Science, University of Ljubljana, Tržaška 25, SI-1000 Ljubljana, Slovenia
{igor.rozanc, bostjan.slivnik}@fri.uni-lj.si

Abstract - One of the most important results of the Bologna reform at the Faculty of Computer and Information Science at the University of Ljubljana is a much larger number of elective courses. The elective courses of the academic study programme were grouped into modules while the professional study programme permits selection of individual elective courses under minor restrictions. Both approaches fulfilled the reform requirements. However, the academic study programme is more rigid and thus easier to carry out than the professional one but students prefer the flexibility of the latter. In this paper we analyze the actual data of elective course selection by students of both programs. First, we examine the link between different module selections to optimize the distribution of elective courses into modules. This examination is possible as a number of students, i.e., the most successful ones, were allowed to select courses regardless of modules. Second, we investigate whether some commonly selected groups of elective courses of the professional study programme could be identified in order to simplify the programme and its implementation. The results show some important improvements could be introduced.

I. INTRODUCTION

In recent years the Slovenian higher education adopted the principles of the Bologna declaration [1]. The main goal of this reform is to define clear requirements for specific academic study levels. These requirements specify a standard three level classification of the study programmes. With implementation of the Bologna principles, the comparison between the programs is clearer, and the recognition of the academic degrees between the European states became much simpler [2, 3, 4, 5]. The reform is aimed to enhance migration of educated young people providing better communication and the basis for better economic outcome in EU [6].

The declaration is specified as a compromise using the "least common divisor" logic, thus uniting sometimes quite diverse solutions that are currently implemented in higher education of different European countries. Consequently, implementation of the declaration was not equally difficult to implement everywhere. In some states the reform implementation caused radical changes of the framework and the length of study, or introduction of previously not implemented principles (i.e., elective courses). This caused diverse reactions ranging from approval in states where similar principles were already in power to substantial opposition in other states.

As the Bologna declaration was a solution of an important issue, most European countries accepted it with the idea that "any solution is better than no solution". Our country was among them as the Slovenian Ministry of Higher Education decided all undergraduate and graduate educational programs to meet Bologna criteria by 2010 and to gradually implement them afterwards.

Faculty of Computer and Information Science at University of Ljubljana started to offer Bologna programs in the academic year 2009/10. The adaptation of undergraduate programs was gradual. In 2009/10 only the first year of those programs was taught in the new way, and in 2010/11 the second year students experienced the changes. In 2011/12 all three 1st level Bologna programs were defined, and by 2013/14 the second level study programmes were introduced as well.

The required level of changes was substantial as our previous academic study followed somehow different (German) study model, while the Bologna declaration enforces more English-like study model. Practically, the Bologna reform influences teaching in two ways. The first part of Bologna changes is about the program structure: the length of study programmes, a program framework consisting of the obligatory and elective (free choice) courses, the uniform course definition, and the conditions for promotion based on course credits (ECTS), etc. The second part advises different teaching style. The main goal of this part is enhanced efficiency of practical work with students i.e., by using new tools [7] or motivating approaches emerged from practice [8, 9, 10]. Those should be evaluated during the course and not only at the end of it.

After the Bologna reform implementation several interesting issues emerge [6]. Several papers address this experience in different European countries and institutions [2, 3, 4, 5, 11], some pointing out specific evaluation methods for the Bologna process [12, 13]. In 2011 we conducted a research on the effect of the Bologna reform on specific programming course at our faculty as well [14]. With more experience (and more data) a more general question of optimal introduction of elective courses is addressed in this paper. Before the Bologna the electiveness of courses was limited and consequently the pre-Bologna study programmes were without elective courses. Still, two study programmes had three variations of curriculum students could choose between. The new programmes introduce electiveness in two ways: moderately through definition of modules, and freely by enabling large number of individual course selection. The advantages and problems of both approaches are pointed out, and some interesting propositions resented.

The issue was studied on three the 1st level study programmes: the academic program of Computer Science and Informatics (CSI), the academic program Computer Science and Mathematics (CSM), and professional programme Computer Science and Informatics (CSI). Their direct pre-Bologna predecessors were the academic undergraduate program of Computer Science and Informatics, the academic undergraduate program Computer Science and Mathematics, and professional undergraduate programme Computer Science and Informatics. All programmes follow the IEEE Curriculum [15] which defines a series of courses that Computer Science students should attend during their studies.

In the second section all 1st level study programmes taught at our faculty are presented, pointing out the electiveness issues. The third section introduces the analysis background, namely the data and methodology used to study the issues. The fourth section presents the four analysis, and discuss the results obtained. Finally, the last section concludes the paper.

II. THE STUDY PROGRAMMES DESCRIPTION

As mentioned above, the three 1st level study programmes offered by the Faculty of Computer and Information Science were taken into consideration; the rest were left aside. There are three main reasons for this. First, the vast majority of all students are enrolled in one of these three programmes. Second, sufficient amount of data to make analysis viable is available for these three programmes only. Third, these are the three study programmes that the students enrolled in the second level study programmes are recruited from.

A. The two academic study programmes

Prior to the Bologna reform, the principal 4-year academic study programme of Computer and Information Science was more or less rigid. After completed the second year of the programme, students were required to choose one of the three subjects (hardware, software, or informatics). Each subject defined a set of courses the student enrolled in, but the rules of the study programme didn't permit choosing courses of two different subject. Although a student could apply for a special permission to do so, such cases were extremely rare since the student had to pass all curses of the chosen subject regardless of an extra course that he applied for.

The structure of the new 3-year academic study program of Computer and Information Science [16] is more flexible. As it is shown in Fig. 1, a student must pass 20 obligatory courses, 9 elective courses and prepare the undergraduate thesis (considered as another course).

However, out of 9 elective courses, a student can select 1 professional and 2 general courses freely; the remaining 6 elective courses are determined by choosing 2 out of 8 modules each consisting of 3 courses. To further enhance the programme's flexibility, students with average grade higher that 8 out of 10 (5 is negative) need not choose two modules but are permitted to freely choose the remaining 6 elective courses from the pool of courses modules are made of.

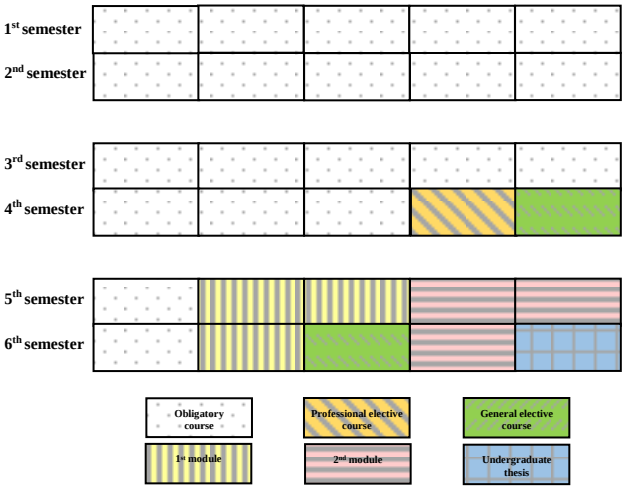


Figure 1. The structure of the 1st level academic study program of Computer and Information Science.

More than two thirds, i.e., 20 out 29 courses, of this study programme are still made of obligatory courses. There are restrictions regarding the selection of elective courses, more of them for average and less for better students. Based on the pool of general and professional elective courses offered in the academic year 2013/14, there are 2604 different alternatives available to average students and more than 12.5 million different alternatives available to their more successful colleagues.

The 3-year academic study programme Computer Science and Mathematics is less flexible. There are no elective courses in the first two years. In the third year, one module of the Computer and Information Science programme must be chosen and 1 professional elective course out of 9 mathematical courses. Additionally, a student must gain further 10 ECTS by enrolling in courses of any academic study offered by the University of Ljubljana. Because of the latter, the total number of reasonable alternatives is hard to estimate since, for instance, taking a course Obstetrics and gynecology might seem attractive to (male) students but can hardly be permitted in practice.

B. The professional study programme

Prior to the Bologna reform, the 3-year professional study programme of Computer and Information Science was more or less similar to and thus as rigid as its academic counterpart except that it was one year shorter. However, the remodeling of this programme was more radical - far more flexible system with more elective courses were introduced.

The structure of the new 3-year professional study programme [17] is shown in Fig. 2. Each of the six concentric rings in Fig. 2 represents one semester where the innermost ring (actually a circle) represents the first semester and the outermost one represents the last semester. The white circles represent 13 obligatory curses, mostly in the first two semesters but including industrial practice and undergraduate thesis in the sixth semester as well. All other circles represent 34 elective

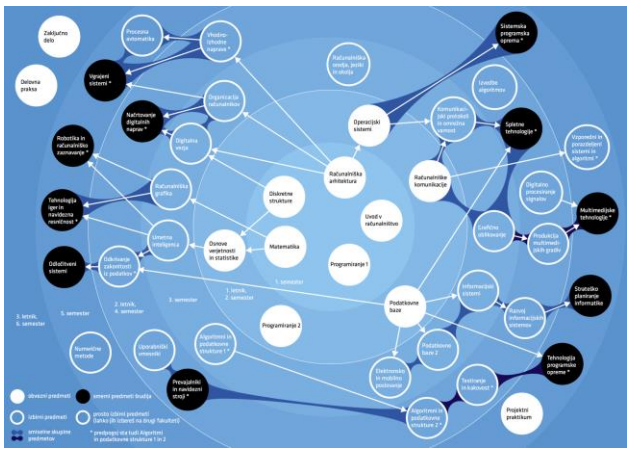


Figure 2. The structure of the 3-year professional study programme of Computer and Information Science [courtesy of the Faculty of Computer and Information Science, University of Ljubljana]

courses. 11 elective courses represented by black circles are target courses - each student must select at least two target courses. Arrows in the Fig. 2 indicates preconditions: an arrow leading from one course to another implies that the course the arrow starts at must be selected in order to select the course the arrow points to.

If the list of courses picked up by an individual student of the academic study programme includes more than two thirds of obligatory courses, almost two thirds of all courses, i.e., 17 out of 27, taken by a student of the professional study programme are elective courses. Taking preconditions into account, there are more than 43 million alternatives for every student.

The dark glue between certain courses in Fig. 2 indicates recommended groups of courses a student may select together. Note, however, that these groups are recommended but in no way enforced - together with preconditions and target courses (both specified by the study programme) groups are provided merely to help students find out reasonable assortments of courses.

III. ANALYSIS BACKGROUND

Five years after the implementation of the Bologna reform compliant study programmes it is appropriate time to check the suitability and quality of both solutions. The complete analysis is definitely complex as several perspectives should be taken into consideration. Most importantly, the level of student knowledge at the end of the study should be objectively assessed as well as the overall cost of such study programme performing, the invested staff effort and the satisfaction of students and teachers to name just a few ones. Thus several diverse metrics involving a large quantity of data should be taken in to consideration [18, 19].

In real world it is impossible to get a complete set of diverse data to conduct such comprehensive analysis, and the usual approach is to use what you've got. In other words, the analysis is driven by the information available to the researchers. In our case adequate information on actual student knowledge after the graduation is not available. Similarly, the objective measurement of the invested staff effort or student satisfaction is not obtainable. But we are lucky enough to have the access to the data of the old web student information system e-Študent.

The web student information system e-Študent provided electronic support for student selection and enrolment, the student course selection management, detailed management of examination records and grades, the alumni records management, and several student data analysis executions [20, 21]. For more than a decade it was used at several faculties (including the Faculty of Computer and Information Science) of the University of Ljubljana. In 2008 the University of Ljubljana decided to replace e-Študent due to the introduction of the Bologna reform. Unfortunately, the first solution did not meet the expectations, and a considerable effort was invested to adequately upgrade and maintain the old system in last few years. Finally the valuable domain knowledge was extracted from it and implemented into to his successor [22, 23]. Finally, in academic year 2013/14 the new information system was deemed adequate and thus replaced old e-Študent.

After the implementation all data were copied from the old system to the new system, but they were not removed from it. The old system still contains the original of all data for the era before academic year 2013/14 and those were used to perform the analysis. The quality of this data could be a problem as the upgrades implemented in the last stage (to meet the new Bologna requirements) changed the structure of recorded data as well. Nevertheless, due to careful maintenance the system preserved the consistent data structure and contents.

Three interrelated groups of data from the old e-Študent system were used for the analysis:

- **the enrolment data** describing the student enrollment into the specific year of the study programme in selected academic year,
- **the curriculum data** defining all (obligatory and elective) courses comprising the year of certain study programme, which is performed in selected academic year, and
- **the elective courses data** specifying the elective course and students, who selected them from the curriculum of the study programme a student is enrolled into.

The Bologna programmes were first introduced in 2009. In this year the 1st year of both academic studies was introduced, but no data were suitable for analysis as all courses were obligatory. On the other hand all three years of professional study were introduced at once and last two years had a lot of elective courses in curriculum. Still, only the elective courses for the 2nd year enrollment were taken in consideration, because for the 3rd year special rules were applied. In the following academic years enrollments with selection of elective courses became frequent. Till academic year 2013/14 586 such enrollments in both academic study programmes were recorded, and 611 enrollments in the professional programme, respectively.

To prepare a consistent base for further research, the data had to be cleansed. The majority of possible elective

courses in the curriculum are defined in the same study programme, but the student is allowed to select a few elective courses from a different programme or even from a different faculty. Those selections are recorded in the data as well, but they can be considered obsolete and thus removed. In practice the data related to a predefined set of courses, i.e., 27 courses in case of academic study programmes, and 37 courses in case of the professional programme, were retrieved only.

Most of the elective courses data should be consistent with the curriculum rules, like the summa of ECTS points of selected elective courses (per enrollment in a certain year and study programme) should be approximately the same. Due to the introduction of the new study, a number of students were considered special cases and the rules were not strictly enforced. Additionally, in some cases diverse rules were applied to diverse groups of students, i.e., in case of enrollment in academic study programmes students with average mark above 8.0 were allowed to individually select courses, while the others had to select entire modules. The data in such cases may seem strange, but it was still taken into consideration if it was related with the predefined set of courses.

IV. ANALYSIS AND DISCUSSION

We would like to determine the hidden rules that can be resolved by the analysis of preferences in selection of the elective courses by students. More precisely, three different curriculum models, i.e., the rigid model without elective courses as in the pre-Bologna programmes, a moderate model with modules as in the Bologna academic course, and the free model with a large number of elective courses like in the professional programme, could be compared by different aspects: the ease of implementation and student satisfaction.

Even without objective analysis we might guess the rigid model is simpler and its performance demands less organizational effort than the model with a large number of individual elective courses. On the other hand the last model is more adaptive and thus it better suits the student's needs. But is it really so? Do the students really select so many different combinations of elective courses, or could the study implementation become simpler and thus perhaps even cheaper?

The characteristics of data obtained allow some analyses in this direction. In the following subsections we present the results of several analyses we performed. Additionally, a short discussion is provided to point out the important conclusions.

A. The number of elective courses

The analysis was performed to determine the typical number of elective courses per student's enrollment. The results for academic (AP) and professional (PP) study programmes are presented in Table I.

The analysis shows that the usual number of elective course at academic programmes is 1 or 2 (in case of enrollment in the 2nd year), 3 or 4 (in case of enrollment in the 3rd year CSM) or 6 or 7 (in case of enrollment in the 3rd year CSI). For professional programme, the students enrolled in the 2nd year select 4 or 5 elective courses, while the students enrolled in the 3rd year select 9 or 10 elective courses. The summa of numbers in each row equals the total number of enrolled student (586 in case of academic and 611 in case of professional programme).

TABLE I. THE NUMBER OF STUDENT ENROLLMENTS WITH SPECIFIC NUMBER OF ELECTIVE COURSES

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
A	344	72	12	16	9	31	101	1							
P			1	33	159	7		8	101	280	14		5	1	2

B. The favorite courses

This analysis reveals most frequently selected courses. Additionally, a rule for finding out courses popular enough is formulated. The number of courses selected in at least 50%, 30% and 10% as many times as the course selected most frequently, i.e., the one with the maximum number of selections, is shown in Table II.

TABLE II. THE COURSES WITH HIGHEST NUMBER OF SELECTIONS

	Max. number enrollments per course	**%**	**Number of courses with more than 50% max**	**Number of courses with more than 30% max**	**Number of courses with more than 10% max**
AP	198	33.8	6 out of 27	10 out of 27	18 out of 27
PP	410	67.1	11 out of 37	18 out of 37	28 out of 37

The analysis shows the student's favorite courses and the number of courses of reasonable popularity using diverse assessments. In the optimal situation, all courses would be approximately equally popular, thus they would be between the courses with more than 50% (or at least 30%) of the maximal number. This is not the case and the number of popular courses is in the best case a half of all courses offered (in case of the professional programme) or a third (in case of the academic programme).

Quite large (between a quarter and a third) is the number of courses with less than 10% of selections in comparision with the maximal number. These courses might be considered considered obsolete. The number of such courses is quite large (between a quarter and a third). Surprisingly, a larger number of obsolete courses was found at academic programmes, even though the requested limit, i.e., the number of enrollments, was lower than in case of the professional programme.

C. The favorite sequences of selected courses

The previous analysis was upgraded to study the sequences of selected courses. First, the courses in each selection were ordered by the course ID to avoid repetition of results. Next, we analyzed the data to find out all different sequences of certain length. Finally, the sequences were checked to determine the most popular ones. The results in Table III were measured for the length of sequences up to the largest number of selected courses per enrollment.

TABLE III. THE SEQUENCES OF COURSES WITH HIGHEST NUMBER OF SELECTIONS

Number of courses in sequence	Max. number of one such sequence (academic prog.)	%	Max. number of one such sequence (professional prog.)	%
2	110	18.8	332	54.3
3	106	18.1	275	45
4	58	9.9	158	25.9
5	31	5.3	105	17.2
6	29	4.9	69	11.3
7	16	2.7	39	6.4
8	1	0.2	27	4.4
9			12	2
10			4	0.6
11			2	0.3
12			2	0.3
13			2	0.3
14			1	0.2
15			1	0.2

In the academic programmes there are many sequences of three courses with similar frequency as those which constitute the module. Still, none of such sequences is too popular. On the other hand the professional programme preserves some popular sequences up to the six courses without modules. Thus, some popular combinations could be easily determined.

D. The number of different sequences

In the next analysis we checked the number of different ordered selected course sequences of certain length and compared that number with the theoretically maximal number of sequences of the same length (calculated as a simple number of combinations disregarding the preconditions). Additionally, the number of important sequences (with the number of enrollments higher than 10% of maximum number) were counted as well. The results are presented in Tables IVa and IVb.

The results per both types of study programmes prove the number of actual different course sequences to be radically lower than it could be. There are obvious reasons for that: the academic programmes have modules, and some other limitations, i.e., professional elective course, while in the professional program these are preconditions in case of some courses. Additionally, the number of enrollments is too low to yield so many sequences.

TABLE IVA. THE NUMBER OF DIFFERENT SEQUENCES OF SELECTED COURSES FOR ACADEMIC PROGRAMME

Number of courses in sequence	Max. possible number of different sequences	Actual number of different sequences	Number of different sequences with number of courses more than 10% max
1	27	27	18
2	702	238	72
3	17550	758	103
4	421200	1176	162
5	9687600	951	137
6	213127200	389	67
7	4475671200	66	21
8	89513424000	1	1

TABLE IVB. THE NUMBER OF DIFFERENT SEQUENCES OF SELECTED COURSES FOR PROFESSIONAL PROGRAMME

Number of courses in sequence	Max. possible number of different sequences	Actual number of different sequences	Number of different sequences with number of courses more than 10% max
1	37	37	28
2	1332	623	174
3	46620	4100	546
4	1585080	13655	1495
5	52307640	30415	2175
6	1673844480	44660	2470
7	$5.18 * 10^{10}$	44030	2552
8	$1.56 * 10^{12}$	31314	1213
9	$4.51 * 10^{13}$	17712	1251
10	$1.26 * 10^{15}$	8525	8525
11	$3.41 * 10^{16}$	3419	3419
12	$8.87 * 10^{17}$	1053	1053
13	$2.22 * 10^{19}$	228	228
14	$5.32 * 10^{20}$	31	31
15	$1.22 * 10^{22}$	2	2

Still, these are the complete data and if carefully used these data point out important rules. By using information of the first analysis (A) only certain lengths of sequences are relevant as we have enough enrollments with that number of selected courses. If we consider the important sequences only (the 10% rule), the number of different sequences becomes very small in cases of all programmes. It is smaller in case of academic programme, but the numbers for professional programme are close to that value. Thus, we can conclude so much electiveness in latter case does not yield results, and a more rigid system could do the job too.

V. CONCLUSION

The study presented in this paper provides some empirical results indicating the popularity of courses within the context of Bologna study programmes with different degrees of flexibility regarding the elective courses. Furthermore, it includes some preliminary analysis and opinions.

The results of the analyses confirmed the expected differences between the academic and professional programmes. The rigidness of the first is shown in smaller number of different sequences experienced and more uniformly distributed selections between them. The professional programme yields a larger number of different sequences and larger discrepancy between popular and non-popular courses. Still, data proves only a small part of the flexibility provided by the professional programme is used, and a more rigid solution could work as well.

In both cases the results reveal unpopular courses. Using a careful analysis of past elective courses selections different modules could be proposed in case of the modified academic programmes, and some coursees could perhaps be removed from the professional programme.

However, one must not confuse the popularity of a course with its significance within the syllabus. The core courses are quite often less popular but should not be removed from the study programme at any cost, and this study should not be used as a justification for that. Even though it has not been explicitly mentioned so far, the authors are aware that the popularity of a course might be significantly influenced by the lecturer's popularity. At present, unfortunately, the lecturer's popularity cannot be taken into consideration due to the Law on Protection of Personal Data.

REFERENCES

[1] "The Bologna Declaration" – Joint declaration of the European Ministers of Education, 19 June 1999, http://www.bologna-bergen2005.no/Docs/00-Main_doc/990719BOLOGNA_DECLARATION. PDF

[2] T. Malan, "Implementing the Bologna Process in France", European Journal of Education, Volume 39, Number 3, pp. 289-297, 2004.

[3] A. Veiga and A. Amaral, "Survey on the implementation of the Bologna process in Portugal", Higher Education, Volume 57, Number 1, pp. 57-69, 2009.

[4] Åse Gornitzka, "What is the Use of Bologna in National Reform?", Creating the European Area, Higher Education, Higher Education Dynamics, Volume 12, pp. 19-41, 2006.

[5] J. Kettunen and M. Kantona, "The implementation of the Bologna Process", Tertiary Education and Management, Volume 12, Number 3, pp. 257-267, 2006.

[6] G. Winckler, "Universities in the 21st century", Information Services and Use (Selected Papers from the International Conference Academic Publishing in Europe: The Impact of Publishing, 20-21 January 2009), Volume 29, Number 2-3, pp. 65-72, 2009.

[7] L. Fürst, V. Mahnič, "A cooperative development system for an interactive introductory programming course". World transactions on engineering and technology education, Volume 10, Number 4, pp. 1557-1584, 2012.

[8] L. Fürst, V. Mahnič, "Introductory programming course : motivating students with prior knowledge". World transactions on engineering and technology education, Volume 11, Number 4, pp. 400-405, 2013.

[9] V. Mahnič, "Teaching Scrum through team-project work : students' perceptions and teacher's observations". The international journal of engineering education, Volume 26, Number 1, pp. 96-110, 2010.

[10] V. Mahnič, "A capstone course on agile software development using Scrum". IEEE transactions on education, Volume 55, Number 1, pp. 99-106, 2012.

[11] P. Vilar and M. Žumer, "The Bologna reform at the department of library and information science and book studies, university of Ljubljana", Journal of Information Retrieval, Volume 12, Number 2, pp. 102-116, 2009.

[12] U. Fullerm A. Pears, J. Amillo, C. Avram, L. Mannila, "A computing perspective on the Bologna process", Working Group Reports on ITiCSE on Innovation and Technology in Computer Science Education ITiCSE-WGR '06, ACM SIGCSE Bulletin, Volume 38, Numbe 4, pp. 115-131, 2006.

[13] A. Kö, A. Gábor, R. Vas, I. Szabó, "Ontology-based Support of Knowledge Evaluation in Higher Education", Proceedings of the 2008 Conference on Information Modelling and Knowledge Bases XIX, IOS Press, pp. 306-313, 2008.

[14] I. Rožanc, "One year later: the effect of the Bologna reform on Algorithms and Data Structures Course Teaching" in Proceedings of the WSEAS International Conference on Recent Researches in educational technologies, pp. 221-226, Korfu, Greece, 2011.

[15] Joint Task Force on Computing Curricula, "Software Engineering 2004: Curriculum Guidelines for Undergraduate Degree Programs in Software Engineering", IEEE CS and ACM, 2004, http://sites.computer.org/ccse

[16] "Študijski program Računalništvo in informatika (Dodiplomski - univerzitetni)", N. Zimic et. al., Faculty of Computer and Information Science, University of Ljubljana, 2008.

[17] "Študijski program Računalništvo in informatika (Dodiplomski – visokošolski strokovni)", B. Slivnik et. al., Faculty of Computer and Information Science, University of Ljubljana, 2008.

[18] L. Cappellari and C. Lucifora, "The Bologna Process and college enrollment decisions", Labour Economics, Volume 16, Number 6, pp. 638-647, 2008.

[19] P. O. Duarte, M. B. Raposo, and H. B. Alves, "Using a Satisfaction Index to Compare Students' Satisfaction During and After Higher Education Service Consumption", Tertiary Education and Management, Volume 18, Number 1, pp. 17-40, 2012.

[20] V. Mahnič, S. Drnovšček, "Introducing agile methods in the development of university information systems" in Proceedings of the 12th International Conference on European University Information Systems, pp. 61-68, Tartu, Estonia, 2006.

[21] V. Mahnič, I. Rožanc and M. Poženel, "Using e-business technology in a student records information system" in Proceedings of the 7th WSEAS International Conference on E-Activities (E-Activities'08), pp. 589–594, Cairo, Egipt, 2008.

[22] I. Rožanc and B. Slivnik, "Producing the platform independent model of an existing web application" in Proceedings of the Federated Conference on Computer Science and Information Systems, pp. 1385-1392, Wroclaw, Poland, 2012.

[23] I. Rožanc and B. Slivnik, "Using Reverse Engineering to Construct the Platform Independent Model of a Web Application for Student Information Systems". Computer Science and Information Systems, Volume 10, Number 4, pp. 1557-1584, 2013.

Data Mining of Computer Game Assisted e/m-learning Systems in Higher Education

Petar Jurić*, Maja Matetić** and Marija Brkić**
* Primorsko-goranska County, Department of Information Technology Services, Rijeka, Croatia
** University of Rijeka, Department of Informatics, Rijeka, Croatia
petar.juric@pgz.hr, majam@inf.uniri.hr, mbrkic@inf.uniri.hr

Abstract - One of the major challenges in the implementation of e-learning systems is motivating students to use it. E-learning systems are mostly unadapted to mobile platforms because they do not use responsive web design and progressive enhancement based on browser, device or feature detection. This paper presents the evolutionary cycle, current status and potential development of new platforms for e/m-learning in higher education. Educational computer games that serve as a motivational element in e/m-learning are presented, as well as the data structures that can complement the existing models of e/m-learning. The University of Rijeka uses MudRi e-learning system, which is based on Moodle open source software. This system collects data on the activity of participants in the educational process, which is suitable for data mining. A preliminary research on the correlation between the activity in forum discussions and the course pass or fail grade is carried out. The possibility of using data mining based on the data structure of the Moodle system and the models of e/m-learning and learning through play is described, aiming at finding optimal ways of learning and optimal learning results. Finally, a model of integrated system user interaction on which our future research will be based is presented.

I. INTRODUCTION

Technologies available through new generation mobile devices, such as smartphones and tablets, are increasingly used as a primary platform for Internet access. These devices are suitable for learning on the go, and their capabilities often surpass desktops and laptops, primarily because of the sensors such as accelerometers, gyroscopes, magnetometers, and GPS which desktops and laptops lack. Using them for learning can be fun and available wherever there is mobile signal.

Unfortunately, to the best of our knowledge, no statistics exists on the number of students in Croatia who use these devices. Therefore, in our future work we plan to investigate this issue thoroughly.

Mobile learning (m-learning) systems can directly and indirectly generate and store large amounts of data. Direct data refer to the learning materials such as video lectures, presentations, forum posts, etc. Indirect data are user activity logs, as well as user locations when mobile phones are used for accessing the system.

Motivational element is an important part present in all forms of the educational process. The main motivation of students in the educational process of higher education is based on collecting the required number of points with regard to the course requirements, regardless of the content and the form. Since learning with reduced motivation results in lower exam achievements, reaching the highest level of creative knowledge in students is not possible without high motivation.

The use of serious and educational computer games can increase the level of interactivity in e/m-learning systems. When students master educational content via computer game elements, the level of their motivation is raised because learning becomes fun. Learning through computer games is an upgrade of e-learning. Therefore, the same advantages and disadvantages presented in the subsequent section relate to them, but with the positive effect on motivation, and negative effect on the time needed to create the system.

Data mining techniques are used to analyse data obtained through these systems. By performing data mining in e-learning systems, new methods for specific types of information related to education are developed [1]. In addition to the data contained in the user profile and system check-in, actions taken by the student based on the available content can also be monitored.

II. E-LEARNING SYSTEMS

There is a number of e-learning definitions. E-learning is often seen as an upgrade of distance learning [2] and computer-assisted learning. All of these categories are characterized by the lack of face to face interaction between the student and the teacher and by some kind of technology (usually a computer) that is used for communication as well as for accessing educational content [3]. Advantages and disadvantages of e-learning are described in Table I [4]. From our point of view, e-learning systems can be described as:

- *classic* – fixed (desktops) or portable (laptops, which are portable, but not mobile in the sense that in general they cannot be used while walking or standing)
- *mobile* – can be used while moving.

We strongly believe that e/m-learning on mobile devices can provide equivalent or even better content experience compared to conventional devices, i.e. desktops and laptops, ever since 2008 onwards, or more precisely, since the advent of smartphones and tablets.

TABLE I. ADVANTAGES AND DISADVANTAGES OF E-LEARNING SYSTEMS

Student		**Teacher**		**Educational institution**	
Advantages	***Disadvantages***	***Advantages***	***Disadvantages***	***Advantages***	***Disadvantages***
Interactivity	Alienation	Automatic system adjustment to the student on the basis of prior knowledge	Less personality and possibilities for intervention	Unlimited number of students	Not all students and teachers have IT skills
Availability	Need for self-organization	Availability	Less activities involving education and more involving moderation	Availability	Dependence on IT
Independence from group	Poor motivation	Besides IT resources, other material resources are less necessary	Communication with the student is primarily written	Lower cost in the long run	Poor influence on usage and motivation
Self-assessment and progress tracking	Rare physical contact with teacher	Students' feedback	Rare physical contact with teacher	Quick course adjustment	Long-term development

A. Classic systems for e-learning

Classic systems are adapted for use on devices with large screens and high resolution. They are most commonly supported in web environments (an Internet browser is sufficient for use) of the following operating systems: Windows, Linux and Mac OS. The keyboard and the mouse, and lately the touchscreen, are used as primary input units for interaction with these systems.

B. Systems for mobile learning

With the invention of PDAs (personal computer assistants) and mobile phones with small screen resolution, m-learning comes into existence and makes it possible to use the e-learning system anytime and anywhere.

In the past six years m-learning has become linked to smartphones. A smartphone is a mobile phone with a mobile operating system that provides standardized interface and platform for application development. It has a built-in camera, a web browser, location services support, and the ability to play multimedia. The most common mobile operating systems are Android and iOS, followed by Windows Phone, Symbian, BlackBerry OS, Bada and others. The touchscreen is used as a primary input unit for interaction with the smartphone system. A tablet uses a mobile operating system and is basically a smartphone with a giant screen; the possibility of making telephone calls is set aside in favour of the facilitated interaction with the device.

With regard to the education system, m-learning is more represented in higher education because most of the students have their own mobile devices. Up to this point of time, courses have not yet been developed in accordance with the capabilities of technology. Most of the applications of m-learning available at universities include news, and calendars with lecture schedules, instead of the content to be learned [5].

Although using m-learning on devices that support new technologies is conducive to students in comparison to the conventional e-learning, exam results might be weaker [6]. The reason for this is usage on the move or in situations where the environment can adversely affect concentration. Therefore, it is necessary to consider additional elements, such as learning through computer games, in order to gain more attention and put more focus on the content. Furthermore, the fact that all new mobile operating systems support multi-touch (up to 10 fingers simultaneously) and identifying the orientation of the screen needs to be taken into account. Hence, it is necessary to use new methods and technologies in developing systems for mobile e-learning.

To achieve greater unification in developing e-learning systems, the technology such as HTML5 ecosystem can be used, which tends to reduce the problem of fragmentation among and within mobile and classic devices.

III. COMPUTER GAME-BASED LEARNING

Computer games are interactive software applications that have fun as the basic objective while users engage in them. By using multimedia, networking and other technologies, they lead users to solve problems and achieve goals in a virtual environment [7]. Unlike computer games, whose primary purpose is fun, learning through computer games should include the following:

- motivational elements but not necessarily fun,
- active participation in problem solving,
- providing feedback,
- adjustment of the level of knowledge,
- clarity in the presentation of the objectives and learning content,
- presentation of content that can be replicated in the virtual world, while drawing parallels with the real world,
- scalability to a large number of simultaneous users.

Computer games with the above listed elements are called educational games. In addition to educational, there are also serious computer games with fun as a motivational element. These are primarily regular games with educational elements, e.g. simulations. The two areas are often intertwined in order to achieve good balance of educational content load and fun.

According to Prensky [8], the new generation of students born in the last quarter of the 20th century is developing different cognitive styles in comparison to the previous generation (Table II).

TABLE II. NEW PREVAILING COGNITIVE STYLES IN STUDENT LEARNING: GENERATIONS WHICH GREW UP WITH COMPUTER GAMES

New cognitive styles	Old cognitive styles
Twitch speed	Conventional speed
Parallel processing	Linear processing
Graphics first	Text first
Random access	Step-by-step access
Connected	Standalone
Active	Passive
Play situations	Work situations
Payoff	Patience
Fantasy	Reality
Technology-as-friend	Technology-as-foe

Related studies point to the benefits of designing educational content through computer games, because students have shown several times higher interest, i.e. the time spent using the content has been recorded as follows: 7-8 times higher compared to the textual content, and 3-4 times higher compared to the audio/visual content [9]. The disadvantage is reflected in the demanding and time-consuming preparation.

According to students, motivational elements to use computer games in higher education sorted by relevance are the following [10]:

- challenge – appropriate level of difficulty, achievement of multiple objectives for the coming victory, constant feedback, unpredictability,
- curiosity – optimal information dosing,
- cooperation – helping each other to achieve common goals,
- competition – comparison of results with other users,
- control – ability of choice and perception of consequences,
- recognition – sense of satisfaction with the achievements,
- imagination – identification with the role.

Learning based on computer games can have a positive impact on learning [11]. The content is interesting not only because students have grown up with computer games [12], but also because it is an alternative way of learning which can be more efficient than the classical one [13, 14].

Educational computer games have an action-based scenario. The most common question is how to adjust the game 'plot' when the virtual world needs to suit the real one.

In certain scenarios students can work together to solve problems, while in others they can play against the computer or compete with each other through competitive games in order to achieve better results [15].

Computer games can teach computer science students to develop and apply algorithms for solving problems, and to simulate, evaluate and detect logical errors [16].

Previously mentioned studies on using computer games for learning, although of a newer date, do not take into account the variety of mobile platforms, and are therefore restricted to a specific range of compatible devices. This means that the problem presented at the end of the previous chapter manifests itself here as well, and it can be largely solved in a similar manner, i.e. by the development of HTML5 compatible games run in an Internet browser.

IV. SYSTEMS FOR COMPUTER GAME E/M-LEARNING AND DATA MINING

Data mining forms the main component and an integral part of the process of knowledge discovery in large amounts of data [17]. Knowledge discovery in data comprises the following tasks: data cleansing (deviation and inconsistency detection), integration (in cases where data from multiple sources can be combined), data selection (only data relevant for the analysis), data transformation (consolidation in the appropriate form, and related operations), data mining (detection of patterns in the data), pattern evaluation (determination of patterns that represent the desired result, i.e. the knowledge that wants to be revealed), and knowledge presentation (using visualization and representation techniques).

E/m-learning and computer game learning are based on the interaction between the student and the application, and thus generate a variety of data suitable for data mining. Although e-learning systems hold large amounts of data, they are primarily designed to support learning rather than the analysis of the stored data [18]. Moodle System (Modular Object Oriented Development Learning Environment), on which MudRi system of the University of Rijeka is based, is one of the most used open source systems for e-learning. Moodle version 2.5 includes over 200 tables in its database structure [19]. These tables store configuration settings, user profiles, courses, access and activity data, etc. Forums rely on eight tables. In Moodle, user authentication is performed by password. The problem with the data from e-learning systems is their interdependence which violates the assumption that majority of machine learning algorithms are based on. For example, forum answers can build on previous ones. Previous studies on data mining in e-learning systems [20] include the following techniques:

- regression – predicting the time students will spend logged onto the system, or predicting to what extent students will be satisfied with the educational institution,
- grouping – establishing models of students who work in similar conditions and recommending activities the student has not yet used which would help him or her in the acquisition of knowledge depending on the predisposition,

- classification – predicting students' final grades, according to their activity in the system,
- association rules – recommending elective courses.

By taking into account the data produced by modern m-learning systems, the database can also record geolocation to find out how often the system is used with regard to the location. Furthermore, if geolocation is recorded periodically, at short intervals, it can be verified whether the system is actually used on the go. Such systems may include all of the data structures from e-learning systems. By using the available APIs, social networks such as Facebook and Twitter can be integrated into the system. Therefore, current conventional communication through forum discussions can be partially or completely replaced. Data mining of texts can be used for analysing student activity on social networks and its effect on the results. Interaction model of the system for computer game e/m-learning and data mining is presented in Fig. 1.

It is evident that the new m-learning systems tailored for smartphones and tablets are related to the classical e-learning systems in a way that e-learning has become a subset of m-learning, while with the old mobile devices and associated systems this relation is reversed (Fig. 2). Accessing e-learning systems via mobile devices makes it possible to identify and track how users access content through certain platforms, and therefore makes it possible to offer appropriate design (a single column or two columns instead of multiple columns, etc.) or adapt to the screen resolution.

In educational computer games places where most mistakes are made can be annotated by following the paths of information and interaction with the interface [21]. The start time, exit time before the end of the game (the student has not passed the whole content) and end time (the student has passed the whole content) can be recorded. Various data can be studied: how many times has the game been played, has the student who passed the whole game come back to it again, etc.

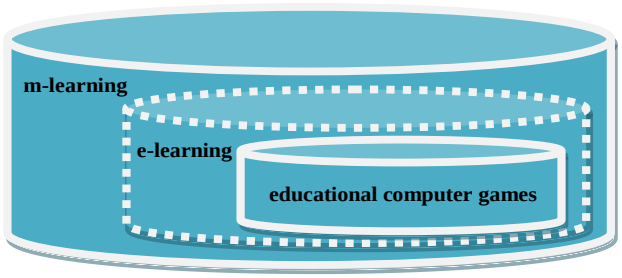


Figure 2. Display of data systems for learning with computer games compared to systems for e/m-learning

Students can be distracted in using the system when accessing it with mobile devices which have the incoming call feature (tablets do not usually fall into this category), SMS, MMS, e-mail, and other messages.

Due to its complexity, data mining may be performed by system administrators or by teachers who create e-courses [22]. The results may help teachers (to find ways to improve the level of acquired knowledge of their students, and to improve the development of e-courses), students (if the system recognizes their level of knowledge and their behaviour, more time-efficient learning can be offered or learning can be supplemented with additional content sources and activities), and administrators (to achieve optimal server load and network resources parameters).

Knowledge discovery in different data systems that form the basis of m-learning can be particularly challenging and complex in initial preparatory activities. The reason for this is the diversity of data which is mostly unstructured (they require special processing methods and techniques, including natural language processing which can be particularly challenging in cases of microblogging service where the number of characters is limited and messages frequently use combinations of abbreviations and sign language).

A further problem lies in various purpose services (e.g. SMS, MMS, geolocation) which are adapted to aggregate data and integrate it into a single unit.

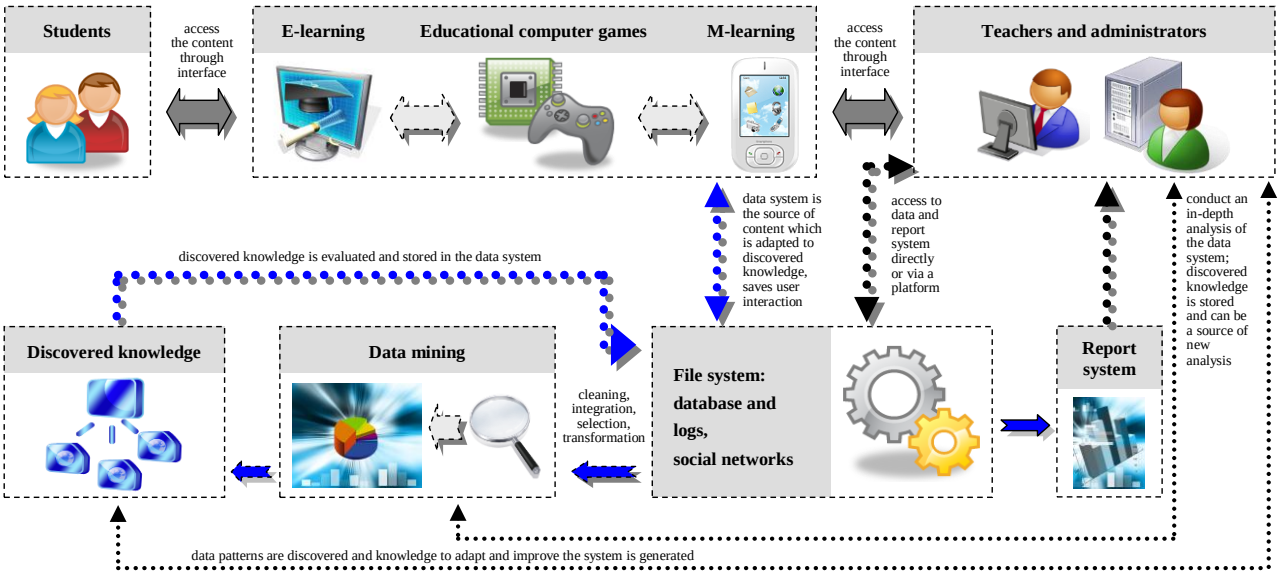


Figure 1. Representation of the user interaction model with data mining of computer game assisted e/m-learning in higher education

V. PRELIMINARY RESEARCH

The analysis presented in this paper includes student activity in forum discussions of the MudRi system and their final pass or fail grade. The simplest way to measure user activity relates to the measurement of the number of posts of a particular user. Mostly users write a post, which is mainly a query, and others give answers to the query. The number of posts can often lead to false conclusions. For example, a user can write only two posts, one may be a detailed answer to a complex problem, and the other only a thank-you note. Using the previously mentioned method, these posts would be rated equal, so a correction is applied which takes into account the average size of posts and the total number of characters in the post. This in turn can again lead to false conclusions because one may copy and quote the entire post as an additional source of information, while someone else may only add a link. Users who give answers to the questions are considered experts in a particular area, and those who ask questions are considered seekers of knowledge. The user type can be observed through the time of posting, so forums can be classified according to their prevailing activity time: morning, weekend, night, etc. This approach can be combined with the aforementioned post counts.

This study analyses the data from two generations of students enrolled in the course Programming 1 at the undergraduate study of Informatics at the University of Rijeka. The research related to the generation 2010/11 includes 72 students, and the one related to the generation 2011/12 includes 76 students. Students who have posted at least once to a forum, i.e. who have been actively engaged in the discussion are further on referred as Forum members. The number of posts is 145 in 2010/11, and 86 in 2011/12. Forum members are those students who have written at least one post in one or both of the discussion forum categories. Namely, the observed forum discussions can be grouped by content, and described as follows:

- Forums with lectures and exercises handle notifications of problems and seeking help to solve problems, and issues related to the content of the script, homework and exams. Generation 2010/11 includes 79 posts, i.e. 54 % of the total, and generation 2011/12 includes 21 post, i.e. 24%.
- Forums with student tutorship mainly relate to the confirmation of arrival to the tutorship. They include 66 posts in 2010/11, i.e. 46% of the total, and 65 posts in 2011/12, i.e. 76%.

To calculate the correlation between students who actively post on forums and those who do not in regard to whether they pass the course, a statistical Chi-square test with one degree of freedom is employed (1) using the data from Table III.

$$\chi^2 = \sum \frac{(f_0 - f_t)^2}{f_t} \tag{1}$$

TABLE III. COURSE PASS/FAIL GRADE COUNTS OF FORUM MEMBERS AND NON-MEMBERS

Passed course	**YES**		**NO**	
	2010/11	***2011/12***	***2010/11***	***2011/12***
Forum members with lectures and exercise	18 $a_{1\text{-}2010/11}$	8 $a_{1\text{-}2011/12}$	6 $b_{1\text{-}2010/11}$	1 $b_{1\text{-}2011/12}$
Forum members with student tutorship	22 $a_{2\text{-}2010/11}$	24 $a_{2\text{-}2011/12}$	10 $b_{2\text{-}2010/11}$	6 $b_{2\text{-}2011/12}$
All Forum members	28 $a_{3\text{-}2010/11}$	26 $a_{3\text{-}2011/12}$	13 $b_{3\text{-}2010/11}$	6 $b_{3\text{-}2011/12}$
Non-members	13 $c_{2010/11}$	20 $c_{2011/12}$	18 $d_{2010/11}$	24 $d_{2011/12}$

By replacing the cell values in Table III with letters a, b, c and d, we obtain:

$$\chi^2 = \frac{N \cdot \left(|ad - bc| - \frac{N}{2} \right)^2}{(a+c)\cdot(b+d)\cdot(c+d)\cdot(a+b)}. \tag{2}$$

The null hypothesis H_0 set in the Chi-square test assumes that there are no differences in passing the course between different categories of Forum members and non-members. The working hypothesis H_1 assumes the dependence between these values and categories.

The level of significance is set to P=0.05 with Yates's correction, due to the small number of posts in some categories, i.e. less than 10.

According to the results shown in Table IV, the null hypothesis is not validated in five out of six monitored cases. To sum up, there is a statistically significant correlation between the course pass or fail grade and the activity in forum discussions.

The working hypothesis is not validated only in the case of Forum members with student tutorship in generation 2010/11. The reason for this may lie in the purpose of the forum for student tutorship; it is not clear whether the students attend tutoring because they need supplemental tuition or advanced tuition.

Generation 2010/11 has an additional motivating factor: they are invited to post the errors identified in the script, and to propose corrections. Thereby, the number of students actively participating in forum discussions is higher in that generation. Text mining in that group might identify and differentiate students who provide assistance and solutions from those who seek help.

TABLE IV. CHI-SQUARE TEST RESULTS OF FORUM MEMBERS AND NON-MEMBERS

Forum members' categories	χ^2		**P**	
	2010/11	***2011/12***	***2010/11***	***2011/12***
Forum members with lectures and exercise	4.744	4.048	0.0294	0.0442
Forum members with student tutorship	3.564	7.456	0.0591	0.0063
All Forum members	3.985	8.494	0.0459	0.0036

VI. CONCLUSION AND FUTURE WORK

The evolution of e-learning and m-learning systems is in progress since the Internet is increasingly accessed via smart mobile devices (phones and tablets). Simultaneously with the development of e/m-learning systems, various studies are conducted regarding how to increase the use of these systems, since distance learning reduces direct communication among participants in the educational cycle. Consequently, this situation is demotivating in most cases. Different methods are used to increase the motivation; one of the most present in today's field of research is the application of educational and serious computer games.

With data mining in e/m-learning a personalized approach to each student can be achieved in a way that it can optimally adapt to both, students who possess sufficient knowledge as well as students who have difficulty in acquiring new material. This is done either by accelerating the path of learning or by offering additional learning sources. In addition to the impact on the quantity and quality of content, this enables new experience in the acquisition of knowledge, which is particularly interesting to new generations of students.

Further research in this area might involve the following stages of development: creating the application and the associated data system in HTML5 ecosystem for easier use on smart mobile devices, processing certain topics (e.g. programming) through computer games and knowledge acquisition through levels, and integration with the existing data from the MudRi system.

Data mining can provide us with a range of answers on the relation between students' activity in the system and their final grade. The efficiency of the optimized learning model can be verified by using the model for processing different generations of students, as well as different courses. Previous studies on data mining in e/m-learning systems mainly relate to a particular educational institution or a course. This raises the question of generalization. Hence, it would be interesting to conduct a study that would synergistically involve both, similar and different universities in Europe and worldwide.

ACKNOWLEDGMENT

This research has been supported under the Grant No. 13.13.1.3.03 of the University of Rijeka.

REFERENCES

[1] C. Romero and S. Ventura, "Educational data mining: A survey from 1995 to 2005," Expert systems with applications, vol. 33, no. 1, pp. 135-146, July 2007.

[2] J. L. Moore, C. Dickson-Dean, and K. Galyen, "e-Learning, online learning, and distance learning environments: Are they the same?," Internet and Higher Education, vol. 14, no. 2, pp. 129-135, March 2011.

[3] M. Ally, "Foundations of educational theory for online learning," in Theory and practice of online learning, 2nd ed., T. Anderson, Ed. Athabasca, Canada: AU Press, 2008, pp. 15-44.

[4] "E-learning: advantages and disadvantages," Innovative learning technologies [Online]. June 13, 2011. Available: http://www.innovativelearningtechnologies.com/?p=80. [Accessed: January 27, 2014].

[5] J. Cheon, S. Lee, S. M. Crooks, and J. Song, "An investigation of mobile learning readiness in higher education based on the theory of planned behavior," Computers & Education, vol. 59, no. 3, pp. 1054-1064, November 2012.

[6] F. Martin and J. Ertzberger, "Here and now mobile learning: An experimental study on the use of mobile technology," Computers & Education, vol. 68, pp. 76-85, October 2013.

[7] S. Tang, M. Henneghan, and A. E. Rhalibi, "Introduction to games-based learning," in Games-based learning advancements for multi-sensory human computer interfaces: Techniques and effective practices, T. Connolly, M. Stansfield, and L. Boyle, Eds. Hershey, PA: IGI Global, 2009, pp. 1-17.

[8] M. Prensky, "The games generations: How learners have changed," in Digital game-based learning, Minnesota: Paragon House, 2007, pp. 35-65.

[9] F. Fong-Ling, S. Rong-Chang, and Y. Sheng-Chin, "EGameFlow: A scale to measure learners' enjoyment of e-learning games," Computers & Education, vol. 52, no. 1, pp. 101-112, January 2009.

[10] T. Hainey, T. Connolly, M. Stansfield, and E. Boyle, "The differences in motivations of online game players and offline game players: A combined analysis of three studies at higher education level," Computers & Education, vol. 57, no. 4, pp. 2197-2211, December 2011.

[11] L. S. Sica, A. D. Veneri, and O. Miglino, "Exploring new technological tools for education: Some prototypes and their pragmatical classification," in Methodologies, tools and new developments for e-learning, E. Pontes, Ed. Rijeka, Croatia: InTech, 2012, pp. 108-128.

[12] D. Oblinger, "The next generation of educational engagement," Journal of interactive media in education, vol. 8, pp. 1-18, May 2004.

[13] I. Kovačević, M. Minović, M.. Milovanović, P. O. Pablos, and D. Starčević, "Motivational aspects of different learning contexts: My mom won't let me play this game...," Computers in Human Behavior, vol. 29, no. 2, pp. 354-363, March 2013.

[14] N. J. Whitton, "An investigation into the potential of collaborative computer game-based learning in Higher Education," PhD Thesis, Edinburgh Napier Univ., Edinburgh, Scotland, 2007.

[15] R. Lawrence, "Teaching data structures using competitive games," IEEE Trans. Edu., vol. 47, pp. 459-466, November 2004.

[16] C. Kazimoglu, M. Kiernan, L. Bacon, and L. Mackinnon, "A serious game for developing computational thinking and learning introductory computer programming," Procedia – Social and Behavioral Sciences, vol. 47, pp. 1991-1999, February 2012.

[17] J. Han and M. Kamber, Data mining: Concepts and techniques, 2nd ed. San Francisco, CA: Morgan Kaufmann Publishers, 2006.

[18] A. Kruger, A. Merceron, and B. Wolf, "A data model to ease analysis and mining of educational data," in Educational Data Mining 2010, 3rd International Conference On Educational Data Mining. Pittsburgh, PA, 2010, pp. 131-140.

[19] "Database schema introduction," Moodle [Online]. Available: http://docs.moodle.org/dev/Database_schema_introduction. [Accessed: January 31, 2014].

[20] C. Romero and S. Ventura, "Educational data mining: A review of the state of the art," IEEE Trans. Syst. Man Cybern. C, Appl. Rev. vol. 40, pp. 601-618, November 2010.

[21] C. S. Loh, "Information trails: In-process assessment of game-based learning," in Assessment in game-based learning, D. Ifenthaler, D. Eseryel, X. Ge, Eds. New York: Springer, 2012, pp. 123-144.

[22] C. Romero, S. Ventura and E. Garcia, "Data mining in course management systems: Moodle case study and tutorial," Computers & Education, vol. 51, pp. 368-384, August 2008.

E-education 2.0: students' digital identity and online learning activities

Snježana Babić, Mario Krešić and Karla Kućel
Polytechnic of Rijeka/Business Department, Rijeka, Croatia
{sbabic, mkresic1, kkucel}@veleri.hr

Abstract - In general, the implementation of second generation web technologies enabled the user to have an active role in creating and sharing online resources on global scale with the ability to self-manage the personal online reputation and create different digital identity in the virtual world. Privacy, responsibility, trust and security are just some of the problems of the new technology. Unlike the implementation of e-learning 1.0 in higher education - which enables students to manage their online identities within the virtual learning environments created at one of the traditional platforms (e.g. learning management system Moodle) - in e-learning 2.0 there is a need for directing the students in the process of online identity self-management. Even though certain technical solutions can be found on the market (e.g. new generation e-learning systems which offer access to numerous Web 2.0 tools), it is considered that the digital identity is the important component of e-learning 2.0 in organizational, pedagogical and social sense. This paper describes an example where online service Google Drive has been used for executing online learning activities, and which preceded the study of student general opinion about digital identity and online reputation in e-learning 2.0.

I. INTRODUCTION

The term "*e-education 2.0*" stands for the area of e-education in which the second generation of web technology is used.

In numerous works, Bubaš et al. [1] confirm many advantages of introducing Web 2.0 tools to online learning activities (e-tivities): "*fostering collaborative and peer-to-peer learning, enrichment of learning experiences and the development of ICT related competencies"*. Furthermore, Zanamwe et al. [2] assert that social network technologies help students in developing social skills in different forms of online communication.

Along with the positive effects in e-education, Bubaš et al. [3] list potential disadvantages of using Web 2.0 technology: *students' concern about privacy, security and misuse of their personal information*.

On the other hand, Dalsgaard i Paulsen [4] claim that the possibility of higher transparency of collaborators' personal data has a positive effect on the quality of educational work. Therefore, the authors [4] point out the importance of the possibility of choosing the *transparency levels* with regard to privacy of personal information and participants' contribution and the problem of finding the optimum level of transparency.

Higher education teacher perceives the usability of Web 2.0 technology in the learning process through its educational usability, however, seen from the organizational aspect of higher education institution there is a problem of controlling the access to new technologies [5].

In the first generation of e-education there was a significant developmental improvement in e-learning systems (e.g. Moodle, Blackboard and others) which enabled the controlling of the educational process and users by the higher education institutions. In the closed learning communities the possibility to manage and control the user access is enabled. Although in traditional e-learning systems there is a possibility of using the Web 2.0 tools (e.g. wiki), the new generations of e-learning systems have been created which offer better choice of Web 2.0 tools (e.g. *Elgg*; see Orehovački et al. [6]).

Even though Web 2.0 technology enabled a more open and flexible approach to e-education, Strawbridge [5] states that many higher education institutions restrict the use of Web 2.0 technologies or limit their original purpose by user access control.

In general, Web 2.0 tools with educational potential are innovation in higher education practice and its implementation makes the higher education teacher one of the most important figures in the educational process with certain level of responsibility in organizational, pedagogical and technical sense.

Based on the above mentioned, this paper poses the following question: *to which extent is the teacher responsible for the way the Web 2.0 tools are implemented in the educational process and what is the responsibility of students while creating their online identities and using them in online learning activities?*

The paper continues with the overview of the researched literature, followed by the description of online learning activity using the Web 2.0 technology which preceded the research. Subsequently, the paper describes methodology and results of the research conducted on the full-time first-year students from different departments of the Polytechnic of Rijeka.

II. LITERATURE REVIEW

A. Digital identity and online reputation

Oztok et al. [7] state that „*identity is simultaneously an individual, social, and cultural phenomenon and that identity is constructed at the intersection of these locations*“. Costa and Torres [8] inspect digital identity, as the new phenomenon, through 2 macro areas:

- *presentation*: the way in which an individual publishes personal information, interacts and communicates in the virtual communities;
- *reputation*: other participants' perception of the individual's online behavior.

Mansouri and Mrabet [9] single out the possibility of monitoring digital identity as the mixture of the following "digital footprints":

- *passive "digital footprints"*: databases containing personal information (such as name, phone number, address, interests, profession, political affiliations, etc.) without any intentional intervention from the user;
- *active "digital footprints"*: voluntary and intentional posting and sharing of the content (articles, reviews, videos, photos, opinions, tweets, "the likes", comments and all the sharing on private and professional social networks).

Furthermore, it is important to point out that the possibility of posting "*positive and negative*" content by other people in virtual environment may influence the development of individual's digital identity [9].

As an example, viewed from the security aspect of social networks, De Paula [10] determines the key characteristic of social identities as "*the ability of the individual to share different social information with different groups*". Online reputation management and tagging technologies can help the users in estimating the trustworthiness of online information about other users (according to [10]).

B. E-education 2.0 and the problem of digital identity

The implementation of Web 2.0 technology in educational practice generates the possibility of creating multiple digital identities of all its participants.

De Paula [10] emphasizes that Web 2.0 applications (e.g. blogs, wiki, forums, social networks) enable each user to publish a great number of information simultaneously in different online locations. Therefore, the new technology leaves the users to choose the behavior and the manner in which they will „*build online trust and facilitate the identification of resources within a trusted social network*“ [10].

In her work, Costa and Torres [8] points out that even though creating the digital identity is not a direct goal of Web 2.0-based educational activities, they do represent the ways of creating digital identities in the very beginning of their use.

Students in different forms of educational practice in e-education 2.0 have the ability to create new identities presenting incorrect personal information in various virtual environments, which indicates the possibility of greater diversity of digital identities [11]. For that reason, Bati and Atici [11] introduce the term "*Identity 2.0*" which they consider to be crucial for the new type of socialization process.

Oztok et al. [7] confirmed there is a significant influence of individuals' different identities on their online learning practices, i.e. the influence of self-representation on individuals' identities creation.

It is important to point out that Microsoft in conjunction with *Data Privacy Day 2011* conducted a research in 5 countries (United States, Canada, Germany, Ireland and Spain) whose results confirm that only 44% of adults „*actively contemplate the consequences of their online activities on how other people see them*“, and mere 38% of interviewees „*think about the effect their online activities may have on the reputations of others*“ [12].

On the sample of 414 students, Bati and Atici [11] examined the effect of the Web 2.0 technologies on their identity constructions and found that 82.4% „*use at least one of the Web 2.0 tools*“; 67.1% of students „*do not distinguish between online and offline and do not have another personality online*“; and 71.7% „*think that others can find their personality totally in Web 2.0 environments*“.

Mansouri and Mrabet [9] conducted a research on the sample of 500 students and found that „*83.8% of the respondents use Facebook to share information, 54.4% use it to share videos, and 45.2% use it to share pictures*“ with the limited access to the content (only 26.4% of them), 42% of the students interviewed state that they never show their true identity, 50.8% of the students assert that they know about the "*online reputation*" concept. In this research the authors concluded that the national strategy influences the student perception of the importance of online reputation management.

Unlike the physical traditional classroom, digital identity management is particularly prominent in virtual learning environment 2.0, according to [8]. Therefore the authors stress that the participants of educational process have the task of finding the optimum "*model behavior*" while the teacher plays the key role.

C. The example of online learning activity: creating the project task using Web 2.0 tool – Google Document

During the academic year 2013/2014, the full-time students enrolled in different departments (Information science, Transport, Entrepreneurship) were using a free *Google Docs* office suite within *Google Drive* service for creating, sharing and posting educational content as a part of the course "*Basics of computer science*".

Students were supposed to create educational content based on a certain IT topic through a collaborative work. The online activity was assessed as a part of the educational element "*in-class activity*". Each student had the ability to independently choose an educational topic previously posted in the online course "*Basics of computer science*" which was created in the University's LMS Moodle. The students could not see the list of other student applicants while choosing the members of their group. There was a possibility of student collaboration in physical or other virtual environments outside the educational process.

After selecting the topic, the students were shown the online list of their group members and had to choose the group leader. The group leader needed to determine the method of collaboration for creating the new document which involved creating the online document within Google Drive and sharing the document with team members, including the teacher (shown in Fig. 1).

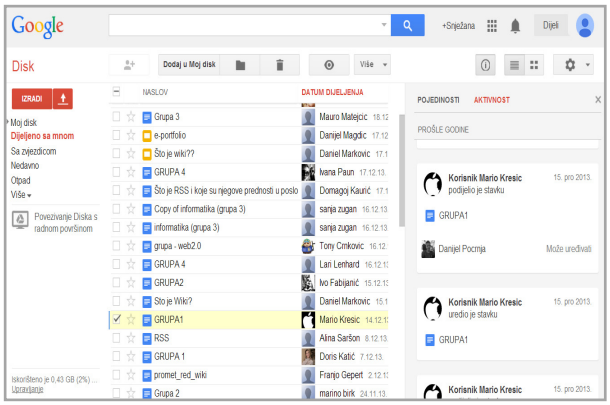


Figure 1. Application of Google Drive (Google Docs) while working on project tasks as part of the educational element "in-class activity" of the "Basics of computer science" course

The teacher had access to the collaborative teamwork using the features of Google Drive service itself. Additional internal communication among students took place on Facebook where teacher had no access. In the traditional classroom all group leaders presented their collaborative work.

The instructions about work methods were given to students in traditional classroom, as well as in online course created in LMS Moodle e-learning system. However, students were not given directions on the choice of user account they will use in an online educational activity using the Google Drive tools, which is closely related to their digital identities.

The practice has shown that students used the existent user accounts which they use privately. Also, it is important to mention that most of the students did not have experience with collaborative work method while creating the document using Google Docs, and that they did not fully understand the way teacher will supervise their work. This can indicate different levels of students' computer literacy, but it can also demonstrate a degree of awareness regarding monitoring their online activities by other people in virtual environment, which is related to personal management of online reputation.

In this case, the practice illustrates the diversity in creating digital identities by students, both in the sense of posting personal information as well as leaving „*active digital footprints*" [9] which is in accordance with the findings of the previously researched relevant literature.

III. RESEARCH METHOD

The basic problem of this research is to examine to which extent are the students familiarized with the online reputation management and their awareness of the importance of digital identity and online reputation in e-education 2.0. The research was conducted on the Polytechnic of Rijeka, after some time has passed since their involvement in online learning activity described in the previous chapter. It is important to emphasize that the students had already been assessed for their effort in the project.

For this research the method of opinion poll was used. It was prepared as the paper form. The data selection was carried out in student groups, according to their departments, and took place in a traditional classroom. After initial instructions were given, the surveying was performed by the examiner (addressed to the student). A new opinion poll form was created with 19 questions, and in its creation results from previously researched relevant literature described in this paper ([2], [3], [7], [8], [9], [10], [11], [12]) were used. In the first part of the questionnaire, besides basic demographical information (age, gender, status, department), there were a more closed-type questions related to the use of social networks and creation of user accounts, as well as self-evaluation about the knowledge of online reputation management which is closely connected to digital identity (according to [2], [9], [10], [11], [12]). Furthermore, the second part of the questionnaire included 10 questions concerning student evaluation of the importance of digital identity and online reputation in e-education 2.0 which was measured using Likert scale (1=strongly disagree to 5=strongly agree).

The data was collected from 194 students at the *Polytechnic of Rijeka*, among whom 19 did not complete the question form correctly. The final number of interviewees in this research was 175 students whose age was between 18 and 23, among them 74 women and 101 men. All examinees were full-time first-year students of *Information science* (31%), *Entrepreneurship* (22%) and *Transport* (47%).

IV. PRESENTATION OF RESULTS

As shown in Fig. 2, the interviewed students have evaluated their knowledge of online reputation management in e-education 2.0 as good (57%) with the average grade of 3.2 (SD=0.67). This may indicate that the majority of students is not familiarized enough with the general concept of online reputation management.

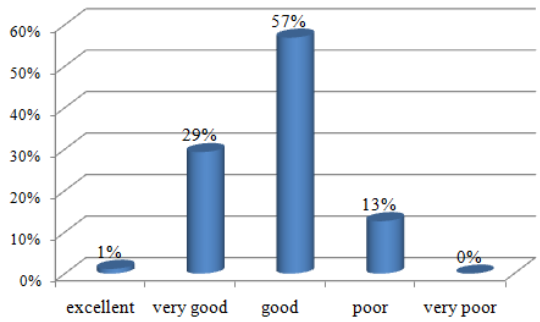


Figure 2. Percentage of responses regarding students' self-evaluation about the knowledge of online reputation management in e-education (N=175)

The research results show that 99% of students have Facebook accounts while other social networks are used considerably less: Google Plus (54%), Twitter (15%), Myspace (5%), LinkedIn (1%), others (19%). It can be concluded from the practice that the number of created user accounts on Google Plus social network has significantly risen because of the need of having that account for the previously described online learning activity. However, the results show that 46% of the students still do not have Google Plus user account which indicates that the students shared the document created using Google Drive via other user accounts.

Regarding the question about the number of user accounts on Facebook, 87% of students answered that they have only one user account which confirms previous findings by other authors indicating that students do not separate work life and private life in virtual environment (see [11]).

It is evident from the Fig. 3 that students use social networks mostly for online communication with other students at faculty (33%), followed by social activities (27%) and education (18%). We can conclude that social networks are used for education (51%), the conclusion also reached by Zanamwe et al. [2] in their study which was used in creating that particular survey question.

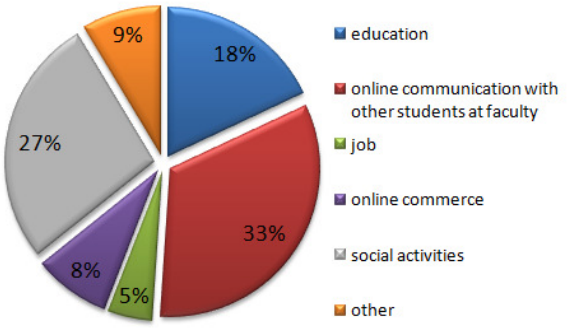


Figure 3. Percentage of student responses regarding the reason for using social networks (N=175)

Student opinion about the importance of digital identity and online reputation in education 2.0 has been examined by their level of agreeing with certain statements which they had to express using the Likert scale from 1 to 5.

Therefore, 68% of students „strongly agree“ and „agree“ with the statement that “*a student's digital identity influences the other people's perception of that student's personality*”, however, 25% have answered that they are “neither agree nor disagree” (see Fig. 4) which may indicate an insufficient awareness about the subject.

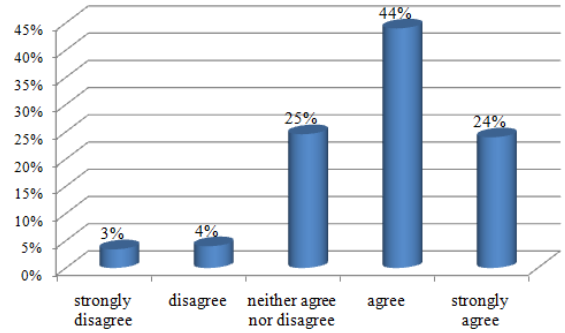


Figure 4. Percentage of student responses regarding the statement about digital identity influencing the individuals from the environment (N=175)

The student opinion about “*...the influence of online reputation on their long-term goals (e.g. during the employment process or after graduation, etc.)*” is very positive (79%). However, 13% of students are unsure regarding the statement, and 8% of students have negative opinion which is considered a significant percentage because they are the ones who can cause negative online reputation of others, which is in accordance with the results from the previously reviewed relevant literature (see [12]).

In this research the students have shown their highly developed awareness of the need to protect personal information in the field of e-education 2.0. This is confirmed by 88% of students who feel it is necessary to take into consideration personal data protection. However, to the statement regarding “*...creating new user account for the educational purposes to protect student privacy*” 28,6% of students are “neither agree nor disagree” and 16% disagree with the statement (see Fig. 5). Such result may indicate the fact that students trust Web 2.0 technology regarding the protection of their personal data.

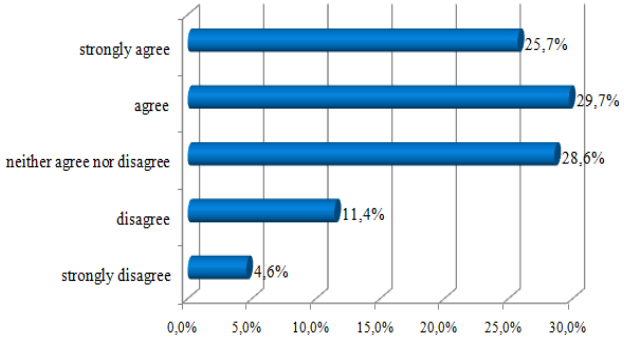


Figure 5. Percentage of student responses regarding the importance of creating new user account for the educational purposes to protect their personal data (N=175)

Furthermore, 46% of students have the negative opinion about the colleagues who post inappropriate data and information about themselves and others (text, images, video, audio, etc.) on social networks (e.g. Facebook), while 31% opted for the response "neither agree nor disagree". The gathered results could be connected to the fact that Facebook is the most commonly used network and there is a thin line between physical and virtual environment where individuals act in different ways, which is in accordance with the researched relevant literature in the paper (see [11]). At the same time, 50% of the students "strongly agree" and 27 % "agree" with the statement that "*…while publicly posting personal data it is necessary to consider appropriate names*" which can also be related to the results of the previous statement and highly developed awareness of the importance to protect personal data.

Since the students took part in the online learning activity which preceded this research, they could use that collaborative experience as basis for their evaluation of how "*…individual responsibility while posting personal data and creating mutual trust during working in online social communities influences the quality of collaborative work*". The results show that 61% of students agree with the mentioned statement. It is interesting that only 22% of students "strongly agree" with the claim which can indicate that the students perceive their colleagues on the friendly basis as well, while tolerating their behavior in the virtual environment (see Fig. 6).

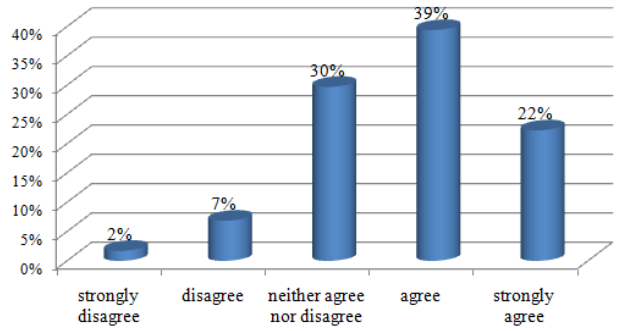


Figure 6. Percentage of student responses regarding the statement about the influence of digital identity on the quality of collaborative work in an online environment (N=175)

To the statement about "*...guidelines in the process of managing online reputation while using social software in educational process*" 68% of students expressed positive opinion (see Fig. 7), which indicates the high level of awareness of potential problems regarding the integration of the second generation web technology in the educational practice. It can also be concluded that in e-education 2.0 students have certain expectations from the teachers in terms of managing online reputation, which is in accordance with the researched relevant literature in this paper (see [9], [12]).

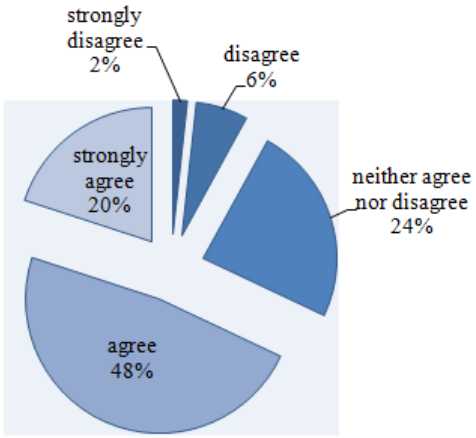


Figure 7. Percentage of student responses regarding the statement about the need for guidelines on how to manage personal online reputation in the educational process (N=175)

V. CONLUSION

Based on the reviewed relevant literature it can be concluded that the *digital identity* and *online reputation management* are important components of *e-education 2.0*, whose influence can be monitored on the level of a certain participant in the educational process, institutional, national and global level. Also, it is important to emphasize that in the past few years the effort has been made to develop individual awareness of protecting the online privacy and online responsibility on national and global level which is confirmed by the research results in this paper. Second generation web technology enables students to be highly creative in constructing their digital identities, but at the same time to create a number of different identities as well.

However, unlike the physical world, in the virtual environment individuals – with their online activities – can take part in creating other people's digital identities (see [2], [11], [12]). The research results in this paper confirmed that students use social networks mostly for educational purposes and that majority have only one user profile. Also, they take into consideration the amount of personal data they post thus protecting their privacy. It is a valid conclusion that students believe it is sufficient to protect their personal information within the service being used. On the other hand, it can be asserted that students do not actively think about the possible influence of others on the creation of their online image, and vice versa, which is in accordance with the existent results in the reviewed relevant literature.

While introducing new web technology into the educational practice, the higher education teacher takes a part of the responsibility in guiding the students towards reaching the learning goals which include the expected "*behavior model*" (see [8]). This can be affirmed viewed from the aspect of teacher experience in educational practice while using Google Document online tool, which is described in this paper. On the other hand, viewed from the student aspect, the same is confirmed by research results – 68% of students want to be guided in the process of managing their online reputation for the educational

purposes. Most of the gathered results are in accordance with the findings of the existent studies. The future studies should encompass broader range of questions.

REFERENCES

[1] G. Bubaš, A. Ćorić, T. Orehovački, „Strategies for implementation of Web 2.0 tools in academic education“, 17th European University Information Systems International Congress, EUNIS 2011, Dublin, Irelan, 2011.

[2] N. Zanamwe, T. Rupere and O. Kufandirimbwa, “Use of Social Networking Technologies in Higher Education in Zimbabwe: A learners’ perspective”, International Journal of Computer and Information Technology,2(01), 2279-0764, 2013.

[3] G. Bubaš, A. Ćorić, T. Orehovački, “The Evaluation of the Use of Online Community Tool Ning for Support of Student Interaction and Learning”, Proceedings of the 21st Central European Conference on Information and Intelligent Systems, pages 171-178, Varaždin, Croatia, 2010.

[4] C. Dalsgaard and M. F. Paulsen, “Transparency in cooperative online education”, The International Review of Research in Open and Distance Learning, 10(3), 2009.

[5] F. Strawbridge, “Is there a case for web 2.0 in higher education? do the benefits outweigh the risks?”, Assignment for Introduction to Digital Environments for Learning, 2010.

[6] T. Orehovački, M. Konecki, D. Radošević, “Web 2.0 i evolucija e-obrazovanja, Zbornik radova stručno - znanstvenog skupa "E - obrazovanje", str. 145 - 155., Varaždin, Hrvatska, 2007.

[7] M. Oztok, K. Lee and C. Brett, “Towards Better Understanding of Self-Representation in Online Learning”, In T. Bastiaens & G. Marks (Eds.), Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2012(pp. 1867-1874), 2012., Chesapeake, VA: AACE. Retrieved February 10, 2014 from http://www.editlib.org/p/41879.

[8] C. Costa and R. Torres, “To be or not to be, the importance of Digital Identity in the networked society”, Educação, Formação & Tecnologias-ISSN 1646-933X, 47-53, 2011.

[9] Z. Mansouri and Y. Mrabet, “Moroccan University Students’ Online Reputation Management”, International Journal of Education and Literacy Studies (IJELS), Australian International Academic Centre PTY. LTD, Vol 1, No 1, 45-54, 2013.

[10] A. M. G. de Paula, “Security aspects and future trends of social networks”, In Proceedings of the Fourth International Conference of Forensic Computer Science (pp. 66-77), 2009.

[11] U. Bati and B. Atici, “Web 2.0 or Identity 2.0: The Roles of Web 2.0 Tools on the Identity Construction of Turkish Youth”, In Z. Abas et al. (Eds.), Proceedings of Global Learn 2010 (pp. 3683-3689). AACE. Retrieved February 10, 2014 from http://www.editlib.org/p/34456., 2010.

[12] Microsoft, “Online Reputation Management: Parents and Children 8-17 (November 2011, United States, Canada, Germany, Ireland, and Spain)” http://www.microsoft.com/security/resources/ research.aspx#onlinerep [20.01.2014.]

Usefulness of Prezi and PowerPoint presentation

A. Špernjak
Faculty of Natural Sciences and Mathematics, University of Maribor, Maribor, Slovenia
E-mail: andreja.spernjak@uni-mb.si

Abstract - In everyday teaching practice great helping to the clearness and visualization of the content are presentations. Two decades ago, the blackboard notes were replaced with traditional slides and overhead projector. With the availability of computer hardware and software are traditional slides converted to electronic slides. In our area the most frequently are used PowerPoint slides, but the problem is payable package of the ´Microsoft Office´. Users looking for free of charge and useful presentation programs. For teachers and students is Prezi a free online presentation service provider, where the applicability in four years grow up to more than 500 million users. The article presents the results of the usefulness of PowerPoint and Prezi presentations among university biology students. They mostly use the PowerPoint presentations because generally they do not know Prezi presentations. Students also expressed the advantages and disadvantages of one and the other presentation program.

I. INTRODUCTION

Computer is an indispensable element of everyday life. It became a must in terms of the education and instruction system. In courses it helps teachers to be more efficient in auditory. For more visual senses and in order to achieve permanent, effective and fluent teaching most teachers use the PowerPoint, which, included in the ´Microsoft Office´ package is a powerful presentations tool. Originally PowerPoint was developed for commercial and business purposes, but it has quickly penetrated the scientific and educational circle as well.

Why teachers use the instructional presentations, for entertaining or education?

Science employs an iconic sign system [1] and have visualisations [2] and argumentation deeply rooted into the study material. In science classes, these visual representation enable scientific meaning to be constructed and make meanings that no teacher or student can readily convey in a different mode of communication or in one mode alone. These visualizations would affect the shaping of knowledge in a positive way [3]. Especially science teachers should therefore use graphic organizers in presentation during which they display the explicit structure of the information covered by previous classes [4]. Some students and many teachers reported that presentations might prove to be more diversion than useful tool, because the slideshows presentations capture students' interest – pay attention or helps them ″stay awake″ in class [5].

Many studies have examined the effect of PowerPoint presentations on motivation, self-efficacy and the academic successes of students, and many studies have compared PowerPoint with traditional classes. Different results have been reported. Reference [6] argued that the self-efficacy of students is high and their attitudes are positive in classes taught with PowerPoint presentations. Reference [7] found that the attitudes of students towards PowerPoint presentations were positive, and that PowerPoint presentations have an impact on short-term memory when designed appropriately. In contrast, reference [8] compared classes using PowerPoint presentations with teacher cantered classes, and found no statistically significant difference between the two methods according to students' academic success. PowerPoint has a negative effect on students' exam performances, although students state that they learn more in the courses taught with PowerPoint [9]. The results [10] of study show that most of the students become more successful when they are taught with traditional method. However, PowerPoint has neither a positive nor a negative effect on students' academic success.

Many studies have compared effect of PowerPoint presentation with traditional classes. PowerPoint presentation has already received much criticism from researchers, but just a few researchers examined effects between PowerPoint and some other presentation like Xerte, Keynote, VoiceThread or Prezi. Reference [4] compered effect of PowerPoint slides with Prezi presentation. They found out there is no significant differences in learning outcomes, self-efficacy, cognitive load and two motivational variables (attitude toward behaviour, intrinsic motivation) between PowerPoint and Prezi presentation. Prezi is relatively new, so the technology is still unfamiliar to researchers and educators, especially in the field of social work. Table I represents general software options of PowerPoint and Prezi presentations.

TABLE I: Comparison of general software options of PowerPoint and Prezi presentation.

General options of	PowerPoint	Prezi
linear sequence	✓	✓
a free-flowing presentation – effect of zooming	×	✓
animations between slides	✓	✓
creating own templates	✓	✓
capability of integrating text	✓	✓
capability of integrating images	✓	✓
capability of integrating animation	✓	✓
capability of integrating audio	✓	✓
capability of integrating video seamlessly into a single presentation	✓	✓
autoplay a single animation or slide	✓	✓*
printing slides	✓	×
free of charge	×	✓
need of high-speed Internet connection	×	✓
publishing presentations online	×	✓
shared, allowing multiple people to	×	✓

collaborate on a single presentation		
support for different plug-in (mathematic, chemistry, physic, …)	×	✓

*possibility just three different times of autoplay (4, 10 or 20 sec)

The question is why use the Prezi if with studies found out no differences between presentations? Traditional presentation software like PowerPoint is requires preparing a linear story line using a storyboard approach (Figure 1).



Figure 1: Linear sequence of PowerPoint presentation.

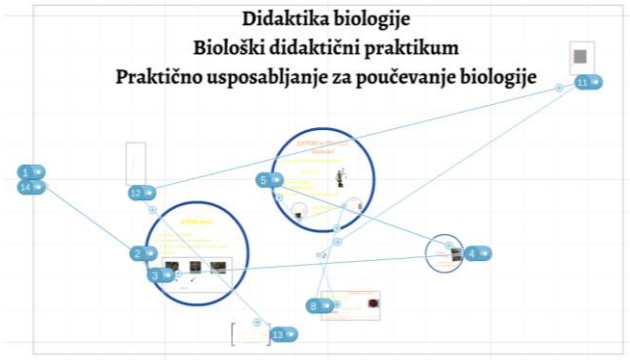


Figure 2: Free-flowing Prezi presentation.

Prezi, on the other hand, allows for both a linear and a free-flowing presentation of a story line (Figure 2). The user creates a presentation on a large blank workspace called the canvas, where all the elements of a presentation are visible. A story line is then created by arranging the elements on the canvas [11]. With Prezi user can group the visual elements info frame and zoom in and out on different objects. It allows making multi-dimensional presentation. The creation of a Prezi presentation does not require the user to have advanced in information communication technology skills, so they can interact on-line for creating highly impactful virtual portfolios [4] Moreover, Prezi encourages using graphic organizers (concept maps, mind mapping, non-linguistic representation) and these could be an alternative to the linear sequencing in PowerPoint presentation [4]. This is also an opportunity to revitalize the interest in graphic organizers, which are not frequently implemented in classrooms [13].

The aim of the current study was to survey biology students´ opinions on and attitudes to PowerPoint and Prezi presentation usage. We wanted to find out which one they use more frequently in reasons why.

II. METHODS

The participants for this study were student biology who volunteered for this study. They were from the 3th and 4th university years at the Faculty of Natural Sciences, University of Maribor in Slovenia. They completed a questionnaire prepared for the purpose of the study. They completed it in electronic form. The quastionare was prepared with Google Drive. We collected 63 questionnaires from 47 women (74.6 %) and 16 (25.4 %) men.

The questionnaire contained three parts. The first part of the questionnaire solicited demographic data such as gender. The second part concerned students' attitudes about usefulness of different presentations (Prezi and PowerPoint). The answers were measured by a 5-point Likert scale. The scale questionnaire was as follows: 0 = I have no base for answer; 1 = definitely disagree; 2 = disagree; 3 = agree; 4 = definitely agree. Students were to value percent's of their university professors' usage of presentations during lectures. The scale questionnaire was as follows: 1 = 0 – 24 %; 2 = 25 – 54 %; 3 = 55 – 74 %; 4 = more than 75%; 5 = all professors. The third part was an open question about advantages and disadvantages of Prezi and PowerPoint presentation. Similar answers were grouped in to four complexes which describe the same answers.

The reliability of the questionnaire about usefulness of different presentations (Prezi and PowerPoint) was measured by the Cronbach coefficient. The Cronbach reliability coefficient was 0.94 for the 7-item scale, which can be considered satisfactory. Microsoft® Excel 2010 was used for data input. The analyses were performed with the statistical package SPSS 21.0, where we used descriptive statistics; One-way ANOVA was used to examine differences in opinions by student gender why they use presentations. The Chi-Square test (χ^2) was not used for comparison of differences between results from the student opinions about the advantages and disadvantages of PowerPoint and Prezi presentation because we have no results of Prezi usage.

III. RESULTS AND DISCUSSION

In a primary or secondary school 36 students (57.1 %) learned about presentation, 20 students (31.7 %) learned about presentation at university classes and just 6 students (9.5 %) claimed that they learned about presentation by them own across investigation. No one took the special lessons or training to learn about e-presentation. PowerPoint presentation use 57 students (90.5 %) and just 6 students (9.5 %) use PowerPoint and Prezi presentations. We have no students to use just Prezi. Students mostly do not use the Prezi because they do not know it and almost the same results are at professors' usage of Prezi presentation. We asked students who many professors at lectures use presentations and 53 (84.1 %) students claimed that more than 75 % of professors during lectures use the presentations but less than 10 % of them use Prezi presentation. Across my researching I found at that Prezi use just one professor and his field is computer science.

With questioner we asked students about advantages and disadvantages of Prezi and PowerPoint presentations but we could not made s comparison of differences between results because just 6 students know the Prezi presentation and these results cannot be relevant. Students who use Prezi claimed that it is good because it is free of charge and because of zooming in and out it is more interesting than PowerPoint. Why students do not use other presentation than PowerPoint is clear out of those statements:

- ″I saw Prezi presentation. It is very interesting but I do not use it because I do not know how″.
- ″I do not use the Prezi because all lecture room has no Internet connection″.
- ″I use PowerPoint presentation because I am familiar with it, it is accessible and mostly use it″.
- ″I will not use Prezi because it is too trumpery″.
- ″Prezi has no support for different plugins for Chemistry, Physic, Mathematic … but PowerPoint has″.
- "I use PowerPoint because I do not know the other presentation programs.″
- ″PowerPoint meets my all presentations needs and therefore I do not use other applications″.

PowerPoint is a user-friendly package [14] and our study also confirms it. 54 (85.7 %) our students said that it is easy to use PowerPoint presentation. There are no significant differences at statement of using PowerPoint presentations by gender. Students mostly use it for the creation of visually clear. Three students (4.8 %) use it because the professors request. Even 6 (9.5. %) students thing just of herself; they use PowerPoint to make their oral presentation easier. These students are not interest in audience and how to effective to represent the contents but just to do it and by easier way. Out of this item these students while preparing PowerPoint presentations do not follow the principles that need to be followed. Hence, presentations which are prepared without following the appropriate principles have a negative effect on learning [14]. According to our study students missing the lectures about appropriate presentation principles and they are not sure how to make a good and effective presentation.

Students need more instructions how effective make presentations. At the lectures they see finished presentations but they can to make it effective if they do not kwon how or if they do not follow the principles that need to be followed to make effective presentation.

IV. CONLUSION

Out of our study we can conclude that our students are no familiar with Prezi presentations although it is free of charge. They mostly use a powerful presentation tool, the PowerPoint software as like as those university professors. They use it first of all for better visualization of contents and self-efficacy. Out of the early findings is for motivation, self-efficacy or the academic successes not important which software for presentation is used; PowerPoint, Prezi or which other else. Most important factor for effective presentation is ″preparing the presentation with an appropriate design″ [15]. For development of effective presentation slide as multimedia instructional materials, teachers should conceptually understand design principles regarding what makes high- or low-quality presentation slides [16].

In further research we would like to represent Prezi presentation to our students and make them to use it. After while we would like to find out which one of software presentation; Prezi or PowerPoint; they would like to prefer.

REFERENCES

[1] K. De Westelinck, M. Valcke, B. DeCraene and P. Kirschner, ″ Multimedia learning in social science: Limitations of external graphic representations″ Computers in Human Behavior, vol. 21 (4), pp. 555 – 573. 2005.

[2] Y. Kali and M. Linn, ″Designing effective visualisations for elementary school science″, Elementary School Journal, vol. 109 (2), 181 – 198. 2010.

[3] G. Kress, C. Jewitt, J. Ogborn and C. Tsatsarelis, ″Multimodal Teaching and Learning. The Rhetorics of the Science Classroom. London, New York: Continuum. 2001.

[4] J. Casteleyn and A. Mottart, ″Presenting material via graphic organizers in science classes in secondary education″. Procedia – Social and Behavioral Sciences, vol. 69, pp. 458 – 466. 2012.

[5] A. Hill, T. Arford, A. Lubitow and L. M. Smollin, ″I´m Ambivalent about It: The Dilemmas of PowerPoint″. Teaching Sociology, vol. 40, pp. 242 – 256. 2012.

[6] J. E. Susskind, ″PowerPoints power in the classroom: Enhancing students' self-efficacy and attitudes″. Computers & Education, vol. 45 (2). 203 - 215. 2005.

[7] H. Nouri and A. Shahid. "The Effect of PowerPoint Presentation on Student's Learning and Attitudes." Global Perspectives on Accounting Education, vol. 2, pp. 53-73. 2005.

[8] G.T. Rosenthal, B. Soper, R. Mcknight, J. Barr, J. Wilkinsin and C. W. Bergen, ″Multimedia. It's how you use it.″ Computers in the School, vol. 26, pp. 77-86. 2003.

[9] R. A. Bartsch and K. M. Cobern, ″Effectiveness of PowerPoint presentations in lectures″. Computers & Education, vol. 41, pp. 77-86. 2003.

[10] A. V. Can, N. Karaca, N. Akyel and S. D. Demirci, ″Evaluating the Fitness of Lecturing with PowerPoint Presentation for Accounting Education – research at Sakarya University″. Social and Behavioral Sciences, vol. 55, pp. 128 – 137. 2012.

[11] B. E. Peron and A. G. Stearns, ″A Review of a Presentation Technology: Prezi″, Research on Social Work Practice, vol. 21 (3), pp. 376 – 377. 2011.

[12] T. D. Chicioreanu and G. C. Oproiu, ″New ways to present the information in the teaching activity″, The 8th International Scientific Conference eLearning and software for Education Bucharest, April 26 – 27, pp 90 – 95. 2012.

[13] I. Kinchin, ″If concept mapping is so helpful to learning biology, why aren´t we all doing it?″ International Journal of Science Education, vol. 23 (23), pp. 1257 – 1269. 2001.

[14] C. Uz, F. Orhan and G. Bilgiç, ″Prospective teachers´ opinions on the value of PowerPoint presentations in lecturing″, Procedia Social Behavioral Sciences, vol. 2, pp. 2051 – 2059. 2010.

[15] J. Holzl, ″Twelwe tips for effective PowerPoint presentations for the technologically challenged″. Medical Teacher, vol. 19(3), pp. 175-179. 1997.

[16] Y. H. Cho and S. E. Lee, ″The role of co-explanation and self-explanation in learning from design examples of PowerPoint presentation slides″. Computers & Education, vol. 69, pp. 400 – 407. 2013.

An On-line Solution Verification System for Programming Competitions

A.S. Dorofeev*
* National Research Irkutsk State Technical University, Irkutsk,
Institute of Cybernetics, Department of Computer Engineering, Russian Federation
Dorbaik2007@mail.ru

Abstract – An application of an on-line solution verification system for programming competitions is considered in the paper. The system has been developed at in National Research Irkutsk State Technical University and used within the competition events. It has client-server architecture and has been implemented on the base of a cluster using parallel schemes of verification algorithms. The system supports several programming languages, such as C++, Pascal, Java. The verification algorithm can also be used to check the results of laboratory works of "Programming" disciplines. The system allows teacher to automate routine activities of students' algorithms correctness verification.

I. INTRODUCTION

Today the student competitions on different disciplines are an integral part of the educational process. The competitions are becoming more and more popular. They let the faculty stuff check the acquired competencies, allow finding talented young people, develop creative abilities, increase an interest in the research work, *etc.* [1]. In addition, for the competition winners, wide career opportunities are opened. Often, the outstanding students become senior developers and even technical directors of software development firms just after only two or three years after the graduation [2].

Every year, a number of programming competitions are held in Department of Computer Engineering of the National Research Irkutsk State Technical University [3]. The winners of regional and All-Russian competitions participate the quarterfinal and semifinal competitions of the world team championship on programming under the auspices of ACM.

To involve more students to the competitions the faculty stuff improves the quality of the competition organization, in particular, minimizing the time period of the task solution verification and expert assessment. Earlier, the process was organized as follows: department staff prepared the terms of the problems to be solved, their correct solutions and test cases before the competition. The competitors solved the problem using personal workstations, and placed their solutions in a specific directory on the local disk of the workstations. At the appointed time, the organizers collected participants' solutions (programs) and distributed them among inspectors. Each inspector executed a program in an environment prescribed the by the programming language, verified its operability and executed the prearranged test case. The test case usually consisted of 10-15 tests, and covered boundary value checks, the productivity of the proposed solution, and other aspects of the solution quality. The verification of the solutions took from about 3 to 5 hours for a large number of the participants.

The results of the test cases execution were recorded in a protocol. The protocol data included the number of tests passed. The results were aggregated by the participants and by the problems, and the winner and two medalists in the individual competition have been determined. If a competition was of an intercollegiate or All-Russian character, the winning team is determined by three best participants of a university paying no attention to their original team memberships.

The next day was dedicated to work of appellate commission; the participants could file requests for reconsideration of their solutions. Then, the final results of the competition are determined after possible appeals.

Naturally, this approach did not exclude the human factor. Mistakes made during the verification and in the final standings could appear. In addition, it was difficult to estimate the productivity of a solution, the number of the program versions passing the test cases. As a result, the probability of the fact that some participants of the competition will occupy the same place is increased. In this case one must pay more attention to the assessment of the quality and the productivity of the program code, *i.e.*, the results became depended upon the expert's intuitive activity.

In order to rise the organization quality of the competitions we needed to solve the following problems: construct a storage, protocols and programming interfaces for automatic solutions accumulation; create an on-line verification engine of the solution variants of a competition participant (an participant can suggest several variants of the solution, and all of them must be checked); develop a time tracking and penalty time accounting subsystem. So, we decided to develop a programming system for the routine activities automation, which is also aimed at exclusion of the human factor in the distributions of the places in the final standing.

II. A PROGRAMMATIC SOLUTION

The first step towards the problem solution was to write auxiliary programs as batch scripts, which allow

verifying the solution on a set of predesigned test cases. These scripts conserved greatly the time taken to check the solutions. The next step was to create a system of automatic solution verification of the competition problems. The system has been developed by collective of the department of Computer Engineering. The system has a client-server architecture and is written in Java.

To get acquainted with the procedure of solution submission before the main round of the competition, a trial round is being conducted with one simple problem, *e.g.*, addition of two numbers. After the familiarization with the system of on-line solution verification, the main round is started, and all the participants receive the problems at the same time. Average number of problems varies from 6 to 8. The complexity level also varies: there are always one simple encouraging problem and at least one complex. The variety of the complexity levels develops tactical skills of the participants. Many solutions at the competition are based on the knowledge of algorithms of discrete and other branches of mathematics. The time of the main tour is usually from 4 to 6 hours depending on the level of the competition.

The solution is accepted if it is solved completely. The completeness of the solution is determined by the number of the test cases passes. The solution to be verified must be represented in one of the following programming languages: Java, C, C++ and Pascal; The following IDEs also supported: Eclipse, Code::Blocks, Free Pascal IDE and Delphi.

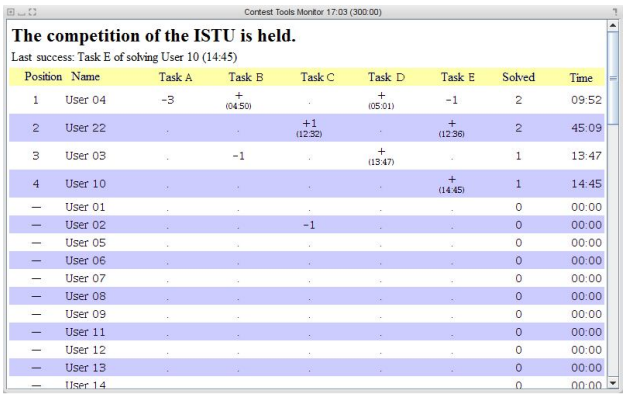
A competition participant submits the solution by means of opening the corresponding file or copying its text from the clipboard. The solution is accompanied with the language name; it is selected by the participant before submitting the solution. All submitted programs are queued on the server side. The verification results are displayed to each participant with a message and to all participants with a standings positions in an application.

Possible responds of the verification system are as follows:

- The solution is accepted.
- Wrong output: the algorithm is probably wrong.
- Data representation error: the program produced malformed output due to, *e.g.*, the incorrectness of the solution, or the task is incorrectly specified.
- Time is out. Possible causes of the respond could be low performance of the submitted algorithm; an error that led to a program hang; the program waits user input after producing output.
- Exceeded the process memory limits. The possible cause is that the program allocates memory, but does not free it.
- Runtime error. This can be produced due to an error in the program, the absence of `return 0` statement in C/C++ programs; the solution could be incorrectly specified; the program consumes too much memory; Java class contains mo method `main()`, *etc*.
- Compile error. Probable causes are as follows: a syntax error in the program; incorrectly specified the implementation programming language, a class having `public static void main (String [])` method not found in a Java program.
- Access violation. The program attempted to execute an action prohibited by the rules of the competition. In this case, a disqualification of the participant could be initiated by jury.

As mentioned above, the solutions of each problem are verified on predefined test cases. The tests are executed in a random order. The test cases are executed by programming system "Checker" that had been developed by competition organizers. Solution of the problem is accepted if it passes all tests. If a test case failed the system returns the number of the test case to the author of the solution. The results of some test cases are presented with the problem statement as examples of input and output of the program to be written. The verified solution is accepted with the time spent by the author to develop the solution: it is assumed that the solution has being developed since the very start of the main round. A so-called penalty time is calculated as a sum of the elapsed times for all accepted problem solutions for each of the competitors. Unsuccessful attempts of a solution submission results in penalty time increase by 20 minutes for each unsuccessful attempt if the solution is finally accepted. If a solution is not accepted, its penalty time is not accounted. So, the penalty time consists of the time spent to produce a solution for each accepted problem and a penalty time for each failed attempt of a solution submission if it is finally accepted.

An hour before the end of the main round, monitoring the place of a participant in the standings is stopped. However, the team leaders and jury can watch the results and the places of the participants during the full time of the round on a special workstation. The final standings positions are distributed according to a number of solved problems, and the penalty time if the number of tasks solved is the same. At the team competitions the places are distributed similarly by taking into account the total number of problems solved by top three members of a command and their total penalty times (Fig. 1).


Contest Tools Monitor 17:03 (300:00)

The competition of the ISTU is held.

Last success: Task E of solving User 10 (14:45)

Position	Name	Task A	Task B	Task C	Task D	Task E	Solved	Time
1	User 04	-3	+ (04:50)	.	+ (05:01)	-1	2	09:52
2	User 22	.	.	+1 (12:32)	.	+ (12:36)	2	45:09
3	User 03	.	-1	.	+ (13:47)	.	1	13:47
4	User 10	.	.	.	.	+ (14:45)	1	14:45
—	User 01	.	.	.	.	.	0	00:00
—	User 02	.	.	-1	.	.	0	00:00
—	User 05	.	.	.	.	.	0	00:00
—	User 06	.	.	.	.	.	0	00:00
—	User 07	.	.	.	.	.	0	00:00
—	User 08	.	.	.	.	.	0	00:00
—	User 09	.	.	.	.	.	0	00:00
—	User 11	.	.	.	.	.	0	00:00
—	User 12	.	.	.	.	.	0	00:00
—	User 13	.	.	.	.	.	0	00:00
—	User 14						0	00:00

Figure 1. Interface of the stangings monitor application

At the end of the competition a rating is figured out, the monitor restarts showing the places of the participants in the final standings, but at this time it shows the real names and universities of their origin instead of the encoded name. After that the prizes are distributed.

The next step was implementing the system according to a parallel scheme (Fig. 2) to verify the solutions on multiple workstations [4] simultaneously; the workstation form a horizontal cluster of undedicated nodes. The usage of the cluster significantly reduced the response time with respect to verification of the queued solutions on one server workstation. The participant might wait for a response about 3-5 minutes. The cluster reduced the time to less than one minute that is very important at the very end of the competitions.

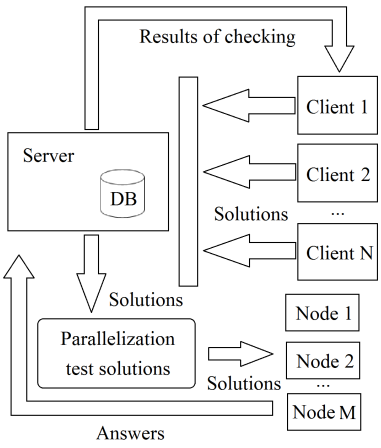


Figure 2. Distribution of the verification tasks between cluster nodes

Now an idea of the system utilization in the class of programming appeared. Viewing the source codes of solutions of student's laboratory work is a rather time-consuming procedure, and it is not always possible to find logical errors in the program. Also it does not always provide an objective monitoring and assessment of the learning outcomes.

In the laboratory works, the programming tasks is often required to be written as a program in one of the programming languages that reads input data in a specific format, processes the data in accordance to the task formulation; the result must be displayed in a particular format as output. This form of user interaction of the laboratory program is typical for the programming competitions. So the developed system is suitable for automatic inspection of the solutions obtained in laboratory works. The test cases and an obviously correct solution are prepared preliminarily.

III. CONLUSION

The benefits of the presented automated approach to the organization of the competition events are obvious: there are less number of routine activities; effectiveness of the final standings results formation; impartiality of the verification; the rating is calculated with formal rules; more detailed registration of the results of the competitions, in particular, all submitted solutions are stored in the database in the sequence of their submission, fixing the submission time within one second; the probability that some participants will take the same place is decreased; to announce the results of verification and the final standings monitor device is used, forming and issuing rankings more efficiently.

Thus, the paper presents the experience of an application of an automatic verification of solutions of problems stated at student programming competitions. The automation reduces the routine activity of the competition organizers on viewing and verification of the problem solution represented as program source code. The system provides an independent and objective verification. It provides teachers and students a convenient means of viewing quality statistics of the submitted solutions. In addition, the system can be used to demonstrate the testing techniques for the students studying the information technologies.

REFERENCES

[1] A.V. Tukhmanov and M.I. Kovalenko, “On the role of programming competitions in the self-determination of youth“, Proceedings of participants IX scientific conference-exhibition “Information Technology in Education“. Rostov-on-Don: Rostizdat, 2009, pp. 133-134 (in Russian).

[2] A.A. Shalyto, “Why do we need champions Programming?“ in Computerra, vol. 14, 2008, pp. 22-24 (in Russian).

[3] A.S. Dorofeev AS, E.V. Ryazanov and N.N. Egorova, “Experience in conducting programming competitions using the system on-line cheching of the solutions“, Proceedings of IV All-Russian Conference “Wiener's reading“. Irkutsk: Irkutsk State Technical University, Part III, 2011, pp. 67-71 (in Russian).

[4] A.S. Dorofeev AS and E.V. Ryazanov, “Verification system solutions programming tasks in real time“, Proceedings of the X International Scientific-Practical Conference “Science and creation - 2013/2014“, vol. 34 “Modern information technologies“ in Prague: Education and Science, pp. 52-55 (in Russian).

Implementation of information competencies as key employment skills at the Faculty of Humanities and Social Sciences

Špiranec, Sonja*, Ivanjko, Tomislav * and Pečarić, Đilda *·
* University of Zagreb, Faculty of Humanities and Social Sciences,
Department of Information and Communication Sciences
sspiran@ffzg.hr, tivanjko@ffzg.hr, dpecaric@ffzg.hr

Abstract - In order to improve the quality of their study programs, universities/faculties today are facing increasing demands to include employment skills into their curricula in order to maximize the potential of their students for a successful career. One of the key aspects in developing such skills is increasing students' employability - a set of skills, understandings and personal attributes that make graduates more likely to gain employment and be successful in their chosen occupations, which benefits themselves, the workforce, the community and the economy. While there is a broad consensus amongst key stakeholders in the process (HEI, employers, students) on the importance of addressing employment skills within higher education, there are still open questions on how this can be achieved. While universities/faculties tend to put more emphasis on developing specific knowledge and skills within their field of education, a number of reports issued by employers urged them to make more explicit efforts to develop the "transferable" or "generic" skills needed in many types of high-level employment. In an ever increasing world of information, as one of the core components of these generic skills is developing information competencies – skill set needed for effective gathering, evaluating and communicating information in all its forms and contexts. This paper aims to explore how these information competencies are implemented at the Faculty of Humanities and Social Sciences, University of Zagreb by conducting research across its study programs on the masters' level. Following that approach, this research will provide a snapshot of current situation and give recommendations on possible optimizations in including information competences into the curricula at the faculty level.

I. THEORETICAL BACKGROUND

In the wake of rapid growth in higher education and the increase in global market competition experienced by many employers, universities came under intense pressure to equip graduates with more than just the academic skills traditionally represented by a subject discipline and a class of degree. A number of reports issued by employers' associations and HEI organizations urged universities to make more explicit efforts to develop the 'key', 'core', 'transferable' and/or 'generic' skills needed in many types of high-level employment [1].

Head et al., [2] list studies have been published about employers' perceptions of information literacy needs throughout the workplace [3] [4]. In these studies, most employers were unfamiliar with the phrase "information literacy", yet they agreed it was important for employees to be able to find, evaluate, and use information. Employers frequently had a difficult time distinguishing information literacy and research skills from other skills such as presentation, writing, and technology competency. Authors conclude that these studies show that employers place a higher premium on graduates' information competencies, but the results do not say which information competencies are important in the workplace and why [2].

In defining what those skills should be some authors are using the term "core" for discipline-specific skills, and "generic" to represent "the so-called transferable skills that can support study in any discipline" [5]. Developing those generic skills is increasing students' employability - a set of skills, understandings and personal attributes that make graduates more likely to gain employment and be successful in their chosen occupations, which benefits themselves, the workforce, the community and the economy [6]. For some employers, even the degree subject studied is not as important as the graduates' ability to handle complex information and communicate it effectively and that graduate recruiters want a variety of other skills, personal and intellectual attributes, rather than specialist subject knowledge [7].

In an ever increasing world of information, as one of the core components of these generic/transferrable skills for graduate students is developing information competencies – skill set needed for effective gathering, evaluating and communicating information in all its forms and contexts [8]. In this context, becoming information literate is considered a product that can be measured against a set of educational standards. Claims emanate from the library and education sectors about the generic and transferable nature of skills gathered under the term "information literacy". The concept of information literacy as a skills-based process is increasingly being adopted by training colleges and universities as a "graduate outcome" [9]. *Information literacy competency standards for higher education* [8] provide a framework for identifying what an information competent/literate individual is able to do:

- Determine the extent of information needed
- Access the needed information effectively and efficiently
- Evaluate information and its sources critically
- Incorporate selected information into one's knowledge base
- Use information effectively to accomplish a specific purpose
- Understand the economic, legal, and social issues surrounding the use of information, and access and use information ethically and legally

Although these competencies are related to information technology skills, they have broader implications for the individual, the educational system, and for society, because "the information literacy of young people has not improved with the widening access to technology: in fact their apparent facility with computers disguises some worrying problems" [10].

In the tertiary education landscape the generic qualities of information literacy are emphasized. Information literacy is viewed as a suite of skills which constitute a competency, developing student information skills which can be applied across a range of academic disciplines and which prepare students for continued formal learning throughout life [11].

To provide an overview how these information competencies are implemented at the Faculty of Humanities and Social Sciences at the University of Zagreb, a research across its study programs on the masters' level is conducted. Following that approach, this research will provide a snapshot of current situation and give recommendations on possible optimizations in including information competences into the curricula at the faculty level.

II. METHODOLOGY AND SAMPLE

After consulting key documents on the topic, the *Information Literacy Competency Standards for Higher Education* (ALA, 2000) was selected as a joint framework for research on which the terminology and the definition of key components of information competencies were based. Following the information competencies present in the Standards, research across study programs [1] at the masters' level of the Faculty of Humanities and Social Sciences, University of Zagreb was conducted in order to explore how these information competencies are implemented in the projected course outcomes.

At the faculty students can study 4 master studies from the area of Social Sciences (Information and Communication Sciences, Pedagogy, Psychology, Sociology) and 30 master studies from the area of Humanities[2].

A total of 34 master study programs were analyzed with a total of 1752 courses. Each course description and outcomes were analyzed and investigated for potential inclusion of generic information competencies in its curricula.

III. RESULTS

From 34 master programs analyzed, in 9 of them no generic information competencies were found. All of 9 studies belong in area of Humanities. Since these studies, all except one can be studied only as double major we assume that students other major could provide courses with generic information competencies.

Out of 1752 courses in 128 (7.3%) courses in 25 study programs generic competencies were identified which makes an average of 3.8 courses per study.

Further analysis of generic information competences was based on those 128 courses in which they were found.

A. *Social science vs. Humanities*

When comparing the presence of courses with generic information competence in Social Sciences and in Humanities, we found that all 4 studies (100%) in Social Sciences have generic information competencies, and 21 (70%) out 30 studies in Humanities have courses that include generic information competencies.

Number of courses with generic information competencies in Social science was 31 (13%) out of a total of 236. Since one of studies in Social Science is Information and Communication Sciences we assumed that it will have more courses with generic information competences than other studies. Analysis confirmed that assumption showing that generic competencies were found in 16 (52%) courses in Information and Communication Sciences, which is more than all other 3 Social Sciences studies combined (pedagogy, psychology, sociology) where they were found in 15 (48%) courses.

Number of subjects with generic information competences in Humanities was 97 (6.4%) out of total number of 1506.

Distribution of total number of courses per study, number of course with generic information competences (GIC) and percentage of courses with generic information competences (GIC) of total number of courses in study is shown in Fig. 1.

[1] Graduate programs at Faculty of Humanities and Social Sciences, University of Zagreb. URL: http://www.ffzg.unizg.hr/programi/diplomski.html (20.12.2013)

[2] Anthropology, Archaeology, Art history, Comparative literature, Croatian studies, Czech language and literature, Dutch studies, English studies, Ethnology and cultural anthropology, French language and literature, German studies, Greek language and literature, History, Hungarian studies, Italian language and literature, studies of India-based languages, Latin language and literature, Linguistics, Philosophy, Phonetics, Polish language and literature, Portugal language and literature, Romanian language and literature, Russian language and literature, Slovak language and literature, South Slavic languages and literature, Spanish language and literature, Swedish language and culture, Ukrainian language and literature, Turkish studies.

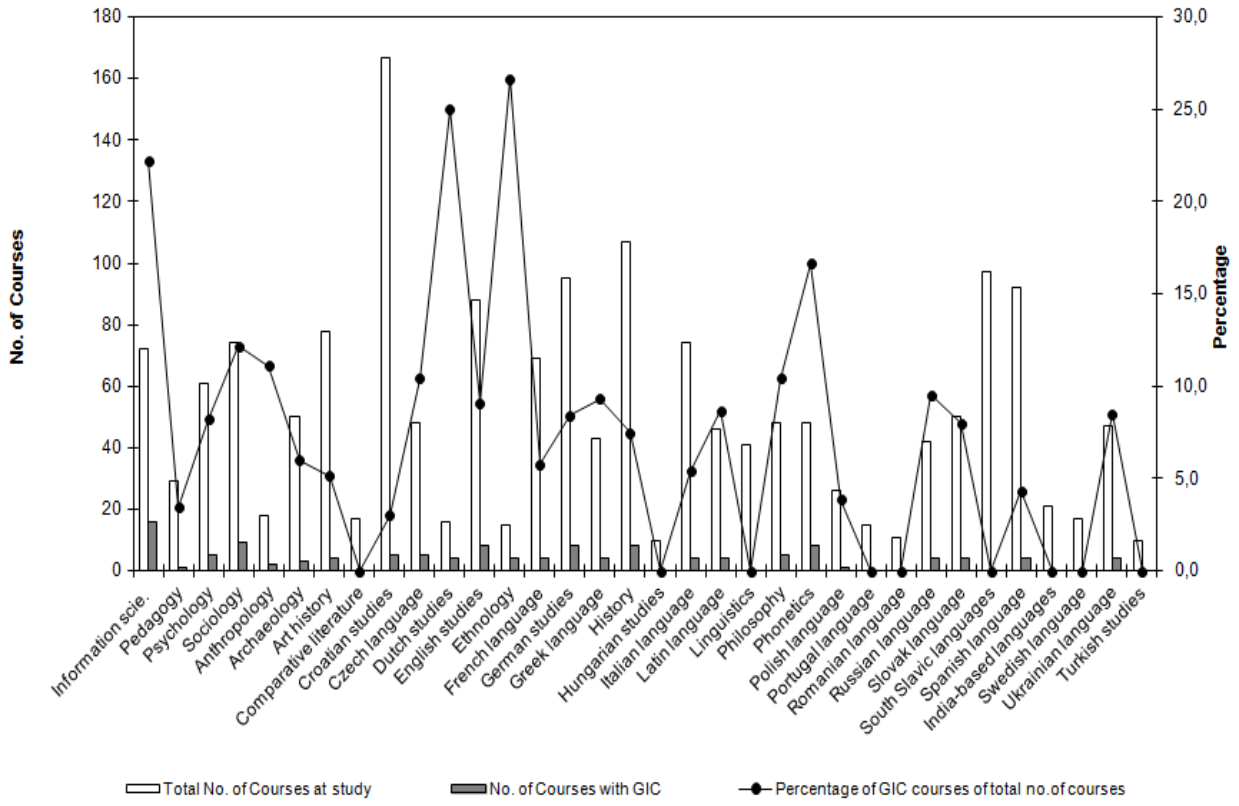


Figure 1. Distribution of courses with generic information compentences by studies

B. Mandatory vs. Elective Courses

Since all of the studies analyzed have mandatory and elective courses, it was important to further analyze the presence of generic competencies based on those criteria. If competencies are present in any of the mandatory courses we can say that all students from that study will be obliged to acquire them. Elective courses, however students can pick and chose, which means that students will adopt those competencies only if they chose to enroll in that course.

It was shown that only 9 studies have courses with generic information competencies in its mandatory curricula. Out of 9 studies 2 of them are Social science studies and 7 are Humanities studies.

TABLE I. DISTRIBUTION OF COURSES WITH GENERIC INFORMATION COMPENTENCES BY TYPE OF COURSE

Area	**Number of courses**				
	Mandatory courses		*Elective courses*		**Total**
		Teaching module		Teaching module	
Social sciences	14	2	11	4	**31**
Humanities	11	17	17	52	**97**
Total	**25**	**19**	**28**	**56**	**128**

Out of 128 courses with generic information competencies 44 (34.4%) courses are mandatory and 84 (65.6%) courses are elective courses (table 1).

Social sciences have 16 (52%) mandatory courses out of 31 courses with generic information competencies, and 15 (48%) elective courses. On average social sciences have 4 mandatory courses per study and 3.8 elective courses per study in which generic competencies were found. Most of the mandatory courses 11 (69%) are courses from Information and communication Sciences, while other 5 are all from Sociology.

Humanities have 11 (29%) mandatory courses out of 97 courses with generic information competencies, and 69 (71%) elective courses. On average Humanities have 0.9 mandatory courses per study and 2.3 elective courses per study in which generic competencies were found.

C. Teaching module

Each of 34 master studies can have they own specialization. For example English study has 4 specializations in graduate program: Literature and Culture, Linguistics, Translation and Teaching. Each student chooses only one specialization in order to achieve graduate degree. In all 34 master studies we found 80 different specializations across studies.

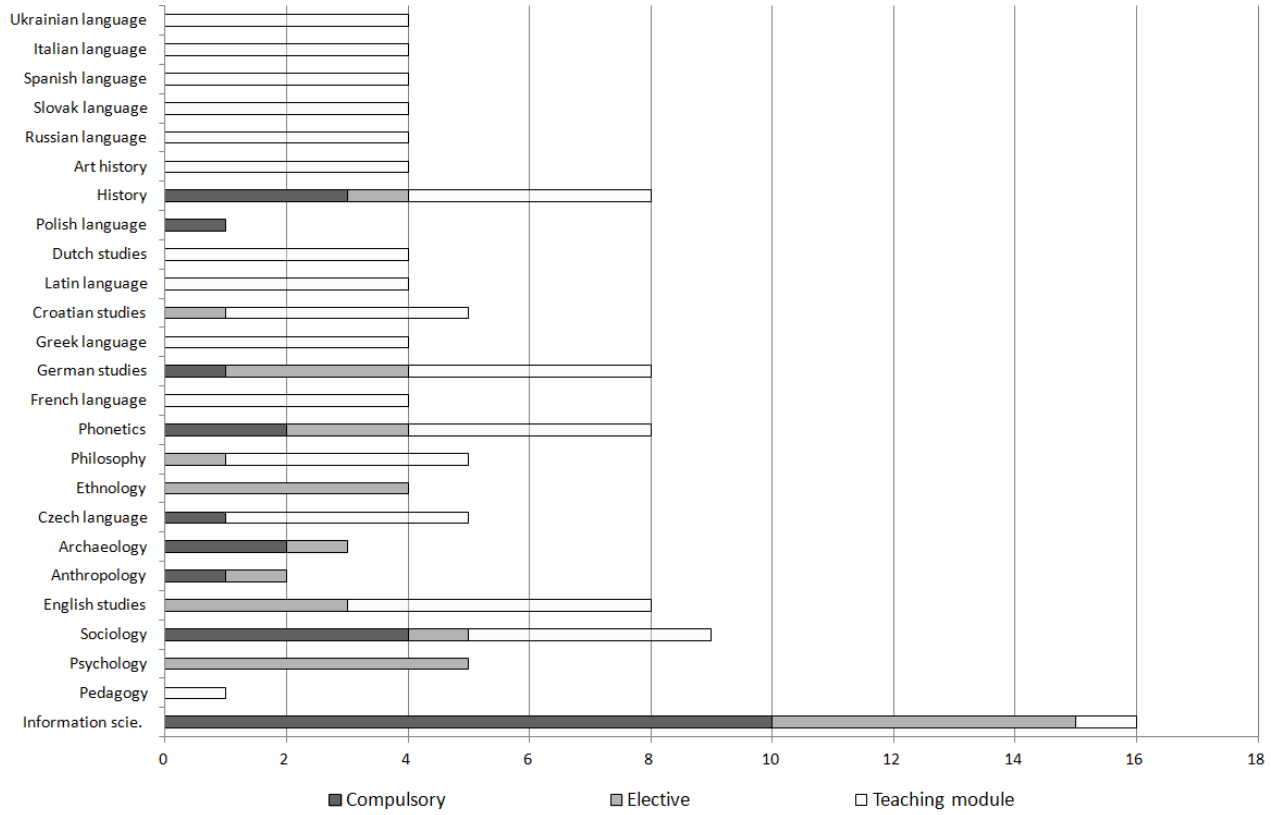


Figure 2. Distribution of courses with generic information compentences by type of course

Analyzing 128 courses with generic information competencies, we found that a big part of those courses (75 courses or 58.6%) belong to the teaching module. Only students interested in obtaining a teaching degree have to enroll in that module.

From 75 courses in teaching module only 19 (25.3%) are mandatory courses and 56 (74.7%) are elective courses. In social science 50% of courses in teaching module are mandatory courses. While in Humanities 25% of courses in teaching module are mandatory courses (Fig. 2).

Also, 11 studies offered courses with generic information competencies that are included only in teaching module: 1 (25%) in social science and 10 (33.3%) studies in humanities.

In we reexamine data from mandatory and elective courses by including courses from teaching module also into electives we can see that most of the courses with generic information competencies will be elective courses (Fig. 2). Out 128 courses with generic information competencies 25 (20%) courses are mandatory and 103 (80%) courses are elective courses i.e. elective courses and courses from teaching module. Social sciences have 14 (45%) mandatory courses out of 31 courses with generic information competencies, and 17 (55%) elective courses. Humanities have 11 (11%) mandatory courses out of 97 courses with generic information competencies, and 86 (89%) elective courses (Fig. 2).

IV. CONCLUSION

This paper provided an overview how information competencies are implemented across study programs on the masters' level at the Faculty of Humanities and Social Sciences at the University of Zagreb.

A total of 34 master study programs were analyzed with a total of 1752 courses. The results have shown that 9 study programs from the area of Humanities do not teach their students generic information competence, and only in 128 (7.3%) courses in total generic competencies were identified.

Analyzing 128 courses with generic information competencies, we found that a big part of those courses (75 courses or 58.6%) belong to the teaching module which only students interested in obtaining a teaching degree have to enroll in, an in addition only 19 (25.3%) of them are mandatory courses.

Lowden, et al. [12] argue that developing graduate employability skills and attributes should be included in HEIs' strategic and faculty/departmental level planning. Universities need to promote employability skills and attributes in their mission statements, learning and teaching strategies, course framework, strategic documents and practical guidance. The analysis has shown that most of the study programs rely on either elective courses or teaching module to teach information competencies to students, with the logical exception of Information and Communication Sciences study program. Being as it is means that a large majority of students will adopt those competencies only if they chose to enroll in

some of the elective courses offered at the faculty, but their study program mostly offers only "core", discipline related skills.

We can conclude that information competencies as a vital part of the employability skills aren't strategically incorporated at the Faculty level and that the implementation of information competencies is still not recognized as a vital part of the curricula being only sporadically present across different study programs.

REFERENCES

[1] G. Mason, G. Williams, and S. Cranmer. "Employability skills initiatives in higher education: what effects do they have on graduate labour market outcomes?", Education Economics, vol. 17, no. 1, 2009, pp 1-30.

[2] A. J Head, M. Van Hoeck, J. Eschler, and S. Fullerton, S. "What information competencies matter in today's workplace?", Library and Information Research, vol. 37, no. 114, 2013, pp 74-104.

[3] L. Klusek and J. Bornstein. "Information literacy skills for business careers: matching skills to the workplace", Journal of Business & Finance Librarianship, vol. 11, no. 4, 2006, pp 3–21.

[4] J. Sokoloff. "Information literacy in the workplace: Employer expectations", Journal of Business and Finance Librarianship [online], vol. 16, no. 1, 2012, pp 1–17.

[5] N. Bennett, E. Dunne, and C. Carré. "Skills development in higher education and employment". Buckingham: SRHE and Open Univerisity Press, 2000.

[6] M. Yorke. „Employability in higher education: what it is-what it is not", in Learning and Employabillity, vol. 1. York, England: Higher Education Academy, 2006

[7] P.T. Knight and M. Yorke. „Learning, Curriculum and Employability in Higher Education". London: Routledge, 2004.

[8] American Library Association (ALA). "Information literacy competency standards for higher education", 2000. Available: http://www.ala.org/acrl/standards/informationliteracycompetency (14.2.2014)

[9] A. Lloyd, A and M. Somerville. "Working information", Journal of Workplace Learning, vol. 18, no. 3, 2006, pp 186-198.

[10] CIBER. "Information behaviour of the researcher of the future", January 2008. Available: http://www.jisc.ac.uk/media/documents/programmes/reppres/gg_final_keynote_11012008.pdf (14.2.2014)

[11] A. Lloyd. "Information literacy landscapes: an emerging picture", Journal of Documentation, vol. 62, no. 5, 2006, pp 570-583.

[12] Lowden, K., Hall, S., Elliot, D., & Lewin, J. (2011). "Employers' perceptions of the employability skills of new graduates", 2011. Edge Foundation, University of Glasgow, SCRE Center. Availa ble: http://www.edge.co.uk/media/63412/employability_skills_as_pdf_-_final_online_version.pdf (14.2.2014)

Learning Management Systems: Trends and Alternatives

L. Abazi-Bexheti[*], M. Apostolova-Trpkovska[**], A. Kadriu[***]
[*]SEE University, Faculty of CST, Macedonia
[**]SEE University, Faculty of CST, Macedonia
[***]SEE University, Faculty of CST, Macedonia

l.abazi@seeu.edu.mk, m.apostolova@seeu.edu.mk, a.kadriu@seeu.edu.mk,

Abstract - One of the main difficulties in Learning Management Systems (LMS) is how to maintain the balance between learning and management. The primary aim of LMSs was to assist educators to manage learning. As a result of this way of thinking, features such as managing content, roster students, and similar, mostly were promoted. As time goes by, the accent of research is more in student experiences and efficacy of the tools. On the other hand some educators argue that as a result of evolving Web 2.0 applications, students can be better served by an LMS alternative, a toolbox of web resources that could contain social bookmarking tools, document sharing applications, social networking applications, timeline tools, and media options offered in the cloud. Primary to this approach is the idea that students should become more experienced with today's technology tools since these skills will be practical in the workplace. As a result, some institutions have begun to offer LMS alternatives, and some teachers are using these options to support their students' learning.

The paper analyses both sides of the latest approach toward LMSs and presents their implications for teaching and learning.

I. INTRODUCTION

The use of computer technologies in teaching and learning started several decades ago at the dawn of the information revolution. Though in the beginning the role of computers for education was an object of discussion, in the 90's the explosive use of personal computers, and the World Wide Web placed computer-assisted teaching and learning at the center of the educational process.

This trend has initiated a rapid development of Learning Management Systems (LMS) – learning platforms with a set of teaching and learning tools designed to improve student's learning by including computers and Internet in the learning process. An LMS covers almost all aspects of managing a course, from administrative functions to course content delivery and assessment. Actually an LMS makes it possible to the teachers to develop and deliver course resources and activities in an online learning environment and in the meantime students can access course content and participate in online learning activities.

From the perspective of LMS deployment in Higher Education institutions, the following three approaches are identified: commercial, open source and in-house development. While a few commercial vendors have attained considerable market share, development of in-house software and open source software are significant trends in higher education. The trend of in-house and open source often lies in apparent inadequate functionality or pedagogic limitations of commercial offerings, despite platform functionality becoming increasingly customizable. At the same time many higher education institutions show interest to maintain institutional autonomy over the development of systems that are key tool in teaching and learning, especially as they can represent valuable intellectual property [8], [10], [12].

Presently multiple LMS solutions for higher education can be purchased from the commercial sources such as the Blackboard Corporation, the leading provider in the United States with more than 50 percent of the educational marketplace, or obtained from the open-source community without any licensing cost. At the same time, European countries rely more and more on free open-source learning management systems based on their commitment to knowledge as a community advantage which should be available to all regardless their economic status.

II. CURRENT TRENDS AND NEEDS

Currently, many higher education institutions have deployed some form of Learning Management System. But the problem is that LMSs usually have been acquired without a comprehensible plan for what they are aimed or how they would impact the learning experience. The initial wave of LMS implementation to a large extent was done locally, with many faculties and other entities within the higher education institution deploying different systems [1], [2], [15].

As e-Learning becomes a mainstream activity for most institutions and for some fundamental to their strategy, the importance and profile of the LMS has increased. However, the existing LMS practices present a significant experience in both technological and teaching aspects [3], [13]. These practices, among other things, show that it is the LMS experience itself that enables many of the users to distinguish the possibilities in the technology and to ask

more of an LMS, by what they become the key drivers towards the LMS change and enhancement. Hence one of the main concerns in the LMS deployment in an institution is to identify user requirements [14]. Based on these requirements, many evaluation and assessment criteria for such systems can be developed.

It is remarkable that with a relative maturity of the LMSs, as the latest report on LMS trends from Brandon Hall Group[1] suggests, almost half of the institutions that have some LMS deployment are trying to improve their LMS or to move from their previously acquired LMS to a new product (Fig.1).

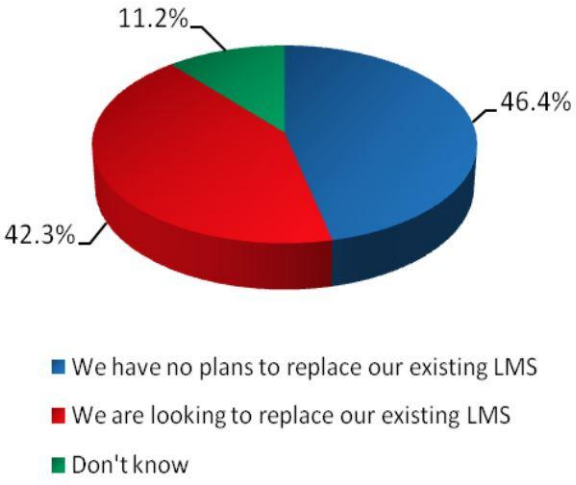


Figure 1. The percentage of institutions looking for a new LMS [7]

This approach is generally motivated either by an attempt to reduce the cost of licensing and maintaining the system or by dissatisfaction with the existing LMS. The dissatisfaction from the existing LMSs is mainly due to the lack of reporting features, difficultness of the system to use and poor support from the vendor. As shown in Fig.2 the requirements from the new LMS mainly include ease of use the system, robust reporting features and ability to integrate with other university software [7].

A recent trend in the LMS market is the rise of cloud-based LMS solutions that offer eLearning services in a distributed computing environment on the Web. There is one noteworthy case among them OpenClass, a cloud-based learning management system that educational publisher Pearson has offered for free as a joint venture with Google. An additional trendy cloud-based tendency in eLearning is the use of massive open online course (MOOC) offerings by either individual higher education institutions or 'consortium' platforms like Coursera [11]. The MOOC phenomenon could be placed within the framework of open educational resources movement focused on creation and distribution of free and open-source educational materials and supporting software [10].

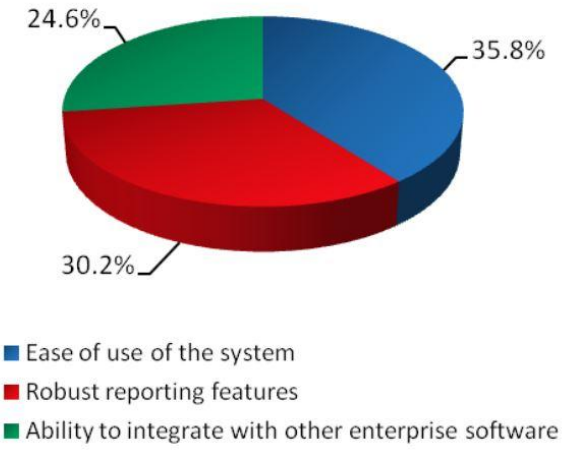


Figure 2. Top features for the new LMS [7]

With the advance of Web 2.0, contemporary LMSs moved towards integration with multiple Web 2.0 enhanced technologies such as Facebook, Twitter, visual and audio tools, and ePortfolio supporting software. Learning management systems also increasingly support mobile devices such as smart phones and iPad-like tablet devices.

III. WEB 2.0 IMPLICATIONS ON LMSS

Web 2.0 is principally accepted as a trend in the use of World Wide Web technology that aspires to increase creativity, information sharing, and particularly, collaboration among users. These concepts have led to the development and evolution of Web-based communities, such as social-networking sites, wikis, and blogs. One of the key principles here is the fact that users have the opportunity to add value and further harness collective knowledge. Consequently, the technology should be set in a manner so that it encourages this participation [6]. As far as the development of Web 2.0 applications is concerned, the following strategies are suggested [17]:

- Release early and release often. This statement of the open source development model is now a critical success factor for Internet-based software. It promotes the use iterative development methodologies to wrap up bug fixes and enhancements into incremental releases that respond to user feedback.
- Engage users as co-developers and real-time testers. The best way for assessing new product is through real-world user behavior. The nature of Web-based applications and the developers' possibility for active monitoring the software usage is a dramatic shift which gives the opportunity for more statistics and controlled experimentation used for informed product decisions.

[1] Brandon Hall Group is a preeminent research and analyst firm that provides world-class research, data and expertise in Learning & Development, Talent Management, Leadership Development, Talent Acquisition, and Human Resources.

- Instrument the product. In the development process, is important to plan and implement not only the customer-facing application but also a framework for capturing how the product is used. Usually, what users do inform more than what they say.
- Incrementally create new products. The product should change and evolve through frequent releases, user feedback and instrumentation.
- Use dynamic tools and languages. The development model would meet best the above mentioned strategies if the appropriately flexible development tools and languages are used. The use of platform-independent, dynamic languages will enable better adaptability to change, speed, and productivity.

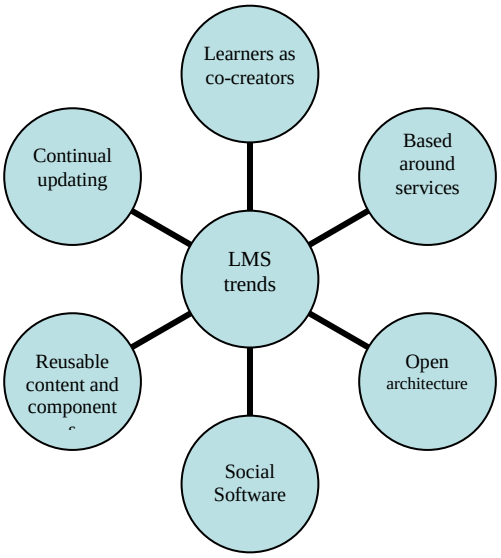


Figure 3. LMS trends

Based on the above mentioned Web 2.0 tendency, one could consider what the implications for e-Learning are and try to foresee the challenges for the future LMSs (Fig.3).

The primary challenge is how a LMS should be constructed and what are the technology aspects of LMS. Accordingly, there are two perspectives for the future LMS issue. On one side it is about the technological adjustments, and on the other side about the educational issues and how we approach and develop the modules for student use. So far the trend was on developing modules that meet the particular needs of a course or group of students. The accent nowadays is on developing modules that can be reused or joined in various ways. In order to achieve this, based on the above mentioned Web 2.0 strategies, new LMS modules initially should be available to certain group of users. And further on, based on its evaluation, the module is released for broader usage among all students [15]. In this manner the LMS develops into a medium for new technologies and good experience that can be dispersed across higher education institutions.

Although many of the Web 2.0 principles are in-line with open source approach towards LMSs, yet a future LMS can be a combination of open source with other existing in house systems in the institution [2],[19]. The key issue is that it is an open architecture, based on standards, and designed in a manner that can support the easy combination of modules mentioned above. Some of these modules may be open source and others in – house solutions. From their content point of view, many of these modules would be from the educational sector, but it is very possible that the future LMSs will incorporate other modules outside the educational sector such as Flickr or Google tools [18].

In context of Web 2.0's impact on educational experiences, the most important principle is that of the collective intelligence. It could be accomplished in many ways, but its main aim is to engage students as co-developers. An example of achieving this is by posting a certain course content on a blog or a wiki and giving the students the opportunity to comment, modify the course content, create shared bookmarks, etc.

Another issue, which is natural in the Web 2.0 mode of functioning, is reuse. It can be approached from two points of view regarding LMSs. On one hand reuse is important in terms of application with meaning reuse of tools and various components. On the other hand reuse can be applied to content as well. For instance, there are many learning object repositories and open content initiatives which aim to promote reuse.

IV. LMS ALTERNATIVES (ADVANTAGES AND WEAKNESSES)

Another interesting impact of the web 2.0 is in finding alternative applications for the traditional LMSs. Actually, some teachers argue that because of evolving Web 2.0 applications, students can be better served by an LMS alternative, a toolbox of web resources that might include social bookmarking tools, document sharing applications, social networking sites, timeline tools, and media options available in the cloud [4],[9].

Fundamental to this approach is the idea that students should become more familiar with contemporary technology tools because these skills will be crucial for them in the workplace. As a result, some higher education institutions have begun to offer LMS alternatives, and some teachers are using this choice to support their students' learning. The framework of an LMS alternative may offer the user a coordinating center with a dashboard or other interface that gives easy access to selected web-based tools according to course and project requirements.

LMS alternatives cover a wide range of tools and functions. One option might be a complex system built in-house at a higher education institution, designed to perform many of the functions of a traditional LMS while giving access to outside applications. Another one might be a mash up of web applications collected by an individual teacher and hosted from a blog platform or a social networking site.

The tools that faculty members select as LMS alternatives usually are free, easy to learn and use, and robust enough to support students learning without any difficulties. Ideally, the LMS alternative might integrate with applications already in use in the institution. In such a design, students could select from among the offered applications to complete their assignments.

The reason for seeking LMS alternatives is also the lack of the services from the traditional LMSs. In this sense, any faculty or higher education institution that supports blogging or uses a collaboration tool like Google Drive in a way is employing an LMS alternative.

The tools obtainable in the cloud offer students various advantages over those in the traditional LMS since web tools are designed to be used easily. The students can learn them rapidly, use them for multiple projects, and share them among their peers. Furthermore, the variety of the tools available makes it possible for teachers and students to choose the tool that best fits a learning exercise.

The weakness of the LMS alternative is that it is a time consuming process for the teachers and the students to evaluate new tools and match them with appropriate assignments or projects. The LMS control is not in the institution and this fact raises concerns about interruptions of service, the security of data and personal information, and the effect on an institution's reputation that could result from a security break.

From another point of view, the LMS is also a tool that makes it possible to monitor the course delivery and to generate various reports on the course content and teachers/student performance [20]. With LMS alternatives the institutions are limited and out of the course schema of a certain course. Also, institutions have little or no option if the alternatives are poorly maintained, shut down for repairs, or ceases to exist. In addition, any alternative LMS could raise issues about a lack of technical support, a lack of faculty support, an increasingly fragmented student learning environment.

However, regardless the drawbacks of the LMS alternatives, they have a positive impact on the teaching and learning process [6]. Actually the use of LMS alternatives encourages teachers and students to perceive learning as an evolutionary process where effective tools are emerging constantly. Such an environment prompts learners to become familiar with a variety of tools rather than any specific one [19]. At the same time, if teachers start adopting tools outside the LMS it may induce institutions to take a new better look to the LMS service they offer. In this manner, the institution may discover the advantages in providing access to a wide variety of tools that enable strong support for teaching and learning.

The use of LMS alternatives may hold the promise of a more student-centric approach, one that encourages students to reach across the limitations of academic terms and learning disciplines and to see their education as a coherent process that they can maintain using a range of applications. By going outside the LMS to use tools that provide more space for student engagement, more effective collaboration, and more active learning in general, teachers could establish new expectations for the LMS. While some of the features they seek from the Web 2.0 world might never make it into the teaching tool's collection of most of the teachers, they may still help push the LMS to its next iteration.

V. CONLUSION

The great interest for original pedagogical-educational practices is a reaction to the community needs for educational change. Such needs come out as a result of the huge request and massification in to higher education, the need to raise competitiveness, mainly through the raise of the human potential and the necessity to take into account new approaches to learning.

These original innovative practices in education presently are mainly available through e-Learning and the most pervasive e-Leaning tools in higher education currently are the Learning Management Systems. LMS have been present for about two decades in higher education. In spite of this they are still changing and evolving due to the technological advancements as well as the change in approaches towards teaching and learning in higher education institutions.

The latest reports show that LMSs have their common role, where a few standard features for supporting teaching and learning are still the most used features [5]. In parallel with their common usage there is a line of new trendy features present in current LMSs. These trendy features are usually blogs, wikis, and similar social and connective Websites that facilitate elements to uncover and share new information.

Among the important challenges to LMSs mentioned by researchers, is a necessity to include constructivist instructional design methods and pedagogy, particularly emphasizing active and shared learning and personalized attention to all students.

The ability of modern LMSs to integrate multimedia and Web 2.0 tools made a significant impact on instructional design and teaching practice. In particular, Web 2.0 applications may be used to enhance student interaction and collaboration. Simulations and serious games are also finding their way into eLearning environment. Finally, a LMS must be adapted by using technologies in order to merge teaching, research, IT, the changing environment, as well the financial accountability and managerial models.

REFERENCES

[1] Abazi-Bexheti, L. LMS: Design, development and Implementing Experience. *Computational Methods and Applied Computing*, Instanbul: WSEAS Press, 2008, pp. 367-371.

[2] Abazi-Bexheti, L., & Dika, Z. A study of evaluation papers and surveys about Learning Managment Systems. *Proceedings of the 2nd WSEAS International Conference on Computer Engineering and Applications.* Cambridge: WSEAS Press, 2008, pp. 158-162.

[3] Abazi-Bexheti, L., Development of a Learning Content Management System. International Journal of Systems Applications, Engineering & Development, 1, 2 (2008), pp. 1-5.

[4] ACPET,"Vic Government elearning program 2012, Alternative DeliveryTools. http://www.slideshare.net/yumproductions/alternatives-to-lms-and-content-development

[5] Aberdeen Group. Research brief: The rice of Larnining Mangament Systems in Talent Managament Systems. Boston: Abeerden, 2008.

[6] Bates, A., & Poole, G. Effective Teaching with Technology in Higher Education: Foundations for Success . San Franciso: Jossey-Bass, 2003.

[7] Grebow D., Werder C., LMS Trends-Webinar, Brandon Hall Group, 2014.

[8] Drozdik, S., Kovács, G. L., & Kochis, P. Z. Risk assessment of an open source migration project. *Proceedings of First International Conference on Open Source Systems* Genova: OSS 2005 Proceedings, 2005, pp. 246–255.

[9] Education for a digital world: advice, guidelines, and effective practice from around the glob, Commonwealth of Learning, Hirtz, Sandy, Harper, David G. Issue Date: 2008

[10] Geser, G. (Ed.). (2007). Executive summary of open educational practices and resources: OLCOS Roadmap 2012. Retrieved on November 16, 2012 from: http://www.olcos.org/cms/upload/docs/ol-cos_roadmap_summary.pdf

[11] Hill, P. Online educational delivery models: A descriptive view. 2012. http://www.educause.edu/ero/article/online-educational-delivery-models-descriptive-view.

[12] Horton, W. (2012). E-learning by design (2nd ed.). San Francisco, CA: Pfeiffer.

[13] Lyman, P., & Varian, H. R. *How much information?* University of California, 2003. http://www2.sims.berkeley.edu/research/projects/how-much-info 2003/printable_report.pdf.

[14] Naidu, S. *E-learning: A Guidebook of Principles, Procedures and Practices.* New Delhi: Commonwealth Educational Media Center for Asia, 2006.

[15] Nantel, R. *LMS Pricing Trends.* Retrieved March 02, 2009, from Brandon Hall Research: http://brandon-hall.com/richardnantel/2007/03/28/lms-pricing-trends/

[16] OECD Center for Educational Rresarch and Innovation. *E-learning in Tertiary Education: Where Do We Stand.* Organization for Economic Cooperation & Devel, 2005.

[17] O'Reilly Radar. Web 2.0 Principles and Best Practices: An O'Reilly Radar Report, 2006.

[18] Siemens, G. Learning or Management System?A Review of Learning Management System Reviews. Learning Technologies Centre: University of Mantoba, 2006.

[19] Weller, M. VLE 2.0 and future directions in learning environments. *Proceedings of the First International LAMS Conference 2006: Designing the Future of Learning*. Sydney: LAMS Foundation, 2006, pp. 99-106

[20] Wise, L., & Quealy, J. *LMS governance project.* University of Melbourne, 2006.

Information Competences - Students' Perspective

K. Pavlina*, S. Špiranec* and A. Pongrac Pavlina*
* Department of Information And Communication Sciences
Faculty of Humanities and Social Sciences
University of Zagreb
kpavlina@ffzg.hr

Abstract - Modern society is characterized by an increasing complexity of tasks which require broad general knowledge and on the other hand a narrow specialization. Because of these conflicting principles modern education recognized the need for development of generic competences on the one hand, and information competences that match the modern information society on the other. In addition to education, which has recognized the need to develop generic and information competences, employers also pay great attention to these competences while hiring. In modern information society information competences become one of the fundamental basis for the development of other generic competences. Information and generic competences are directly linked because information is a fundamental element of learning which is based on the collection, organization, search and interpretation of acquired information. Development of modern information environment is characterized by growing complexity in quantitative and qualitative terms that require additional research of information competences required for the transition from the tertiary education sector to the labour market. This paper presents a survey of students' attitudes about the development of information competences during the study and their opinions about necessary information competences for their future professional careers.

I. INTRODUCTION

Contemporary understanding of information literacy is observed within two complementary directions; a list of competencies and a sociotechnological phenomenon [1]. Information literate people are those who possesses competences to effectively find and use the necessary information [2] and these competencies should be developed in spite of the development of technology [3]. Information literate people are "... those who have learned how to learn. They know how to learn because they know how knowledge is organized, how to find information, and how to use information in such a way that others can learn from them. They are people prepared for lifelong learning, because they can always find the information needed for any task or decision at hand. To be information literate, a person must be able to recognize when information is needed and have the ability to locate, evaluate, and use effectively the needed information" [4]. Doyle [5] complements this definition by integrating new found information into an existing corpus of human knowledge in order to create critical thinking.

Sociotechnological phenomenon rests on the impact of social and technical development. The information literate person adapts to social change by aligning personal goals with economic, social and technological changes [6], [7]. Technical innovations brought by development of information and communication technology greatly improved information searching for appropriate information [8]. Shapiro and Hughes [9] describe information literacy as the new liberal art because of the synergy of using a computer for the purpose of finding the appropriate information that is then critically evaluated from the perspective of their own cultural, philosophical, social and technical context .

Information literacy follows sociotechnological development of society and is seen as a foundation that brings together other kinds of literacy; basic literacy, library literacy, technological literacy, media literacy and visual literacy [10]. The increasing influence of technology leads to strengthening links between information literacy and other digital literacies. Mishra and Mishra [11] argue that information literacy unites seven different types of literacy characteristic for 21st century; media, computing, networking, digital, science, visual and critical literacy. Media literacy is the kind of literacy which aims to communicate messages structured into appropriate format (audio, visual, textual) [12]. Computer literacy refers to the knowledge of IT infrastructure visible in the proper use of hardware and software, while network literacy enables use of World Wide Web [11]. Bawden [13] observes digital literacy in terms of understanding of hypertext, which includes understanding of the text, images and sounds in online environment or in digital format. He gives great importance to critical thinking and forming informed attitude towards found information. Showalter (paraphrased in [14]) thinks that scientifically literate individual possesses the scientific knowledge which is evident in knowledge and application of scientific theories, laws and principles in solving problems which is aided by the use of technology. On that basis scientifically literate individual is able to create a general knowledge that characterize scientific experts [15]. Visual literacy is a kind of literacy which refers to the use of images and visual objects with the intention of conveying information to others, but also receiving and interpreting visual information transmitted by other people [16]. Critical literacy involves deeper reflection based on the analytical skills of writing, reading and speaking on the basis of which it develops critical thinking [17].

SCONUL 's model of information literacy in higher education [18], [19] is intended for students from undergraduate to postgraduate research level. This model covers three academic levels of study (undergraduate, graduate and postgraduate), starting from the lowest levels of the undergraduate program where students are taught fundamental library skills. The foundation is identifying information need and knowledge of the different types of printed and digital resources to be able to choose the resource that best suits requested information query. ICT skills in this model are required to be able to find information, compare information from different sources, organize and synthesize information with intention of creating new knowledge. At the level postgraduate study, students should possess skills in all seven pillars of information literacy.

Because of such understanding information literacy has become basic literacy for 21st century that unites the seven different types of literacy. Modern educational process in higher education should harmoniously complement these types of literacy for contemporary flows of information and communication age. Higher education should develop all these types of literacy because they are fundamental competences expected for professional work in contemporary world.

II. METHODOLOGY

The research was conducted by survey on students of Faculty of Humanities and Social Sciences at University of Zagreb. Respondents were asked to rate the level to which certain information competences are developed during their studies at university and how important are certain information competencies for future work in their profession. The development of information competence is measured on a scale from 1 (not developed) to 5 (maximum development) and the importance for future work in their field was rated on a scale from 1 (not important) to 5 (most important). This research used a list of information competences defined by Australian and New Zealand Information Literacy Framework [20]. A questionnaire was filled out by 128 respondents using Google Forms service. The sample by gender comprised of 96 females (75%) and 32 males (25%). One of the goals of this research was to investigate the attitudes of more experienced students about the importance of certain information competences. So 20 students were on 3rd year of undergraduate study (16%), 40 students were 1st year of graduate study (31%) and 68 students 2nd year of graduate study (53%). Research also aimed to investigate possible differences in the attitudes of information professionals and general population, so sample included 54 students in information science and 74 students of other studies on Faculty of Humanities and Social Sciences.

III. RESULTS

The results in Table I. show that students highly valued importance of information competencies for their future professional work (4.11). It is interesting to note that at the same time they hold that during their study their information competencies were moderately developed (3.33).

TABLE I. SOURCES OF INITIAL INTRODUCTION TO OPEN SOURCE SOFTWARE

Source	Developed	Importance
1.1. Define and articulate the information need	3.33	4.14
1.2. Understands the purpose, scope and appropriateness of a variety of information sources	3.45	4.17
1.3. Re-evaluate the nature and extent of the information need	3.36	4.06
1.4. Use diverse sources of information to inform decisions	3.41	4.25
2.1. Select the most appropriate methods or tools for finding information	3.25	4.19
2.2. Construct and implement effective search strategies	3.16	4.19
2.3. Obtain information using appropriate methods	3.39	4.14
2.4. Keep up to date with information sources, information technologies, information access tools and investigative methods	2.73	4.17
3.1. Assess the usefulness and relevance of the information obtained	3.63	4.47
3.2. Define and apply criteria for evaluating information	3.33	4.11
3.3. reflects on the information seeking process and revises search strategies as necessary	3.13	3.86
4.1. records information and its sources	3.56	3.86
4.2. organises (orders/classifies/stores) information	3.36	4.14
5.1. compares and integrates new understandings with prior knowledge to determine the value added, contradictions, or other unique characteristics of the information	3.38	4.06
5.2. communicate knowledge and new understandings effectively	3.52	4.42
6.1. acknowledge cultural, ethical, and socioeconomic issues related to access to, and use of, information	3.20	3.94
6.2. recognises that information is underpinned by values and beliefs	3.30	3.88
6.3. conforms with conventions and etiquette related to access to, and use of, information	3.67	4.13
6.4. legally obtains, stores, and disseminates text, data, images, or sounds	3.08	3.91
AVERAGE	3.33	4.11

Students of information science evaluated the importance of information competencies almost the same (average grade 4.11) as other students (4.12) . Larger difference is visible at a level of development of information competence at university, where students of information science, as expected, assess the development with higher average grade of 3.47, while other students assess the development with an average score of 3.22. It is interesting to notice the difference in the assessment of the most important information competencies for future work. Students of information science as the most important

competency evaluated *3.1. Assess the usefulness and relevance of the information obtained* (4.44), the other students also highly valued this competence (4.49), and they rank it second. Second most important competence by information science students is *4.2. Organises (orders/classifies/stores) information* (4.41). Unlike information science students, other students evaluated this competence with average importance of 3.95, so it ended up ranked in 17th place. The third most important competence by students of information science was *1.4. Use diverse sources of information to inform decisions* (4.37), other students ranked importance of this competence for future work on 9th place with average grade 4.16.

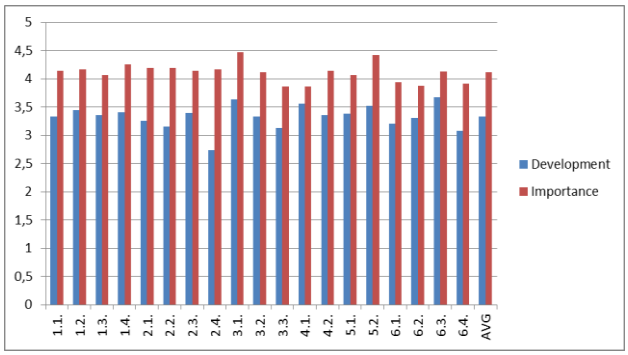


Figure 1. Development and importance of information competences

The results in Figure 1. show that as best information competency developed during the study, students assessed *6.3. Conforms with conventions and etiquette related to access to, and use of, information* (3.67). It is followed by competence *3.1. Assess the usefulness and relevance of the information obtained* (3.63) which is interesting to point out that students perceived it as the most important competency for future work. Although it is one of the most developed competence it can be noted that there is a big difference in the level of development of competence during the study and expressed importance of future work in the profession (0.84). Third best developed competence is *4.1. Records information and its sources* (3.56) in which students also expressed the slightest difference in importance for future work and development during the study (0.30).

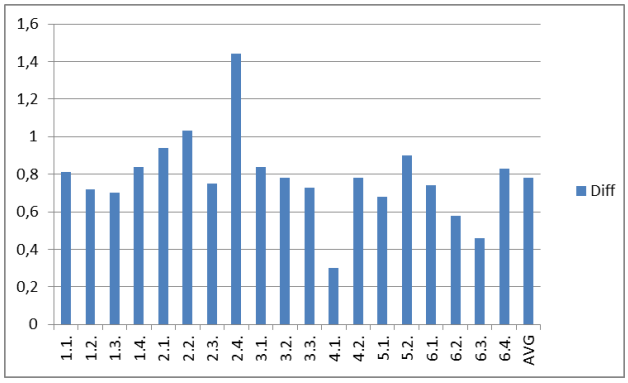


Figure 2. Difference between develoment and importance of information competences

The biggest difference between the importance for future work and development during study was detected at competence *2.4. Keep up to date with information sources, information technologies, information access tools and investigative methods* (1.44), where it can be concluded that teachers should take a greater effort to track with the development of information resources, technologies, tools and methods (see Figure 2.).

IV. CONCLUSION

The results of this research show that students attribute high importance to information competencies for their future professional work. It is necessary to improve the curricula contents that develop information competences, because there is a significant difference between the perceived level of importance of information competencies for future professional work (4.11) and the extent to which they are developed during the study (3.33). Curricula should regularly be updated in order to track the development of information technology, resources and search method, and thus enable quality education for future professional work.

REFERENCES

[1] S. Špiranec and M. Banek Zorica, Informacijska pismenost : teorijski okvir i polazišta. Zagreb : Zavod za informacijske studije, 2008.

[2] L. Burchinal, „The communications revolution : America's third century challenge“ in The future of organizing knowledge. Texas A & M University Library's Centennial Academic Assembly, 12 pp, Sept. 24, 1976. URL:http://personalpages.manchester.ac.uk/staff/drew.whitworth/burchinal_the_communications_revolution.pdf

[3] W. Demo, The idea of „information literacy“ in the age of high-tech. New York : Tompkins Cortland Community College, 1986. URL: http://files.eric.ed.gov/fulltext/ED282537.pdf

[4] American Library Association, Presidential committee on information literacy : final report, 1989. URL: http://www.ala.org/acrl/publications/whitepapers/presidential

[5] C. Doyle, Outcome measures for information literacy within the national education goals of 1990: final report of the National forum on information literacy, summary of findings. Washington, DC:US Department of Education, 1992.

[6] The Alexsandria proclamation on information literacy and lifelong learning. Beacons of the information society, 2005. URL: http://archive.ifla.org/III/wsis/BeaconInfSoc.html

[7] S. Webber and B. Johnston, „ Conceptions of information literacy: new perspectives and implications“ in Journal of information science, vol. 26, pp. 381-397, 2000.

[8] S.J. Behrens, „A conceptual analysis and historical overview of information literacy“ in College & Research Libraries, vol. 55, pp. 309-323, July 1994.

[9] J. Shapiro and S. Hughes, „Information literacy as a liberal Art“ in Educom Review, vol. 31, March/April 1996. URL: http://net.educause.edu/apps/er/review/reviewarticles/31231.html

[10] B. Ferguson, „Information literacy : a primer for teachers, librarians and other informed people, 2005. URL: http://bibliotech.us/pdfs/InfoLit.pdf

[11] R.N. Mishra and C. Mishra, „Relevance of information literacy in digital environment“ in Journal of emerging trends in computing and information sciences, vol.1, pp. 48-54, 2010. URL: http://cisjournal.org/journal/index.php/computing_information/article/viewFile/19/18

[12] R. Hobbs, „The seven great debates in the media literacy movement“ in Journal of Communication, vol. 48, pp. 16-32, 1998.

[13] D. Bawden, „Information and digital literecies; a review of concepts“ in Journal of documentation, vol. 57, pp. 218-259, 2001.

[14] R. Laugksch, „Scientific literacy:a conceptual overview“ in Science education, vol. 84, pp.71-94, 2000.

[15] P. DeHart Hurd, „Scientific literacy : New minds for a changing world in Science education, vol. 82, pp. 407-416, 1998.

[16] E. Sims, R. O'Leary, J. Cook and G. Butland, „Visual literacy : what is it and do we need it to use learning technologies effectively?“, Appears in Winds of change in the sea of learning : proceedings of the 19th Annual Conference of the Australasian Society for Computers in Learning in Tertiary Education (ASCILITE): 8-11 December 2002.

[17] K. Brown, Developing critical literacy. Sydney:National centre for language teaching and research, 1999.

[18] SCONUL advisory committee on information literacy, „ Information skills in higher education“, The Society of College National and University Libraries, 1999. URL: http://www.sconul.ac.uk/sites/default/files/documents/Seven_pillars2.pdf

[19] The SCONUL seven pillars of information literacy, „Core model for higher education“, 2011. URL: http://www.sconul.ac.uk/sites/default/files/documents/coremodel.pdf

[20] Australian and New Zealand Information Literacy Framework. Adelaide: Australian and New Zealand Insitute for Information Literacy, 2004.

Development and Analyses of "ProSysEngine" M-Learning Course Management System

Jovica Saveski*, Majlinda Fetaji*, Bekim Fetaji*, Mirlinda Ebibi** and Lejla Abazi*
* South East European University, Contemporary Sciences and Technologies, Ilindenska bb, 1200 Tetovo, Macedonia
** International Balkan University, Information Technology, Skopje, Macedonia
js17035@seeu.edu.mk, m.fetaji@seeu.edu.mk, b.fetaji@seeu.edu.mk, m.ebibi@ibu.edu.mk, l.abazi@seeu.edu.mk

Abstract - The research study investigates the development approach and analyses of m-learning course management system developed as case study. The contribution of the research study is the analyses of the task based instructional strategy and conceptual definition of the assessment methods regarding the level of mobile learning outcomes while assessing different software engineering aspects of development mobile application for Windows phone 8. The analysis also covers the development with primary focus on mobile learning using ASP.NET MVC 5 technology, and design issues for responsive techniques. Finally insights are stated, results of the analyses are provided and recommendations and discussions of the approach are described.

I. INTRODUCTION

Because the constant changing of the technologies, how the learners learn and interact, their demands are constantly growing for different types of learning materials [2], [3]. For example we might say that a professor has shared youtube.com video URL and the professor demands of the student to know the context of the particular video. The older management system did not allow the users to share easily video URL's [9]. Then the students had to copy the link and transfer it to a web browser, and the content can be easily accessed by a single click, and be opened by a mobile browser quickly and from anywhere at any time.

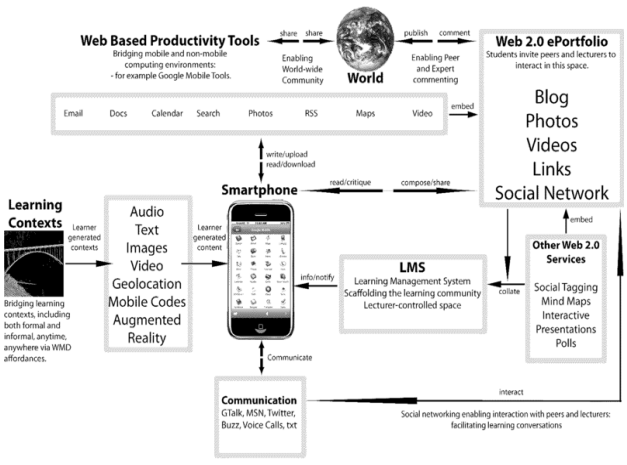


Figure 1. Mobile Web 2.0 Concept Map[2]

This interoperability has inflicted the need to create new and improved learning systems, that not only could assist the learning of contents but also speed up the process of interaction, allow different ways of conducting it, and allow for it to be accessed literally from any device or any place.

II. ANALYSES OF M-LEARNING PROJECTS AND TRENDS

An example of mobile learning system called GoLearn from the initial Merrill Lynch pilot is given by ADL Mobile Learning Team [9]. Standards for delivery on the BlackBerry were established in design, technology, security and privacy. The goals of the pilot included proving the access, usage and the effectiveness of learning delivered via the BlackBerry to the global population. Additionally they sought to:

- Deliver training with no degradation to learning effectiveness
- Achieve 25% of eligible participation
- Achieve a comparable average score to the control groups, and
- Obtain a 10% higher completion rate in 10% less time

Over a seven-week period, the learning materials were wirelessly pushed to over 2,100 investment bankers and select support staff.

The outcomes exceeded the goals. Higher scores were obtained in half the time. Bankers who completed the training did so in 54 less minutes and tested higher on the final assessment tests than the remainder of the firm. Mobile users also completed their training twenty days earlier than those who trained via MLU.

According to [8] VPs and higher leveraged the mobile materials the most. Of the 2100 eligible employees, 61% launched the content at least once. 317 people completed 704 courses. Overall the mobile learners obtained a 12% higher completion rate in 30% less time than the control group.

170 employees responded to a survey indicating:

- 99% felt the format and presentation supported the learning
- 100% would complete more training in this format

• More than 75% praised the benefits of convenience, time management and training with no distractions

With this successful pilot, Merrill Lynch moved into the next phase with additional training topics such as onboarding for new hires, ethical decisionmaking, performance management, market abuse, and sexual harassment.

According to [4] there is an increasing number of colleges and universities who are adopting mobile wireless technologies as teaching and learning tools. More than 90% of public universities and 80% of private universities in the US have some level of mobile wireless technologies, such as mobile wireless devices and networks.

According to [4] Task based learning as instructional strategy has been shown as the most appropriate for mobile learning compared with inquiry, project and problem based learning. The user filling the 4 task defined each semester week out of 15 weeks is guided and not left alone but instructed and forced to work regularly and in exact time periods.

An important advantage of the handheld computer and mobile devices in general is that can be accessed from anywhere at any certain time, providing learning resources to be articulated in the learning and other actions [4]. There are advantages in using a mobile device offering individual, private and learning at own pace and learning within specific contexts which can provide 'reliable cultural and environmental indications for understanding the uses of information which may enhance encoding and recall and enable learners to access relevant information when and where it is needed'. The learners and institutions that consume m-learning systems because of the changing dynamics and can benefit from it.

III. RESEARCH METHODOLOGY

The study focused on fundamental research on m-learning, curent projects and trends, exploratory research regarding the conceptual design and afterwards action research to buidl a software solution to test the insights gained from previous phase. In order to evaluate the results the solution was evaluated from different participants. Participants were students, teaching and administrative staff of the Contemporary Sciences and Technolgies Faculty of the South East European University.

IV. CONCEPTUAL DESIGN

The main purpose of the web application is to provide learning system to students, professors and educational institutions with option to be accessed fryom anywhere at any time. In this regards the web application is available for any internet capable device such as mobile phone devices, mobile touch devices (tablets and slates) and television sets with internet connection through a browser. This application at this first version primarily focuses on knowledge assessment for students with given administration rights to the professors folowing the guidelines from [11].

The M- Learning aspect of this application gives it a huge importance in today's dynamic world where distance learning and constant connectivity to internet are a regular state. On the other side the usa ge of the ASP.NET MVC 5 framework gives it a huge advantage because it is the latest technology and the easiest framework available to maintain in regards of updating the business logic or the design.

Primarily the idea was to make just a quiz generating application, but during the research study process we realized that what this world lacks of is more learning assessment systems based on mobile learning and using latest and best technologies. Learning assessment systems now are hugely needed not only to evaluate the knowledge of the final user, but also to specify the direction in which the learning process should go and declare where the final user did not pay attention.

This application also has a big advantage in the aspect of scalability, and for development, testing and presentational purposes, will be hosted on the Cloud, a cloud hosted by Microsoft called Windows Azure [13].

In essence the main research questions is: "how to develop a good and easily changeable and maintainable m-learning web application accessible from any device?"

Below is given the Activity diagram in which are represented the activities in a general form for the web application. Represented is the flow of the application. At first the client needs to access the website where he has only access to data that is available to all users. If he wishes to interact with the web application he needs to verify his credentials. He sends his credentials information's to the ASP.NET identity membership provider [10]. He is then verified by the application using the Azure MS SQL Server.

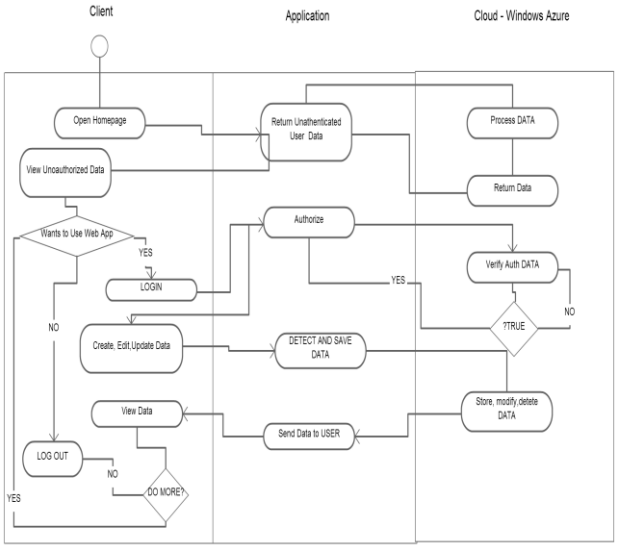


Figure 2. Activity diagram of the developed solution

If the user gets verified correctly he has access to the web application. If the user wishes to interact with the application, according to his role in the system the user gets different options. In general the user can create, edit and update data on the web application. For example, if the user has a role of administrator he has access to every

functionality of the web application except for solving quizzes which is a functionality of the web application only accessible to users with the role student [1].

The data is stored on Microsoft SQL Server on Windows Azure Cloud [5], [6], [7], [10], [12]. When the user is finished, if he wants to delete his session, cookies and exit the application he can easily log out.

On figure below is shown the sequence diagram for an example scenario when an administrator is creating a course. Here we have four objects which literally present the Model View Controller software architecture [13]. When the user first demands a website resource he is firstly encountered in the background of http process with the routing rules of the application. In MVC the paths, URL's or routes are defined in the routing module [12]. The application decides where it will send the request to, to which controller it will be directing the user. The controller gets the data from the model which takes the data from within itself connecting to the database and applies its methods and presents the data to the end user using the view [5].

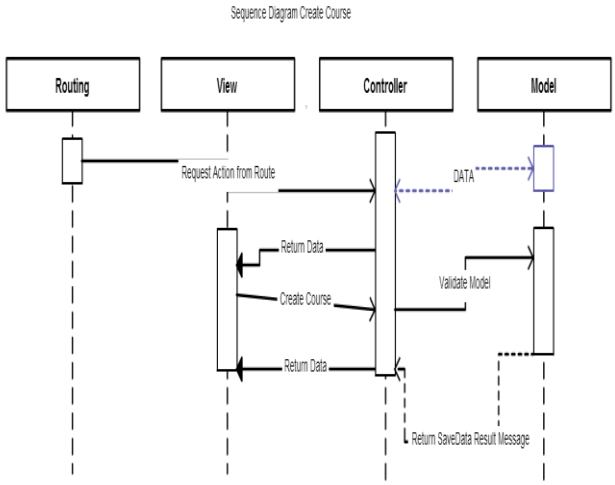


Figure 3. Sequence diagram of the developed solution

Created the following classes: Course, Enroll, Quiz, Question, Answer, StudentQuizTake and studentAnswers. We are using these classes in order to organize my web application in a way that the system has courses which have students enrolled.

In each course created quizzes with questions and answers. Afterwards when the student goes's to solve a quiz he creates a new instance of the StudentQuizTake class as an object.

The user fills the StudentAnswers class with his answers to the appropriate questions in the quiz. In order to differentiate users from one another will be using the new ASP.NET identity membership provider.

The ASP.NET identity uses the following classes: UserRole class, ExternalLogin class, LoginViewModel class, User class, UserSecret class, RegisterViewModel class, Role class, UserLogin class and ManageUserViewModel class.

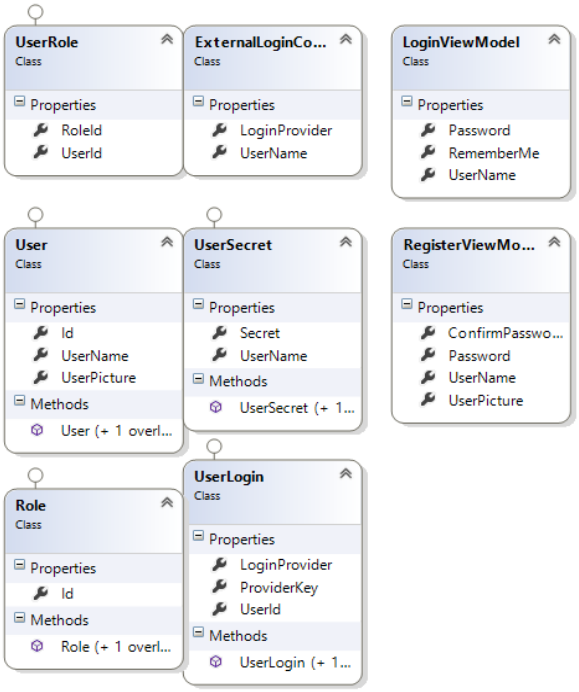


Figure 4. Class diagram identity

V. Development Aspects and Approach

The name of the developed solution is Prof Sys Engine, and it is developed in ASP.NET MVC 5.

We have used Microsoft Visual Studio Ultimate 2013 Preview, .NET Framework 4.5 folowing the guideliness from [11] and [13] .

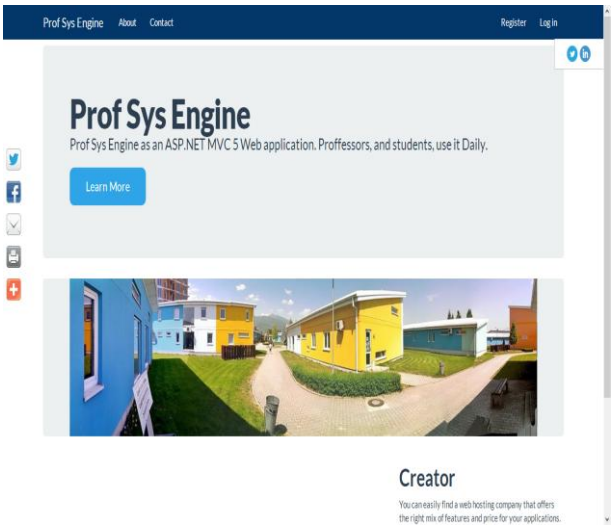


Figure 5. The screenshot of the developed solution

According to [13] Microsoft defines the ASP.NET MVC framework as follows: The Model-View-Controller (MVC) architectural pattern separates an application into three main components: the model, the view, and the controller. The ASP.NET MVC framework provides an alternative to the ASP.NET Web Forms pattern for creating Web applications. The ASP.NET MVC

framework is a lightweight, highly testable presentation framework that (as with Web Forms-based applications) is integrated with existing ASP.NET features, such as master pages and membership-based authentication. The MVC framework is defined in the System.Web.Mvc assembly.

The developed solution is a cloud application; it is being hosted on Windows Azure Cloud platform by Microsoft. It is using a SQL Server on the cloud. This feature of the web application gives it enormous advantage because it is scalable, flexible and cost friendly. In one in a million case it could happen that the page in its first month of being published online to get thousands of users. Having the web application on the cloud allows me to scale the servers up and get new processors for serving the website, expand database space or the other way around.

Another of its best features is that the security of the application is being provided by a cloud service. Also if the server fails the cloud system transfers the applications to another server and in seconds time it is up.

In order to differentiate the web application users we have created a filter that is used just as an attribute in a declaration of controller classes. The code for the authorization filter follows:

```
namespace Mvc5website1.Filters
{
   [AttributeUsage(AttributeTargets.Class |
AttributeTargets.Method)]
   public class AuthorizeRedirection : AuthorizeAttribute
   {
      private const string IS_AUTHORIZED =
"isAuthorized";
      public string RedirectUrl =
"~/Error/UnauthorizedUser";
      protected override bool
AuthorizeCore(System.Web.HttpContextBase
httpContext)
```

This filter allows us to easily restrict access or allow access to a particular role in the controllers. So just by using "[AuthorizeRedirection(Roles = "student")]" The code communicates with the ASP.NET MVC Framework that will only allow access to the particular action or controller to the users which have a role of "student".

Also listed all of these classes in the DbContext class called MyDbContext. The DbContext class is used by Entity Framework to reference the already declared classes in the project, in order to allow Entity Framework to understand what the developer wants its business logic/ model work like. Upon starting the application, if the DbContext classes in instantiated the referenced classes or DbSet's in DbContext are created as entities in a database which connection string have declared in Web.config.

Entity Framework analyses the defined classes and looks at the data types, the relationships declared between the class properties of different classes. If the class contains Data Annotations, Entity Framework uses them to constraint the values, define foreign keys or just change the display name [12].

We have enabled Entity Framework Migrations in the web application project. With the enabling of Entity Framework Migrations allowed automatic migrations, and allowed Entity Framework to update the database model automatically based on the changes made in the classes declared as DbSet's in my DbContext class. So in the case when wanted to add user pictures to the user class and save that information to the database Entity Framework upon calling detects the changes in the model and applies the changes to the database.

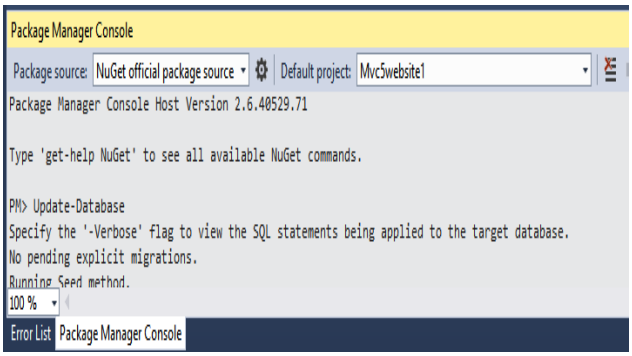


Figure 6. Database migration

The "Update-Database" shown in figure above tries to detect the changes in the model and finds out there is no pending explicit migrations, and finally executes the Seed method. Created a Seed method that populates the database with startup users end roles as seen in the use case diagram there are three user roles and those are administrator, professor and student. For each role assigned one startup user.

On the screenshot below is shown the management panel where the administrator of the web application can change the roles of the users registered to the web application.

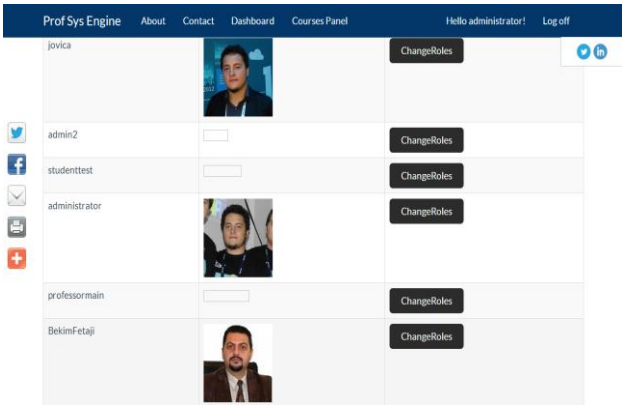


Figure 7. Administrator options available

The following screenshot displays the view for creation of a course. Only the administrator has the right to create courses. Also only he has the privilege to edit or delete them. Professors who are owners of a course, can only edit the information about the course they own, and if they try to change a course that they do not own, the

Course Controller in the web application will redirect them to an error “Unauthorized”.

Only the administrator has the right to delete a Course from the web application.

Upon Creating a Quiz, the Professor or Administrator will declare the duration of the Quiz. So if a Quiz is declared by the professor with 20 minute mark, the student will only get results for the answers he gave in that period.

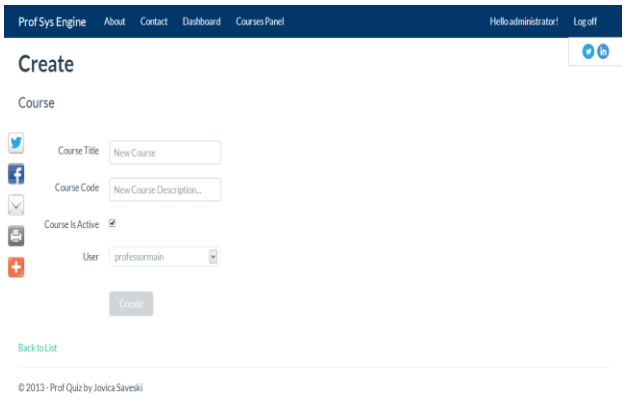


Figure 8. Course creation functionality

The only place, that only a student has access to, it is the student controller which is used to display his quiz data and direct the user to solve the quiz. When a student starts to solve a quiz this system notes the time when the user has started solving the quiz. In the meantime for each answers submissions the application stores the student’s answers. After the student finishes the quiz, the result is given to him.

VI. NATIVE MOBILE CLIENT ON WINDOWS PHONE 8

The project also includes a native mobile client on Windows Phone 8. This native mobile application gives the user who in this case is the student the options to check in the system in which courses he is currently enrolled in, the option to view the quizzes that he needs to solve (he has not solved them yet) and view the names of the quizzes he has already solved. In the case if the user decides to solve another quiz or just visit the web application he can use the button to go to the web application, in which case he is transferred to default browser.

Used a Microsoft HTTP client libraries in order to bring the common functionality of the HttpClient class.

```
c1 = new HttpClient();
c1.BaseAddress = new Uri("{URL}/api1/");
```

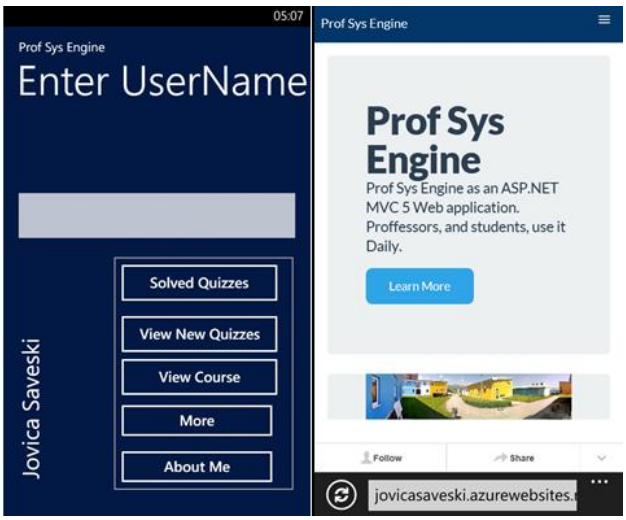


Figure 9. Native mobile aplication for Windows Phone 8

VII. TESTING

In order to check if the existing and planned functionality of the application is all right, conducted tests on four kinds of display.

First and foremost tested the functionality, usability and design on a Windows 8 (15” display monitor). Afterwards tested it on a 10” android tablet. After summarizing the results of the previous two devices conducted another two tests. One on a 4” Windows phone HTC 8s and finally tested the application on a 42 display using a television set.

The results were satisfying in the means of that in every device the user could access the needed data, functionality and resources. The front end of the web application was equally responsive to the display size and type on every devise.

Engaged and had the help of 10 collaborating testers, who tested the web application and gave me feedback. All of the feedback that was received was positive, and helpful. The beta testers during the time of testing and feedback giving gave advice on some other features that might be implemented in the future.

They successfully managed to access to web application, also it was easy for them to create courses and manage them. They used the panels to create quizzes, and even made a small competition among them taking the quiz.

Five advanced users were among the beta testers which had some kind of knowledge of the process of creating applications. And five of them were regular users which interests were not developing applications, but using it for their own needs and for sharing an expanding the knowledge among people.

Different beta testers got different users roles, some of them were students, some of them were professors and some were administrators.

VIII. CONLUSION

For the research study and project used ASP.NET MVC 5 and although at the time it is a very new framework it has flawlessly worked during the creation, testing debugging and publishing the project. The web application works in the defined parameters and specs and is equally functional and tested its responsive on a Windows 8 laptop (Asus k52jt), on Android Tablet (10" Quart QP1070), and on HTC 8s Windows Phone (4" mobile phone).

The web application after the testing is evaluated a useful to the students as it is for the professors; it allows interaction between the standing parties. The best feature of the web application is that it is available equally among all capable devices. It saves the precious time of the professors, giving immediate results to the student, without the need for the professor to check each and every one of the quiz takes, if the professor needed to check all of the quiz results by hand. That is another advantage, if the student base expands, the professors cannot manually manage the learning assessment in the same time as they did before.

A great feature is that it can be easily changed, in a way that the market demands. The web application can easily be upgraded to become a big learning management system, with features like no other. Feedback from users can quickly be heard and implemented, thanks to the MVC software architecture pattern.

In the future we can implement a fully grown API, for all kinds of developers, and make it more available to even more devices, such as Raspberry PI, Windows Kinect, and even an electronic news feeder tool.

If this web application is implemented in another institution, it can quickly adapt it to its needs. For example if this web application is implemented as an assistance tool for conferences guests' lottery, we can add this functionality fast using Entity Framework Migrations.

For the future work intended is to add more common functionality, such as course based real time chat, messaging, file sharing, library and wiki panel.

REFERENCES

[1] Christian Gross (2008), Beginning C#: From Novice to Professional, 2008

[2] Dominic Betts, Alex Homer, Alejandro Jezierski, Masashi Narumoto, Hanz Zhang (2012), Moving Applications to the Cloud 3rd Edition, 2012

[3] Dominic Betts,Alex Homer, Alejandro Jezierski,Masashi Narumoto,Hanz Zhang (2012), Developing Multi-tenant Applications for the Cloud 3rd Edition, 2012

[4] Fetaji Majlinda., and Fetaji, Bekim., (2008) "Universities go mobile: Case study experiment in using mobile devices" – ITI IEEE conference, Dubrovnik, Croatia, 2008

[5] Jeffrey Palermo, Jimmy Bogard, Eric Hexter, Matthew Hinze, and Jeremy Skinner (2012), ASP.NET MVC 4 in Action, 2012

[6] Jess Chadwick, Todd Snyder, and Hrusikesh Panda (2012), Programming ASP.NET MVC 4, 2012

[7] Jon Galloway , Phil Haack, Brad Wilson, K. Scott Allen (2012), Professional ASP.NET MVC 4, 2012

[8] Judy Brown and Jason Haag (2011), Mobile Learning Handbook, 2011

[9] Kim, S.H., Mims, C., & Holmes, K.P. , 'An introduction to current trends and benefits of mobile wireless technology use in higher education'. AACE Journal, 14(1) ,2006

[10] Mahesh Chand, (2002) A Programmer's Guide to ADO.NET in C#, 2002

[11] O'Malley, UoN , G. Vavoula, UoB, J.P. Glew, UoB, J. Taylor (2003) OU,M. Sharples, UoB,P. Lefrere, OU, WP 4 – Guidelines For Learning/Teaching/Tutoring In A Mobile Environment, 2003

[12] Simon Robinson,Christian Nagel,Jay Glynn,Morgan Skinner,Karli Watson,Bill Evjen (2004) Professional C# Third Edition, 2004

[13] Tom Dykstra (2012), Getting Started with the Entity Framework 4.1 Using ASP.NET MVC, 2012

Investigation of Natural User Interfaces and their application in gesture-driven human-computer interaction

Aleksandar Petrovski*, Bekim Fetaji*, Majlinda Fetaji*, Mirlinda Ebibi**
* South East European University, Contemporary Sciences and Technologies, Ilindenska bb, 1200 Tetovo, Macedonia
** International Balkan University, Information Technology, Skopje, Macedonia
ap102578@seeu.edu.mk, b.fetaji@seeu.edu.mk, m.fetaji@seeu.edu.mk, m.ebibi@ibu.edu.mk,

Abstract - As contribution the research study as outcome developed a software application and programmed several new recognition gestures and presented a set of tools supporting the spectrum of software engineering natural user interfaces to be used in University promotion and marketing. This research study tries to foster several contributions. Firstly, this system is built on basis of separate library used for the gesture recognition part of the application, which can also work separately from the system and can be used in different contexts. It is very extendable and allows the users to implement new gestures on top of the library, thus enriching its context and provides easy start for the user allowing him to concentrate to the business logic instead of the low level programming of the gesture recognition algorithms. Secondly, in addition to the basic gestures, created are several other gestures that add additional value to the library itself. New gestures that are added to the library are DownSwipe, UpSwipe, ClockwiseCircle, CounterClockwiseCircle, Push gestures, ManualDoorOpen gesture, AutomaticDoorOpen gesture and WalkingGesture. Thirdly, this research also contributes to researchers and professionals who are dedicated to a research in the area of natural user interfaces and their efforts in finding an effective algorithm for human gesture recognition. This research investigates different approached to gesture recognition, such as Record and recognize approach, Neural Networks approach and Gesture composition approach, presented which are their advantages and weaknesses and presented why chosen is the Gesture Composition approach.

I. INTRODUCTION

The research study investigates the effectiveness and efficiency of the several gesture recognition approaches.

Natural user interfaces as discussed in [5] are continually becoming important part of the contemporary human-computer interaction. They are introduced in almost every new device [2] easing the interaction between the user and the device itself. According to [6] from touch-sensitive displays to 3D cameras and infrared sensors, natural user interfaces make the electronic devices usage much easier, faster and more intuitive. Some of the natural user interfaces require minimal contact between the human and the device [2].

The most advanced interfaces don't require any contact and they combine different kinds of cameras and sensors to use the human body as an input source [1].

II. BACKGROUND RESEARCH

Since the beginning of the computer era, many computer engineers were searching for a solution to improve and ease the way people interact with the computer systems [1]. Good user interface will improve user experience with the computer and other mobile devices [3] and will make the interaction as discussed in [12] easier, accessible and more optimal.

One of the most important devices [5] that take the NUI usage on the next level is the appearance of the Microsoft Kinect for Xbox and Microsoft Kinect for Windows [8]. This device allows the player to control the system using his/her body movements. The user doesn't have any direct interaction with the system, but his body is captured and gestures are recognized using one RGB and one Infrared camera. Kinect quickly became interesting device for the programmers worldwide. Many people have found a lot of interesting applications of Kinect in the real world environments.

The concept of Natural User Interface (or NUI) is firstly introduced by Steve Mann [8]. Steven Mann is a researcher and inventor best known for his work on computational photography, particularly wearable computing and high dynamic range imaging. According to [10] Mann holds degrees from the Massachusetts Institute of Technology (PhD in Media Arts and Sciences '97) and McMaster University, where he was also inducted into the McMaster University Alumni Hall of Fame, Alumni Gallery, 2004, in recognition of his career as an inventor and teacher. In the period of 1970s – 1990s he developed many different devices which were meant to replace the standard command line interface and graphical user interface. For his projects he had used several terms, including 'Natural User Interfaces', 'Direct User Interfaces' and 'Metaphor-Free Computing' [7].

According to [8] August de los Reyes, Director for Surface Computing in Microsoft firstly mentioned the Natural User Interface as the next phase in the user interface evolution. Although natural user interfaces can't replace graphical user interfaces, since most of the NUIs

use graphical user interface to show the result, natural user interface tries to replace the WIMP (window, icon, menu and pointing device) principal. Later, on several conferences, representatives from Microsoft talked about the importance of the natural user interfaces and their efforts in providing best experience to the users using NUI features. According to [6] gesture recognition system is natural user interface which tries to follow the body movement and produce input out of this information. According to [11] there are two kinds of gesture recognitions systems: systems which require physical contact with the body and systems which can follow the body without physical contact.

In 2012 Microsoft released a new Kinect device meant to work only with personal computers. Unlike the older version which was originally published for Xbox 360, the new version had better resolution of the RGB camera and the depth infrared sensor.

Together with the new Kinect for Windows device, Microsoft released a new version of the software development kit. The new version of the SDK included improved skeleton recognition in different positions, better stabilization algorithm and additional tools for easier development of Kinect enabled applications.

According to [7] Kinect device is composed of several parts. Main parts of the device are the RGB camera and the infrared sensor.

The RGB camera is retrieving RGB video of the objects in front of the device. It supports resolution up to 1280x960 at 30 FPS. The resolution that will be used is programmable using the software development kit.

Next to the RGB camera is located the infrared sensor [7]. The infrared sensor is consisted of infrared transmitter and infrared receiver. When the infrared wave is transmitted, it is reflected from the objects in the front, and it is returned to the receiver. The time that the wave needed to return to the sensor can determine the distance between the object and the Kinect device [7].

Each Kinect device also has 4 microphones located on different places on the device, 3 in the front and 1 in the back of the device.

Using the 4 microphones, Kinect supports advanced audio features, such as beam formation to detect the audio source, noise suppression and echo cancelation and easy integration with Microsoft's speech recognition platform.

Kinect also has a motorized tilt, which together with the gyroscope is responsible for vertical angle position of the Kinect sensor.

III. RESEARCH METHODOLOGY

The research study is fundamental research focusing on exploratory research and then action research developing a software solution to implement the insights gained from the fundamental research. The research investigation is focused on the Kinect sensor, programming new gesture recongnitions and their features. The Kinect sensor allows users to interact with the system without any physical contact with it. The user stays in front of the sensor (2-4 meters away). The Kinect uses its RGB and Infrared cameras to capture the user position in the space. This information can be retrieved easily from the sensor, but what is missing in the Kinect software is the gesture recognition part.

The research study ivestigates the effectiveness and efficiency of the several gesture recognition approaches. The primary focus of research is the test for accuracy, speed and correctness of the different algorithms. The final system is implemented in Microsoft .NET technologies. Researched is how the gesture recognition system can be implemented in the .NET framework and be reused in many different projects. Participants were students, teaching and administrative staff of the Contemporary Sciences and Technolgies Faculty of the South East European University.

IV. GESTURE RECOGNITION APPROACHES

Gestures and gesture recognition systems are main power driver for the Kinect device and Kinect enabled applications. Kinect for Windows SDK provides full skeleton information about the people in front of the device including location information about each joint on the skeleton, but the real power and value of the Kinect is fulfilled using the gestures and gesture recognition systems.

Since its beginning and first appearance on the market, Kinect became main topic of research of many computer professionals and enthusiasts all around the world [7]. Many of them researched how it can recognize a gesture and do specific action for each gesture. Few different approaches were discovered, each of them with several advantages and disadvantages.

A. Record and Recognize Aproach

According to [6] one of the very first approaches to gesture recognition systems was the 'record and recognize' approach. With this approach each gesture is recorded once and is memorized by the underlying application. After the gesture is memorized as a valid gesture, then it can be used to recognize other movements and perform an action if the specified gesture is recognized.

The advantage of this algorithm is that it can be easily implemented [4]. The complete procedure consists of two phases. One phase is to follow the gesture, convert it to computer-understandable information and save the gesture in the memory or the database. The second phase is to record the algorithm, convert it to computer-understandable information using the same algorithm as in the first phase and search for that information in the data store. If the algorithm is recognized, perform the desired action.

This approach has several disadvantages when applied to a real application. First, this approach has no fault tolerance. Even very small distortions will prevent the algorithm from recognizing the gesture. Several error and distortion correction algorithms can be applied to this approach to minimize or reduce the error, but since the

data is very sensitive and not deterministic, it will still contain different data and the algorithm will fail.

Second problem with the approach is that when one gesture is recorded and saved in the data store, it would hardly be possible to use the same gesture model for different person. People have different body structure, size and natural movements and the gestures that are naturally performed by one person, are not same for another person.

B. Neural Networks Approach

According to [9] Neural networks approach was also used as one of the first algorithms for gesture recognition. Neural networks are used for different pattern recognition algorithms, including optical character recognition and are proven to be reliable and intelligent approaches to different real world problems.

Neural networks approach for gesture recognition is reliable and proven mechanism. Its advantage is that after sufficient training with different data sets it can recognize simple gestures. However, although this approach is good for academic and research environments, it's not very useful in a production, application environments, when the training time should be avoided.

Main disadvantage of the neural networks approach is that they require big input, training dataset [4]. The dataset should contain different gestures from big number of people, with different body structure, different size and different cultures because they have different natural movements. This is hardly achievable, and besides research environments it is hardly usable and effective in application environments.

Another disadvantage is the initial effort time for creating the neural network. The developer will need a lot of time to research and conclude which type of network and network structure is more suitable for the gestures that are needed to be recognized by the system.

C. Gesture Composition Approach

According to [6] gesture composition approach was introduced in its initial version by a group of employees at Microsoft Corporation in United Kingdom. Right after the appearance of the Kinect for Windows Software Development Kit, they had made internal research about what is missing as part of the development kit. They found out that in order to make a Kinect enabled application, one of the main parts of the system should be gesture recognition system. They provided a simple gesture recognition approach which was based on creating a composition or a model of the gesture upon which the application is able to recognize the gesture.

Gesture Composition Approach is a complete and reliable approach that is proven to work with different people in the same way. We tested this approach on 50 different employees in our company. We found out that each of them would require 5-10 minutes in order to learn and practice the gestures that we have made in our system. After the learning time, each user was able to efficiently use our application and play our games.

During our analysis we found out that the gesture is made of parts. Each gesture could be effectively divided in different parts. Each part presents a position in which specific body parts or joints are located in the space and relative to the other body parts which are relevant for the gesture. When these parts of the gesture are executed in a specific order in a specific time frame, we can say that the gesture is performed.

V. DEVELOPMENT OF NEW GESTURE RECOGNITIONS

Gesture Recognition Library is developed in .NET Framework, which is the framework on which Kinect for Windows SDK works and on which most Kinect enabled applications are developed. In its base it uses the libraries and APIs from the Kinect for Windows SDK in order to communicate with the Kinect, send commands to the device and get the input streams for the RGB and depth sensing cameras [9].

Skeletal recognition function enables the Kinect to recognize the human body (skeleton) in front of the device and track its movements. Kinect recognizes the skeleton and provides the position of the skeleton together with the body joints to the programmer. There are several body joints provided by default: AnkleLeft, AnkleRight, ElbowLeft, ElboxRight, FootLeft, FootRight, HandLeft, HandRight, Head, HipCenter, HipLeft, HipRight, KneeLeft, KneeRight, ShoulderLeft, ShoulderCenter, ShoulderRight, Spine, WristLeft, WristRight [10]. For each joint its position in meaning of x, y and z coordinates in the space is provided [12].

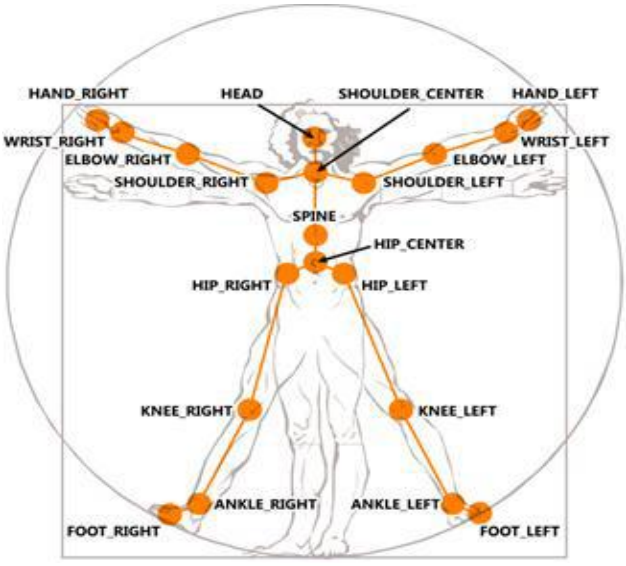


Figure 1. Body Joints that are recognized by the Kinect SDK Framework

Within the research study developed are several gestures that were needed for experimenting and for the development of the University Exploration System.

All these gestures are included in the library and are presented in the following parts. However, its design is very extensible and anyone with no huge effort can develop additional gestures that can be added to the library and very easily integrated.

The library is consisted of several main classes, such as KinectManager and GestureManager which are the main components of the library. Apart from that, there is class for each part of a gesture that is defined in the library.

The current implementation within the research study of the GestureType contains 13 gestures which are already implemented in the library: WaveRight, WaveLeft, DownSwipe, UpSwipe, ClockwiseCircle, CounterClockwiseCircle, PushLeft, PushRight, ShortLeftSwipe, ShortRightSwipe, ManualDoorOpen, AutomaticDoorOpen and Walking.

VI. CONCEPTUAL DESIGN OF THE UNIVERSITY EXPLORATION SYSTEM

One of the applications developed on top of the gesture recognition library is the university exploration system. This application is targeted to the new students and all the people who are interested in the university structure, study programs on different faculties and all additional information about the university itself. Using natural user interfaces as primary input source, this application provides extraordinary user experience, especially for the future students. Providing natural user interface as a first contact between the student and the university, it's promoting the university in the best way to the future students and introduce them to the advanced topics of the information technology.



Figure 2. GUI of the University Expolaration system

The University Exploration and Presentation System is implemented in .NET framework version 4.5 and works in the same runtime environment together with the gesture recognition system. It utilizes all the features that the gesture recognition library provides and is complete example of the usage of the library.



Figure 3. Facylty Exploration System

The application is implemented as a Windows Presentation Foundation (WPF) application. The main application object is the MainWindow which implements the IPageSwitcher interface. There is one Switcher static class which is responsible for switching the content of the main application window. The idea behind this approach is to have one MainWindow which can display any WPF user control in its content. In this way there will be single window application which can just switch between different components and only show the currently active component. Every component, such as menu component, panorama component and HTML view component is implemented as separate WPF user control.

Complete configurability of the application allows extensibility and easy maintenance of the application during its lifecycle. The backend user is allowed to change the structure, content, texts and image resources at any time, and with one quick restart of the application complete new system state could be achieved.

Modularity and configurability of the application allow the application to be distributed to different locations and configured remotely form centralized system. The university is able to set different points on different locations that will perform individually and at the same time the administrator is able to control and configure the content and the behavior of the application.

A. Implementation of Menu Component

Menu component is one of the main components and the entry point of the application. Upon starting with the usage of the application, the user is able to select from one of the three official languages of the university as a main language and to continue to use the application with the selected language.

This component provides easy and intuitive navigation through the menu using swiping gestures with the hands and selection of the menu items with performing push gesture with the right hand or performing a walking gesture. Every menu item is configurable and different image can be assigned per item. The action is also configurable and based on the configuration it can open another menu, open a HTML view component or can start different component, such as panorama view component.

The configuration of the component is done using XML files stored in the same directory with the application executable. This allows the application to be easily configurable and extendable with new content and resources. The XML file contains configuration for the current menu and can also contain link to different XML file which defines another menu which is specified using the action attribute of the item configuration.

Location of the main menu or entry menu configuration for the application is specified in the applications app.config file, which is XML configuration file intended for a .NET application. When the application is started, it looks for the location of the entry point configuration, loads the configuration and renders the menu (shown in figure 6.1). One simple menu configuration is shown below.

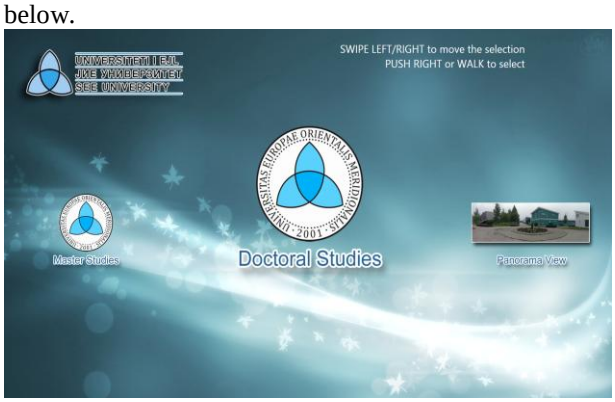


Figure 4. Menu component

B. *Implementation of the Web View*

Web View component presents any HTML content to the user, which is able to navigate in the content using hand swipe gestures. This component is very useful for presenting rich content to the user, including text resources, images, animations and any other objects which can be displayed using Hypertext Markup Language.

The user is able to navigate or scroll through the content of the HTML view using the hand swipe gestures. The advantage of this component is that the content is configurable and it can be stored locally or if the client has internet access it can be stored on any place on the web. This allows real time content update and centralized content location for all distributed clients. In the background, this application uses Internet Explorer engine for rendering the HTML content, so it is able to show any content that is compatible with Microsoft Internet Explorer web browser.

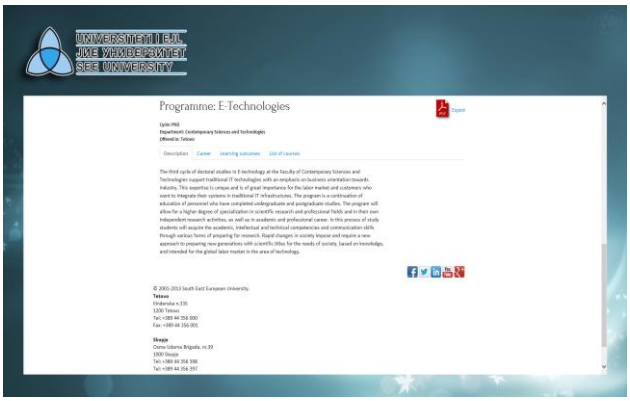


Figure 5. Web View Component Implementation

HTML view component is widely used in the University Exploration and Presentation application for showing different type of content. Study programs, text resources and additional information for different topics are all shown using the HTML view component. It allows easy and consistent rendering of the information equally to all the clients. Additionally, the feature for showing HTML content from a web page allows some part of the information to be directly get from the university website (www.seeu.edu.mk), in that way providing easy update of the application and keeping the content consistent between the website presentation and application presentation of the university.

C. *Panorama View Component*

The third component of the base application components is the panorama view component, which shows panoramic images to the user in interactive way.
The user is able to move the image to the left or right using hand gestures and move the panoramic view to the left or right, thus allowing the user to explore the space in 360 degrees view.



Figure 6. Panorama View Component Implementation

This component allows the user to see the objects nearby his current location and get more information about each object in the current view. Easier navigation is achieved using the hand gestures which allow moving from one object to another one.

VII. RESULTS

After the completion of the system, different kinds of tests were performed in order to ensure the system performance, functionality and stability.

Usability testing that was performed on the system ensured that the system is usable and responds to the user behavior as expected during the design phase. Testing of the system was performed by people from different age ranges with different habits in order to test the system for its stability and acceptability.

The results from the testing show that the users learn the gestures very fast and they start to effectively use the system in a very short amount of time.

TABLE I. Time needed for people from different age to start to effectively use the system

Age Range	Up to 2 minutes	2-7 minutes	7+ minutes
< 15	40%	52%	8%
15-20	77%	19%	4%
20-25	68%	23%	9%
25-30	63%	24%	13%
30-35	45%	39%	16%
35-40	22%	40%	38%
> 40	14%	34%	52%

From the table above we can see that 78% of the people from before-university grade will start to effectively use the application features in less than 2 minutes. Also, most of the people in the age range of 15-30 years will master the system in less than 2 minutes.

From all the people younger than 30 years which tested the system, more than 90% of them learnt how to use it in less than 7 minutes, while more than 50% of all the tested people achieve the same goal in less than 7 minutes.

WaveLeft and WaveRight gestures also have high percentage of recognition, about 90%. They are natural to the user but are composed of more gestures and are more complex for the system. At the bottom of the list are ClockwiseCircle and CounterClockwiseCircle which are the most complex gestures. They are consisted of 4 different positions which should be performed in a specific order. These movements are not so natural to the user, they are not present in the everyday movements and they are not so easily performed by the person.

This concludes that the gestures which are composed of smaller number of states are much easier for performing. The gestures should also be natural to the user and intuitive enough so the person doesn't need much time for learning the gesture. Before composing a new gesture, user behavior should be monitored and taken in mind, the gesture should be designed with as little states as possible and it should be natural to the performer.

VIII. CONLUSION

This research tries to make several contributions in the development and application of Natural User Interfaces and university promotion and presentation providing a new approach.

First, it presents different approach for presenting the university and any of its internal institutions to the outside people and future students. Besides the traditional brochure and TV/radio marketing campaigns, with this system the university is able to present itself to the public in different, remarkable and distributed way throughout the whole university year, with little or no costs at all. The highly interactive way of communication between the user and the system will leave high impression and great mark at the user experience.

Second, this system is built on basis of separate library used for the gesture recognition part of the application, which can also work separately from the system and can be used in different contexts. Since this application thesis will remain in the university library, it can be basis for any other future research papers and systems that utilize the Kinect device for gesture recognition systems. It is very extendable and allows the users to implement new gestures on top of the library, thus enriching its context and provides easy start for the user allowing him to concentrate to the business logic instead of the low level programming of the gesture recognition algorithms.

In addition to the basic gestures provided by the employees at Microsoft department in United Kingdom, created several other gestures that add additional value to the library itself. New gestures that were added to the library are DownSwipe, UpSwipe, ClockwiseCircle, CounterClockwiseCircle, Push gestures, ManualDoorOpen gesture, AutomaticDoorOpen gesture and WalkingGesture. After additional testing and small code improvements, the new gesture library is planned to be sent to Microsoft team working on the Kinect SDK and also to publish it online as an open source contribution to the community of programmers who are working on natural user interfaces improvements..

For the future work intended is to add more common functionality, such as course based real time chat, messaging, file sharing, library and wiki panel. Also will try to introduce new gestures and implement new components.

The future research will focus on head-driven gestures, i.e. gestures which are performed only with movements of the head. One of the main challenges is also configuration of the gestures using XML files instead of the programmatic approach which will also improve the extensibility of the application and new gestures could be added without modifying the application code.

REFERENCES

[1] Brin, Sergey (2012) Project Glass Demo at Google I/O – Google I/O conference

[2] De los Reyes, August (2008) Predicting the Past – Conference Presentation

[3] Fetaji Majlinda., and Fetaji, Bekim., "Universities go mobile: Case study experiment in using mobile devices" – ITI IEEE conference, Dubrovnik, Croatia, 2008

[4] George R. and Joshua B. (2010) Objects, Containers, Gestures, and Manipulations: Universal Foundational Metaphors of Natural User Interfaces

[5] Jetter H., Gerken J. and Reiterer H. (2010) Natural User Interfaces: Why We Need Better Model-Worlds, Not Better Gestures – University of Konstanz

[6] Kavanagh, Sam (2013) Facilitating Natural User Interfaces through Freehand Gesture Recognition – University of Auckland

[7] Kean, S., Hall, J., Perry, P. (2011) Meet the Kinect: An Introduction to Programming Natural User Interfaces – Apress

[8] Knies, Rob (2010) New, Natural User Interfaces – Microsoft Research Paper

[9] Liu, Zhaochen (2013) Design a Natural User Interface for Gesture Recognition Application – University of California – Berkely, EECS Department, Technical Report No. UCB/EECS-2013-101

[10] Moore, Christian (2006) New community Open – Natural User Interface Foundation Community Article

[11] Widgor, D. and Wixon, D. (2011) Brave NUI World: Designing Natural User Interfaces for Touch and Gesture – Morgan Kuaffman

[12] Xu, Wenkai and Lee, Eung-Joo (2013) A New NUI Method for Hand Tracking and Gesture recognition Based on User Experience - International Journal of Security and Its Applications, Vol. 7, No. 2, March, 2013

ICT supported laboratory testing of installation circuit breaker

Davor Gadže, Milivoj Puzak, Dragutin Haldek
Department of Electrical Engineering
Polytechnic of Zagreb, Zagreb, Croatia
davor.gadze@tvz.hr, milivoj.puzak@tvz.hr, dragutin.haldek@tvz.hr

ABSTRACT - Circuit breakers are one of the most common safety devices in electrical circuits. Presentation of their characteristics to the students will be shown in this article. Characteristics could be shown in three different ways. First way as convenient experiment with standard instruments and mutual presentation of achieved results. Second, as building model in mathematical modeling software and critical review of corresponding results. And third as fully automated experiment with programmable logic controller for control and data acquisition and visualization software for automated results presentation. All three ways of presentation of these characteristics to the students will be critically reviewed.

I. INTRODUCTION

Circuit breakers must simultaneously satisfy several conditions: as shown in Fig 1, continuously leads full load current - area [1], occasionally leads short overcurrents - area [2], delayed disconnects long overcurrents - area – [3], immediately disconnects short circuit currents – area [4]. All of these characteristics have also fuses, as shown in Fig 2, but not so fine tuned. Similarities are in areas [1], [2] and [3]. As shown in Fig 3, this first part of characteristic is called thermal, and both circuit breaker and fuse have it incorporated.

Thermal part depends on heat dissipation, I^2t, and disconnects supply in case of circuit cabling temperature rises to limit, during hours. Maximum admissible temperature at the conductor is 70°C for PVC type of insulation, and 90°C for XLPE and EPR types of insulation (table 52-4 of IEC 60364-5-52), [4].

Second part of circuit breaker characteristics is called magnetic, instantaneous or slightly time delayed. Magnetic part depends mostly on current, I. This magnetic part enables faster break in case of short circuit. It also disconnects supply in case of circuit cabling temperature limit rises to 125%, during short circuit. It is assumed that during this time, typically fraction of a second, generated heat inside conductor remain in conductor, causing its temperature to rise.

Magnetic part of characteristics also disconnects supply in case of particular fault. During this fault, exposed conductive parts which are not normally alive, has become alive. Protection from such fault is in IEC 61140 called "Fault protection", former "protection against indirect contact". Maximum safe duration of the assumed values of AC touch voltages from 230V to 400V is 0,2 seconds.

Current IM, on Fig 1, for B or Z characteristics of circuit breaker (so called *fast*) is typically 3-5xIn, for C characteristics (*slow*) is typically 5-10In, and for D or K (*extra slow*) characteristics is typically 10-20xIn. Letters B, C and D describes circuit breaker characteristics for domestic use (IEC 60898, 0 level maintenance) and letters B, C, D, Z, K are for industrial use (IEC 60947-2).

These two different characteristics, thermal and magnetic, have different time domains. Thermal is shown in hours, and magnetic is shown in milliseconds. Except different time domains, characteristics also have different current domains. Thermal one covers current ranges from 1xIn, to 1,5xIn, and magnetic one covers current ranges from 3xIn to 50xIn. Due to these reasons, characteristics could be better presented if both axes have logarithmic scales, instead of linear ones.

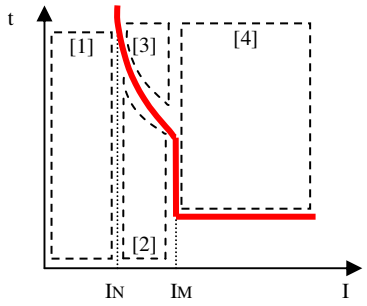


FIGURE 1: CIRCUIT BREAKER CHARACTERISTICS

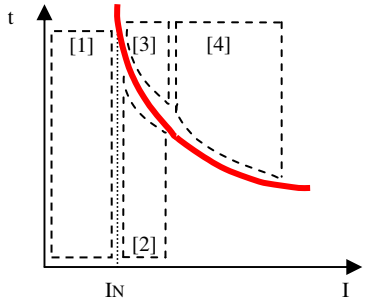


FIGURE 2: FUSE CHARACTERISTICS

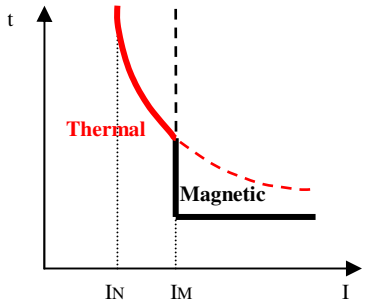


FIGURE 3: CIRCUIT BREAKER TWO PART CHARACTERISTICS

This introduction is held at lectures, before exercises, and gave students knowledge of basic principles in domain of circuit breakers.

II. CONVENINENT EXPERIMENT

As first exercise, students should connect circuit breaker to high short current supply. Circuit breaker is standard miniature circuit breaker, B or C characteristics, nominal current from 6 to 16 Amps. Nominal current is limited through supply capacity. Trip of circuit breaker should be measured at least for two different currents, whose ratio is at least 2 for magnetic part. For C characteristics it means 20xIn (320Amps), and 40xIn (640Amps).

So big ratio is pointless on thermal part, since time limitation in laboratory testing are 20 minutes for each current. This leads to minimum current of about 1,8xIn. A bit smaller currents, causes timeless experiments, e.g. 1,45xIn is 1 hour, causing impossible to carry experiment out on 1 hour lectures.

Experimental results of 3 pole short circuit on 3x400V 50Hz network with circuit breaker C63A are shown on Fig 4 and Fig 5. During short circuit voltage remarkably drops (Fig 4), and currents are remarkably high (Fig 5). Shown experiment lasts for 5,5ms, which could be better read from currents durations.

There is no doubts that, so short tripping time is caused because of magnetic part of circuit breaker characteristics. During first stage of short circuit, during current is rising, voltages drops to value defined only by current and impedance of short circuit. On second stage, during current is falling, voltage is defined by network impedance.

In laboratory experiment such oscillograms wouldn't be measured, due complexity of acquiring instruments. Only effective values of current and voltage will be measured.

Due short trip time, it is impossible to read off the current and voltage values during experiment, so some type of storage on instrument should be applied.

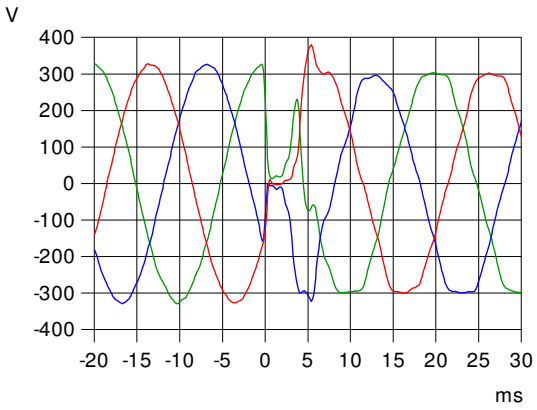


FIGURE 4: SHORT CIRCUIT, PHASE VOLTAGES, U1, U2, U3, [2]

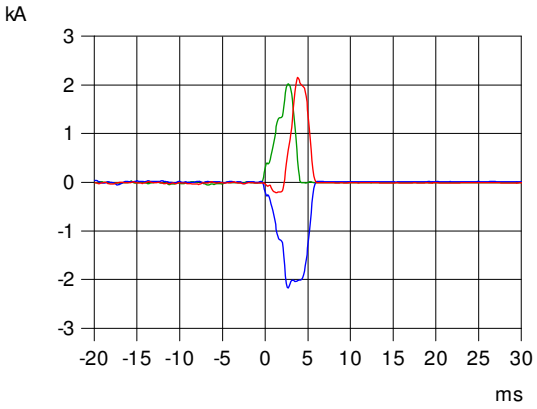


FIGURE 5: SHORT CIRCUIT, PHASE CURRENTS, I1, I2, I3, [2]

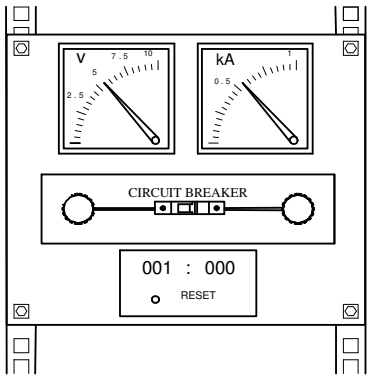


FIGURE 6: 19 INCH MODEL FOR CONVENIENT EXPERIMENT

Thermal part of characteristics lasts remarkably longer, for thousands of periods. With current "I" and breaker voltage "Ub", on thermal part, could be calculated internal resistance and power loss of circuit breaker. They are typically 10mΩ, for both impedance and resistance. Power loss are typically 2-3 Watt per pole, [5].

Students needs to adjust current supply for few current values, each causing breaker trip. When current start to flow, stopwatch is automatically started. This should be done with current sensing relay. When breaker opens

contacts, stopwatch is automatically stopped, with SD auxiliary breaker contact. Circuit breaker warms up during experiment, so current values should be adjusted in descending order, starting with highest value. Stopwatch should measure milliseconds for magnetic part, but also minutes for thermal part of characteristics. For 6 digit stopwatch maximum time limit is 16 minutes.

Advantages of this method to students are simplicity of the experiment, and good perception of physical conditions, leading to breaker trip, especially for thermal part of characteristics. Magnetic part, due its short reaction time, is a bit more difficult to recognize.

This convenient method doesn't have real disadvantages. Only one, according to students, is that experiment isn't modern one.

III. MATHEMATICAL MODEL

As second exercise students should build a mathematical model in Matlab software, Fig 7. They are modeling separately magnetic and thermal part of characteristics. And also, separately test these both parts of characteristics. Finally they incorporated these two parts into one model, and generate experiment report.

Modeling software offers complete test to be carried out, for whole range of currents, or even for whole range of circuit breaker, without limitation of experiment duration. These are obvious advantages of this method.

Disadvantage of this method, in the first row, is neglecting of circuit breaker complexity, [1] and [3], and too simple circuit breaker model. But, more complex model will result that students won't understood matter at all.

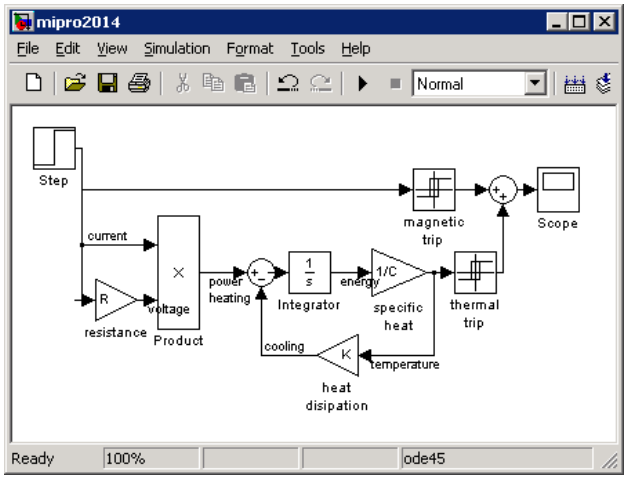


FIGURE 7: CIRCUIT BREAKER MATHEMATICAL MODEL

Second disadvantage is that students haven't good perception of circuit breaker propose. Currents in model are understood as numbers, not real Amps, through conductors become warm. Conductors' temperatures students could feel in other two ways of experiment.

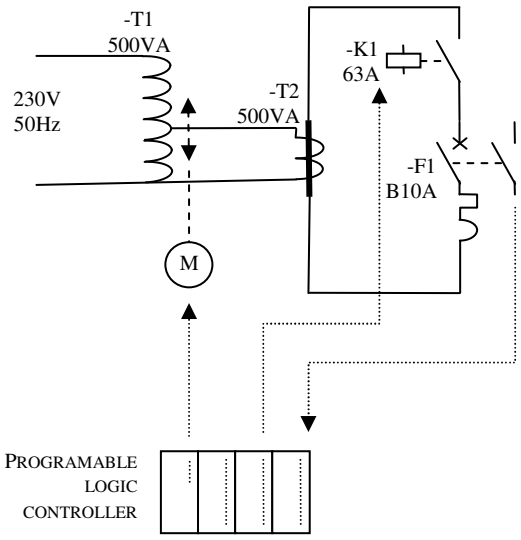


FIGURE 8: AUTOMATED EXPERIMENT SCHEME

IV. AUTOMATED EXPERIMENT

In third exercise students needs to select current range according to circuit breaker nominal current. Experiment is held as traditional one, with automated current set point adjustment, and automated acquiring and visualization of achieved results.

Programmable logic controller automatically defines current set point, according to programmed descending order, and forwards it through two digital outputs (Step up/step down) to high short circuit supply. This supply are two transformers connected in serial, -T1 and –T2, on Fig. 8. Transformer –T2 is oppositely connected current transformer, for this experiment apparent power of current transformer is 500VA, despite that current transformers usually have apparent power to 60VA. In laboratory experiments short circuits are carried out at significantly decreased voltage (few volts). With higher voltages, transformer with higher apparent power must be used, with nominal current at least 10 times then nominal current of tested circuit breaker.

PLC also automatically measures time from closing contactor –K1 till breaker –F1 trips. Contactor closing and breaker trip are recognized through digital 24V input wired on auxiliary contacts. PLC cycle could be as short as 1ms. For described "protection against indirect contact" in first subtitle maximum safe duration is 200ms, so resolution of 1ms is small enough.

Visualization of achieved results is shown on screen, Fig 9. Both results are shown, temporary one, with voltage, current, and time, from present current set point. Shown are also final results, with time vs. current curve.

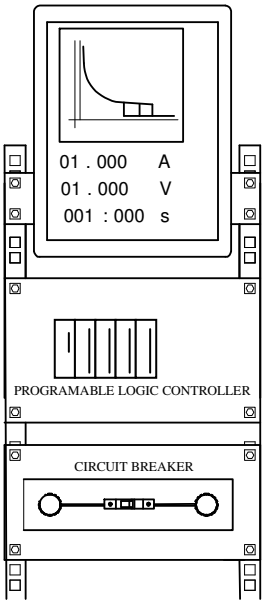


FIGURE 9: 19 INCH MODEL FOR AUTOMATED EXPERIMENT

All time limitations as in convenient experiment are present. This model could be also described as automated traditional experiment, because despite automation of adjusting set point and acquiring and visualization of achieved results, there is no other difference to convenient experiment.

V. CONCLUSION

It is impossible to say that one way is better than the other one. Each approach to the problem has its own advantages, so after introductory lecture, all three ways of experiment should be held to the students. In first one, they are forced to be active on exercise. They have a lot of manipulation, connecting circuit breaker, adjusting current set point, write down archived times. In second one, they have a lot of work on a PC, to build a mathematical model, and to handle all the result data. Finally at third one, after connecting circuit breaker, and choosing right testing procedure, they sit in front of model, and look at experiment as some multimedia presentation, but with real feel about conductors' temperatures, and feel of smell of burned dust on these conductors.

VI. LITERATURE

[1] Analysis of auxiliary DC installations in power plants and substations, Srđan Skok, Ante Marušić, Sejid Tešnjak, Milivoj Puzak, 6. Proceedings CC CIGRE, Cavtat, November 2003.

[2] Short circuit testing of 320kVA UPS, Gorislav Erceg, Davor Gadže, Danko Sirotić, Study, PEO, FER, Zagreb, March 2003

[3] Circuit-Breaker Model for Over-Current- Protection Simulation of DC Distribution Systems - Telecommunications Energy Conference, 1995. INTELEC '95., 17th International

[4] DATA SHEET: BMS_E / 10-2012, SCHRACK TECHNIK GMBH, Seybelgasse 13, A - 1230 Vienna

[5] Electrical installation guide 2013, According to IEC international standards, ISBN: 978.2.9531643.3.6, Schneider Electric Industries SAS, 35, rue Joseph Monier CS30323, F-92506 Rueil-Malmaison Cedex

Web-based control of programmable logic controllers in process control computers education

R. Buinac, D. Maršić and G. Malčić
Department of Electrical Engineering, Polytechnic of Zagreb, Zagreb, Croatia
rade.buinac@tvz.hr

Abstract - Industrial control computers can be connected to the Internet, and equipped with a Web interface. This enables access to process data from any computer with a set of common web-related software applications. This paper deals with laboratory exercises performed as a part of courses regarding process control computers, which include tasks dealing with this topic. Integration of Java and HTML technologies is described, and an example of a simple process and its Web interface is given. PLC computers equipped with an advanced communication modules are connected to a personal computer (PC) via Ethernet. Alongside implementation of Java applets in Web environment, human machine interface design is regarded.

I. INTRODUCTION

Programmable logic controllers (PLCs) are the most popular type of process control computers, used in a variety of industrial automation systems, controlling plants producing almost everything, from food and beverages to steel and cement. PLCs are controlled by programmed code created on a PC and downloaded to memory of the controller. Depending on the state of the inputs and results of programmed functions, the controller will activate proper outputs. Even though a PLC can control a part of a process, or an entire process automatically, it is desirable to have complete diagnostics of the process available for supervision and intervention by the human operator.

A system for operator control and surveillance enables management of an industrial process. Such a system is called SCADA, abbreviated from Supervisory Control And Data Acquisition. Main characteristics of a SCADA system are visualization of the process and control from a central location, intended to simplify surveillance and control of the process. SCADA system installed on a server computer can be controlled by any PC-type computer connected to the server via Ethernet, and a commercial SCADA system can be replaced by a Web-application control system for some types of processes, as described in [1].

PLCs are modular devices and this modularity is often used to provide additional functions, e.g. modules increase the number of available inputs and outputs for the PLC CPU or provide inputs with greater sampling frequency for fast paced processes. A composition of PLC CPU, power supply unit and modules is called a station. Some modules allow connectivity to certain types of networks, and sometimes provide additional, advanced computer networking functions and services.

The modules used in the exercises described in this paper have *Profinet* and gigabit Ethernet connections. Some industrial network specific services are supported, but its' IT functions are of interest in creating a Web-based system for process control [2].

II. WEB AND PLC INTEGRATION

The most demanding part of work in integrating the Web and PLC was done by module manufacturer. IT functions of the module include HTML process control, FTP functions and E-mail sending. Ability to monitor devices and process data using HTML is the fundament of the integration. In order to achieve it File Transfer Protocol functions for file management are required.

In fact, the module can be looked at as a web-server with pre-installed JavaBeans library, connected to the controller and able perform surveillance and control tasks over the process assigned to the PLC, as represented on Figure 1.

Java is a class based object-oriented programming language designed to be platform independent, meaning it can be compiled on one software or hardware platform, and run on another without the need to be recompiled. The JavaBeans concept allows Java programming language objects, most often classes, to be created and linked

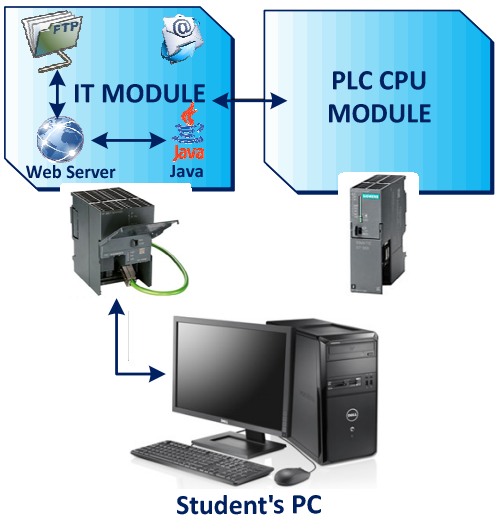


Figure 1. Information movement in the system configuration

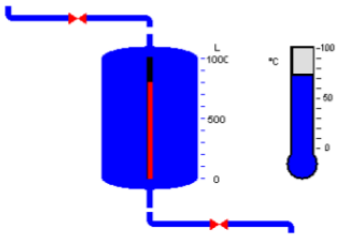


Figure 2. Graphic applets for display of memory variables

simply to executable objects [3].

In HTML process control JavaBeans are used as applets in HTML pages used for web diagnostics and control of the PLC. An applet in general is a small application performing a pointed task as a part of another programming interface, usually much more complex. Java applets are used in Web pages of general use as auxiliary applications, but here they are intended to be used to read and write data to the CPU module of the PLC.

JavaBeans Library pre-installed on the module contains applets specially designed for this type of usage called *S7 Beans Class Library*. The object classes contained in this library can be used for object-oriented access to a variety of information on the PLC CPU and for graphic display of process variables on the web page. Applets are inserted into HTML pages by their names when an applet call is embedded in an HTML page. Besides the name, certain attributes and parameters should be specified when calling the applets. This increases the amount of HTML code required to insert into the web page, but allows easier setup of the applets. These parameters define the usage of an applet, graphic display function, or its' type, etc. For example, some of the most useful applets are:

- *S7IdentApplet* is used to identify a module in a station
- *S7StatusApplet* queries the status information of a specified module
- *S7GetApplet* cyclically reads data or data areas
- *S7PutApplet* receives values entered by the user and transfers them to the PLC CPU [3].

These applets, integrated with additional *S7 beans* are used to create graphical interfaces on the Web pages. Beans used for graphical display vary from simple input fields and buttons, to more complex graphical solutions, such as the ones on Figure 2. Utility Beans performing simple functions such as reading date and time can also be used.

Advanced applications of Java beans could create the need for modification of existing beans or even creation of new ones, which requires builder tools and a Java programming environment. The builder tools are a part of the Java Development Kit, and a development environment can be chosen from one of the most popular integrated development environment (IDE) solutions. An IDE is a software application consisting of comprehensive set of tools whose purpose is facilitation of software development. New beans can be used for Web control after being imported into the library.

The S7 Beans class library provides an open interface allowing extended process data evaluation. This data can be transferred or used in table calculation, or in management information systems [3].

SCADA system is usually installed on a server computer, and any computer connected to the server via Ethernet can be used to control it. If a commercial SCADA system is replaced by a web application, there is still the need to have a central application and/or web application server in order to control the process from a visually acceptable interface [1]. However, HTML process control function and JavaBeans class library remove the need for the controller to be connected to a server computer – they are connected to the module, and the module is connected to an industrial computer network.

In order to use the module, and its' advanced functions and services, one needs to configure the hardware. During the hardware configuration process user accounts for control of the PLC station need to be created. Users can be granted rights to:

- Use the symbol table
- Read variables with absolute addresses from PLCs' memory
- Write variables with absolute addresses in PLCs' memory
- Access the files on the station with FTP
- Send an email
- Query the status of modules on the station
- Query the order number of modules.

Regarding user access level, there are three types of user accounts: accounts with no permissions, restricted permissions accounts, and administrator accounts. For instance, the automation engineer will have administrative account in order to get the system up and running, but the operator will have restricted access. Security is an important issue in all computer networks, but especially on the Internet, and it should be regarded as high priority if industrial processes are managed and controlled over the web. A tool for maintaining a decent security level in this system is definition of user accounts. Regarding security, the module has IP access protection providing security functions, e.g. restriction of communication with the station via the module for specified IP addresses.

III. LABORATORY EXERCISES

The goal of the exercises for the students is to explore remote surveillance and control possibilities given by Advanced IT modules for PLC automation. Additional goal is to give the students insight to Web control of PLCs basics, including introduction to HTML code and simple web pages design, FTP server manipulation and usage of Java applets for PLC control.

At the beginning of the exercise students are given a short introduction to basics of technologies used in the exercise, including Java, FTP, HTML and Web from the server point of view. PLC station configuration and PLC programming are familiar from exercises performed

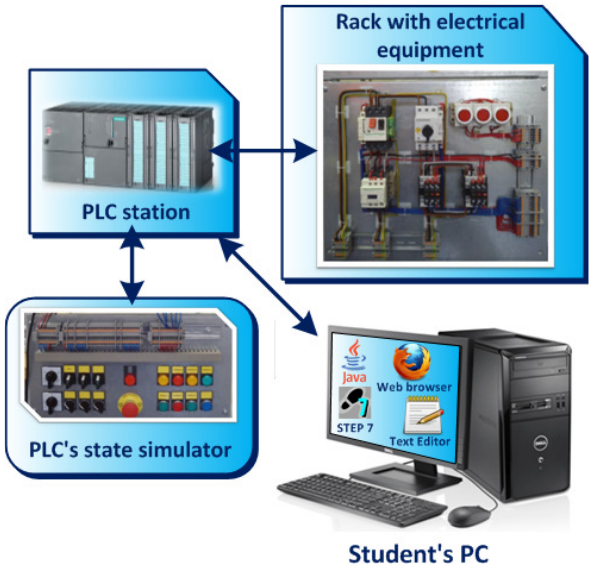


Figure 3. Workplace in the laboratory

earlier in the course. In case the short introduction is not enough, students can look up for more information in the teaching materials written for the exercise they are equipped with at the beginning of the class.

In the exercise manual students can find information needed to complete each step of the exercise. After an introduction to goals of the exercise and technology used to accomplish them there is a short review of the knowledge and skills gained so far. Detailed information and simple instructions can be found in the remaining chapters, following the tasks performed in order to complete the exercise. Educational goals of the exercise are accomplished by fulfilling the technical goal which is setting up a PLC station, programming it to control a simple process and deploying the Web control of the process. Students' workplaces are equipped with PCs connected to Ethernet and the Internet, a PLC state simulator module and racks with PLC stations and electrical equipment often used in automation industry, e.g. switches, buttons, relays, etc. Software installed on the PC is manufacturer's software for PLC configuration and programming, a Java runtime environment, a Web browser, and an advanced text editor for HTML code editing. Although HTML pages can be edited in a basic text-editing program it is much easier to work in an editor that has syntax highlighting capabilities, and it is important in the exercise because it helps students recognize parts of the code. Block diagram of the student's workplace can be seen on figure 3.

First step of the exercise is hardware configuration of the PLC station, performed always at the beginning of a new task for the station. During the configuration, the students should define IP address and the user accounts and passwords. If they are looking to know more, they can explore possibilities given by the software, under guidance of the teaching assistant, in order to modify settings related to Addresses, Diagnostics, IP access protection, etc. Second task is the design of the program for process control. This is done using ladder diagram logic.

Third task is symbols configuration – symbols simplify Web pages design by enabling Java applets to read or write data from a location called by a symbol, not an absolute address. Symbols configuration is basically adding text to a certain memory address of the PLC.

Fourth task is web page design and FTP server upload. During this task students are guided from making a simple web page with text only, to creating a web page including Java applets controlling the PLC. First version of a web page contains only title and a short text, in order to familiarize the students with writing HTML pages. After saving the page, it is uploaded to the module using a FTP client, and then opened by connecting to the module using a web browser. Second version of a web page has different background color and an image inserted, and the third version contains a simple applet. Applets code is copied from HTML files available at the students' computers' desktops. The code is copied in order to spare time, but also to avoid mistyping errors.

As a result of the exercises, student create a web page containing a text, an image and several applets. First applet shows catalog numbers of modules, second one shows status of the CPU module. Third and fourth applet allow modification of memory locations and the last one shows thermometer displaying the state of a certain memory address. The page is shown in Figure 4. After creating a page, students upload the final version to the module, and control the virtual process from the web browser on their PC.

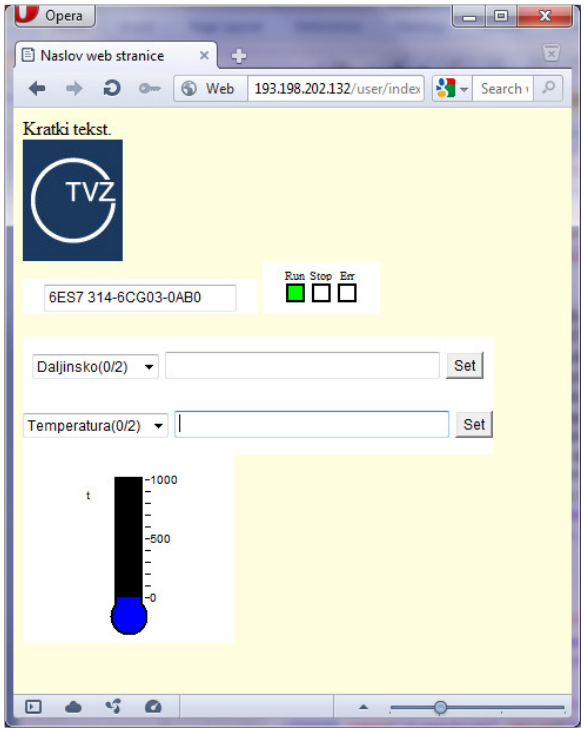


Figure 4. Web page created at the end of the exercise

IV. CONLUSIONS

Students' responses vary depending of their knowledge, IT skills and interests. While few of them find the exercise interesting, most of them find it too difficult. Lack of computer related skills is the greatest difficulty for most of the students. However, an electrical engineer should have the knowledge and skills to implement computer software components, especially if one is to work in the field of industrial automation, which has the need for versatile experts.

An advantage of this exercise is the ability to perform it from a remote location, for example at home. Student only needs appropriate software installed on the PC, and the IP address of the module to connect to in order to complete the exercise.

REFERENCES

[1] S Predanić, D Maršić, and G. Malčić, "Upravljanje procesom preko WEB aplikacije", 31st International Convention MIPRO 2008 : Proceedings ; Vol. 2 : Computers in Technical Systems & Intelligent systems (CTS & CIS), Rijeka, Croatia

[2] Siemens, "Information Technology in SIMATIC S7 with the CP 343–1 IT / CP 343–1 IT GX20 and CP 443–1 IT", Nürnberg : Siemens AG, 2003

[3] Siemens, "SIMATIC NET S7Beans / Applets for Advanced CPs Programming Tips", Nürnberg : Siemens AG, 2008

[4] A. Milinović, "Praćenje i kontrola procesa preko web preglednika pomoću Java appleta", Final thesis, The Polytechnic of Zagreb, Zagreb, 2011

Towards more efficient teaching

Janko Žufić, Tomislava Žajgar
Juraj Dobrila University of Pula / Department of Educational Sciences, Pula, Croatia
jzufic@unipu.hr, tzajgar@gmail.com

Summary: The enclosed study provides analysis and describes experiences from teaching the subject "Audio and video processing" conducted at the Teachers' Academy at the Juraj Dobrila University of Pula. During the first year of the collegium, audio processing was performed using the application "Audacity", and video by Adobe Premier CS4. Processing of the materials was done using examples assigned by a teacher. Some changes were introduced in the following two years of the collegium. Besides the work described above, students were tasked with a creation and development of an idea for a short movie, recording and editing of the material. The final product was a film 3-10 minutes in length. This approach led to a better success rate among the students: from GPA 3.52 (at the scale 1-5; σ = 0.83, N=29; 2 students did not pass; average number of exam attempts per student was 1.79, σ = 1.05) to 4.33 (σ =0.80, N=21; 2 students did not pass; average number of exam attempts per student was 1). Evaluation of teachers by students performed as a part of the official institutional evaluation also changed: "*Classes encourage interaction and team work among students*" went from 3.63 (σ = 1.11) to 4.0 (σ = 0.89); "*My work during the semester was regularly evaluated and graded*" went from 4.07 (σ =0.94) to 4.83 (σ =0.47); "*Tasks corresponded with the contents of the collegium*" went from 3.89 (σ = 1.15) to 4.33 (σ =0.82); "*Grades received during the semester have impact on the final grade at the exam*" improved from 4.25 (0.80) to 5.00.

Key words: Audio processing, video processing, more efficient teaching

I. INTRODUCTION

Students of the primary class teaching program at the Department of Educational Sciences at the University of "Juraj Dobrila" in Pula, after the first year of study opt to select one of the two offered modules: "Informatics" or "English Language". In addition to the basic competencies of the teaching profession, they also achieve additional competencies in one of the two modules. Each of the offered modules represents up to 25% of the total hours.

The module „Informatics" comprises, among the other subjects, of the collegium „Audio and video editing", with 60 hours of teaching. The general idea behind the introduction of this module in the curriculum of the Teaching studies was to enable future teachers for the individual camera recording and audio-visual material editing utilizing computer application and its application in the primary teaching. The intention was to enable future teachers to use information technology as much as possible, as it becomes more accessible and easier to use and obtain.

Similar Teacher Education faculties in the Republic of Croatia (at the University centers in Rijeka, Zagreb, Split and Zadar) do not offer such courses, while the Teacher Education Faculty in Osijek has one elective course („*Graphics, animation and films*" represented with a total of 30 hours), and the program is partially similar to the course „*Audio and video editing*". Similar courses, but with substantially more hours (and even a separate study program), exist at the Academy of Dramatic Art in Zagreb.

During the first two years (2009/2010 and 2010/2011) of the course, students practiced camera recording, editing of the audio and video material and were given home works, but the teacher's impression was that students made minimal efforts and was of the opinion that their acquired knowledge and skills would not be long-termed. In the following two years (2011/2012 and 2012/2013), teacher changed the teaching method.

A. *Project teaching*

Thinking how to change the teaching discourse, the teacher did not want to use traditional forms of teaching, as the outcome of teaching was not entirely accomplished. In the end, he opted for the method of demonstration and a segment from the Theory of Constructivism that focuses on students that are actively involved in the teaching process and come to the personal cognition through the individual experience. Path to learning is unique for each student as they through individual activity, „construct" their own knowledge with an individual effort based on their personal experience. Constructively efficient teaching is founded on the relevancy of the existing knowledge, immediate experience in reality, integral experience, reflectivity of learning and teaching, effective creativity of the inner motivation, role of the teacher and the method of integrity of teaching [1]. Furthermore, some of the characteristics of the contemporary teaching are achieved through learning by discovery, self-initiative of students, individual problem solving, divergent, creative thinking [2]. De Zan emphasizes teaching models such as problem-oriented teaching, learning based on problem solving, creative teaching, research-oriented teaching, project teaching etc. The common thread to all these models is the existence of the problem that needs to be resolved.

The Project Method applied by the teacher is the specific type of learning by discovery [3], and it is considered the most complex method of learning by experience as it implies work of several groups of students on a complex problem set in a real life environment [4]. Teacher needs to put an extra effort due to possible difficulties and uncertain outcome of the project, in spite of that, the method itself is very interesting and motivating for the students. Teacher plays the role of a project manager that supervises students during the entire process, while it is important for students to be able to work in a team environment, as they need to take up parts of the

assignment and complete their part of the tasks in order to successfully complete the project [5]. Project teaching requires more time, but achieved results are better and long-lasting [6].

Using the project method in teaching involves a student-centered partnership between a student and a teacher, an openness to the problem' situations and tasks from the real life, correlation with other areas of science and human activities, methods of cooperation and team-work, development of organizational and communication abilities of students, application of new teaching methods.

A1. Role of a student in the project tutoring

Project tutoring teaches students to learn individually and to be individually responsible. Students in most parts organize their learning processes. The problem is that schools in general do not provide these competencies to students and existing conditions in schools, such as large study groups, defined time frames, limited financial abilities and teachers that obstruct development of individual responsibility in students. These factors make the process of qualifying students for learning via project teaching that much harder [7]. Learning process in project teaching includes: definition of distinct goals, formation of working groups, decision on time frames and adhering to them, work assignment and its coordination, team-work, constructive conflict management, implementation of all activities to the end, presentation of individual results to other coworkers in the team control and evaluation of one's own work and other team-mates work.

A2. Role of the teacher in the project teaching

Teachers are assuming roles that outgrow the traditional roles. Depending on the project phases, roles will be changed in alternation. Teachers need to become familiar with the roles and help students understand what roles they assumed [7]. Roles of the teachers are: A) *Coordinators:* depending on the age group and project experience of students, teachers will at the beginning observe passively development of the project and only if necessary, coordinate the project. In this role, teachers are establishing external contacts thus simplifying the path to these contacts for the students. B) *Moderators:* Teachers observe the course of learning and make sure that all participants in the projects have appropriate conditions for the participation in the learning process. Teachers as moderators will not have any impact on the formulation of the process, but they can open the free space and give wider prospective. C) *Coaches, advisors, assistances:* Teachers are always ready to provide directions and offer a helping hand in case those activities come to the dead-end or if problems arise. D) *Conflict managers:* Teachers need to mediate in situations when conflicts within the working groups arise, but also in the situations that involve external groups and persons that are not part of the school environment. It is necessary to develop several scenarios and methods to bring parties to the conflict to a favorable solution. To prepare for this role, teachers might need to have additional qualifications. E) *Professionals:* In each project and in individual situations, teachers need to present their professional knowledge. F) *Teachers are learning too:* Project themes, in their complexities, sometimes outgrow the teacher's spectra of knowledge. Therefore, teachers on the project often acquire important learning experiences. Presentation of the role of a person who permanently learns is something that needs to be mentioned to the students in order to help them experience learning as a continual process.

A3. Phases in the project teaching

Phases in the project teaching are: preliminary phase, planning, implementation, project presentation and valorization. Each of the phases has several steps [8].

Preliminary phase: a) selection of the project theme – it needs to be interesting, under researched, within the area of the student's affinities; b) Selection of the task, or the project goal, while keeping them feasible.

Planning: a) familiarization of students with the selected theme – directing students to the literature or showing the initial steps that would help them with the project implementation and b) writing a project implementation plan that would incorporate the following: definition of place and time for the work on the project, duration of the project, definition of length of each phase of the project; selection of material and work methods; formation of teams and work division within teams; anticipation of problems and possible resolutions.

Implementation: Following the project plan with a permanent teacher's supervision.

Presentation of the project results: Upon the finalization of the project, results of all teams need to be merged and the implementation documented via recording, posters, web sites, lectures.

Valorization consists of: a) evaluation of the accomplishment of the project outcomes and b) critical analysis of the work on the entire project – difficulties, successes etc.

B. Realization of the project teaching

B1. Preliminary phase

Preliminary phase consisted of selecting goals and sub goals. The main goal was to independently (or in teams in certain elements), on a general level and by the principals of profession, participate in making, recording and editing of multimedia materials. Sub goals were: to get familiar with the functions of a video camera, to set and work with a camera, to work with a tripod and lighting, to master basic framing, interior and exterior lighting, separate audio recording on a video camera, transfer of recorded material to a computer and post-production of recorded material. Additionally, to edit and ad visual effects and prepare and convert material for posting on a web site or store it on a DVD media. Students formed groups for the additional tasks that included finding ideas for a film (documentary, feature or experimental films), writing of a screen-play, treatment and shooting script for the feature and experimental films, i.e. drawing a sketch for the screen-play and select locations, camera positioning, background text and questions, in case of a documentary film. They also had to find actors for the feature and experimental films and obtain all necessary authorizations for the filming locations.

B2. Planning

Planning consisted of: a) preparation of students for the selected theme – each student, or group of students, was introduced to the basic functions of the equipment and students decided how they would use it. For example, students independently chose the positioning of camera, camera angle, exposition, frames for shooting etc for each individual scene; b) writing a project implementation plan that would include (for each phase) selection of place and time for the work start, and also duration of the activities.

Forming groups of students was made in the preliminary phase and division of tasks within the groups was reduced to the decision who and when would use the camera (apart from the personal cameras, there was just one more camera available to all groups as per schedules). This phase involved also anticipation of problems that might occur and ways for their resolutions (obtaining authorizations for filming on certain locations, permission to work with children, transportation to the place of filming etc).

B3. Implementation

Implementation phase consisted of following the plan. After all the work tasks were assigned and students given basic instructions how to approach them, realization started with an active involvement of the teacher and his time equally divided to all groups.

B4. Presentation of the project results

Once a week, each group (or an individual in cases of writers of screen-plays or shooting scripts) had to present what was done during the previous week. Students from other groups and the teacher commented on what was well done and what needed to be redone. It went on every week for three months. In the end of that period, students completed their tasks. A total of 16 films (during the last two years) were made. Films were shown in a review of films that was organized for that occasion in a public cinema in Pula. Screening was organized before an audience of approximately 100 viewers (mostly students, their friends and families) and it was also mentioned in the biggest daily news magazine „The Voice of Istria“ („Glas Istre“).

B5. Valorization

Valorization consisted of a) evaluation of the percentage of the assigned project goals and b) critical analysis of the work on the entire project – difficulties, successes etc. An important fact needs to be mentioned and that is that none of the students had any pre-existing knowledge in any element used in their tasks, and the accomplishment is so much greater when compared to their starting level. Teacher was also very satisfied with the achieved results and convinced that knowledge and skills acquired through laborious work would be long-lasting. Feedback from the students was also more than encouraging, and it might be illustrated by the experience of a student.

A learning experience of filmmaking at the Teacher Education faculty – a student's point of view

> *We all know how to comment that a film is poorly made, or that story is bad, but do we know how the films are made? Experience of learning how to make a film at the Teacher Education program is something that changes your approach to the film art and it convinces you that filmmaking is art. None of the students had previous knowledge of camera, but we all recorded something with our cell phones. When we received our tasks, the problems arose, such as how to position the camera – color, contrast, camera angle and what that means. We had theoretical knowledge and knew the terminology from the module „Culture of media“. To apply it in practice was by far more challenging. We learned that filmmaking is a creative process and not only telling a story. This type of art demanded complete dedication and we traveled a long way from the idea to the realization. At the beginning we had many ideas that were, through mutual pondering, slowly building in one that made us say „Yes, this is going to be my film“.*
>
> *We had to gather ample documentation and it was time consuming. Each of us wrote at least 20 pages of text for a 3-minute video film and it took us three months to complete it. Each frame, scene, every movement, sound, facial expression and gesticulation needed to be put on paper prior to filming. Many students found writing shooting scripts problematic, as they were mostly visual types. What each frame means, how it would express tension or relief, which one made sense and which one did not – we all mastered that through the script writing and the film recording itself. Why is the shooting script important? It is one of the most important document in the filmmakinging project. It is based on the screen-play and it visually and technically presents the film prior to its filming. The most important task in the shooting script is the whole film, each scene written (and sometimes even drawn) by frames…*

II. RESEARCH METHODOLOGY

The goal of the research was to determine the presence of statistical deviation in the quality of teaching with the traditional approach and the project teaching in the module „Audio and Video editing“. The tools utilized for this research were: a) records of the exam's marks and number of attempts retrieved from the ISVU system (Information Service of the Institution of the Higher Education) and b) results of the institutional evaluation of all teachers and courses performed annually by the University. The tool utilized for the evaluation was in the form of a questionnaire and it was used by students who anonymously evaluated elements of the course and the teacher in the range between 1 (*Extremely bad*) to 5 (*Excellent*). Questions presented to the students were: *1. I received clear and detailed information on the goals and*

content of the course and what are the expectations; 2. Assignments were presented clearly and comprehensibly; 3. Themes were reasonably selected and in accordance with the assignment; 4. Teacher is drawing parallels between the assignment and other courses and the profession in whole; 5. Teacher promotes analytical, critical and argumented thinking; 6. I was encouraged to actively participate in lectures (participation in debates/ posing questions/answering questions..); 7. Course requirements are in line with the ECTS; 8. Teacher is accessible and open for communication; 9. Overall I am satisfied with the course.

III. RESULTS

Results (marks) of the students are presented in the Table I., and results of the institutional evaluation of teachers are presented in the Table II.

TABLE 1. ACHIEVED RESULTS OF THE STUDENTS

	2009-11	2011-13
Number of students N	29	21
Average mark by students	3,52	4,33
Standard deviation of the received marks	0,83	0,80
Wilcox test with correction z=3,245		
Average number of attempts at passing exams	1,79	1,00
Standard deviation of the number of attempts at passing exams	1,05	0,00
Wilcox test with correction z=3,459		
Did not qualify for the exam	2	2

TABLE II. RESULTS OF THE INSTITUTIONAL EVALUATION OF TEACHERS

No.	**Statement**	***N***		**Arithmetic mean[1]** $\overline{X}$		**Standard deviation σ**	
		2009-11	***2011-13***	***2009-11***	***2011-13***	***2009-11***	***2011-13***
1	I received clear and detailed information on the goals and content of the course and what are the expectations	28	21	4,32	4,50	0,67	0,55
2	Assignments were presented clearly and comprehensibly	28	21	**3,57**	**4,50**	0,96	0,55
3	Themes were reasonably selected and in accordance with the assignment	27	21	3,89	4,17	0,75	0,41
4	Teacher is drawing parallels between the assignment and other courses and the profession in whole	28	20	**3,71**	**4,00**	1,01	0,00
5	Teacher promotes analytical, critical and argumented thinking	27	21	**3,89**	**4,50**	0,89	0,55
6	I was encouraged to actively participate in lectures (participation in debates/posing questions/answering questions..);	27	21	4,04	4,33	1,06	0,52
7	Course requirements are in line with the ECTS	27	21	**4,00**	**4,60**	0,96	0,55
8	Teacher is accessible and open for communication	29	21	**4,38**	**4,83**	1,01	0,41
9	Overall I am satisfied with the course	29	21	**3,55**	**4,50**	1,06	0,55

In addition to the results obtained via marks and institutional research, students also provided verbal feedback that could be abridged in the following:

> *Watch the video films made by previous generations of students and learn from their mistakes.*
>
> *In addition to the program package for video editing Adobe Premiere CSx, learn also Adobe After Efects that is of great help during the post-production of the video material and for addition of special effects in order to make better final product.*
>
> *In addition to the public screening in the end of the school year, organize internal competitions of student's video works.*

IV. ANALYSES OF RESULTS

Analysis of the student's accomplishments reflect significantly higher score (from 3,52 to 4,33 – statistically significant difference - Wilcox test with correction z=3,245 that is much higher than the required 1,96 with p<0,05), and lower number of attempts at passing the exam (from 1,72 to 1 – statistically significant difference - Wilcox test with correction z=3,459 that is much higher than the required 1,96 with p<0,05). In both cases, two students failed the exam (they didn't qualify for entering the exams – they abandoned the program).

Analysis of the institutional evaluation of teachers indicate that students in all nine questions gave higher marks to the teacher (statistical difference was present in 6 of the 9 questions). Wilcox test was utilized in all researches and correction was introduced due to the equality of levels. For all cases, the border was set at z=1,96 i.e. p<0,05.

In total, student's accomplishments are better, students expressed more satisfaction, the teacher was more satisfied, results are long-lasting and the course incorporated elements from other courses (Culture of Media, Literature, Dramaturgy etc) although the effort of students and teachers was increased and more time was consumed.

V. CONCLUSION

The article presents experience gained during the course of „Audio and Video editing“ at the Teacher Education program in Pula, with the application of the project teaching method. Project teaching approach requires more time for preparation of teachers and substantial commitment of students, but the accomplishment is better and long-lasting.

REFERENCES

[1] Podravec, D. (2003.) "Radost učenja, integrirana i projektna nastava u osnovnoj školi" Osnovna škola Franje Viktora Šignjara, Virje

[2] De Zan, I. (2005.) "Metodika nastave prirode i društva" Školska knjiga, Zagreb

[3] Bognar, L., Matijević, M. (2002.) "Didaktika" Školska knjiga, Zagreb

[4] Varošanec, S. (2006.). Učenje otkrivanjem. http://web.math.hr/nastava/mnm1/ucenje.doc (viewed 13.2.2013.)

[5] Bjelanović Dijanić, Ž. (2012.) Računalo u istraživačkom radu studenta u nastavi matematike, Napredak, 153 (2) 203-218

[6] Žufić, J., Cvilinder, H. (2013.) ICT learning experience in the multilingual and multicultural environment of ERASMUS IP TAARLoG project // MIPRO 2013, Computer in Education/ Opatija, pp. 616-621

[7] Auchmann, M., Bauer, L., Doppelbauer, A., Hölzl, E., Winkler, S. (2002.) "Projektna nastava – prijedlozi za primjenu" K-education, Sarajevo

[8] Bjelanović Dijanić, Ž., Tkalec, M. (2010.) "Fotostrip u integriranoj i projektnoj nastavi hrvatskog jezika, francuskog jezika i informatike u srednjoj školi" Bjelovarski učitelj 15 (1-2), pp. 38-45.

The potential of multimedia in Cross-Cultural Education

Zita Buzasi*
* PhD Candidate at Josip Juraj Strossmayer University, Osijek, Croatia
Work: Faculty of Management, Sremski Karlovci, Serbia
zizy84@gmail.com

Abstract - This paper presents a research study that was conducted in the aim to investigate the perception of students and teachers about the potential of multimedia in teaching cross-cultural values. Our aim was to provide a helpful insight into the challenges of cross-cultural collaboration and the benefits of multimedia aided teaching. Critical thinking as a process that allows the average citizen to not absorb all the information as a sponge but learn how to filter all the information critically.

Key words: Multimedia, Cross-Cultural, Education, Critical thinking.

I. INTRODUCTION

The overall purpose of this research is to determine whether multimedia can contribute to promoting and acquiring cross-cultural values. The questions we hope to address can be formulated as follows:

Does multimedia aided learning of cross-cultural values have potentials?

What are their believes of cultural diversity?

We hope that our findings will help us assess their relationship and define activities strategies.

II. INTRODUCTION TO CROSS-CULTURALISM

Awareness of cultural diversity is present since the time of ancient Greece, moreover, with the word barbaros (βάρβαρος) greeks identified anyone who was *not Greek or was a stranger*. Acquiring cross-cultural skills is of increasing importance. Hofsteede defines culture as "*the collective programming of the mind that distinguishes the members of one group or category of people from others*". [1] Beside the term Cross-Cultural in contemporary literature we find confusing terms like transculturalism, interculturalism, multiculturalism and seemingly they all have a similar meaning. However, each has its own meaning that differs. In multicultural environments cultures are living side by side, in cross-cultural environments certain cultures are related to others, while in intercultural environments there are exchanges between different cultures.

Hammer et al. use the term ‘‘intercultural sensitivity’ to refer to the ability to discriminate and experience relevant cultural differences, and (...) intercultural competence to mean the ability to think and act in interculturally appropriate ways.” [2] Hofsteede for instance implied that “*culture is more often a source of conflict than of synergy. Cultural differences are a nuisance at best and often a disaster”* [3] Hammer et al. suggested in their work that “cross-cultural research has posited the importance of intercultural competence in both global and domestic contexts” and Bennets earlier work “has additionally suggested an underlying theoretical framework, which he calls the Developmental Model of Intercultural Sensitivity (DMIS), for conceptualizing intercultural sensitivity and competence.” [2]

For understanding the complexity of culture is often used Hall's "iceberg" analogy of culture, according to which only 1/10th of the iceberg is visible, while the larger part is hidden under the water. Visible, like art, literature and folklore are easier to recognise to the outsider, but underlying beliefs, values, thought patterns, tastes... are recognisable only with time. [4] Situations and occupations that require certain cross-cultural skills are more common nowadays than a century ago.

Welsch also states that “Modern societies are multicultural in themselves, encompassing a multitude of varying ways of life and lifestyles.” [5] In the EU for instance the fundamental principle of free movement of workers entitles EU citizens among others to look for a job in another EU country and work there without needing a work permit. [6] It has simplified the migration within the EU, insomuch that “*at the end of 2012, 14.1 million EU citizens were residing in another Member State*”. [7] Other situations that require cross-cultural skills are for instance working in a multinational company, where the individual is in a situation where he/she often works daily with other employees from other cultures. Sometimes he/she is being transferred to other countries or even continents and being the only member of a culture can trigger cultural shock. Oberg Calervo in the 1960s introduced first the concept of Cultural Shock (Kulturschock) which implied “an occupational disease of people who have been suddenly transplanted abroad. (...)

Cultural shock is precipitated by the anxiety that results from losing all our familiar signs and symbols of social intercourse (...) when to shake hands, what to say when we meet people, when and how to give tips, (...) when to take statements seriously and when not." [8]

In 1992 Bhawuk and Brislin developed for example The Intercultural Sensitivity Inventory (ICSI) which was used to measure an individual's ability to modify behaviour in culturally appropriate ways when moving between different cultures. The results of their research was that that participants with three or more years of cross-cultural experience showed a greater degree of intercultural sensitivity .[9]

Bennet observed that as people became more interculturally competent they moved from ethnocentrism to ethnorelativism. With the term "ethnocentrism" he refered to the experience of one's own culture as "central to reality." He suggested that the beliefs and behaviours that people receive in their primary socialization were unquestioned.The term "ethnorelativism" is the opposite of ethnocentrism. (...) The more ethnorelative worldviews are ways of seeking cultural difference, either by accepting its importance, by adapting perspective to take it into account, or by integrating the whole concept into a definition of identity. [10]

A typical situation when cross-cultural values are essential is for example working in a multinational corporation or team. One can experience the benefits of multinational team work in the form of different approaches, working styles or working relationships. These teams are nowadays often virtual teams with co-workers who rarely if ever meet in person. Godar and Pixy Ferris states that "virtual teams may be especially vulnerable to communication miscues because of their reliance on electronic rather than F2F communication and because team members may speak different languages or come from different cultural backgrounds, where communication styles vary." [11] Managers who are aware of cultural diversities (for example how different cultures react in the same business settings) can avoid subtle miscommunication and cultural miscues.

III. MULTIMEDIA

In this paper we will examine the potentials of using ICTs, namely multimedia in teaching cross-cultural values. Considering that multimedia applications enable the synergy of media and simultaneous effects on multiple sensors (speech, sight, and sound), their use in education has many advantages:

- Engaging students in the teaching process,
- Eliminating the barriers between different objects,
- Supporting individual learning
- Enhancing creativity, moreover convergent and divergent thinking is evolving.

The results are reflected in increasing cognitive abilities of students, as well as reducing the time required to master the same material.

With interaction the students actively participate and receive feedback on the quality of their work; they have an interactive experience. Considering multimedia can contribute to successful training only if it is didactic and pedagogically optimised we emphasize the importance of **multimedia didactics**. The approaches of teaching with integrated information and communication technologies represent a relatively new way of teaching that pedagogically differs from classical teaching methods.

Some authors like Johnson believe that the key to successful diversity education is integrating visual arts into the curriculum. She believes in the "capacity of visual arts to inform and reveal the history, the political and social ideals, the literature, the folklore, and the underlying psychosocial dynamics of a culture" also states that "art not only provides genuine insight into particular cultural experience, but also expresses the universal truths of our shared collective genealogy. [12]

Gonzalez identified that students need to "develop critical thinking skills" accordingly she recognized the importance of following skills:" [13]

- Analytical Thinking & Critical Analysis
- Associative Thinking
- Creative Thinking
- Critical Thinking
- Evaluative Thinking
- Hypothesis Problem Solving
- Perceptual thinking – Interpretation & Summarizing
- Questioning
- Synthesizing"

The author also agrees with Gonzales' idea, that exercising all nine of the above skills (implemented in multimedia) can contribute to learning cross-cultural values.

The aim is to develop user friendly interface design and collaborative multimedia application. Considering, web-based software and applications provide superior accessibility and collaboration our multimedia should also be in the cloud. The advantages of **web based multimedia** are communication and interaction in wider circle, access from any web-enabled computer, interactivity-active participation and also cost efficiency.

A. Overcoming cultural stereotypes

As introduced above, our main aim is overcoming cultural stereotypes that are unfortunately often too reduced to its essentials and consequently easily

recognizable, but on the other hand often are hidden in everyday life for example in media (newspaper, TV, internet...). Therefore it is important to have the ability to question the sources of information. In connection to this Hobbs defines 5 questions for teaching media analysis that enables students to "*merge their understanding of construction and intention together to consider their own personal response to media or a media message"* [14]*:*

- Who is sending the message and what is the author's purpose?
- What techniques are used to attract and hold attention?
- What lifestyles, values and points of view are represented in this message?
- How might different people interpret this message differently?
- What is omitted from this message?

These "literacy" skills/abilities will help to encode and decode symbols and to synthesize and analyse messages. Anaïs Nin notes that "We don't see things as they are, we see them as we are."

B. Possible activities

Based on the above, we suggest the inclusion of the following activities:

- Work in groups of 15-20 participants
- Online scrapbooks, creating interactive online posters and wall charts that will allow our group to work together.
- Analysing artworks. Naming the differences between European and Asian art, learning about folk-art and folk tales
- Analysing media with asking questions.
- Digital storytelling. Finding similar folk tales in different cultures.
- Introducing other aspects of multinational and cross-cultural teamwork, stressing its advantages and disadvantages, backed with concrete examples and case studies.
- Sharing experiences with cultural miscues and language barriers for example acting out misunderstandings through role plays and express their views.
- Participants can be asked to identify the dimensions of culture provided by Hofsteede: "Individualism – Collectivism, Masculinity – Femininity, Power distance, Uncertainty avoidance, Long term – Short term orientation, Indulgence – Restraint."[3]

IV. RESEARCH FINDINGS

Guided by various instruments found in modern literature measuring (global) intercultural competency (for example Bennett's Intercultural Sensitivity Index (ISI) or Bhawuk's and Brislin's Intercultural Sensitivity Inventory (ICSI)) we developed our preliminary and simple questionnaire. The aim was to determine if our idea would be environmentally, technically, socially acceptable and sustainable and whether our participants would be interested in this kind of learning. Our sample size is 60 participants. Based on the fact that our region is cross-cultural we chose our participants randomly with equal probability of being included in the sample. We targeted members of higher education; therefore data for this study were collected primarily from teachers employed by universities (Vojvodina, Serbia) and also from undergraduate/ postgraduate students.

Besides gathering descriptive data on our participants (Questions 1 to 3) in the next section of the questionnaire the aim of the conducted study was to assess **1. numerical data** on: a) at what level was multimedia used in education (Questions 4 and 6), b) did they have positive experiences in multimedia aided learning (Question 5) and whether they ever used multimedia for learning a foreign language and **2. open-ended questionnaires** with questions such as: did they think it was useful (Question 7 and 9), c) what were their thoughts of using multimedia to teach cross cultural values (Question 13), d) whether they believed in cultural stereotypes (Question 11) and were asked to name the advantages of recognizing cross-cultural values (Question 10) as well as situations when knowing them would be advantageous *(*Question 14*),* also they were asked to identify how much knowing a foreign language could help them understanding different cultures (Question 8) which was measured using a 5-point Likert scale.

As a result of this study it was indicated that

- 80% of participants are females.
- 80% belong to the age group of 25-35
- 44 participants (73,33%) are employed as teachers. When asked whether they used multimedia in their class 90% of them responded affirmatively as showed in the following figure.

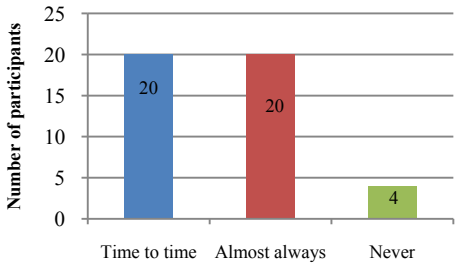


Figure 1. The use of multimedia in teaching.

- 100% of participants thought that multimedia can aid in learning a foreign language. 80% of participants had experiences in learning a language with multimedia. The remaining 20% despite stating that they had never used multimedia to learn a foreign language had thought that it could be useful;
- 100% of participants thought that learning a foreign language can help bridging differences with other cultures. This particular question indicates their attitude towards openness and that our participants recognized that learning a foreign language implies learning about that culture. Through the study of other languages individuals gain knowledge of the culture of that certain language, for example they learn ways of how to address people culturally appropriate. We believe that through foreign language learning in school, all our participants "met" with a different culture.
- 73% participants reported believing that multimedia can help promoting cross-cultural values.
- In general 73% of our participants didn't believe in cultural stereotypes.
- 67% thought it would be useful to implement cross-cultural training into education. As presented in Table 1, among the most common reasons are a) building tolerance, respect, b) living and working in multicultural environment, c) learning cultural values of other cultures.

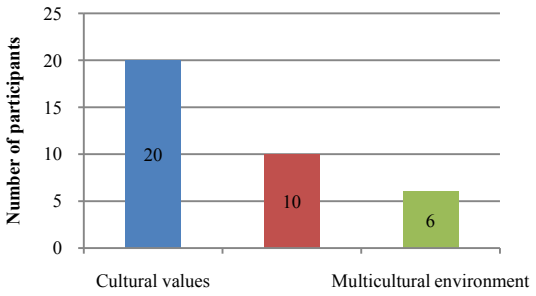


Figure 2. Benefits of implementing cross-cultural training into education

- In question 14 the participants were asked to put into words and describe in a few sentences their thoughts and name situations when cross-cultural values would be of service. With open-ended questions we wanted to give them the freedom to express themselves with their own words. We believe that our targets need to develop the ability to think and act outside of their particular cultural circumstance; hence we filtered out the most common recurring situations they identified (not limited to only one situation) and presented them in Figure 3.

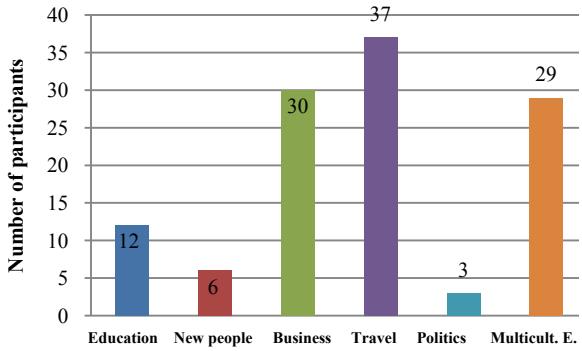


Figure 2. Situations of importance for assessing cross-cultural values

V. CONCLUSION

The purpose of this preliminary study was to assess the awareness of our citizens about cross-cultural values, stereotypes and their opinions and experiences about multimedia aided learning., our targets in higher education need to think

On the question whether multimedia aided learning of cross-cultural values have potentials we came to the following conclusion:

Based on the positive comments we received from our participants we consider that this research proves there is a legitimate need for involving multimedia in teaching cross-cultural values.

Their beliefs about cultural differences show us that they are open to other cultures and they identified that assessing cross-cultural values can help them in building tolerance and respect, deeper understanding of their multicultural environment. We appreciate their different views however we consider that the number of 16 participants who believe in cultural stereotypes is disturbingly high. We feel that considering 12 of them are teachers it should be lower.

We also concluded that the age of respondents had not influenced the responses.

Also, further understanding of how multimedia can support learning cross-cultural values is needed.

VI. POSSIBLE OUTCOME

The phase of realisation can be divided into three parts:

1. The first step is defining the indicative budget.
2. Finding partners. Considering that creating multimedia for learning cross-cultural values requires interdisciplinary skills, namely design skills, software development skills and detailed knowledge in culture, interdisciplinary teamwork is proposed.
3. Finding financial support for the realization of our idea.

VII. REFERENCES

[1] G. Hofstede, "Dimensionalizing Cultures: The Hofstede Model in Context". Online Readings in Psychology and Culture, vol 2. issue 1, 2011.

[2] M. R. Hammer, M. J. Bennett, R. Wiseman, "Measuring intercultural sensitivity: The intercultural development invetory", International Journal of Intercultural Relations, Vol 27, pp. 421-443, 2003.

[3] http://www.masterinternationalcommunication.eu/links/geert-hofstede-consortium/professor-dr-geert-hofstede/.

[4] E, T, Hall, "Beyond Culture", Anchor Books Editions, New York 1976.

[5] W. Wolfgang, "Transculturality - the Puzzling Form of Cultures Today", Published in M. Featherstone, S. Lash eds.,"Spaces of Culture: City, Nation, World." Sage, London, p. 194-213, 1999.

[6] http://ec.europa.eu/social/main.jsp?catId=457.

[7] http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2013:0837:FIN:EN:PDF.

[8] K. Oberg, "Cultural Shock: Adjustment to New Cultural Environments", Practical Antropology, Vol. 7, p.p. 177-182, 1960.

[9] D. P. S. Bhawuk, R. Brislin, "The measurement of intercultural sensitivity using the concepts of individualism and collectivism", International Journal of Intercultural Relations. vol 16, p.p. 413-436, 1992.

[10] M. J. Bennett, "Becoming Interculturally Competent", Published in J. Wurzel ed., "Toward multiculturalism: A reader in multicultural education", 2nd ed., Newton, MA: Intercultural Resource Corporation, p.p. 62-77, 2004.

[11] S. F. Godar, Sh. Pixy Ferris eds. "Virtual and collaborative teams: process, technologies and practice", Idea Group Publishing, London, 2004. ISBN 1-59140-268-9

[12] Johnson, "Art-Centered Approach to Diversity Education in Teaching and Learning", Available at: http://cspl.uis.edu/illaps/aboutus/facultyandstaff/documents/ArtCenteredDiversityEducation.pdf.

[13] D. Gonzales, "Teaching Cross-Cultural Values: 50 Interactive Critical Reasoning Skills Development Activities", Weekly Reader Press, 2004. ISBN: 0-595-33708-2

[14] R. Hobbs, "Reading the Media", Teachers College Press, New York, 2007.

Computer network applications, practical implementations and structural control system representations

Nderim Zeqiri*

* Faculty of Applied Sciences-Mechatronics – State University of Tetova, 1200 Tetovo, R. Macedonia, office, Mechatronics
nderimzeqiri@gmail.com

Abstract – The computer network play an important position for practical implementations of the differently system. To implement a system into network above all is needed to know all the configurations, which is responsible to be a part of the system, and to give adequate information and solution in real-time. So if want to implement this system for example in the school or relevant institutions, the first step is to analyze the types of model which is needed to be configured and another important step is to organize the works in the context of devices, as a part of the general system. Often before configuration, as important point is descriptions and documentations from all the works into the respective process, and then to organize in the aspect of problem-solving. The computer network as critic infrastructure is very specific so the paper present the effectiveness solutions in the structured aspect viewed from one side, and another side is, than the paper reflect the positive aspect in the context of modeling and block schema presentations as an better alternative to solve the specific problem because of continually distortions of the system from the line of devices, programs and signals or packed collisions, which are in movement from one computer node to another nodes.

Keywords: Computer network, implementations, modeling, structural representations.

I. INTRODUCTION

Often the first step in attempting to design a new part of software is needed to create firstly an algorithm which can be used to implement any practical solution. The main purpose of the paper is to bring information to the reader in general networks [5]. In this paper I have tried to restrain some formal rules and structural work in order to have a better content and be much more manageable and understandable to the reader. At the beginning of this paper, we present a brief introduction to some features of computer networks. The main part of the paper is to specifically include the purpose of the use and problems [2]. Also here is to clarify the reasons for the use of nets in relevant institutions, businesses (companies), while continuing to use them at home, then, we present the development and advancement of computer networks while being part of control systems [6]. A very special part is their use in social terms and their use by mobile users. Another objective of this work is to make the analyses and formulate mathematical model responsible and to develop algorithm for latter use or latter upgrade system [10]. So, this phase is an important step for extracting and gathering the requirements, such is, Figure.1 [12].

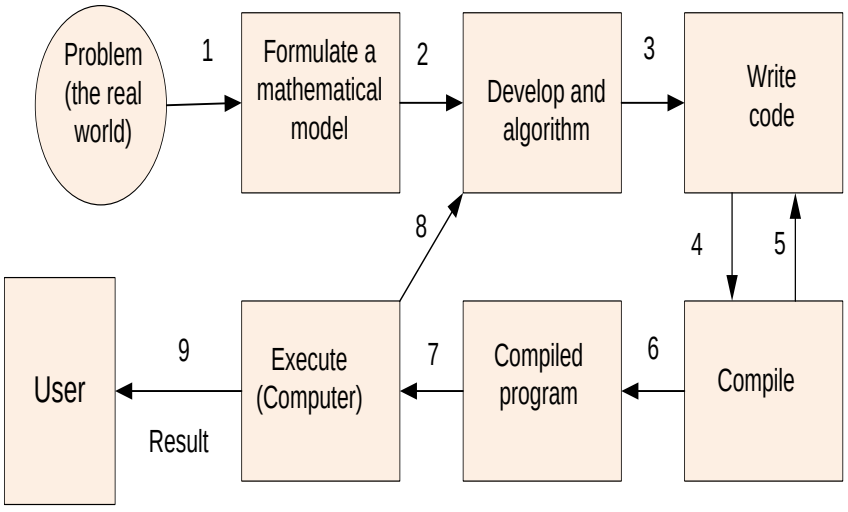


Figure 1. The process of problem-solving and respective elements

II. PROGRAM CODE AND COMPUTER PROBLEM-SOLVING THE CONTEXT OF TIMESET

In the next step is given program code in C#, to set up start time and end time for computer solving problem, and to return any value as true or false [10].

```
private bool TimesOverlap(System.DateTime dt1a,
System.DateTime dt2a,
 System.DateTime dt1b, System.DateTime dt2b)
{
   // technically the end time and start time of any back-
to-back
   // programs will fall within the same second
   dt2a = DateAndTime.DateAdd(DateInterval.Minute, -
1, dt2a);

dt2b = DateAndTime.DateAdd(DateInterval.Minute, -1,
dt2b);

   if ((dt1a <= dt1b && dt2a >= dt1b) || (dt1a >= dt1b
&& dt1a <= dt2b))
   {
      return true;
   }
   return false;
}
```

The method takes the input values of timeset1 start time, timeset1 end time, timeset2 start time, and timeset2 end time. It will return true if there is an overlap [12].

III. INTERNET ACCESS, AND THEIR INTERCONNECTIONS BASED ON RESPECTIVE ALGORITHM

In the next step is given general form, how to control any system, man-machine [8], LAN network, internet access, and all of them to be connected with Control system (to access an output object) such as robots in general context based on respective algorithm [13], Figure.2.

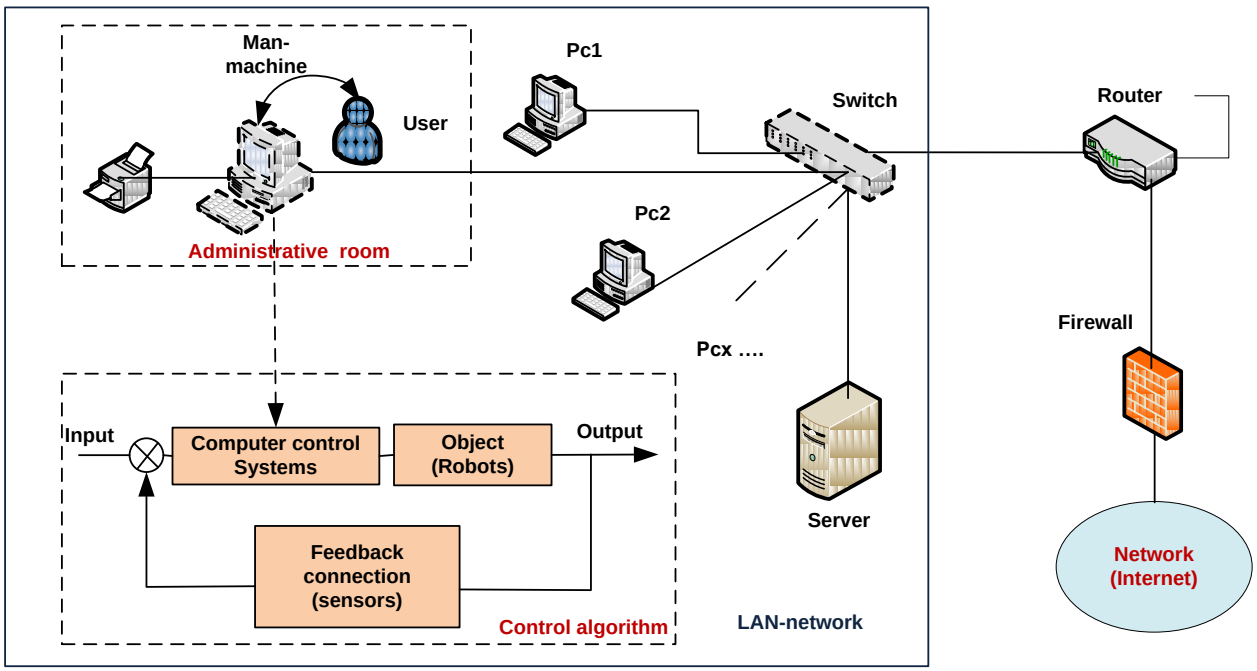


Figure.2.The control system (to access an output object) such as robots in general context, man-machine, LAN network, internet access, and their interconnections- based on respective algorithm

Very important part for above algorithm is the control algorithm, where can be installed in differently system, and also managed by human machine-interface [7], Figure 3.

In the next step, Figure 4, is given the part of the control system [2], where can be incorporated differently secured system, above all the system can be secured by smart card (with microcontroller) where can be stored the cryptographic algorithm to increase the security of the system [8].

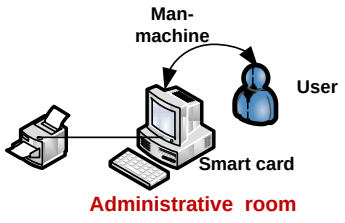


Figure 3. Man-machine, user, smard card as part of administrative room.

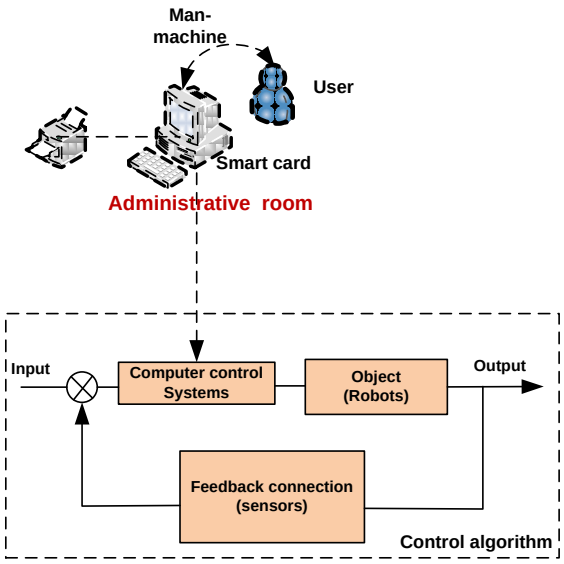


Figure 4.Control system as part of computer network, and administrative rooms, with smart card security.

IV. USING COMPUTER NETWORKS

Before we start to examine in detail the technical issues, it should be noted that although people are interested in computer networks and why use them. Ultimately, if no one were interested in using them, this means that fewer of them will be built. We will start with the company uses, and then we move to the use of home networks and develop capital last about mobile users and will conclude with social issues [4].

The basic elements of computer networks: for every computer network are required:

- At least two computers
- Every computer should have:
  - Network card
  - Cable, as intermediaries in communication
  - Passive and active device network
  - Operating Systems: Microsoft windows, Novell Netware etc.

Networks were developed as a result of business development. Business needed a solution on how to communicate effectively, how to avoid duplication of equipment and resources, and selection was carried out by networks of computers that increased productivity and saved money. Any manufacturing company creates hardware and software based on standards. As a result many companies were incompatible networks [4][5]. Network devices need to implement new technologies. One solution was the creation of standards for LANs, Figure 5. As a result compatible devices made by companies. New technologies definitely needed to distribute information between businesses, as far and as fast. Requirements of companies have a number of computers [4]. For example, a company might have a computer for every employee to view the products you design, to print brochures, and to do list for the payment of workers. The goal is to make all programs, equipment and especially data available to the entire network without having to connect the physical location, source or user. A clear example is the widespread and when some office workers share a common printer [1]. Small or large companies are vitally dependent on information. Most companies have customer data, product information, inventories, etc. financial statements [3] etc.

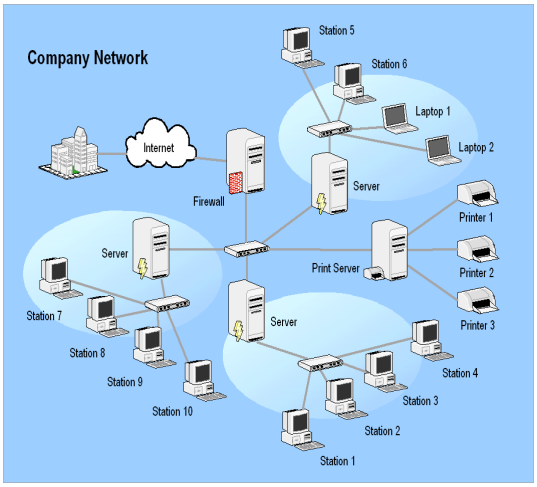


Figure 5.Company network (for example use a same printer-shared into the network)

Using nets and social issues: Social networks, message boards, information sharing sites allow people to share their views, beliefs, attitudes and above all their goals. Analytical-researchers of this practice argue that peer-to-peer networks and other content should be treated in the same way because they are all just "stream of bits" of network packets.

Foremost is the most used network is the Internet, which has enormous number up to the latest development of wireless networks. They use wherever you go people, resulting in a patchwork of coverage in cafes, in hotels airports, educational institutions, trains, planes. Anyone with a laptop and wireless modem can be connected to the Internet [7]. This approach increases the risk quickly to different situations created in the context of detection, the password and the performing harmful actions. Therefore, in Figure 2, is given the possibility of using the control system [8], where you can put a smart card for increased level of security through the use of different cryptographic algorithms [13]. However safety is always the threat from various factors, therefore the algorithm shown in Figure 2 enables flexibility to make updates, and generating respective insurance values.

V. THE PERFORMANCE OF THE COMPUTER NETWORK SYSTEM

For example, when is needed to compare the performance and to compare some characteristics of the computer Control system -to control robots movement and the role into the computer network system, firstly is needed to consider, which of the works are affirmative to be analyzed and to be in systematizations for dedicated procedures. If we suppose four tasks, so, if we have to

analyses the procedures of processing with differently structured systems, based in our (own) block structured system, here we may to give interesting results, with our (own) systems priority [11][12].

Computer Control System	Task1 (time-seconds)	Task2 (time-seconds)	Task3 (time-seconds)	Task4 (time-seconds)
Our (own) network -Tx	**3**	**3.5**	**2.5**	**3.5**
Network 2-Ty	**3**	**4**	**3**	**4.5**
Network 3-Tz	**4**	**4.3**	**3**	**4.8**
...	**...**	**...**	**...**	**...**

Table1. Performance verifications, for three differently system, and where is viewed the priority of own network control computer system

Tx-total time for (own) network Control system -to control robots),$T_{total\,x} = 12.5$; $T_{total\,y} = 14.5$; $T_{total\,z} = 16.1$.
So the better performance has our constructed system because, $T_{total\,x} < T_{total\,y} < T_{total\,z}$. So the better performance has the system that can be ready to solve any problem with short time of executions.

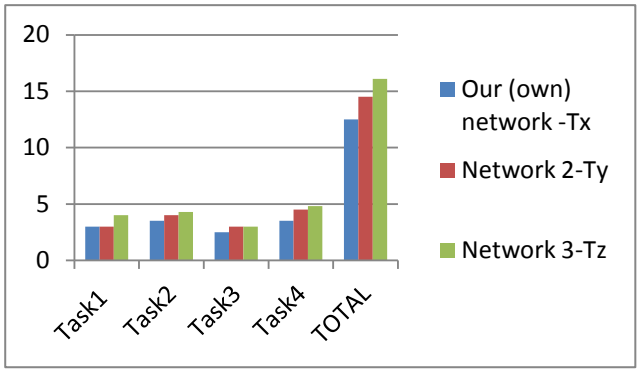


Figure.6. Graphical performance presentations, where can be seen, the own Control system -to control robots with duration time Tx-has a short time of executions, and so this system is with better performance.

VI. SOFTWARE ELEMENTS ANALYSIS

The most important task in creating a software product is extracting the requirements. Customers typically know what they want, but not what software should do, while incomplete, ambiguous or contradictory requirements are recognized by skilled and experienced software engineers [5]. Frequently demonstrating live code may help reduce the risk that the requirements are incorrect [7].
Specification: Specification is the task of precisely describing the software to be written, possibly in a rigorous way. In practice, most successful specifications are written to understand and fine-tune applications that were already well-developed.

Application Software:

- General business productivity applications: software program that perform a specific function directly for the end user, examples include: office applications.
- Cross-industry application software: software that is designed to perform and/or manage a specific business function or process that is not unique to a particular industry; examples include professional accounting software, human resources management [9].
- Utilities software: a small program that performs a very specific task. Examples include: compression programs, antivirus, search engines, font, file viewers, voice recognition software, etc [10].

VII. CONCLUSION

In this paper is presented computer network system and is given algorithm, to create a good space and meet the optimal requirements for the realization of the relevance of the use of computer systems. It is also made to find an adequate solution in the context of the domain, is simplified way of approaching the problem by improving access control and time, when using the respective operations. By determining the function of the carrier complex systems (systems with complex structure) used so far laid down rules which apply to the combination of dynamic elements. Also is given the Control system -to control robots, man-machine, LAN network, internet access, and their interconnections- based on respective algorithm [2]. Looking forward the rules are the basic rules of the control theory; also here can implement all the rules and methods of simplification, for dynamic systems and all complex structures can be transformed into basic and simple structure [13]. If our viewpoint is extended so we can create, a system, which can be interacted with others to increase the level of the observations. This level is very important to be stable system in the context of the practice because this may be a case to create the bridges for differently interfaces such as man-machine, field elements, remote control etc. Foremost is the most used network is the Internet, which has enormous number up to the latest development of wireless networks. They use wherever you go people, resulting in a patchwork of coverage in cafes, in hotels airports, educational institutions, trains, planes. Anyone with a laptop and wireless modem can be connected to the Internet. This approach increases the risk quickly to different situations created in the context of detection, the password and the performing harmful actions. Therefore, in Figure 2, is given the possibility of using the control system, where you can put a smart card for increased level of security through the use of different cryptographic algorithms. However safety is always the threat from various factors, therefore the algorithm shown in Figure 2 enables flexibility to make updates, and generating respective insurance values.

REFERENCES

[1] I.J. Nagrath, Control Systems Engineering, New Age International, Jan 1, 2006 - 858 pages

[2] Potashnik, Michael, and Adkins, Douglas. Cost Analysis of Information Technology Projects in Education: Experiences from Developing Countries. Education and Technology Series, vol. 1, no. 3. Washington:World Bank, 1996.

[3] Stephan Holzer, Yvonne Anne Pignolet, Jasmin Smula, and Roger Wattenhofer. 6th International Workshop on Algorithms for Sensor Systems, Wireless Ad Hoc Networks and Autonomous Mobile Entities, Bordeaux, France, July 2010.

[4] http://learnlink.aed.org/Publications/Sourcebook/chapter4/Computers_in_Schools_modelofuse.pdf

[5] M. Rosenblum, E. Bugnion, S. Devine, and S. A. Herrod, "Using the SimOS machine simulator to study complex computer systems," ACM Trans. Modeling Computer Simulation, vol. 7, no. 1, pp. 78–103.

[6] Y. Zhang and G. B. Adams, "An interactive, visual simulator for the DLX pipeline," IEEE TCCA Newslett., pp. 9–12, Sep. 1997.

[7] J. L. Bechennec, "ASF: A teaching and research object-oriented simulation tool for computer architecture design and performance evolution," in Proc.Workshop Computer Architecture Education (WCAE-98), Barcelona, Spain, Jun. 27, 1998, pp. 93–96.

[8] Paulo Tabuada, Approximate simulation relations and finite abstractions of quantized control systems.

[9] A. Bicchi, A. Marigo, and B. Piccoli. Feedback encoding for efficient symbolic control of dynamical systems. IEEE Transaction on Automatic Control, 51(6):987– 1002, June 2006. 14

[10] E. M. Clarke, O. Grumberg, and D. Peled. Model Checking. MIT Press, 1999.

[11] A. Girard and G. Pappas. Approximate bisimulations for constrained linear systems. In Proceedings of the 44th IEEE Conference on Decision and Control, Seville, Spain, 2005.

[12] Ames, S. R., Jr.; Gasser, M.; and Schell, R. R. 1983. "Security Kernel Design and Implementation: An Introduction." Computer 16(7):14-22. Reprinted in Advances in Computer System Security, vol. 2, ed. R. Turn, pp. 170-77. Dedham, Mass.: Artech House (1984).

[13] Lipner, S. B. 1982. "Non-Discretionary Controls for Commercial Applications." In Proceedings of the 1982 Symposium on Security and Privacy, pp. 2-10. Silver Spring, Md.: IEEE Computer Society. Proposes a way to use military-style mandatory security controls in a commercial environment.

The Significant Progress of the Microprocessors and Microcontrollers Course for Computer Science Students

Sasko Ristov, Nevena Ackovska, Vesna Kirandziska, Darko Martinovikj
Ss. Cyril and Methodius University, FCSE,
Skopje, Macedonia
Email: {sashko.ristov, nevena.ackovska, Vesna.kirandziska}@finki.ukim.mk, darko.martinovikj@gmail.com

***Abstract*—The objective of the Microprocessors and Microcontrollers course is the students to meet the general idea and concepts of a low-level assembly language and controlling the hardware devices with embedded microcontrollers. Since this is the hardware based course, which requires a lot of efforts by both the teachers and students, the students avoid to learn its material for gaining knowledge, but just to pass the course and continue with their studies. Apart of the courses bad reputation, it is usually among the courses with the lowest throughput. Starting from 2010, the courses syllabus was changed each year towards making the course more practical exercises and exams where the students could use their software skills. In this paper we elaborate all changes that improved not only the course throughput, but also the courses reputation, average grade, curiosity to learn hardware, etc, which are presented through years.**

***Index Terms*—Education, Hardware, Teaching, Learning**

I. Introduction

A general impression for all undergraduate teachers is that hardware oriented courses are not the favored by the computer science students. Therefore, awaking the students interested in hardware courses is a great challenge. Teaching how hardware works requires interactive exercises where instructor's efforts are the most important part of the whole hardware teaching process. Starting from the Computer Architecture and Organization course, through the Microprocessors and Microcontrollers (MM), and finishing in Robotics and Modern Computer Architectures, the computer science students are usually stacked somewhere before graduation. Most of the students hate these courses and learn them only to pass, without keeping some knowledge after taking the exam.

This motivated us to make many radical changes of the curriculum and program during the several study years in order to improve the MM course. Our goal was to make this course more attractive to our students, to alleviate student learning process and increase student curiosity about hardware. This paper describes the progress the MM course through years and analyzes the main students parameters.

The rest of the paper is organized as follows. Section II presents the related work in the area of teaching the hardware based courses. Section III presents the MM course organization, while Section IV introduces the improvements made in the course through the years. The results from our experience are presented in Section V. The latest course's changes for further improvements are presented in Section VI. The final Section VII is devoted to conclusion and future work.

II. Related Work

Different methodologies exist for teaching the computer science students about hardware concepts. One teachers prefer teaching through working on real hardware equipment, others teachers propose the usage of modern visual simulators, while there are new proposal for distance learning and using the remote engineering in virtual laboratory.

The teachers must introduce the appropriate hands-on exercises, homework assignments and projects, apart from the lectures in order to make the course more interesting and to provoke the students to dive more deeply to learn how the microprocessors and microcontrollers works and can be programmed to control some task. Liang [1] performed a nice survey of hands-on assignments and projects.

A. Practical work on real hardware

The most important part of the teaching process is the practical work. It is preferable for students to actually have their hands on real hardware [2], [3], [4]. Working on real hardware in the laboratory is also very important [5], [6], [7], which requires a lot of effort by both the teachers and the students.

Wang [8] and Lee [9] proposed FPGA-based configurable processors to be used in laboratory exercises. In this case, the students can develop, implement and monitoring processor systems on real hardware.

B. Hands-on Exercises: Using Simulators

Introducing the appropriate visual simulators in the laboratory exercises lightens the teaching process and can significantly improve the students' interest in hardware generally [10], [11]. Due to lack of real hardware, appropriate simulators can be used for practical exercises [12], [13]. Ristov et al. [14] presented how their EDUCache visual simulator can be used in the computer architecture and organization course.

C. Virtual laboratories and distance learning tools

The simulators gathered with web online tools can share the laboratory equipment, thus removing the obstacles of cost,

time-inefficient use of facilities, inadequate technical support and limited access to design and laboratory resources [15]. Gusev et al. [16] presented the e-Learning and benchmarking platform for learning the Parallel and Distributed or other related courses.

D. Teaching Methodologies

Several authors presented methodologies to teach the hardware courses in general for computer science students. Reinbrecht et al. [17] present a methodology that integrates functionally verified ASIC (Application-Specific Integrated Circuit) soft cores into an FPGA in order to allow the students to learn the fundamentals of hardware and its designs challenges, not only development, but also verification and physical implications. Hatfield and Jin [18] designed and developed many laboratory exercises to design and implementation of an operating model of a pipelined processor.

Da [19] elaborated the common methods of classroom teaching and experimental teaching and some efficient methods for classroom and laboratory teaching. He [20] addressed five problems related to computer architecture education in multi-core era and suggested a possible solutions.

III. About the course

This section describes the courses material and organization.

A. The course material

The MM course is obligatory course for computer science students in their third year of studies. The course is lectured with 2 classes of theoretical lectures, 1 class of theoretical exercises and 2 classes of practical exercises per week. The practical classes are conducted in computer laboratories with a group of at most 20 students.

The teaching material is divided in two parts, by one for each midterm. The first midterm covers the x86 microprocessors internal architecture, its instruction set, the interrupt handling system, BIOS and system routines. The material for this part is based on [21] and various source code libraries.

The material for the second midterm is focused on analyzing the organization of several types of microcontrollers, their instruction set. Additionally, modern peripheral and embedded systems are also covered.

B. Grading scheme

Students are graded based on their activity on the practical exercises, their projects and the results from the midterms or final exams. Student's activity on the practical exercises includes solving some given problems, which should be solved during the laboratory classes.

Student projects are more complex programming assignments that students usually do at home. They require additional research in the areas not covered in the course classroom. There are two midterms during the semester and there are three more sessions for the final exam. The exams are divided in two parts theoretical and practical. They should properly test student knowledge, so they are based both on the theoretical part of the course and the more practical one.

C. The MM course objectives

In the previous curriculum for MM the programming assignments were done in an inappropriate programming and simulation environment. Microsoft Macro Assembler [22] along in the Programmers Work Bench (PWB) environment. Since many national contestants and good students had problems with the course, it was obvious that the environment did not give students the appropriate sense for programming in a low-level language.

Maybe one problem was that the students did not have the opportunity to make real world applications. Real application and working on real hardware should make the curse material less abstract and more practical which is one of the most valuable objectives of this course.

Another problem this course had was the disjointed material between lectures, theoretical and practical exercises. Following this methodology, students had much more material that had to be learned, but the differences made it impossible to grasp the most important parts of the lectures. This course need a more practical approach where the lectures and the theoretical exercises are both support for the most important practical exercises in this course.

Therefore, the courses organization and curriculum had to be changed. The decision was to migrate a hardware-based course such that it will become as close as the software-based one. That is, to create a course that will awake the computer science students curiosity about the hardware-based courses by challenging their software and high-level programming skills.

IV. Improvements through years

This section describes the changes made in the MM course each year, starting from 2010. These courses radically improved the course reputation, students grades, throughput, etc.

A. Awake the computer science students curiosity about the hardware - New course curriculum

The starting point for the changes was the year 2010, when the first task was to awake the students curiosity to learn the hardware, especially for the theoretical and practical exercises for both parts of the course.

The old DOS based Microsoft Macro Assembler (MASM) simulator was abandoned and replaced with a new modern visual simulator. This simulator enables students to watch registers, parameter values, memory locations, the program stack etc. Also it enables step-by-step program execution.

The other simulator for 8051 was also abandoned, since the students had a huge problem to learn similar, but not the same assembler. A totally new microcontroller PIC16F887 together with its development board was introduced. We decided to use Peripheral Interface Controller (PIC) platform mostly because of the large number of implementation and the availability of embedded software for a direct connection with PC and the availability of Flash memory [23].

B. Enforce the students to work on real hardware - New grading scheme

A new grading scheme was applied in order to properly reflect the changes in the curriculum previously explained [24]. The points for the practical part of the exam were increased from 50% to 80% (50% practical exam, 20% laboratory exercises and 10% for the mandatory project). Theoretical part was reduced from 50% to 20%. Additionally, a non obligatory project was introduced, which was mandatory a hardware project, i.e., working on real PIC. They can achieve maximum 20% for this project. Students that will decide to work on an additional project would have the opportunity to put their hands on PiC16F778. They faced an additional problem here, that is, lack of RAM or ROM memory, or both. Usually this is one interesting and a bigger student project. By doing this project students have a potential to achieve higher grades. Introducing the new grading scheme has also increased the students' curiosity about learning the hardware, as well as the desire to work on a real hardware and face with real problems.

C. Adaptive learning

Usually, the computer science students' structure is very discrepant, starting from the students that study computer science because they want to get a job easier to the students that are excellent in mathematics and computer science and are national or international contestants and even winners. This fact enforces us to introduce an adaptive learning, that is, to introduce the incrementally weighted exercises, as well as opportunities for bonus scores. It enables the students to group based on their gained knowledge and interest for a particular level of learning. Basic students need to work on the mandatory exercises that cover all learning objectives, while the advanced students are allowed and encouraged to work on harder tasks, since they had finished the mandatory exercise earlier. With this adaptive learning, the advanced students are more challenged and focused through the entire course.

D. Utilize the students' software skills - New hands on exercises

We introduced new practical assignments for the exercises that involve the visual simulators [25]. Doing this we abandoned the old hardware oriented practical exercises and introduced more software exercises. The new exercises include arrays and matrix, strings, procedures, macros etc., which are well known concepts of the programming courses.

The practical exercises for the material of the second midterm (microcontrollers) were also changed. They are based on mikroC programming language, the high-level language that the students are used to program in. They could use many additional hardware components with only a few lines of code. Components like alphanumeric LCD displays, ports for serial communication, EEPROM memory, Keypads and others were studied on the practical exercises. Using mikroC as a programming language was quite positively accepted with the students because of their previous experience in C.

E. Busy students - OER Approach

We introduced new practical assignments for the exercises that involve the visual simulators [EDUCON2013]. Doing this we abandoned the old hardware oriented practical exercises and introduced more software exercises. The new exercises include arrays and matrix, strings, procedures, macros etc., which are well known concepts of the programming courses.

The practical exercises for the material of the second midterm (microcontrollers) were also changed. They are based on mikroC programming language, the high-level language that the students are used to program in. They could use many additional hardware components with only a few lines of code. Components like alphanumeric LCD displays, ports for serial communication, EEPROM memory, Keypads and others were studied on the practical exercises. Using mikroC as a programming language was quite positively accepted with the students because of their previous experience in C.

F. Enforce the iterative learning Prepared students in the laboratory

As many education researchers reported, the laboratory still remains the place where the students can learn the hardware based courses and working on the visual simulators is not enough. Therefore, our new teaching methodology enforces the students to come prepared on the laboratory exercises. Tutorials with examples and solutions were announced a week before the corresponding laboratory exercise. The students then work on problems during the laboratory exercise and have to complete their activities during the laboratory classes. A minimum of 50% of all points must be achieved by a student in order to allow him / her to go on the final exam. Otherwise, he / she must reenroll the course in the next study year.

V. The results

This section presents the results of the students that were enrolled on the MM course through several years, starting from 2009 with the old course curriculum, and 2010 with the improved one.

A. The course's tasks completition

The percentage of completed students in each year has is depicted in Figure 1. This percentage has decreased for around 10% in 2011, and kept its value around 80%. This is the exact period of the iterative learning enforcement implementation, as described in Section III.F.

B. Improved the course's throughput

Figure 2 depicts the course throughput through years. It shows some paradox, that is, the percentage of the failed students increases starting from 2010, when the radical changes are implemented. Figure 3 depicts better the decreasing of the percentage of the passed students through years.

However, Figure 4 disproves this paradox, since it depicts that the percentage of the passed students that completed all tasks during the semester is increasing due to the implemented improvements. The 45% of the passed students in 2013 is

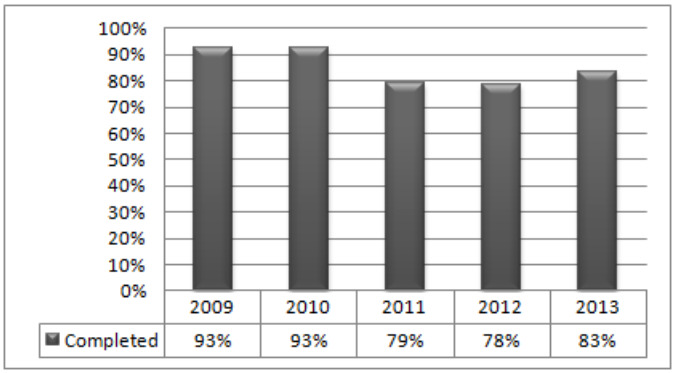


Fig. 1. Percentage of the students that had completed all tasks during the semester

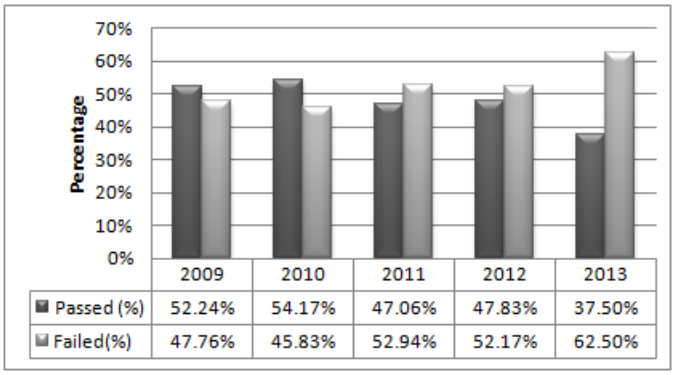


Fig. 2. Percentage of the passed students of all enrolled students

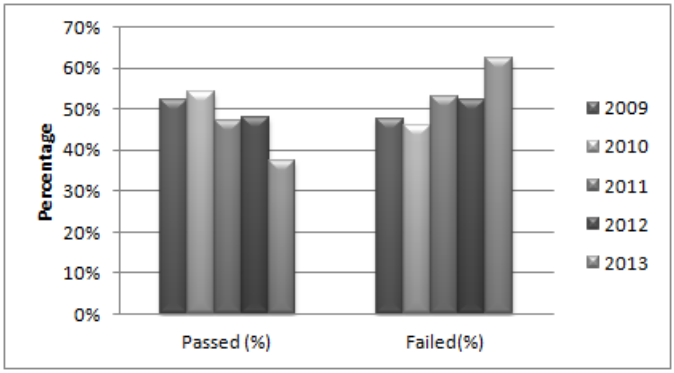


Fig. 3. Comparison of the percentage of the passed students of all enrolled students

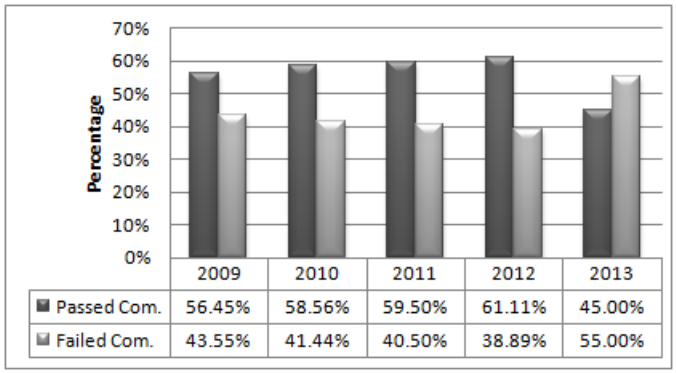


Fig. 4. Percentage of the passed students that had completed all tasks during the semester

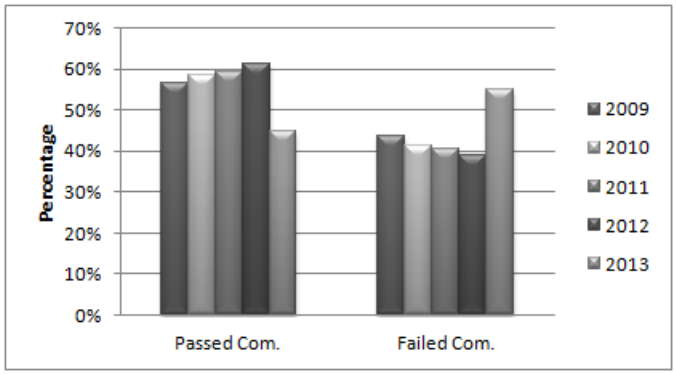


Fig. 5. Comparison of the percentage of the passed students that had completed all tasks during the semester

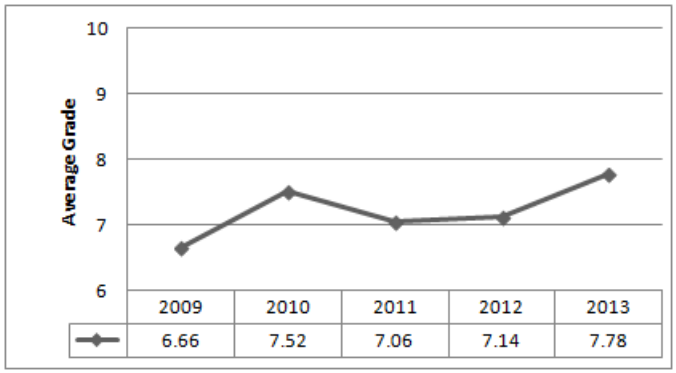


Fig. 6. The students average grades through years

low since only one session have passed and two additional session exist, where the authors believe that this percentage will be the greatest compared to prior years. Figure 5 depicts the improvement of the percentage of the passed students that completed the tasks through years even better.

C. Improved students' average grade

Another improvement is reported in the students average grade through the years after the implementations of the improvements, which is depicted in Figure 6. The average grade has increased for more than 1, that is, from 6.66 to 7.78 in 2013. This graph explains the student accomplishment through the years in general.

In 2010 a small jump in the average grade can be noticed. The average grade in 2010 is 7.52 which is higher than in 2009 and in 2011. The explanation is that the year 2010 was the first study year with the new revolutionary curriculum, and also 10 of 120 students that were enrolled in 2010 were honorable students, who later were chosen as demonstrators at our faculty.

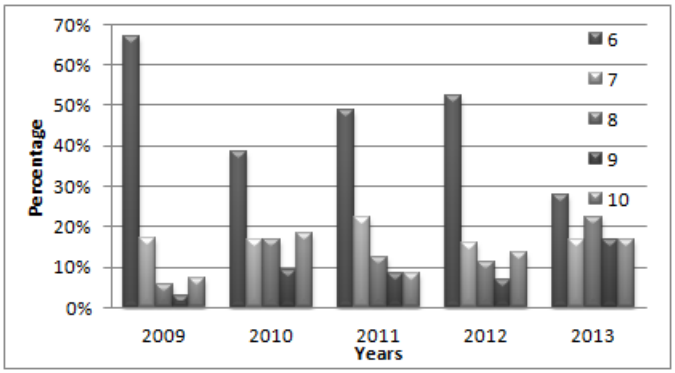


Fig. 7. The grade distribution per year

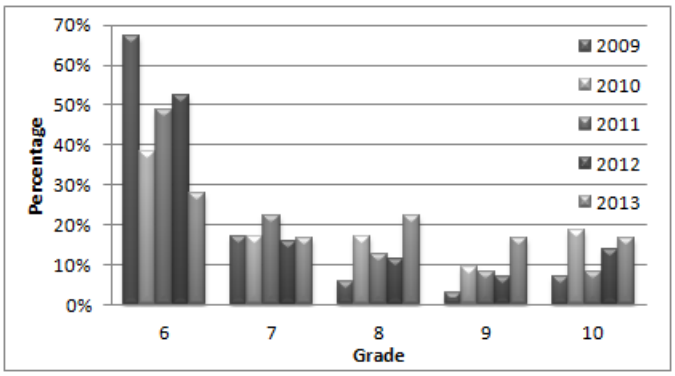


Fig. 8. The grade comparison through years

D. Improved grade distribution

The grade distribution through years is depicted in Figure 7. Let us examine firstly the year 2009, which is the year prior the improvements. More than 2/3 of the students achieved the minimum grade 6, or 84% of the passed students had the grade 6 or 7. The grade distribution is significantly improved starting from 2010, which is emphasized in 2013. However, probably similar grade distribution will be achieved after finishing this year course, since our analysis reports that the students achieve lower grades in the last two sessions.

E. Improved grades through years

Figure 8 compares the achieved grades through years. This figure shows that the grades 8, 9 and 10 are much greater after the implementation of the improvements, the percentage of the students that achieved grade 7 are constant, while the 6s are reduced significantly.

VI. Further course changes for additional improvements

Apart of the presented course changes and achieved improvements, there is still place for further changes. This section presents the changes of the course made this year.

A. Student Collaboration - Developing simple games

Starting from this year, collaboration is implemented in the nonobligatory projects. That is, the projects represent some interesting two-men game, such as battleship, X/0, checkers, memory etc., where the first student develop the game configuration and the artificial intelligence, while the other student develops the part for gaming. Programming games could be interesting and productive activity for students, where they can implement their knowledge gained in the previous artificial intelligence courses.

B. Separate the exam in two parts

Another change is the enforcement to have minimum 40% of the points for each part of the course, i.e., the microprocessors (assembly language) and microcontrollers (mikroC). Since this would made the course heavier, we allow the students to take the course parts separately.

The students can choose whether they can or like to take one or both parts of the exam in an exam session. This approach will enable the students to concentrate more on just one part of the exam in each session, thus enabling them to pass the exam with a higher grade.

VII. Conclusion and Future Work

This paper presents the changes that were made of the MM course several study years: the new curriculum, new grading scheme, adaptive learning, iterative hands-on, OER approach for busy students, and already prepared students for laboratory exercises. These changes had improved the course bad reputation, as well as the students' grades and throughput. But the most important thing is that the students finally started to love the course and to learn the hardware.

We believe that the last two curriculum changes, i.e., the student must passed each part of the course with more than 40% of the practical exam (the microprocessors and microcontroller) and the collaborative learning will improve even better both the students and the course. We will survey the students about the changes and continue to improve the course to improve the results od the students even more, without lowering the criteria for passing the course.

The other teachers could implement our proposed changes into their related based courses for computer science students in order to make the hardware based courses more software-based.

References

[1] X. Liang, "A survey of hands-on assignments and projects in undergraduate computer architecture courses," in *Advances in Computer and Information Sciences and Engineering*, T. Sobh, Ed. Springer Netherlands, 2008, pp. 566–570.

[2] D. Kehagias and M. Grivas, "Software-oriented approaches for teaching computer architecture to computer science students," in *Journal of Communication and Computer*, 2011.

[3] M. N. J. E. I. Kastelan, D. Majstorovic and M. Katona, "Laboratory exercises for embedded engineering learning platform," in *Proceedings of the 35th International Convention MIPRO*, 2012, pp. 1113–1117.

[4] S. S. Z. Y. Q. Jiangb, W. Rangding and X. Zhijun, "Simplifying and integrating experiments of hardware curriculums," in *3rd IEEE International Conference on Computer Science and Information Technology (ICCSIT)*, 2010, pp. 610–614.

[5] I. Kastelan, D. Majstorovic, M. Nikolic, J. Eremic, and M. Katona, "Laboratory exercises for embedded engineering learning platform," in *MIPRO, 2012 Proc. of the 35th Int. Conv.*, 2012, pp. 1113–1117.

[6] J. Qian, R. Wang, S. Shi, Y. Zhu, and Z. Xie, "Simplifying and integrating experiments of hardware curriculums," in *Computer Science and Information Technology (ICCSIT), 2010 3rd IEEE International Conference on*, vol. 9, 2010, pp. 610–614.

[7] D. Kehagias and M. Grivas, "Software-oriented approaches for teaching computer architecture to computer science students," *Journal of Communication and Computer*, vol. 6, no. 12, pp. 1–9, Dec. 2009.

[8] X. Wang, "Multi-core system education through a hands-on project on fpgas," in *Frontiers in Education Conference (FIE), 2011*, 2011, pp. F2G–1–F2G–6.

[9] J. H. Lee, S. E. Lee, H.-C. Yu, and T. Suh, "Pipelined cpu design with fpga in teaching computer architecture," *Education, IEEE Transactions on*, vol. 55, no. 3, pp. 341–348, 2012.

[10] B. Nikolic, Z. Radivojevic, J. Djordjevic, and V. Milutinovic, "A survey and evaluation of simulators suitable for teaching courses in computer architecture and organization," *Education, IEEE Transactions on*, vol. 52, no. 4, pp. 449 –458, nov. 2009.

[11] S. Ristov, M. Stolikj, and N. Ackovska, "Awakening curiosity - hardware education for computer science students," in *MIPRO, 2011 Proceedings of the 34th International Convention, IEEE Conference Publications*, Opatija, Croatia, may 2011, pp. 1275 –1280.

[12] J. D. B. Nikolic, Z. Radivojevic and V. Milutinovic, "A survey and evaluation of simulators suitable for teaching courses in computer architecture and organization," in *IEEE Transactions on Education*, 1009, pp. 449–458.

[13] C. Wick, "Teaching embedded computer systems with a windows-based simulator," in *Proceedings of the 26th Annual Frontiers in Education*, 1996, pp. 242–245.

[14] S. Ristov, M. Gusev, B. Atanasovski, and N. Anchev, "Using educache simulator for the computer architecture and organization course," *International Journal of Engineering Pedagogy (iJEP)*, vol. 3, no. 3, pp. 47–56, Jun. 2013. [Online]. Available: http://online-journals.org/i-jep/article/view/2784

[15] D. Pop, D. G. Zutin, M. E. Auer, K. Henke, and H.-D. Wuttke, "An online lab to support a master program in remote engineering," in *Proceedings of the 2011 Frontiers in Education Conference*, ser. FIE '11. USA: IEEE Computer Society, 2011, pp. GOLC2–1–1–GOLC2–6.

[16] M. Gusev, S. Ristov, G. Velkoski, and B. Ivanovska, "E-learning and benchmarking platform for parallel and distributed computing," *International Journal of Emerging Technologies in Learning (iJET)*, 2014, in press.

[17] C. Reinbrecht, J. Da Silva, and E. Fabris, "Applying in education an FPGA-based methodology to prototype ASIC soft cores and test ICs," in *Programmable Logic (SPL), 2012 VIII Southern Conference on*, 2012, pp. 1–5.

[18] B. Hatfield and L. Jin, "Improving learning effectiveness with hands-on design labs and course projects for the operating model of a pipelined processor," in *Frontiers in Education Conference (FIE), 2010 IEEE*, 2010, pp. F1E–1–F1E–6.

[19] L. Da, "Computer hardware curriculums, curriculum contents and teaching methods," in *Computer Science Education, 2009. ICCSE '09. 4th International Conference on*, 2009, pp. 1506–1511.

[20] L. He, "Computer architecture education in multicore era: Is the time to change," in *Computer Science and Information Technology (ICCSIT), 2010 3rd IEEE International Conference on*, vol. 9, 2010, pp. 724–728.

[21] "Microsoft macro assembler," 2011. [Online]. Available: http://www.masm32.com/history.htm

[22] R. Hyde, *The Art of Assembly Language*, 2003.

[23] "Microchippic16f887," 2013. [Online]. Available: http://www.microchip.com/wwwproducts/Devices.aspx?dDocName=en026561

[24] M. Stolikj, S. Ristov, and N. Ackovska, "Challenging students software skills to learn hardware based courses," in *Information Technology Interfaces (ITI), Proceedings of the ITI 2011 33rd International Conference on*. IEEE, june 2011, pp. 339 –344.

[25] N. Ackovska and S. Ristov, "Hands-on improvements for efficient teaching computer science students about hardware," in *IEEE Global Engineering Education Conference (EDUCON), 2013 IEEE*. Berlin, Germany: IEEE, March 2013, pp. 295–302.

State of e-learning projects in Croatian companies

Tea Dragičević, Maja Ćukušić and Mario Jadrić
Faculty of Economics, University of Split, Split, Croatia
{tea.dragicevic, maja.cukusic, mario.jadric}@efst.hr

Abstract – The purpose of this research was to analyse the current state of e-learning projects in business settings in Croatia through the prism of organisations' technological and methodological approaches, which (along with the characteristics of the organisation) constitute the basic framework for the implementation of e-learning projects. The study was conducted in early 2013. The survey participants were employees of large Croatian companies with particular interests and working focus in e-learning area. The results indicate that Croatian organisations keep up with global trends, but there is certainly lot to improve. This kind of analysis has not been conducted so far. Other than bringing forth the analysis of the state of corporate e-learning projects in Croatia, the contribution of this pilot-study represents the basis for future research in the area of e-learning in business settings.

I. INTRODUCTION

The traditional concept of classroom training is gradually losing its dominant role in business settings, while e-learning is considered to be one of the newer and more and more implemented concepts that it is certainly breaking on the Croatian market. The propagation of e-learning use in business settings is based on a massive expansion and price drop of internet services, as well as the enormous time and cost savings and flexibility. All above mentioned variables encourage immediate knowledge sharing among all users, collaboration and computer-based assessments as some of the application areas within both SMEs and large organisations.

With regards to what e-learning in companies actually stands for providing the frame for the subsequent interpretations, we use the definition brought forth by HELIOS project [1]. Labelling different "territories of e-learning" e-learning in the workplace is defined as the use of ICT in the corporate (business) environment, public administration and agencies. The difference in the extent and approach to the implementation of e-learning between corporate and government environments largely depends on the organizational structure and practices and policies related to human resources management. Given the nature of the underlying systems for learning management and management of day-to-day activities related to human resources within business organisations the overarching e-learning programs are often implemented in smaller chunks as projects with definite scope, time frame and resources. Terms e-learning programs and e-learning projects are interchangeably used in the paper.

Today's workforce must process and master more and more information within limited time frame. This fact largely contributes to importance of e-learning area in business settings [2]. Technology is the mean that allows the management and efficient utilization of information and learning process [3], with its main purpose to ensure that the employees of whether small or large organisations can learn faster and "consume" new information more often [4] and at a convenient time. Ultimately, the main goal of e-learning in business settings is to serve corporate goals and needs in order to increase competitiveness, profit, efficiency and effectiveness [5].

In short, the mission of e-learning in business settings is to deliver up-to-date and cost-effective educational program to all employees [2]. One of the characteristics of future e-learning solutions in business settings is the sole focus on employees - in particular the spotlight will be set on a user-centric model implying that knowledge workers must be provided with all the appropriate services and techniques [6].

Although the technology enables and drives the implementation of e-learning in business settings, its use is largely shaped by social factors, including the objectives and priorities of those who specifically decide on the implementation of e-learning and manage the process [7]. It should be noted that cautious consideration of the methodological aspects, the way in which e-learning is organised, is one of the most important prerequisites for the successful implementation of e-learning [8]. Overlooking the methodological issues or the failure to meet the so-called "soft" requirements is the second most frequently listed reason for unsuccessful implementations of e-learning solutions in business settings; this is especially with regards to two causes [9]: the lack of focus on people and the inappropriate use of e-learning concepts and related tools.

Often, the main reasons for failure of e-learning projects are generally listed as [10]:

- problems with the technology,
- lack of management support,
- poorly designed courses and
- disregard for individual learning preferences.

It is clear that within the abovementioned causes, major problems are related to pedagogy/andragogy, i.e. poor design of learning courses and disregard for learning

preferences of learners. With regards to the lack of focus on people, one of the problems and reasons may be that businesses/corporations do not perceive their employees as learners [11].

This paper presents a short pilot-analysis of current scenario of use of e-learning concepts in business settings in Croatia. The results are presented through two different aspects – namely, the questions relate to both technological and methodological issues as well as to some organisational characteristics. The theoretical implications of e-learning implementations in business settings are presented in the second chapter. The third chapter brings forth results of the study and the fourth chapter concludes the paper.

II. E-LEARNING IN BUSINESS SETTINGS

A. *Different aspects of e-learning in business settings*

Generally, employees have a positive attitude towards workplace learning. According to a recently conducted study in the USA, 85% of employees prefer learning during working hours and 62% prefer learning in the office [12]. According to the same author, more than half (52%) of employees are advanced computer users and have a very positive attitude towards e-learning. The statistics show that 80% of learners (employees) can effectively "consume" learning in all forms and therefore the majority can safely, without any issues learn using the Internet.

Corporations have recognized this point of view as well as the positive perception towards e-learning by their employees, so it is no surprise that in the USA 95% of corporations use at least some kind of form of e-learning [13]. Therefor the majority does so in order to deliver a training or professional development courses to their employees [2]. The paradox today is that people need to learn a lot more but have less time for that purpose, so the question is: how do we know what we need to know it in the exact moment when we need it [14]?

One of the important motifs for use of e-learning in business settings is that all organisations do not have the capacity to organise face-to-face learning sessions [12]. Sometimes, face-to-face learning courses cannot be carried out because large organisations are dispersed in the form of multiple geographical locations and even multiple time zones while employing a large number of people with different levels of knowledge [2]. Evidently, there are major opportunities for learning during working hours, but the e-learning solution conveys even more convenience for the stay-at-home employees who live in remote areas or cities [12].

Learning in business settings is interdisciplinary and especially task-oriented [15]. What principally distinguishes majority of business organisations from learning organisations is that their primary goal is not learning itself [5]. Workplace learning comes down to use of e-learning applications and affects/results in changes to working practices [16]. One of the main characteristics of e-learning in business settings is cooperation, since knowledge sharing practices with working/learning peers is a very important part of the learning process itself (ibid).

Social networks are one of the new features and trends in organisational e-learning endeavours [17]. In fact, it was only a matter of time before social networking tools "migrate" to a professional environment with the purpose to increase the overall company knowledge-level [18]. Some predict that the development of e-learning tools and projects in business settings could progress straight to full integration with social networks [19]. The starting point for such predictions is the fact that people are progressively using more and more social networks and have gained confidence by sharing a variety of data, information and documentation over the Internet [18].

Most of the tools that should and could be used for e-learning in business settings already exist, but the only question is whether the companies are willing to use them [19]. Some of those tools which portray new forms of Learning Management Systems (LMSs) – the social LMS are (ibid): (1) calendar, (2) chat, (3) events, (4) articles, (5) survey, (6) blogs, (7) shoutbox, (8) pictures, (9) documents, (10) forums and etc.

Another trend in e-learning for businesses is the adoption of mobile learning technologies, or m-learning [17]. The increasing usage of mobile information exchange is a relevant signal for managers, to adapt and to develop an architecture for delivering learning material via smart phones and tablets [20] and more recently phablets.

Considering other technological trends in e-learning solutions, it is worth noting that LMSs should develop in the direction of becoming Personal Learning Environments (PLEs) [6]. To sum up, the approach to (workplace) learning should be socially-oriented, decentralized and above all individualised.

B. *The current state in Europe and Croatia*

After this brief literature overview, the question presents itself – what is the current situation in Croatia? Analytical insights of corporate e-learning in European enterprises, emphasizing Italian indicators, are presented in a recent study [21]. Based on the data, from Eurostat, an overview of development of e-learning in business settings for the period 2003 to 2009 is prepared and presented in Table 1.

According to the maximum percentage of e-learning use in business settings in 2009, Croatia takes the 19th place (31%) whilst the average of EU-27 takes the 26th place (24%). When the mean percentages of e-learning use in business settings through all seven years are observed, it can be stated that Croatia (23.86%) is keeping up well with the EU-27 countries (22.17%).

The data in the table is sorted by the progress (the delta) in the latest year (2009) from the start year (2003). For Croatia, the shift puts us in the 16th place overall with the increase of e-learning use in business settings of 12 percentage points. Therefore, a very high progress in comparison with the EU-27 average and other countries can be noted.

TABLE I. PERCENTAGE OF E-LEARNING USE IN BUSINESS SETTINGS 2003 - 2009 (ADAPTED FROM [21])

No	Country	Min	Max	Mean	Delta
1	Cyprus	25	59	46,14	34
2	Malta	22	51	29,86	29
3	Norway	14	43	29,14	29
4	Denmark	8	30	19,14	22
5	Bulgary	20	41	25,86	21
6	Estonia	21	41	32,29	20
7	Poland	22	42	30,71	20
8	Slovakia	31	50	42	19
9	Slovenia	30	49	39,86	19
10	Ireland	23	41	33,43	18
11	Greece	36	53	44,86	17
12	Portugal	19	36	25,86	17
13	France	10	24	15,57	14
14	ltaly	4	18	12,86	14
15	Luxemburg	12	25	17,29	13
16	**Croatia**	**19**	**31**	**23,86**	**12**
17	Finland	31	41	36,57	10
18	Czech R.	30	40	33,14	10
19	Turkey	23	33	29	10
20	Belgium	15	25	21,57	10
21	Hungary	12	22	17,29	10
22	Lithuania	50	59	56,14	9
23	Latvia	34	43	38,86	9
24	Austria	21	30	25,43	9
25	Romania	56	64	59,29	8
26	Spain	27	35	31,29	8
27	Sweden	20	28	25	8
28	UK	20	28	24,14	8
29	Germany	14	22	18,43	8
30	Netherlands	10	17	13,71	7
31	**UE 27**	**20**	**24**	**22,17**	**4**
32	Iceland	17	20	18,29	3

Since there is no detailed state analysis of e-learning in business settings in Croatia, an overview of pilot study is conducted and its results are presented in the following chapter. Such study was deemed important since there are no available studies examining the level of use of e-learning concepts in business settings in Croatia. In particular, it is interesting to inspect whether or not the corporate managers in Croatia recognised the importance and the potential of e-learning. Certainly, their level of recognition would manifest in terms of choosing the appropriate technologies, and in the number of methodological aspects taken in consideration when implementing the e-learning projects.

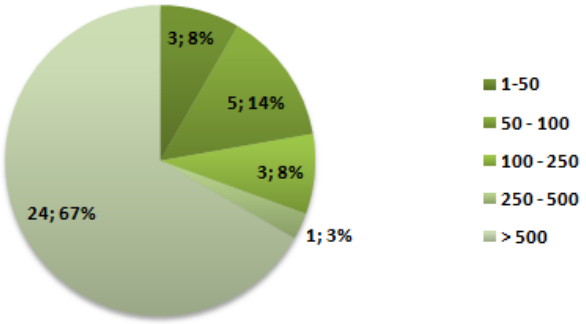


Figure 1. Number of employees in respondent's organisation (N=36)

III. STATE OF E-LEARNING IN BUSINESS SETTINGS IN CROATIA

A. *Research method and participants*

The study was conducted on a sample of business people in February and March 2013. The sample consisted of participants who are employed in large Croatian companies and have particular interests in the field of e-learning, i.e. they are working on or are planning to implement e-learning in their organisations. All the participants took part in the first conference related to e-learning in corporate settings titled "Knowledge in focus" which was held in company Podravka in September 2012.

A questionnaire was sent to 138 email addresses, and the total number of completed questionnaires was 36. It is a relatively small number mostly because activities related to e-learning are still in the early stages of development in Croatia (also illustrated by the fact that this was the first conference dedicated to this topic). A tool used for collecting data was Survey Gizmo. The questionnaire contained 15 questions of which 13 where questions with multiple choice answers and two open-ended questions. The complete questionnaire in Croatian is available at http://tinyurl.com/q7zoe5u. We chose to present the most interesting questions/replies. The questions are presented as the titles of the figures.

From 36 of respondents, 66.7% work in organisations with more than 500 employees (Figure 1). The results must therefore be considered and interpreted as an overview of e-learning state in large business organisations.

Figure 2 represents number of organisations with regards to a year in which they started with implementation of their e-learning projects. Most of the corporations started during the last 5 years (66%). The number of projects that have started in 2003 and earlier (17%) generally corresponds to the number from Bucciarelli's study [21]. The respondents had different roles (Figure 3) – those involved in creating courses (44%), execution of learning related processes (69%), technical support (25%) and process management (19%). Therefore their responses related to technological and methodological aspects of e-learning should be considered as relevant. But, considering the structure of the sample of respondents and the companies, it should be noted that it was not possible to collect data on the specific business requirements that should be aligned with e-learning

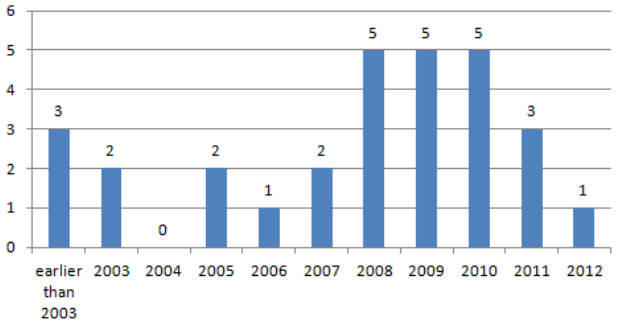


Figure 2. Number of organisations considering the year their e-learning projects started in

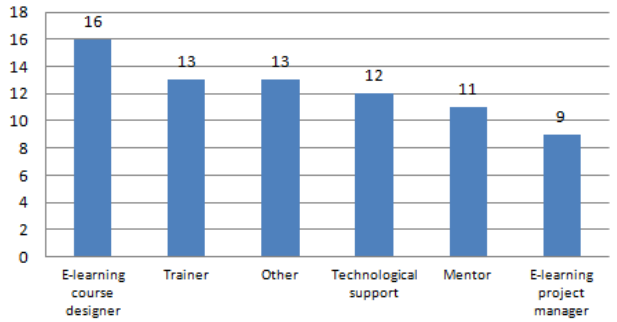


Figure 3. Respondent's role in e-learning processes in organisations

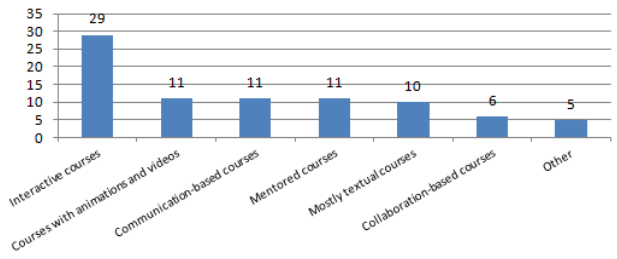


Figure 4. Types of e-learning courses in organisations

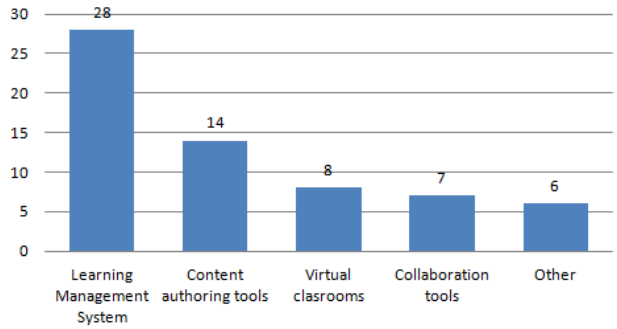


Figure 5. Types of technologies used for e-learning in oganisations

programmes. This is somewhat limiting the contribution of the research results.

Regarding the use of online courses, which organisations implement for various purposes, the most common usage is for improvement/advancement of professional knowledge and skills (72%) and for technical training (44%). Courses that provide training for the new products and services (39%) and jobs (14%) are also commonly implemented.

In terms of learning content (Figure 4), the most common response (81%) is that the courses are interactive, meaning that they contain both learning materials and assessment. In addition, 30% of the courses contain animations and videos while almost the same percent of courses (28%) contain only text. Also, over 30% of the courses are mentored. For over 30% of the courses, participants state that they support communication between the participants (e.g. forum, chat), while only 17% of courses support collaboration between participants (e.g. submission of joint work). The results confirm that the corporations have recognized the importance of methodological principles in structuring courses. Responses to questions such as: who are the participants of the courses, what is to be achieved by implementing the course, what are the organisational capabilities and etc., are all input for the first phase of designing an e-learning course. For example, if potential attendees/learners of an e-learning course are mainly dislocated, fully-online mentored e-learning course could be a better choice than a hybrid approach to e-learning (of course, this does not need to be set as a rule).

In terms of time a relatively small number of organisations provide learning during non-working hours (36%), i.e. employee cannot learn wherever and whenever he/she wants. This can and should be explained from the perspective of organisations, since 72% of the cases organisations enable e-learning activities to only take place during working hours. In some cases business processes, and consequently the learning material, can be considered as sensitive information and in some cases even a business secret. It is understandable that the access to such information may be restricted. In most cases courses are flexible regarding time, but 33% of organisations have scheduled and limited the time frame for e-learning related activities. Time management is considered to be one of the most important characteristics of successful learners whether they are participating in an online course or in traditional learning activities. The most successful learners engage at early stage, they participate frequently and make sure to have enough time in their schedule to ensure success in all aspects of online learning. However, in scenarios when employees have no limits regarding the time and place for accessing the e-learning content they are generally overwhelmed with other work-related obligations so they have to possess a very high level of self-discipline and motivation to conscientiously carry out the tasks in e-learning courses.

Motivation is being emphasized in both traditional and online education, since it is considered to have huge positive impact on learners/employees in e-learning settings. Intrinsic motivation is associated with the employees' desire to learn and acquire new knowledge, while extrinsic motivation is associated with the expectation and reaction to an award or a penalty such as the impact of assessment results, gaining certificates, etc. [22]. When planning and considering employees' motivation for accessing e-learning courses it is justified and recommended to cautiously combine extrinsic and intrinsic motivation mechanisms. However, the extrinsic

factors do predominate in the Croatian organisations, i.e. 42% issue certificates and in 31% of cases employees are required to gain a specific level of credits. But, the results have shown that employees are intrinsically motivated as well with visually attractive and interactive content (33%), by emphasizing the best participants (22%) and various awards (6%). The optimal combination for motivating participants contains moderate use of extrinsic and high intrinsic motivation [23].

The learning activities and e-learning courses are promoted more or less equally through direct communication (64%), intranet (61%) and e-mail communication (newsletters) (61%), while only 17% of organisations organise special events.

In terms of technology, it is important to differentiate between different e-learning technologies before presenting the results. Technological providers can be commercially differed as [24]:

- Technology providers, which are prerequisites to setting up solutions for e-learning in organisations (such as collaboration tools, content authoring tools, knowledge management systems, etc.),
- Service providers (providing service assessment and testing, system integrators, media hosts, etc.) and
- Content providers (offering soft-skill courses, IT courses, generic courses, simulation content, content aggregators, etc.).

The highest percentage of organisations (78%) from the sample use "classic" e-learning management system – LMS (Figure 5). This is followed by 39% of them which use tools for content authoring and 22% who use virtual classrooms. Less than 20% of corporations use collaboration tools, while only 17% do not use any of aforementioned tools.

Considering the implementation of e-learning content, results indicate that large number of organisations (81%) generate their own content while in addition 42% buy generic content (specifically IT and soft skills content). With regards to service providers, 31% of organisations use expert services for learning content creation, while 20% of them buy custom content. Only 3% of the organisations buy very specialized off-the-shelf simulation content. In short, organisations in Croatia use both off-the-shelf and custom solutions and services for e-learning project fulfilment.

Finally, it is important to emphasise the fact that more than half of the organisations (54%) have elaborated a strategy for e-learning. Unfortunately, only 23% of organisations enable mobile access to e-learning courses.

Considering that results have indicated that the majority of organisations initiated an e-learning project in the last five years, it is obvious that organisations are currently at different stages of e-learning implementation. Although there is some empirical evidence that number of organisations are in the higher stages of e-learning development cycle – they motivate employees, use custom content and strategically plan e-learning processes – most organisations are still somewhere in the middle of the development cycle, i.e. they recognize the advantages and best practices, but still have not fully established e-learning procedures.

IV. CONLUSION

Large organisations from all over the world, including the Croatian ones, became aware that by implementing e-learning solutions, they can reduce process costs and increase efficiency. Though both technological solutions and methodological mechanisms are widely available, they should be implemented optimally and in relation to specific business requirements of an organisation. Often, the problem arises when there is a need to integrate several solutions for organisation – organisational learning, knowledge management and business processes management solutions – with e-learning solutions, processes and projects. However, when the successful integration of the above mentioned areas is achieved the results are manifested as easier corporate development, as well as the encouragement of employees for learning and sharing knowledge, promoting innovation and as application of acquired knowledge in adequate way [25].

The paper illustrates that the larger Croatian organisations keep up to date with global trends, but there is certainly a lot of space for improvement. The contribution of this pilot study is primarily expressed in terms of analysis of the current state of corporate e-learning projects in Croatia. As already stated, considering the variations in the structure of the sample of respondents and the companies, the contribution of the research is somewhat limited on giving insight into the situation of a specific number of organisations. Nevertheless, considering this kind of analysis has not been conducted so far, this pilot study represents a good basis for future research in the field of e-learning in business settings. Apart from collecting data on the specific business requirements, that should be aligned with e-learning programmes, we plan to inspect which types of companies are more prone to in-house training, the correlations of e-learning projects' maturity with having research departments or higher percentages of highly skilled or better paid employees etc.

Even though some organisations are still not ready to implement a project of this type, we hope that the results of this study will motivate them to engage in e-learning projects. Starting with e-learning project in organisation is never easy, but it should definitely begin with good formed e-learning strategy. Strategy should be goal-driven, representing what the organisation wants to achieve. Three things should be kept in mind: technology, pedagogy and business priorities. Nowadays there are huge number of guidance and portals offering best solution for implementing e-learning (one of them being http://eucenje.efst.hr), but it should be clear that each project must be implement on individual basis, based on internal variables and limitations in the organization.

On the other hand, if organisations are already applying this concept, the results may encourage managers of human resources to try to form a better e-learning strategy having in mind some of the

recommendations we bring forth with regards to holistic approach, individualisation, motivation and other trends such as integration with social tools so that their needs and goals are fully supported by this concept and, therefore, achieve additional value.

REFERENCES

[1] C. Delrio and T. Fischer, HELIOS: Redefining e-Learning Territories. eLearning Papers, No4. 2007.

[2] T. A Urdan and C.C. Weggen, Corporate e-learning: exploring a new frontier. WR Hambrecht + Co. 2000.

[3] R. Newton and N. Doonga, „Corporate e-learning: Justification for implementation and evaluation of benefits. A study examining the views of training managers and training providers“ in Education for Information, 2007, pp.113-130.

[4] R. H. Wild, K. A Griggs and T. Doning, „A framework for e-learning a tool for knowledge management“ in Industrial management & Data system 102, vol7, 2002, pp.371-380.

[5] A. Welle-Strand and T. Thune, „E-learning policies, practices and challenges in two Norwegian organizations“ inEvaluation and Program Planning 26, 2003, pp.185–192.

[6] M. A Chatti, M. Jarke and D. Frosch-Wilke, „The future of e-learning: a shift to knowledge networking and social software“ In Knowledge and Learning, vol 3, 2007, pp.405-420.

[7] A. Macpherson, M. Elliot, I. Harris and G. Homan, „E-learning: reflections and evaluation of corporate programmes“ in Human Resource Development International 7. vol 3,2004, pp.295-313.

[8] T. Govindasamy, „ Successful implementation of e-Learning Pedagogical considerations“ in Internet and Higher Education, vol 4, 2002, pp.287–299.

[9] S. Carliner and P. Shank. The e-learning handbook. San Francisco: Pfeiffer. 2008.

[10] K. Frankola, Why Online Learners Drop Out. Workforce. 2001.

[11] G. Morgan, „Thirteen “must ask” questions about e-learning products and services“ in Learning Organization, vol 8(5), 2001, pp.203 - 211.

[12] S. Sitnikov, B. Kruk, O.Zhuravleva and N. Chupakhina, „Corporate E-learning Strategy“ in International Journal of Advanced Corporate Learning (iJAC), Vol 3, No 4, 2010, pp.41-44.

[13] R. E. DeRouin, B. A Fritzsche and E. Salas, „E-Learning in Organizations“ in Journal of Management, 2011, pp.920- 940.

[14] M. J. Rosenberg, Beyond e-learning- approaches and technologies to enhance organizational knowledge, learning, and performance. San Francisco: John Wiley & Sons. 2005.

[15] A. Sangster, P. MacLaran, and S. Marshall, „Translating theory into practice: facilitating work-based learning through IT“ in Innovations in Education and Teaching International, vol 37(1), 2000, pp.50–58.

[16] M. Wang, W. Ran, J. Liao and S. J. Yang, „ A Performance-Oriented Approach to E-Learning in the Workplace“ in Educational Technology & Society, 13(4), 2010, pp.167-179.

[17] B. Arth, The Business Impact of Next-Generation e-Learning How Today’s e-Learning Drives Business Results. Bersin & Associates, 2011

[18] K. Young, „From e-learning to 'we' learning“ in Elearning age, 2011, pp.14-15.

[19] J. Roth, Corporate Training Meets Facebook, retrieved November 12, 2012 from: http://interactyx.com/social-learning-blog/corporate-training-meets-facebook/. 2012.

[20] J. Jeevan, GARTNER’S TOP TECH 2012 PREDICTIONS: IMPACT ON LEARNING AT 3 LEVELS, retrieved November 12, 2012 from: http://www.learningcafe.com.au/2012/04/gartners-top-predictions-for-2012-and-its-impact-on-learning/. 2012.

[21] E. Bucciarelli, F. Muratore and I. Odoardi, „Consolidation processes of human capital in modern economic growth dynamics: an estimate based on the role of European corporate e-learning activities“ in Procedia - Social and Behavioral Sciences, vol 9, pp. 622–627. doi:10.1016/j.sbspro.2010.12.207. 2010.

[22] C. Ghaoui, E-Education applications: human factors and innovative approaches. Information Science Publishing, Hershey, USA, 2004.

[23] Y.-G Lin, W. J. McKeachie and Y. C. Kim, „College student intrinsic and/or extrinsic motivation and learning“ in Learning and Individual Differences, vol 13, Issue 3, 2001, pp.251 – 258.

[24] J. Massey, Study of the e-learning suppliers’ “market” in Europe, Brussels, Belgium: European Commission, Directorate General for Education and Culture, 2006.

[25] Y. Zhuab, Z. Niea, S. Zhanga and S. Chenc, „Analysis on Research Framework of Corporate E-learning“ in Procedia Environmental Sciences, vol 11, 2011, pp.525 – 529.

PaintArt 2.0 - Computer Program for Education of Visual Arts in Secondary Schools

Zvjezdana Prohaska, BFA*, Zdenko Prohaska, PhD** and Ivan Uroda, PhD***
* Senior Adviser for Visual Art, Education and Teacher Training Agency, Office Branch Rijeka, Rijeka, Croatia
** Full Professor, University of Rijeka, Faculty of Economics, Rijeka, Croatia
*** Senior Teaching and Research Assistant, University of Rijeka, Faculty of Economics, Rijeka, Croatia
zvjezdana.prohaska@azoo.hr, zdenko.prohaska@ri.t-com.hr, ivan.uroda@gmail.com

Abstract - This article presents a computer program consisting of a paint program and a database of specially prepared forms in bitmap format for education of visual arts in secondary schools.

Since more theoretical issues are prevailing on the secondary school level this computer software, which is actually a paint program, is equipped with different forms in bitmap format, with yes/no questions, multiple choice questions, classic questions demanding full text answers and questions where school boys and girls have to apply the solutions of the questions on specific paintings presented in the questions.

By using this program, secondary school boys and girls can compare their own answers with the provided solutions in PaintArt 2.0 and enhance their knowledge of arts.

I. Introduction

Based on the main characteristics of their computer program PaintArt for primary schools presented at the last MIPRO Conference held in Opatija in 2013, the authors developed their program further to be suitable for teaching visual arts on the secondary school level, where more theoretical topics are usually elaborated. Enhancements are provided on two levels. On the first or technical level the program has now full freehand drawing possibilities, the delete function is enhanced and the windows can be resized not only to minimum and maximum size but also at every variable size desired. [7]

Secondly, this computer program now includes four different forms or screens in bitmap format, where it is possible to ask questions in the form yes/no, multiple choice, ordinary questions demanding answers in full-text or where real application and drawing of the solution is required.

So, if the lesson is about Baroque, the dominant artistic style during the 17th century which is characterized by intense emotion and a sense of movement, the teacher could show a painting of a well known artist and ask to which movement in painting it belongs in the form of yes/no, multiple choice or ordinary questions. It is also possible to present examples in sculpture, architecture etc.

Showing examples in architecture, i.e. a church in Baroque, it could be asked to draw a line in the painting of the church which is characteristic for Baroque. [1]

By presenting this software in the class and by sharing it, pupils can take it home and by comparing own answers with the provided solutions in the software they can verify their knowledge of the topics discussed at school.

II. The Structure of the Program

Before presenting the main menu of the program PaintArt 2.0 it has to be pointed out that in the System menu box it is possible now not only to close, minimize and maximize the size of the program window, but complete moving in all four directions and resizing of the window is also accomplished. [4] Fig. 1

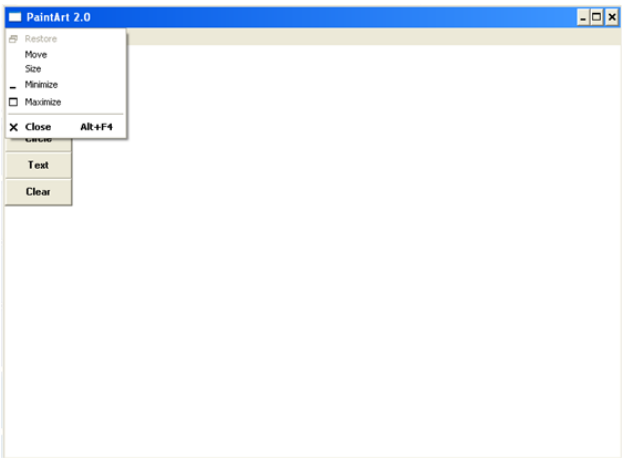


Figure 1. System menu box

As presented in the article of the last Conference the main menu of program PaintArt 2.0 is also divided in seven parts (submenus), i.e. the Load File submenu and the Save File, Drawing Color, Print, Paste from Clipboard, Paste to Clipboard and Exit submenus. Fig. 2 The Load File submenu enables the user to load or open a file in the current directory on the hard disk of the PC or even to change to different directories on the hard disk. To activate scroll bars it is only necessary to click and drag with the mouse. For that reason a complete support for a Microsoft compatible mouse is built in the program. [7]

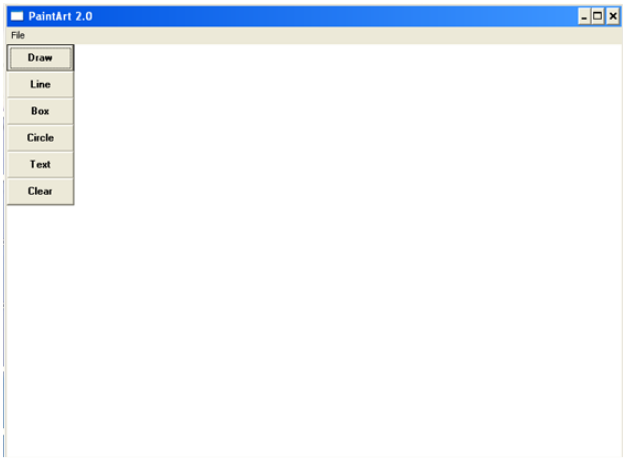


Figure 2. Main menu

The next submenus, as mentioned above, give the user the opportunity to save, print, change colors, paste to and from Clipboard and to exit the program. Fig. 2

The now enhanced graphics or painting capabilities of the computer program PaintArt 2.0 are part of the main window and enable the user to draw straight lines, boxes, circles, text, completely freehand and of course to clear or erase everything again. It is also possible to copy and paste the content or the bitmap image to the Clipboard or to a file. Fig. 3

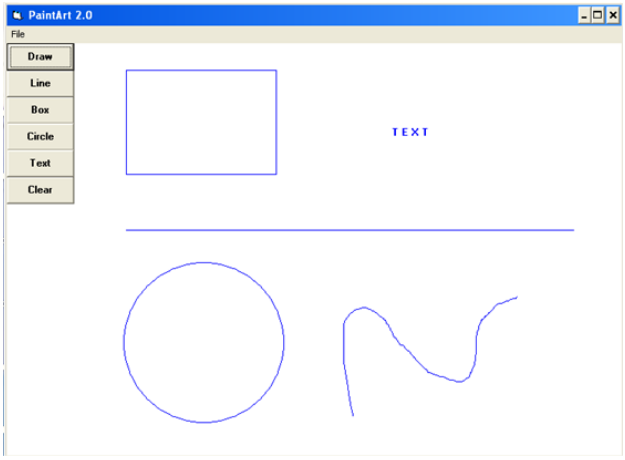


Figure 3. Painting capabilities

After the presentation of general characteristics of the computer program PaintArt 2.0 it is time to illustrate the built-in database of selected questions and answers which can be prepared in four ways or forms. [6] If the teacher of visual arts in secondary schools wants to teach about artistic styles, i.e. Baroque, the dominant artistic style during the 17th century which is characterized by intense emotion and a sense of movement, the teacher could show a painting of a well known artist (Caravaggio, Rembrandt, Velazquez etc.) and ask to which movement in painting it belongs in the form of yes/no, multiple choice or ordinary questions. It is also possible to present examples in sculpture, architecture etc. Showing examples in architecture, i.e. a church in Baroque, it could be asked to draw a line in the painting of the church which is characteristic for Baroque. It is recommended to choose paintings of well known artists, photos of famous buildings or sculptures which are a good example for the topic to be explored.

As part of the computer software, the art database - Baroque includes the following paintings of famous artists: Caravaggio – Taking of Christ, Rembrandt – Self Portrait, Velazquez – Las Meninas etc. Painting in Baroque was using a technique called "chiaroscuro" (light/dark), in which bright light forms a strong contrast with deep shadow, producing an effect like in theatres. [2] The paintings (photos) or bitmap files are numbered in consecutive order as a question (q01.bmp), answer of the pupil (q01a.bmp) and solution provided by the teacher (q01s.bmp). After the first painting or photo consisting of three files, the other paintings or photos follow as well in raising order, i.e. file number 2, 3 etc.

For illustration purposes we will present four forms of different questions all about Baroque. The first question is in yes/no form, i.e. is the statement correct or not. After having chosen the first question (q01.bmp) in the Baroque database the picture of the painting (Caravaggio) with the adequate question appears on the screen of the computer. Then the pupil imports its answer form, marks the right answer (yes/no) and saves it for later use (q01a.bmp). The solution together with the answer is provided in the solution form (q01s.bmp). Fig. 4

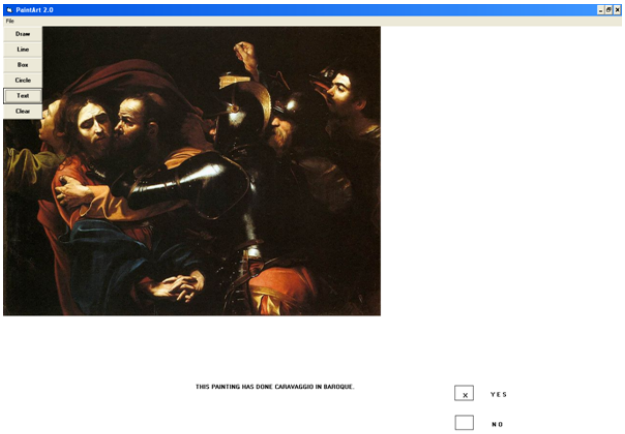


Figure 4. Yes/No form of questions

The second question is about another artist (Rembrandt) and his painting appears in the second file (q02.bmp). This time it is a multiple choice question and four answers about the author and the movement are presented, but only one is correct (a-d). After seeing the painting and having read the question the pupil should mark the appropriate answer in the answer file (q01a.bmp). Having done so the pupil can save that answer to the question in this or another file for later use and switch to the solution of the question the teacher has prepared and saved in another file. To do so, she or he has to open the third file connected to this question or the file containing the solution or the right answer (a, b, c or d) to the imposed question. In this case it is the file called (q02s.bmp). Fig. 5

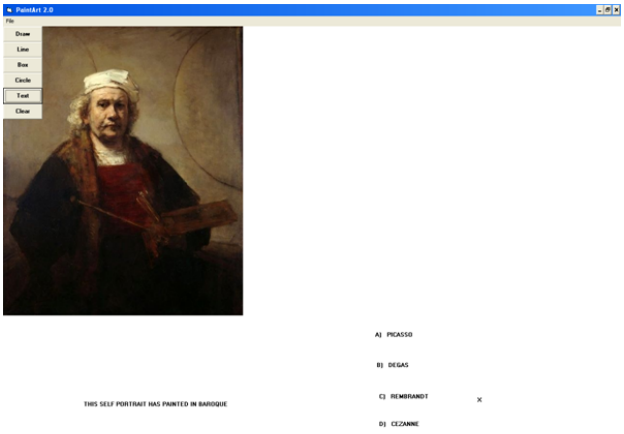


Figure 5. Multiple choice form of questions

The third sort of question (q03.bmp) is in the form of an ordinary question demanding answers in full-text mode about the artist (Velazquez) and the style (Baroque), i.e. the pupil should write down the answer and save it in a file as mentioned above (q03a.bmp). After that the solutions file (q03s.bmp) can be loaded and the answer and the right solution could be compared. Fig. 6

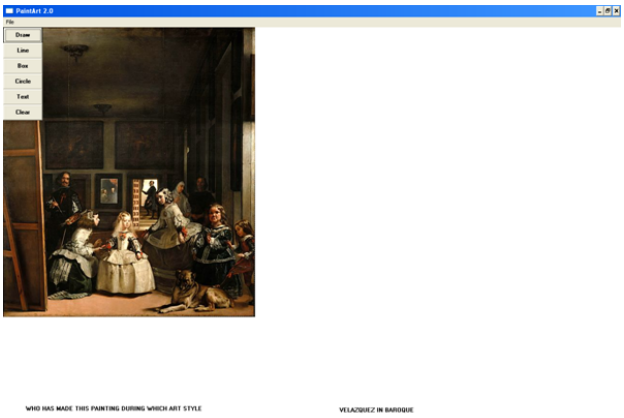


Figure 6. Demanding answers in full-text mode form of questions

The last sort of question could be one where a real application and drawing of the solution is required. A photo of the famous church Santa Maria della Salute in Venice will be displayed (q04.bmp) and the pupil should answer the name and the town of the church, to which art style it belongs and to draw a characteristic line in the photo itself by which this movement in architecture could be described. (A red line in the form of a wave or the letter «S») Then he or she can save the solution in a file (q04a.bmp) and compare it with the solution provided by the teacher (q04s.bmp). Fig. 7

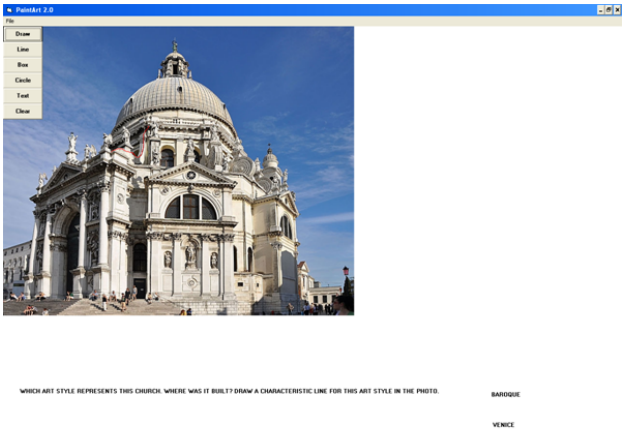


Figure 7. Real application & drawing of the solution form of questions

Additional questions and answers concerning other important artists being representative for this art movement, i.e. Baroque could be added, as well as other buildings, monuments or sculptures characteristic for this style. Having finished the database about Baroque, teachers could easily prepare lectures also for other movements like Gothic or Renaissance style for teaching visual arts in secondary schools. [2] [3]

III. Installation and Start of the Program

PaintArt 2.0 is an educational computer program for teaching visual arts in secondary schools where art database files in graphic format are included, so the installation and setup process is very simple. The program consists of only two application files, i.e. the PaintArt 2.0 program file (part2.exe) and the runtime module (vbrunxxx.dll). The Art database consists of at least 40 files in bitmap format (q01.bmp, q01a.bmp and q01s.bmp, etc.) for every single database (art style) called for example Baroque, Gothic, Renaissance etc.

To install the program PaintArt 2.0 it is necessary to copy all the files to a directory on the hard disk of a PC (for example c:\part2) and to start the program one should double-click the program file or icon (part2.exe) with a Microsoft compatible mouse.

IV. Conclusion

In this article a computer program consisting of a paint program called PaintArt 2.0 and a database of specially prepared forms in bitmap format for education of visual arts in secondary schools is presented.

The forms are containing questions in different forms (yes/no, multiple choice, ordinary questions etc.)

This program is designed and programmed in Microsoft Visual Basic for Windows, so it runs on PCs with Windows operating systems. For use with other operating systems like Linux adequate Windows program emulators could be used.

References

[1] Damjanov, J., Likovna umjetnost, I. dio, Školska knjiga, Zagreb, 2009

[2] Hosack Janes, K., et al., Great Paintings, DK Dorling Kindersley Ltd., London, 2011

[3] Janson, A.F., History of Art, Prentice Hall, NJ, 6th ed., 2004

[4] Microsoft, Microsoft Visual Basic 6.0 Programmer's Guide, Microsoft Press, Redmond, 1998

[5] Preziosi, D., The Art of Art History, Oxford University Press, Oxford, 2009

[6] Prohaska, Z., Teaching and Learning Visual Arts by Using a Computer Program, in 33rd Internationl Convention MIPRO 2010, Conference Book, Computers in Education, Opatija, 2010

[7] Prohaska, Z., Uroda, I., PaintArt – An Interactive Computer Software for Education of Visual Arts, in 36th International Convention MIPRO 2013, Conference Book, Computers in Education, Opatija, 201

Informatics as a basis for a knowledge – based society

Mila Nadrljanski* and Veronika Domitrović **
* Faculty of Maritime Studies – Split, Split, Croatia
** Visoka škola za inspekcijski i kadrovski menadžment u pomorstvu, Split, Croatia
*milamika60@yahoo.com, mila@pfst.hr
**veronika.domitrovic@icloud.com

Abstract - This paper deals with information education in development of a knowledge-based society in Croatia. The basic objective of the paper is focused on systematic comprehension and analysis of the situation related to information education and its tendency. Hence, some concepts and models for IT education in Croatia were propounded. Accordingly, these tasks were fully accomplished. Application of ICT has contributed to education and introduced the possibility of distance education using the Internet and computers, without space and time limits. Nowadays referred to as electronic education, such an education is a trend in society. Acquiring electronic education, incorporated in different levels of educational systems - undergraduate, graduate post-graduate and especially lifelong education – is a demand of contemporary education and Bologna process requirement.

Key words: computer science, information society and knowledge-based society

I. INTRODUCTION

The contemporary world lives in information society in which more and more people work on data collection, transmission and processing of information. Thanks to information technology and global connections today, knowledge is created and shared faster than ever. Information science, as the skill of data management and collection, is still the kind of prevailing knowledge. It implies knowledge by which the small interest group gains dominance and advantage over the majority, becoming the most successful one. Information society, by its nature, has no boundaries. However, most of the world's population will not benefit unless they are not fully able to participate in expanding of information society based on knowledge.

Modern information and communication technologies, services and infrastructure, especially broadband Internet access, play the fundamental role in reaching economy and society of knowledge. These technologies ensure the use of knowledge in business. This opens the possibility of a number of new e-business opportunities, as well as so-called virtual markets - which means that Croatian companies can, without any difficulties, deliver their ideas, products and services anywhere in the world.

II. FACTORS OF KNOWLEDGE SOCIETY AND THE ROLE OF INFORMATICS IN IT

The term 'knowledge society' denotes and accepts the fact that defining characteristics of today's society are human knowledge, skills, science, and intellectual capital [1]. Today's society is often referred to as the "information society", in order to point out the growing impact of information and communication technologies. It is the society where low-cost information and ICT are widely used, thus forming the basis of "society of knowledge", and emphasizing the fact that the best profits comes from the investment in intangible, human and social capital and that the fundamentals of the social prosperity are knowledge and creativity. This new society offers great opportunities: new job opportunities, new tools for education and training, easier access to public services, better involvement of people with special needs and better connections between regions [2].

With the development of technology in the 1960s, many authors came up with scenarios about the future world. American writer and futurologist Alvin Toffler[1] published his theory of the three waves [3]. According to it, in the first wave people lived on the farms and were engaged in agriculture, in the second wave they moved to the cities and started the industrial society which, in the end, wiped out the information and knowledge in the third, information wave.

According to Toffler, we have just entered the fourth stage of the wave, when it is necessary to master new forms of communication and technology for survival. In previous technological waves, the source of survival and power was in assets, the ability of doing simple labour and mastering certain physical skills. In the new age, the ability of using the means of communication is essential, and IT knowledge has become the source of politics and economic power. Toffler's opinion is that individuals and groups, who do not adapt to new knowledge and technologies, are destined to fail [4]. Toffler insinuated the possible consequences of new technology and faster pace of life, and he named them 'change disease'. A disorder caused by the inability to cope and failure to adapt to the

[1] In his last book "Powershift", Alvin Toffler demonstrated that the only solution to all the future problems is - the knowledge society.

new changes are manifested in the information overload and the stress of decision-making [5].

Conceptual approaches, methodology and information technologies that are commonly used while studying this topic are identified by Web browsing, contacts with institutions or individuals, using the e-mail and, finally by detailed evaluation.

According to the World Bank, [2] knowledge-based society consists of four indicators:

- education and training (educated population is a prerequisite of creating, sharing and using of knowledge),
- information infrastructure (from radio to the Internet, in order to spread information),
- legal and economic framework (which will encourage the free flow of knowledge and investment in the economy, information and communication technology, as well as to encourage entrepreneurship),
- innovation system (a network of research centers, universities, private companies and community groups – in order to accept global knowledge and adapted to local needs, and to create new knowledge).

Of the four mentioned indicators, the last three apply knowledge, and first one works on its creation and dissemination. It illustrates the gap between what the knowledge society really is, and what is talked about it. On the other hand, the world experience indicates that for the better transfer of knowledge from science to the economy, one of the most important questions is whether the government policy provides adequate compensation for the work of universities and the public sector. Simply put, the knowledge-based society is a society where knowledge is the main creative force in the personal, economic, social, cultural and material progress. First of all, we advocate for the introduction of performance indicators for the information society, in accordance with the European and international standards and norms. In the knowledge society all individuals have an equal opportunity to gain and share knowledge, and to use it for personal development (applied knowledge). If we compare this definition with the perception of a knowledge – based society in Croatia, we notice the basic mistake: whenever we talk about knowledge society in Croatia, we talk only about the acquisition.

III. HOW TO GET TO THE KNOWLEDGE SOCIETY

Knowledge has become the primary source. It is valuable, but the comprehension of its value is valuable as well. The comprehension of the importance of the knowledge society is invaluable. Knowledge and information must be easily accessible to everyone. In contemporary society, based on e-business and the rules of the new economy, professional knowledge is a prerequisite of economic growth. Knowledge has become the basic economic resource and the main prerequisite for gaining competitive advantage. That is why the modern society is knowledge–based society [6]. Acquisition of knowledge is only one of the four indicators of a knowledge society. Croatia has set a strategic goal to reach the knowledge–based society, but its elaboration is more than annoying, because it is being handled without any order and schedule. The world's most accurate tool for placing countries on the list of the knowledge society is renewed every year, and the total index of the knowledge society is calculated by indexes of four indicators: education, innovation, legal and economic regulations and ICT structure. Besides, the World Bank generates the knowledge index which shows the impact of knowledge on the economic development of a country. Indexes are expressed in figures from 1to10[3].

On the scale of 140 countries, Croatia[4] ranks 38th, with an index of 7.05. If we look linearly, the result is not bad. However, if we perceive Croatia in the context of the EU 28 +2 (28 EU countries together with Norway and Switzerland, countries whose internal legislation and practice are in line with those in the EU), Croatia has only Bulgaria and Romania behind. Even these two countries are catching up, having advanced 10 to 12 places on this scale for the last ten years, while Croatia has regressed six places in the same period (being skipped by Latvia, Lithuania and Barbados).

If we look at the table of the knowledge society[5], Croatia has many reasons for dissatisfaction, being on the 38th place, but may find comfort in the fact that its knowledge index increased from 6.86 (2006) to 7.05.

IV. CONCLUSION

Informatics is very important for the future strategies of development; therefore investment in equipment, teacher training and other facilities for efficient realization of educational plan must take an important position in priorities of the education system.

Program contents in the field of information and communication technologies should provide:

- acquiring skills for using contemporary computers and application software;
- learning the basic principles and ideas that computers and ICT are based on;
- acquiring skills for applying information and communication technology to different areas.

We have also found that skills, basic knowledge and problem-solving are three components of the educational process which can be considered separately. However,

[2] The World Bank Institute has defined the pillars of the knowledge society (in the Knowledge for Development Program) in order to find easier way to the so-called knowledge-based economy for transition countries. (http://www.worldbank.org/kam).

[3] The placement of Croatia on the scale of the knowledge society is shown on the Knowledge Index (KI) of the World Bank. (www.worldbank.org).

[4] Ibid., op.cit..

[5] Ibid., op.cit.

only their interleaving can give participants a good foundation for future life-long learning.

Skills are closely related to the current situation of information and communication technology. With the change of technology, the necessary skills change as well. Managing of those skills is commonly called information literacy. Nowadays, those skills comprise:

- the procedure of turning on a personal computer (connecting hardware, the procedure of starting up)
- using the basic functions of the operating system (installing new applications, deleting unnecessary programs, launching applications, setting some parameters of the system);
- the use of word processing program and document formatting (fonts, layout, text positioning, editing and correcting);
- using of image processing software and drawing tools;
- connecting the computer to the network (connecting to a computer network physically, service provider login);
- using web design programs;
- the use of the Internet for finding information (using the browser);
- using computers to communicate with others (e-mail);
- the use of spreadsheet applications;
- the use of databases;
- the use of instructional materials for further exploring the possibilities of application programs, and mastering new programs (using of documentation, printed manuals and help-file).

REFERENCES

[1] A. Uzelac, "Digital cultural assets in the Information Society between public domain and private property," Media Research no. 2004

[2] Department of Economic and Social Affairs of the UN, „Understanding Knowledge Society," 2005. source: http://unpan1.un.org/intradoc/groups/public/documents/un/unpan020643.pdf

[3] A. Toffler, "Powershift: Knowledge, Wealth and Violence at the Edge of the 21st Century", Bantam Books, 1990.

[4] A. Toffler, "Future Shock", Bantam Books, 1970.

[5] V. Grbavac and J. Grbavac, "Communication systems: theoretical foundations, reality, application, vision", DOIK, Zagreb, 2008, p.419

[6] M. Vidas-Bubanja, "Regional cooperation in the development of information society in Southeast Europe", a presentation, 2009. source: http://www.eukonvent.org/downloads/Regionalna_saradnj_u_JIE

[7] Dj. Nadrljanski, "Cybernetic foundations of system modeling and design," Misao, Novi Sad, 1982.

[8] Dj. Nadrljanski, "Computer Science for teachers," UF Beograd, 1996.

[9] Dj. Nadrljanski, "Information Systems," FIM, Kruševac, 2004.

[10] Dj. Nadrljanski and M. Nadrljanski, "Cybernetics in Education," University of Novi Sad, Sombor, 2005.

[11] Dj. Nadrljanski and M. Nadrljanski, "Fundamentals of Computer Science," University of Split, Faculty of Philosophy in Split, 2007.

[12] Dj. Nadrljanski and M. Nadrljanski, "Computer literacy and computerization of education," Informatologia 39, 2006, 4, pp. 262-263

[13] Dj. Nadrljanski and M. Nadrljanski, "Digital Media - Educational Software," University of Novi Sad, Sombor, 2008.

Informatics education in a knowledge-based society

Mila Nadrljanski* and Veronika Domitrović **
* Faculty of Maritime Studies – Split, Split, Croatia
** Visoka škola za inspekcijski i kadrovski menadžment u pomorstvu, Split, Croatia
*milamika60@yahoo.com, mila@pfst.hr
**veronika.domitrovic@icloud.com

Abstract - This paper discusses the place, role and importance of informatics education in contemporary society. The levels of computer science education are analyzed and explained, from elementary computer literacy, which is the basic requirement for the age of knowledge, to the education of professional IT experts. Implementation of computers in education, as a significant factor of educational development at all levels and areas of society was particularly emphasized. Some important clues for the future development and application of informatics are featured, together with its impact on the overall development of Croatian society.

Key words: informatics education, information technology, knowledge-based society

I. INTRODUCTION

The basic objective of this paper is to point out some relevant factors by which informatization contributes in joining the knowledge-based society.

In this sense, the paper is structured to point out important effects of development that are the basis for achieving the goal. Identifying developmental stages could raise the sensitivity of nation for Information Age as a prerequisite of the knowledge-based society.

Computer literacy should be implemented in widespread intervention through numerous models: through formal IT education (of children and adults), by enabling access to technology for all social groups, by development of online commercial markets and, what we do in Programme e-Croatia, creating conditions for business operations with the government via the Internet.

Basic knowledge is to some extent independent of the current state of technology and its concrete implementation. Being familiar with basic knowledge can facilitate lifelong mastering of new skills which are related to new forms of technology. Basic knowledge has more lasting value than skill does.

In education nowadays, information technology is not only a subject whose contents are studied by participants of educational process. It is the basis which primarily increases the efficiency of learning majority of subjects as well as effectiveness of educational process.

Lifelong education and training should be the focus of public policy and it should emphasize the significance of economic and social standards for the successful development of a knowledge-based society. The use of digital technology in society and a willingness of acquiring new skills that are necessary for the use of highly developed technology, as well as corporations based on knowledge - became the basic support of innovation and sustainable economic development. In return, the knowledge-based society can create new and better jobs and higher standards of living in order to achieve social prosperity. Therefore, the imperative is to reinforce the education that supports and accelerates the process of acquiring the current changes.

The knowledge-based society is built on the impact and development of contemporary information and communication technology and means a fundamental reshaping of the global economy from industrial to information society. [1] Its meaning goes beyond development of Internet companies and dot.com sector. The rapid ICT development has introduced profound changes in the way of work and life. ICT brings organizational, commercial, social and legal innovations.

II. IT EXPERT'S FUNCTION IN A KNOWLEDGE-BASED SOCIETY

The term IT expert implies knowledge and skills - usually a theoretical information systems and technologies, standards, languages and techniques an IT expert must have in order to perform tasks that particular systems require. Although the IT architecture is nowadays considered to be an integral part of web design, IT professionals involved in building and maintaining such complex business information systems value it as much as they value their job. Basic knowledge on web design,

[1] The term "Information Society" is commonly used for the new socio-economic paradigm which is the result of a constantly changing comprehensive process. The information society is not only a consequence of radical technological progress that is created due to "Research and Development (R & D)" or as a result of innovations in specific industrial sectors. It is the product of information technology and telecommunication technology integration that allows new, radical markets.

databases and programming languages, network systems and intelligent systems, therefore represent only necessary basis. On the other hand, the use of these systems (the most important of which is the search method) requires gaining of different skills and knowledge by IT experts, from search techniques, through understanding users' IT training and behaviour, to complete information systems.

The job of IT experts can be determined by the system they work on, or by the specifics of their employment, i.e. whether they are involved in the construction, maintenance or usage. Therefore, an IT architect participates (together with a graphic designer and a computer programmer) in building Web sites, Web portals, complex intranet systems and knowledge management systems. A digital librarian [2] forms digital collections using digital library software. IT specialist works on searching for information in the system, selection and evaluation of sources, and examining the usage of system in, for example, studies of users. Finally, in collaboration with experts of certain areas, a knowledge engineer [3] creates professional systems of these areas. Today it is common for IT expert to assume the role of a webmaster[4] or a web editor at a certain point, in order to perform maintenance and refreshing of the contents of the above systems.

III. ICT IN EDUCATION

Analyzing the influence of information technology on education, we have found that it works in three directions. The first is a growing supplement to classical education contents, its methods and forms of teaching with new information and communication technologies and other various innovations. The second is development of distance education, used at all levels of formal education, particularly at university. There are also different forms of informal education in virtual world, that meet individual needs of people (different courses, "university of the third age", etc.). The third is the extension of education, raising educational level of average people, democratization of education, vertical and horizontal linking of all levels and forms of education, increasing general and professional culture of the youth, and linking education and work.

The use of ICT in education, as well as e-learning technology, should be in every particular case effectively and directly linked to and aligned with the established educational (pedagogical) principles and objectives.

E-learning allows education "anytime, anywhere", which means that it is accessible during working hours and in free time, on work, at home or on a business trip.

The adoption of e-learning should be a strategic decision, and made by the administration of higher education institutions, in the context of a general strategy of actions and development of universities. Necessary resources should be provided to ensure applying of e-learning, and the process of introduction and application should be supported and constantly monitored by institutions and those in charge. The use of technology must be simple and adapted to the real needs of teachers and students. It has to be provided by high-quality, reliable, sustainable and efficient support system for e-learning.

Computer Science in education will aid teachers in acquiring a high level of professional and pedagogical knowledge for teaching in elementary and secondary schools - because that particular profession requires better knowledge of the fundamental areas of IT (operating systems, networks, programming).

The use of IT in education today is accessible even to those who are not computer experts. It takes only a basic ability to use a computer and the Internet for:

- available tools and software;
- making and implementation of education programs.

There are a large number of web pages designed specifically for lecturers, regardless of whether they are teachers in schools or professors at universities:

- tips on how to improve teaching with innovative approach;
- free print materials and their use in the classroom;
- a number of different tools for mastering certain parts of the teaching material.

Although distance learning is not a new phenomenon, with the Internet it gets a completely new meaning. The possibility of transmission and distribution of large amounts of information with the aid of new media turns over a new leaf in the concept of permanent education. The significance of this approach to education is recognized worldwide - especially in the EU countries and the U.S.A. Using their experiences and recognizing the value of distance learning, there is a big target group in Croatia that is interested in this kind of education.

Today's models of distance education almost entirely leave the classroom and move into virtual space. The Internet as a distributed network with millions of connected computers and computer networks is an ideal choice for its structure an content, because it assumes interaction and encourages collaboration. There are already thousands of quality and dynamic online projects in the field of e-Learning. Distance learning is no longer limited to one-way broadcast, but also includes a range of electronically transmitted educational materials and interactions within workgroups.

Lifelong education system is essential for growth and development of society and it is necessary in the era of rapid change of required skills and knowledge.

The educational system of lifelong education is more directed to users. Individuals become active participants in the educational process, which is greatly contributed by the use of information and communication technology. The use of ICT in education offers innovative methods of teaching and learning. Development of lifelong education[5]

[2] http://www. hidd.hr/virref/zanimanja_digitalni.htm
[3] http://www.hidd.hr/virref/zanimanja_inzenjer.htm
[4] http://www.hidd.hr/virref/zanimanja_master.htm

[5] The Cooperation Agreement on Promoting Lifelong Learning was signed on 2nd October 2007 by the Ministry of Education.

in Croatia is necessary because there is increasing number of seniors in total population. Sensibility and concern for this population should be developed by educational content and the media as well. Work, social and family relationships are changing with rapid development of new technologies, so people are expected to keep up with these changes and participate in society actively. One has to learn continuously in order to be able to keep pace with the changes.

The use of personal computer and the Internet, a global network which is an inexhaustible source of various information and knowledge, has enabled an outstanding change in the way of acquiring knowledge in educational process. Computer skills become one of the basic elements of literacy. The introduction of information technology in educational process has two aspects: knowledge about information technology, and ability to use it as a technological starting point in other subjects.

Computer Science should not be merely a content of a particular subject; it should rather be implemented in contents of other subjects and activities at school. There are a number of reasons for this. First of all there are global issues of organization and management processes in education.

New information technologies have become the backbone in application of multimedia in teaching. They provide great opportunities, but also require specific knowledge and skills together with appropriate implementation strategy. Using new technologies efficiently is a challenge for teachers who must acquire relevant knowledge and skills in order to respond to the increasing demands of educational process.

The introduction and application of information technology (IT) in primary and secondary schools have initiated new groups of tasks related to computer and communication infrastructure. Computer-communication infrastructure provides a wide range of applications in education, from finding information on the Internet, to virtual classrooms or schools. The introduction of IT in educational process should be encouraged and supported. It will be necessary to create digital textbooks, organize online courses, create programs for virtual classrooms, hold classes of other subjects in computer rooms, etc. At the same time, we should continue with education of teachers who will create more and more educational materials on the network and use them.

There is an increasing usage of information-communication technologies in the classroom [1]. Teachers often use presentations during classes. However, they should know certain rules for making a presentation in order to make it understandable and interesting for their audience.

The development and improvement of information and communication technologies play a crucial role in all aspects of human activity, as well as in education. Thus certain activities change their structure and functions of individual elements within it. In the teaching process we can observe the influence of technology from the perspective of every element of education: student, teacher, content / program or the context in which those elements are placed. Transformation of the teacher's role is particularly important in this process. With introduction of ICT, the teacher becomes a person who motivates and leads (more like a mentor or tutor) and thus helps a student transform the information into their own knowledge. In this process teachers are expected to find ways to transform their old traditional skills into new ones, to develop new paradigms of teaching, and to adopt new administrative and technical skills.

Bologna concept of studies implies adequate information support in the process of studying (Learning Resources Support). Therefore, it is necessary to install CourseWare applications as an aid in the process of teaching.

The new vision of higher education system[6], in which the student is at the center of the teaching process, opens possibilities of teaching and learning, which assume that teaching methods are tailored to student's needs. The methods of communication, assessment, feedback and overall 'teacher – student' and 'student – student' interaction, all of them are adjusted to individual student's needs.

Possibility of using IT in education nowadays is accessible even for those who are not computer experts. In creation and implementation of various educational programs over the Internet, there are so called CourseWare tools, available on the market.

Education and professional development of teachers are the main factors that affect the level of students' achievements. When talking about improving the quality of teacher education, we should keep in mind not only presumed higher education of teachers, but also their professional development from the perspective of lifelong learning. The aim of both forms of education is lifelong learning and renewal of knowledge and skills of teachers, but still in a function of lifelong learning of their students.

IV. INFORMATION EDUCATION IN CROATIA

The education system should adapt to society. Today's education is partially changed in the countries which had the need to change, whereas in many of them the system was not changed because the society did not change much.

Most of these changes were related to the organization of the system, e.g. education policy, but teaching itself has not experienced any necessary changes.

For present needs this approach to education is not efficient enough, and even the content is not often adequate. Despite new horizons are opened at all levels of education, the educational system is not able to provide what society needs.

All potentials provided by new insights on pedagogy and information technology are currently not used.

In the traditional way of teaching there is a discrepancy between the teaching material which is presented and knowledge or skills needed in the workplace.

[6] Source: http://hlede.net/portal/content/view/53/40/ (April 15, 2007)

In contrast to accomplishments of the second half of the last century, Croatian education system has lost qualities of an export product in demand, with few exceptions. At this point there are no indications of upcoming changes. Opening of private colleges at the level of professional studies has enabled competitiveness with foreign educational institutions. It is a risky venture because foreign educational products have already been used on the Croatian market, such as IBM and CISCO academies or e-learning courses.

It can be concluded that there are further necessary actions required for development of information society in the Republic of Croatia:

- creating appropriate economy initiative and institutional environment;
- creating a society of educated, flexible and creative people;
- making a dynamic information infrastructure;
- creating an efficient innovation system;
- adoption and adaptation of the legal infrastructure for the knowledge-based society;
- promoting the development of a knowledge-based society.

V. CONLUSION

It is well known that Croatia has a long tradition and significant results in primary, secondary and university education, which is a good presumption for development. However, it is necessary but not sufficient condition.The challenges of a global market of education impose imperative of changes on Croatia, changes which are rather radical and will enable global competitiveness of Croatian education, as it is indicated in this paper.

Highly educated and qualified professionals of Croatia, close cooperation of scientific and educational communities with the economy and society as a whole, based on high technologies, together with insistence on competitiveness of knowledge, they are the answers to the question of how to achieve a knowledge-based society in Croatia.

REFERENCES

[1] A. Mee, "E-learning and change in higher education – the policy environment'," European Journal of Open, Distance and E-learning (EURODL), 2007.
http://www.eurodl.org/materials/contrib/2007/Adrian_Mee.htm#Figure_1.

[2] Dj. Nadrljanski, "Cybernetic foundations of system modeling and design," Misao, Novi Sad, 1982.

[3] Dj. Nadrljanski, "Computer Science for teachers," UF Beograd, 1996.

[4] Dj. Nadrljanski, "Information Systems," FIM, Kruševac, 2004.

[5] Dj. Nadrljanski and M. Nadrljanski, "Cybernetics in Education," University of Novi Sad, Sombor, 2005.

[6] Dj. Nadrljanski and M. Nadrljanski, "Fundamentals of Computer Science," University of Split, Faculty of Philosophy in Split, 2007.

[7] Dj. Nadrljanski and M. Nadrljanski, "Computer literacy and computerization of education," Informatologia 39, 2006, 4, pp. 262-263

[8] Dj. Nadrljanski and M. Nadrljanski, "Digital Media - Educational Software," University of Novi Sad, Sombor, 2008.

Visual Verification of the Properties of Pseudorandom Sequences

S. Predanić, G. Vujisić, T. Horvat
Polytechnic of Zagreb, Department of Electrical Engineering, Zagreb, Croatia
stipe.predanic@tvz.hr; goran.vujisic@tvz.hr; thorvat3@tvz.hr

Abstract - This paper describes properties of graphical presentation of pseudorandom number generator. Numerous more-or-less rigorous tests for randomness are available in the literature, and some of them are presented here. It is shown that both graphical and analytical tests give almost equal results, but graphical tests are simpler and closer to human perception. Because of that properties, visual tests combined with simple statistics are proposed as optimal for undergraduate higher education curricula. Mentioned methods are used on multiple courses at professional study of electrical engineering at the Polytechnic of Zagreb.

I. INTRODUCTION

Risen in everyday higher education environment and practice, initiated by students' problems, supported with Knuth [1] and motivated by Wolfram [2], multimedia modelling of telecommunication processes is illustrated with the case study of pseudorandom sequence (PRS) .

Born and grown in the ICT rich environment, students of the present expect multimedia lecturing as common educational standard. Short in time and with complex knowledge to be presented and adopted classic teaching methodology as mathematical formulas and graphs is to be augmented with audio and visual models. Visual randomness tests for the PRS are better because one can immediately conclude from the graph whether a PRS is random or not.

II. A PSEUDORANDOM SEQUENCE

Pseudorandom number generator is a deterministic algorithm that, given the truly random input binary sequence length k, outputs a binary string l>>k appears to be random. The input for the generator is called a seed, while the output is called a pseudorandom sequence (PRS).

One commonly used pseudorandom number generator is linear congruential generator (LCG), devised by D.H.Lehmer in 1948. It uses the following formula:

$$K_{i+1} = (a \bullet K_i + c) \mod m \quad (1)$$

The sequence is always repetitive, but one can get a random looking sequence when integer numbers a, c, m and first number K_0)(often called seed) are well chosen.

Example:

If there is a four-bit computer system that uses binary representation for integer numbers from 0 to 15, K_0 may be any number ("seed") between 0 and 15, in order to present the following formula:

$$K_{i+1} = (5 \bullet K_i + 3) \mod 16 \quad (2)$$

Given K_0 = 1, the PRS generated by (2) is: 8,11,10,5,12,15,14,9,0,3,2,13,4,7,6,1,8,11...

Another commonly used pseudorandom number generator is Lagged Fibonacci generator (LFG), which uses generalisation of Fibonacci sequence with formula:

$$Ki=(K_{i-1} \bigcirc K_{i-2} + c) \mod m \quad (3)$$

where $\bigcirc$ denotes any general binary operation (addition, subtraction, multiplication etc.)

Example:

If there is a four-bit computer system that uses binary representation for integer numbers from 0 to 15, K_0 and K_1 may be any number ("seed") between 0 and 15. If operation is addition, the following formula is used:

$$K_i=(K_{i-1}+K_{i-2}) \mod 16 \quad (4)$$

Given K_0=1 and K_1=4, the PRS generated by (4) is : 1,4,5,9,14,7,5,12,1,13,14,11,9,4,13,1...

III. RANDOMNESS TESTS

Verification is needed to determine whether the generated sequence seems to be random or not. Due to the question of useful randomness, some postulates and theories were devised and from there various tests for checking the randomness of a sequence. Standard tests are given in the Table 1. Their combination provides greater reliability for the randomness evaluation.

TABLE 1 RANDOMNESS TESTS [3]

Frequency (Monobits) Test
Test For Frequency Within A Block
Runs Test
Poker test
Random Binary Matrix Rank Test
Discrete Fourier Transform (Spectral) Test
Maurer's Universal Statistical Test
Lempel-Ziv Complexity Test
Linear Complexity Test
Serial Test
Approximate Entropy Test
Cumulative Sum (Cusum) Test

A. Golomb's randomness postulates

Golomb's randomness postulates were among the first attempts to establish some necessary conditions for periodic pseudorandom sequences that seem to look random, but these conditions are far from being sufficient for such sequences to be considered random.

If s is a periodic sequence of period N, and sequence is binary, Golumb's randomness postulates are as follows:

R1: In the cycle sN of s, the number of 0's and 1's is different by at most 1.

R2: In the cycle sN , at least half the runs have length 1, one-fourth have length 2, one-eight have length 3, and so on as long as the number of runs so indicated exceeds 1. For each of these lengths, there are almost equally many gaps and blocks.

R3: The autocorrelation function C (t) is two-valued. That is for some integer K.

$$N \cdot C(t) = \sum_{i=0}^{N-1} (2s_i - 1) \cdot (2s_{i+t} - 1) = \begin{cases} N, & \text{if } t = 0, \\ K, & \text{if } 1 \le t \le N-1. \end{cases} \quad (5)$$

B. Frequency (Monobits) Test

The purpose of the test is defining the proportion of 0's and 1's for the whole binary sequence. The test estimates the distribution of those close to ½. The number of 0's and 1's in the sequence should be about the same. All subsequent tests depend on passing this test.

C. Frequency Test within a Block

This test is focused on the proportion of 1's within the M-bit blocks. However, its purpose is to determine whether the frequency of the M-bit block is approximately M / 2, as expected under the assumption of randomness. If the block size in this test is M = 1, the test degenerates to Frequency (Monobit) test.

D. Runs Test

The basis of this test is the total number of runs in the sequence, where a run is a continuous sequence of identical bits. A run of length k consists of exactly k identical bits and with opposite values before and after the run. The purpose of the runs test is to estimate whether the number of runs of zeroes and ones of various lengths is as expected for a random sequence. This test determines whether such oscillations between zero and ones are too fast or too slow.

E. Maurer's universal statistical test

The basic idea of Maurer's universal statistical test is that it is not possible to compress significantly (without the loss of information) the output sequence of a random bit generator. So, if a sample is significantly compressed, the generator should be rejected as being invalid. Instead of actually compressing the sequence, the universal statistical test calculates the amount which relates to the length of the compressed sequence.

F. Approximate Entropy Test

The focus of this test is the frequency of overlapping m-bit patterns. The objective of the test is to compare the frequency of overlapping blocks in two consecutive / adjacent lengths (m and m +1) against the expected result for a random sequence.

G. Linear Complexity Test

The purpose of this test is to register the length of the generated feedback, and to determine whether or not a sequence is complex enough to be considered random. Random sequences are characterized by a longer feedback register. A short feedback register implies non-randomness.

H. Serial Test

This test is based on the frequency of overlapping in each m-bit pattern across the entire sequence. The purpose of this test is to determine whether the number of occurrence of 2m m-bit overlapping samples are approximately the same as it would be expected for a random sequence.

Some of these tests are suitable for teaching students in technical studies, because they do not require complex mathematical background. Graphical tests are even more suitable. They will be described in the next chapter.

IV. GRAPHIC / VISUAL REPRESENTATIONS OF RANDOMNESS TESTS

Scilab environment was used to get the graphical presentation of randomness tests. Why Scilab? Because it is free to download from the web and use without paying for it. So students can download it and use it at home to practice. They can develop their own random generators and the picture will tell them immediately whether the sequence is random enough or not.

For graphic analysis histograms and 2D plots of matrix are used. With these two tests it can be determined whether the sequence is acceptable. The following examples will show how to make a histogram and 2D plot of a matrix in Scilab.

A. Histogram

This test is an extension of the Frequency Test. Instead of their individual bits, bytes or their decimal equivalents their data are counted and resulting statistical frequencies are plotted at Figure 1 and Figure 2.

TABLE 2 LFG AND HISTOGRAM PLOTTING

```
h(1)=31;
h(2)=121;
for i = 3:n,
h(i)=modulo((h(i-1)+h(i-2)+13),255);,
end; //  To get LFG sequence
histplot(256,h,normalization=%f)
```

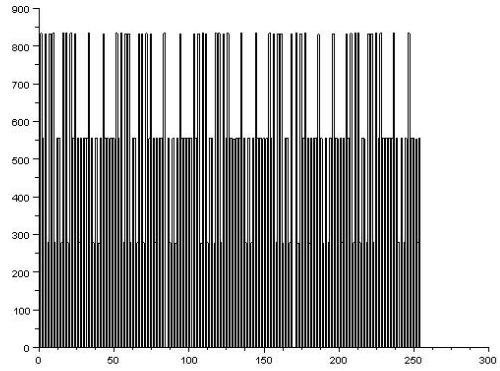


Figure 1 LFG Scilab Histogram

TABLE 3 . SCILAB RANDOM FUNCTION AND HISTOGRAM PLOTTING

```
r = pmodulo(round(10000*rand(N,1)),256);
// To get the random sequence
histplot(256,r,normalization=%f);
// To draw a histogram
```

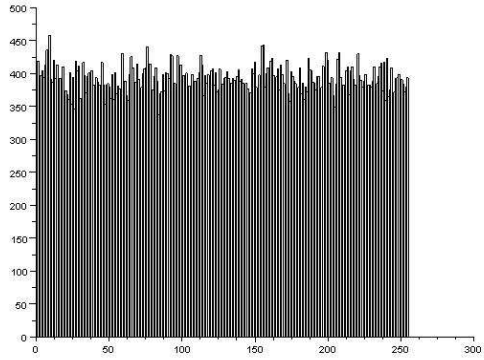


Figure 2 Random Scilab Histogram

In Figure 1 Histogram of Lagged Fibonacci generator (LFG) is shown. It is obvious that the distribution is not uniform. This means that a different random sequence generator should be used to get a better random sequence.

In Figure 2 another random sequence generator histogram is shown. From graph it is visible that the distribution of relative frequencies is close to uniform.

B. 2D Histogram

This test presents the frequency of ordered and overlapped bytes. It is expected that the distribution of pairs (2D) is uniform and that each byte relative to the previous one is of the uniform distribution as well. If not so, two conclusions are possible:

a. Pairs are not of equal probabilities

b. Transitions from one byte to another are not of equal probabilities.

In both cases algorithm is not to be accepted as good as RNG. 2D, 256x256 histogram matrix is plotted in Scilab environment for the LFG (Table 2) and the random function (Table 3).

Scarse point dispersion in Figure 3 indicates that transitions are not of equal frequencies.

In Figure 4., random function's 2D matrix plots is presented. Intensity of white is proportional to the correlation of each byte with the following one. If any pattern can be identified in the graph: circle, rectangle, line, triangle, etc..., that means the sequence is not random enough.

TABLE 4 2D MATRIX AND HISTOGRAM PLOTTING

```
for i=1:N-D,
z(1+r(i),1+r(i+D))=1+z(1+r(i),1+r(i+D));,
end; // Calculating the field
grayplot(1:256,1:256,z); //
To draw a 2D plot of matrix
```

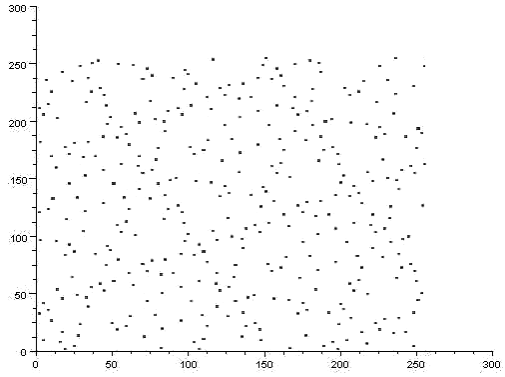


Figure 3 Scilab 2D matrix plot (LFG)

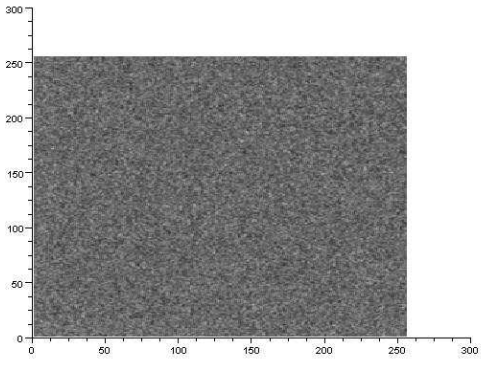


Figure 4 Scilab 2D matrix plot (random)

V. USE OF PSEUDORANDOM SEQUENCE

Pseudorandom sequences made and checked by described methods must be used in situations where true randomness is not necessary and described sequences could be called „random enough“. Simulations and models of communication channels can use pseudorandom sequences even if they don't pass all regular tests, while complex algorithms like those used in cryptography heavily depend on true randomness.

At the Polytechnic of Zagreb pseudorandom sequences are used in courses that have Matlab and Scilab based simulations and models of communication channels, protocols and encoding. For example this is used in laboratory exercises where all students get the same original message, and use a pseudorandom sequence with length of 32768 (15 bit) or 65536 (16 bit) numbers as a noise generator. Only Gaussian noise (white noise) is truly random, so even some repetition or unevenness in pseudorandom sequence can return real results. At the same time students can compare results and acknowledge difference in output depending on channel noise. Another use in modeling communication channel is for message itself when protocols are analyzed: by using a sequence length of 256(8 bit), a simulation of one byte (either binary or ASCII) can be made. Some repetition and unevenness is expected (and wanted) as most protocols use constant headers, and often some values of parameters are more used than others. Message can then be analyzed in comparison to constant noise in transmission channel, or it can be used to present coding and queueing theories.

VI. CONCLUSION

As described, it is easier for students when a large matrix of data is presented visually than numerically. Students can immediately recognize from the graph if the sequence generated is random or not. Basic statistical tests of randomness, combined with some visual tests should give a substantial reliability of the randomness. Both graphical and analytical tests give almost equal results, graphical tests are simpler and closer to human perception. Therefore, the graphic and visual mode is more suitable for the teaching purpose.

REFERENCES

[1] Donald E. Knuth. Art of Computer Programming, Volume 2: Seminumerical Algorithms, Addison-Wesley 2001.

[2] Stephen Wolfram. A New Kind of Science, Wolfram Media, Inc., 2005

[3] Guide To The Statistical Tests, NIST at http://csrc.nist.gov/groups/ST/toolkit/rng/stats_tests.html [08.02.2014]

INTEGRATION OF CONCEPTUAL DATA MODELING METHODS: HIGHER EDUCATION EXPERIENCES

Lj. Kazi, B. Radulovic, I. Berkovic, Z. Kazi
University of Novi Sad, Technical faculty "Mihajlo Pupin" Zrenjanin, Serbia
leremic@tfzr.uns.ac.rs, biljana.radulovic66@gmail.com, berki@sbb.rs, zoltan.kazi@gmail.com

Abstract - One of the most important phases in information system development is data modeling. In higher education and in professional environment, the basic data models are ER model, relational model and object oriented model, i.e. class diagram. Since conceptual modeling is crucial data modeling activity, special concern on methods for conceptual modeling is given in this paper. Case study shows that application of separate conceptual data modeling methods produce different and incomplete models. This leads to the conclusion that it is important to use integration of these methods and apply them within the information system development process. Educational experiences of applying this approach at University of Novi Sad, Technical faculty "Mihajlo Pupin" in Zrenjanin are presented as description of class exercises and exams results.

I. INTRODUCTION

Information system development is based on the crucial phase related to business domain knowledge representation and client requirements specification [1]. Client requirements most often are presented in the form of textual description. Knowledge representation of business processes and data are commonly presented by business process models, data dictionary and data models [2], supported by CASE (Computer Aided Software Engineering) tools. The process of knowledge representation of a particular business domain starts with textual description of business process workflow and transformation to business process models such as data flow diagrams or business workflow diagrams. Within creating these diagrams, data dictionary is filled with data flows and data stores structure elements - data items. This way, business processes model presents a basis of functional aspect and data dictionary is basis for information aspect of an information system architecture [3], i.e. application architecture and information architecture within information system architecture [4]. Our approach for mapping of business process to use cases in application architecture is proposed in [5] and mapping of data dictionary elements to data models elements is proposed in [6].

Most commonly used data models within professional and educational working environments are ER model [7], relational model [8] and object-oriented model, i.e. class diagrams [9]. Scientific environments (as well as some advanced educational environments [10]) in recent years include domain ontologies creation as more complete method for business domain knowledge representation [11], with methods and tools for transformation of ontologies to conceptual data model [12] and relational data model [13].

Conceptual data models (CDM) are supported by CASE tools and widely used within higher educational environments. Automation of transformation of ER, relational and object-oriented models from one type to another is supported within CASE tools. Since ER model (often considered as conceptual data model) could automatically be transformed to relational and class diagram, the crucial activity in data modeling is conceptual modeling.

Our observations on data modeling teaching issues and initiative for improvements started at practical lectures and exams at University of Novi Sad, Technical faculty "Mihajlo Pupin" in Zrenjanin, Serbia (TFZR). These initiatives lead to research on students' data modeling errors and proposing teaching improvement implications [14], as well as development of software tools for automated evaluation of conceptual data models [15, 16]. The approach to the integration of conceptual modeling methods in conceptual modeling (including some new conceptual data models), as well as integration of application of these methods within an information system development were introduced in [6].

Integration of methods application within educational environment was formulated as a new methodology within this paper. This paper presents description of existing and new conceptual data modeling approaches and methods and the integration of these methods. A case study describes using separate methods and shows the necessity of their integration application. Experiences of the methodology application at practical exercises and exams is described with some conclusions. Further improvements suggestions are presented.

II. HIGHER EDUCATION IN CONCEPTUAL DATA MODELING

Conceptual data modeling at University of Novi Sad, Technical faculty "Mihajlo Pupin" in Zrenjanin, Serbia is included in subjects/courses of Databases (at 2nd year of study) and Information systems (4th year of study) at Bachelor study in IT. During Databases course, student

learn separate types of models (ER, relational) and their basic elements. They practice creating simple models by using very detailed (having all elements descriptions available), but simple examples of business domain descriptions. The context of creating a data model within an information system development is introduced at Information systems course.

Research on conceptual data modeling in higher education was conducted upon a small sample of higher educational institutions (Universities schools/faculties of electrical engineering from Zagreb (Croatia) [17], Sarajevo (Bosnia and Herzegovina) [18], Podgorica (Montenegro) [19], Skopje (FYR Macedonia) [20] and Belgrade (Serbia) [21]). At all these higher educational institutions, conceptual data modeling is taught within the Databases course at Bachelor study level, mostly at second year of study. In all these institutions, the topic of conceptual data modeling is tightly related with ER-diagrams. At almost all of these higher educational institutions, foreign literature on databases fundamentals is proposed as a main learning source and that literature do not include all conceptual data modeling methods, or even their integration and related issues. Educational contents of the examined database courses are mainly focused on different data models and technologies, rather than detailed approach to conceptual data modeling methods. Information systems courses (at the same sample schools/faculties) also do not cover all methods of conceptual data modeling.

III. CONCEPTUAL MODELING APPROACHES AND METHODS

In this section a short overview of approaches and methods for M is given as an introduction to the proposed integration approach.

Some *approaches* in conceptual data modeling that were described in related work are: (A) Direct (complete) modeling vs. Sequential (partial) modeling and integration [22, 23]; (B) New models creation vs. Using and modification of design patterns [23]; (C) Forward (direct) modeling vs. Reverse engineering [23]; (D) Data modeling as information system development phase vs. Data modeling within the whole information system development lifecycle [19] and (E) Iterative conceptual data model refinement [3].

Some *methods* that were described in related work are described in Table 1.

The new methods (from [6]) are presented in Table 2.

IV. THE METHODOLOGY FOR CONCEPTUAL MODELING METHODS INTEGRATION

The approach of integration of conceptual data modeling methods within an information system development is introduced in [6]. The proposed approach include two main aspects of integration:

1. Integration of conceptual data modeling methods - creation of models by application of different modeling methods and their integration to complete model;

TABLE 1. METHODS OF CONCEPTUAL MODELING IN RELATED WORK

Method	Method description
Grammar analysis of business domain textual description [23, 24]	Each noun is candidate for entity, gerund is candidate for entity, verb is candidate for relationship. Application of this method results in model with entities (from nouns) and relationships (from verbs).
Normalization of relations(tables) in relational model [3, 10, 23]	Each table in relational model should be in 1st, 2nd, 3rd normal form, or other normal forms.
Creating submodels based on: External schema [22], XML schema [23], Primitive process [23]	For each external schema, XML schema or primitive process, submodel is created and integrated to the whole system's conceptual model.

TABLE 2. NEW METHODS OF CONCEPTUAL MODELING [6]

Method	Short description
Assignment of entities to attributes from data dictionary	Data items from data dictionary from data flow diagrams are assigned to existing entities in model or new entities are created according to attribute name.
Data store decomposition by structure notation analysis	Decomposition of data stores by analyzing syntax anotation [25]) of their structure.
Data store normalization	Decomposition of data stores from data flow diagrams by application of 1st ,2nd , 3rd normal form.
UML use case models based partial modeling	Each use case is assigned to a conceptual data sub-model that presents data to be used by the specified use case.

2. Integration of the process of applying conceptual data modeling methods within information system development using artifacts from previous phases of information system development as an input to the process of conceptual data modeling.

In this paper we present the integration of approaches and methods as a new methodology that focuses on application of approaches: Direct (complete) modeling; New models creation; Forward (direct) modeling; Data modeling within the whole information system development lifecycle (by taking artifacts from previous information system development phases; artifacts are: a) business process model with data dictionary, b) use case diagram); Conceptual data model refinement (by using one conceptual modeling model after another, central model becomes more complete).

The proposed methodology integrates application of existing and new methods (described in the previous section) of conceptual data modeling within a particular sequential order. The order of methods application within the proposed methodology is presented in Table 3.

The methodology includes creating "integral model" by integration of results of conceptual data modeling methods application. There are two basic approaches to application of this methodology:

- Creating an „integrated“ model – application of each method separately and integration of all results from each method application;
- Creating a „central“ model – all methods are performed upon a central model, each enriching the central model with new elements and enhances quality of the final conceptual data model.

The methodology also includes methods for evaluation of created integral model by checking correctness and completeness of the created model.

Correctness of the model is evaluated from syntax and semantic aspect by using specific check list created according to Van Belle's framework for general model evaluation [26]. This check list (detailed description in [27]) is used within exercises (at TFZR) for evaluation of working models during classes, as self-checking tool for each student during practical exam and as a check list that is used by a teaching staff for assessment of each student's practical exam work. The proposed checklist includes characteristics in the scope of a whole model as well as each element category (entity, attribute and relationship) characteristics, quality features and metrics.

Completeness of the model is evaluated by checking completeness of entities, attributes and relationships of the central model. Completeness is checked regarding input material usage, i e. how much is the given input material «covered» by elements of conceptual diagram. Completeness is also evaluated by creating conceptual data sub-models for each use case. As an input material to conceptual data modeling process, a table of assignment of primitive processes to use cases, i.e. software functions (introduced in [5]) is also given and is consulted within the process of completeness evaluation of created conceptual data model.

TABLE 3. ORDER OF METHODS APPLICATION WITHIN THE PROPOSED METHODOLOGY

Order of methods application	INPUT - material	Method	OUTPUT – result
1	Textual description of business process	Grammar analysis of business domain textual description	1st draft model with: Entities, relationships
2	List of data dictionary data items	Assignment of entities to attributes from data dictionary	2nd draft model with : Entities with attributes assigned
3	Syntax structure of each data store from data flow diagram	Data store decomposition by structure notation analysis	3rd draft model with Entities, relationships,attributes,: cardinalities of relationships, IS-A hierarchy relationships
4	Data store structure, i.e. list of data items for each data store	Data store normalization	4th draft model with : Entities, relationships, attributes, identifiers, functional dependencies within entities, non-transitive functional dependencies within entities
Integrated/Central draft model			
5	Integrated draft model	Checking correctness and performing corrections	Correct model
6	Correct model, Table (created according to [5]) of primitive processes assignment to Use cases	Checking completeness and corrections	Correct and complete model
Correct and complete model			

V. CASE STUDY

A. Background

Within several recent years, all previously presented approaches and methods, as well as their parial integration were presented and applied within theoretical and practical lectures (exercises) and practical exams at TFZR within the Information systems course of Bachelor studies in IT at TFZR. Only method „Grammar analysis of business domain textual description“ is applied in a simplified form within *Databases* course of Bachelor studies in IT at TFZR.

B. Aim and Research Question

Aim of this case study is to prove the necessity of application of the methodology for integration of conceptual data modeling methods.

Research question: Does application of the methodology for integration of conceptual data modeling

methods give better results, than application of separate methods?

C. Research Method

Research is conducted by comparing results of application of separate methods to the result of integration of these methods application results. The research is performed as follows:

1. Each of the four methods for conceptual data modeling is applied upon material from the same business domain. Application of each method gives particular conceptual data model.
2. Integrated model is created as a result of integration of conceptual models that are results from application of particular methods.
3. Comparison is performed descriptively, based on number of entities, relationships and attributes.

D. Research material

The case study is applied in 2013/2014 school year during winter semester of practical classes in Information systems. The methods of conceptual data modeling are applied by using CASE tool Sybase Power Designer 10. The business domain of the case study is related to simple taxi organization, i.e. phone-service subdomain workflow. Material for conceptual modeling methods application in this CASE study consist of: textual description of business process, process tree, data flow diagrams, list of data items and data types/domains, table of mapping primitive processes from process tree to software functions, use case diagram with software functions.

E. Results

Results of separate application of methods are presented as separate conceptual data models. The first model (Figure 1) is created by application of method „Grammar analysis of textual description“. The second model (Figure 2) is created by application of method „Attribute assignment to entities“.

The third model (Figure 3)) is created by application of method „Syntax analysis of data store structure“. The fourth model (Figure 4) is created by application of method „Normalization of data stores“.

Final model presents integration of separate models (Figure 5).

F. Discussion and conclusion

Results of separate application of methods show that the resulting models are different and incomplete:

- Application of grammar analysis (first method) of textual description leads to incomplete conceptual data model, since it gives only entities and relationships, without attributes. *Incompleteness is in attributes.* It is important to mention that even textual descriptions could be given with more or less details and therefore are not a reliable source of knowledge.

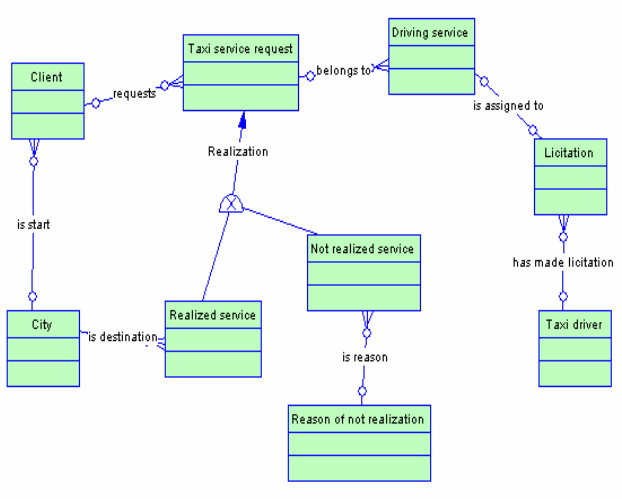


Figure 1. Conceptual data model as a result of the first method

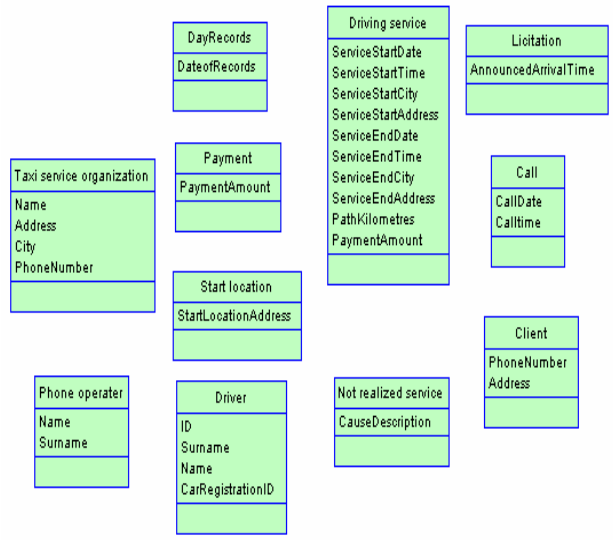


Figure 2. Conceptual data model as a result of the second method

- Application of attribute assignment to entities (second method) gives entities with attributes, with few (if any) relationships. *Incompleteness is in relationships;*
- Third method's *incompleteness is in lack of identifiers. Both methods (syntax decomposition and normalization) of data stores* could lead to non-connected conceptual data submodels (each submodel is related to one data store). *Incompleteness is in relationships among entities from different submodels.*

Conclusion: Comparation of previously presented diagrams shows that that only integrated model (i.e. model created by the proposed integration methodology) has all required CDM elements (entities, relationships and attributes) and could be considered a complete and correct model.

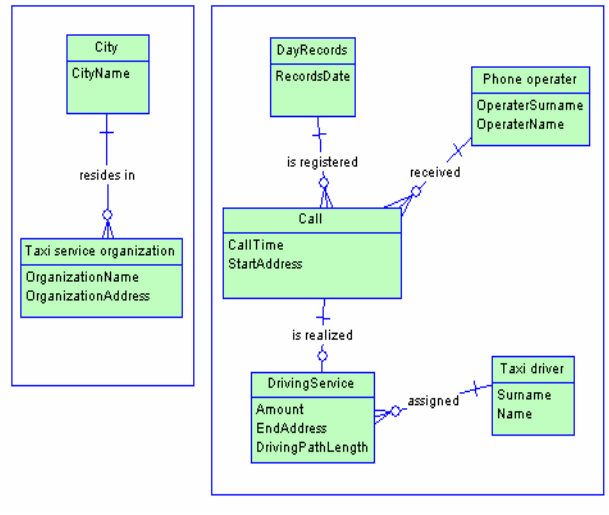


Figure 3. Conceptual data model as a result of the third method

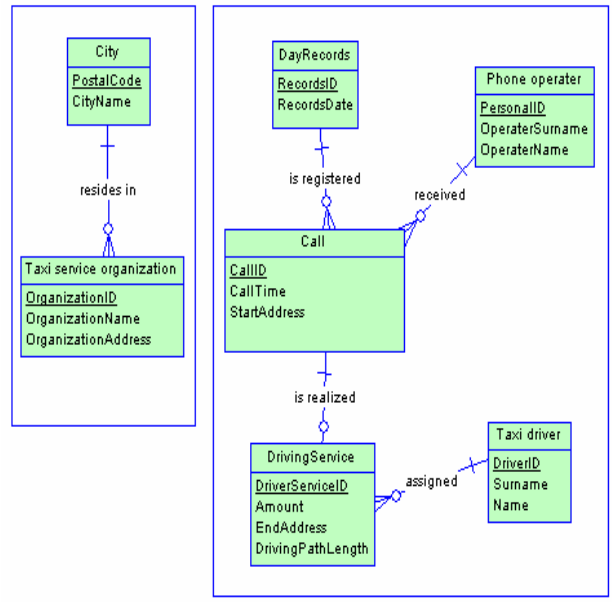


Figure 4. Conceptual data model as a result of the fourth method

VI. ISSUES IN EDUCATIONAL EXPERIENCES

Experiences of from practical classes and exams in conceptual data modeling at TFZR within Information systems course address some issues:

ISSUE 1: In the set of proposed methods, which order of their application to accept and which modality of integration to apply („central“ model or „integrated“ model approach). EMPIRICAL RESULTS (ISSUE 1): A questionaire was given to 56 students at TFZR in 2013/2014 school year regarding attitudes related to preferences on type of integration of methods.

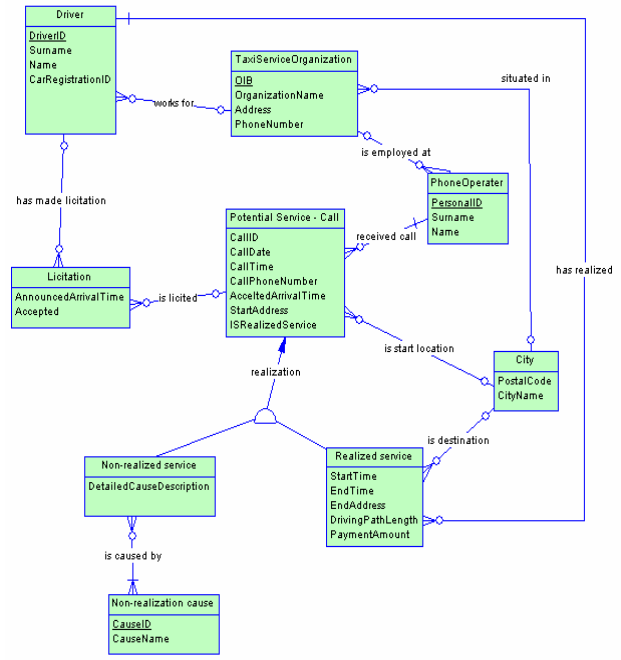


Figure 5. Integrated conceptual data model

In research population of 30 students, 92% of them choose „central“ model to be applied within practical exams, while „integrated“ approach they choose to be used within practical lessons, as a more appropriate approach.

ISSUE 2: Students are used to the “pattern thinking” from the Databases course and should adapt to the Information systems course requirements where more material and methods are given, the elements should be combined, and critical, analytical and synthetical skills should be performed. Students are faced to „similar-to-real-world“ situations in aim to be ready for professional environment. ISSUE 3: An issue for organizing practical exams arises from the problem of insufficient time for creating models of complex systems (real-world or similar to real – world examples) within limited time for practical laboratory exams (3-4 hours). PROPOSED SOLUTION: Students could choose practical exams (with simple examples) or projects (with more complex material).

EMPIRICAL RESULTS (ISSUE 2 and ISSUE 3): In 2013/2014 at TFZR, 20% of students at Information systems course chose to work on project assignments with real business problems. Other students chose to have practical exam with small examples. In past ten years, the trend at TFZR shows rise of students interest for projects [28].

ISSUE 4: Since the insufficient time for practical exams on conceptual modeling, for all methods to be applied, i.e. integrated methodology application, it is needed to focus on smaller set of methods to be used within practical laboratory exams. PROPOSED SOLUTION: Since data stores are concepts that represent data to be stored, the most appropriate method (if we have to choose among all methods) is data stores normalization or data stores decomposition according to the syntax

structure. EMPIRICAL RESULTS (Issue 4): During 2013/2014 school year, the application of the presented integrated methodology is introduced within practical exams at TFZR. There were two groups of students at examination. First group (15 students) applied integrated methodology, while second group of students (15 students) applied decomposition / normalization of datastores. Students from the first group had more complex models, but with more syntax errors. Students from second group created less complex models with less syntax errors.

VII. CONCLUSIONS AND FUTURE WORK

This paper presents new methodology of integration of conceptual data modeling approaches and methods within information system development. Educational experiences are presented within issues and empirical results from practical lessons and exams.

Future work will include further empirical research on applicability of the integrated methodology in higher education environment, research on new methods for conceptual data model quality assessment, pragmatics aspect and optimization, research on application of the proposed methodology within different educational environments, as well as possible using of the proposed methodology within professional environments.

ACKNOWLEDGEMENT

Paper presents results of project TR32044 – „The development of software tools for business process analysis and improvement“ financed by Ministry of science and technology development, Republic of Serbia.

REFERENCES

[1] D.E. Avison, and G. Fitzgerald, “Information Systems Development: Methodologies”, Techniques and Tools, McGraw Hill, 2003.

[2] B. Lazarević, V. Jovanović, and M. Vučković, “Projecting information systems”, (in Serbian: Projektovanje informacionih sistema), Naučna knjiga, Beograd, 1988.

[3] R. Elmasri, and S.B Navathe, “Fundamentals of Database Systems”, 5th edition, Pearson International Edition, 2007.

[4] A. Vasconcelos, P. Soursa, and J. Tribolet, “Information system Architecture Metrics: an Enterprise Engineering Evaluation Approach”, The Electronic Journal of Information Systems Evaluation, vol. 10, no. 1, pp. 91-122, ISSN 1566-6379, www.ejise.com.

[5] Lj. Kazi, B. Radulović, D. Radosav, M. Bhatt, N. Grmuša, and N. Štiklica, “Business Process Model and Elements Of Software Design: The Mapping Approach”, International conference on Applied Internet and Information Technologies ICAIIT2012, Zrenjanin, Serbia, ISBN 978-86-7672-166-5, pp. 17-20.

[6] LJ. Kazi, Z. Kazi, B. Radulović, and O. Stanciu, “Integration Of Conceptual Data Modelling Methods in Information System Development”, International Symposium Engineering Management And Competitiveness EMC 2011, Zrenjanin, Serbia, 2011, ISBN – 978-8687672-135-1, pp. 451-456.

[7] P. Chen, “The entity-relationship model - Toward a unified view of data, ACM Transactions on Database Systems, vol. 1, no. 1, March 1976, pp. 9-36.

[8] E. F. Codd, “A Relational Model of Data For Large Shared Data Banks”, Communications of the ACM, Volume 16, Number 6, June 1970.

[9] G. Booch, J. Rumbaugh, and I. Jacobson, “The Unified Modeling Language User Guide”, Addison-Wesley, 2000.

[10] J. Ullman, H. Garcia – Molina, and J. Widom, “Database Systems: The Complete Book”, Department of Computer Science, Stanford University, Prentice Hall, New Jersey, 2002.

[11] V. Sugumaran, and V.C. Storey, “The Role of Domain Ontologies in Database Design: An Ontology Management and Conceptual Design Environment”, ACM Transactions on Database Systems, vol. 31, no. 3, September 2006.

[12] H. El-Ghalayini, M. Odeh, R. McClatchey, and T. Solomonides, “Reverse Engineering Ontology to Conceptual Data Models”, IASTED International Conference on Databases and Applications DBA 2005.

[13] E. Vysniauskas, and L. Nemuraite, “Transforming Ontology Representation from OWL to Relational Database”, Information Technology and Control, vol. 35, no. 3A, ISSN 1392 – 124X, 2006.

[14] Lj. Kazi, Z. Kazi, B. Radulovic, and O. Stanciu, “Evaluation of Students’ Work on Data Modeling -Teaching Improvement Implications”, Information technologies and development of education journal, vol. 1, no. 1, Zrenjanin, Serbia, 2011.

[15] Lj. Kazi, Z. Kazi, B. Radulović, and D. Letic, ”Using Automated Reasoning System for Data Model Evaluation”, 8th International Symposium on Intelligent Systems and Informatics SISY 2010, September 10-11 2010, Subotica, Serbia.

[16] Z. Kazi, and B. Radulovic, “Software Tool for Automated Analysis of Conceptual Data Model”, 34th International Symposium MIPRO 2011, May 23-27 2011, Opatija, Croatia.

[17] Databases course at Faculty of Electrical Engineering in Zagreb, Croatia, http://www.fer.unizg.hr/fer2/racunarstvo

[18] Databases course at Faculty of Electrical Engineering in Sarajevo, Bosnia and Herzegovina, http://c2.etf.unsa.ba/course/category.php?id=27

[19] Databases course at Faculty of Electrical Engineering in Podgorica, Montenegro, http://www.etf.ucg.ac.me//DOCS/Nastavniplan_IIgodina_SPR.doc

[20] Databases course at Faculty og Electrical Engineering in Skopje, Macedonia, http://www.feit.ukim.edu.mk/fakultet/studii/i-tsiklus-dodiplomski/studiski-proghrami-2004-2011/informatika-i-kompjutiersko-inzhinierstvo

[21] Databases course at Faculty of Electrical Engineering in Belgrade, Serbia, http://www.etf.bg.ac.rs/

[22] P. Mogin, I. Lukovic, and M. Govedarica, “Principles of database projecting” (In Serbian), Univerzitet u Novom Sadu, Fakultet tehničkih nauka, Novi Sad, Srbija, 2000.

[23] B. Lazarevic, Z. Marjanovic, N. Anicic, and S. Babarogic, “Databases” (In Serbian), Univerzitet u Beogradu, Fakultet organizacionih nauka, Beograd, Srbija, 2003.

[24] E. J. Naiburg, and R.A Maksimchuk, “UML for Database Design”, Addison-Wesley, 2001.

[25] B. Lazarević, S. Nešković, “Object oriented transformational development of information systems” (in Serbian), Faculty of organizational Sciences, Belgrade, Serbia, 1993.

[26] J-P Van Belle, “A Framework for the Evaluation of Business Models and its Empirical Validation”, The Electronic Journal Information Systems Evaluation 2006, vol. 9, no. 1, pp. 31-44, www.ejise.com.

[27] Lj. Kazi, Z. Kazi, and B. Radulović, “Information systems 1 and Information systems 2 – practicum for exercises”, Tchnical faculty Mihajlo Pupin, Zrenjanin, Serbia, ISBN 978-86-7672-215-0, 2013.

[28] Lj. Kazi, Z. Kazi, B. Radulović, "Data Warehouse based evaluation of students' achievements in information systems education″, International convention on Information and Communication Technology, Electronics and Microelectronics MIPRO 2012, Opatia, Croatia

Technology Use in EFL Learning

Darija Pešut,
Visoka škola za informacijske tehnologije, Zagreb, Croatia
dpesut@vsite.hr

Abstract - The connection between philology and technology is no longer considered unusual. The use of technology is being adopted in foreign language learning. In this context, the author wishes to present the good practice of using computers in teaching English as a foreign language. The article thus gives detailed examples of this use while teaching different language skills: reading, writing, speaking and listening. In order to reinforce the postulate of the benefits of using technology in foreign language learning, a survey has been conducted among EFL students at the College for Information Technologies, Zagreb. A comparison of survey results from 2010 will also be shown.

I. INTRODUCTION

Technology is being successfully integrated into the traditional foreign language curriculum. As the approaches to foreign language learning have changed significantly over time, today it is hard to define what a traditional approach actually is.

Lixian Jin and Martin Cortazzi define the traditional approach through five versions having five possible definitions shown in Table I [1].

The use of technology in the traditional approach to language learning was very limited. It usually involved the teacher playing cassettes / CDs on the player or using a graphoscope for showing pictures, while some of the teachers also used video / DVD players for showing movies and short films.

When learning any foreign language, it is of great importance to simulate an environment where learners are able to interact in the language of instruction. The language of IT is English, thus, using technology in EFL (English as a Foreign Language) learning is maybe most beneficial. This article presents ways of using technology in an EFL classroom at a higher education institution and also gives results of a new case study from 2013 in comparison to the one in 2010.

II. COMPUTER ASSISTED LANGUAGE LEARNING (CALL)

With the huge breakthroughs in the field of technology, computers are now a tool regularly used in language teaching and learning. Computer Assisted Language Learning (CALL) includes the use of Information and Communication Technology in foreign language teaching and learning [2]. Egbert defines CALL as "learners learning language in any context with, through, and around computer technologies" [3].

TABLE I. DEFINING 'TRADITIONAL' APPROACHES AS FIVE DIFFERENT VERSIONS

Conceptions of 'traditional'	Approximate timescale	Key features
Classical grammar-translation	up to 1900s	Uses of grammatical paradigms, explanations, bilingual vocabulary lists, translation exercises, use of L1 in classroom
Wider grammar-translation	1900s-2000s	Grammatical explanations, bilingual vocabulary lists, translation exercises, also oral elements and other practice exercises
Audiolingual	1940s-1960s	Emphasis on structures, oral skills, mimicry, memorization, substitution and pattern drills
Mainstream EFL	1930s-2000s	Mixed approach, includes presentation, practice, production, communicative activities
"Humanistic" or alternative	1970s-2000s	Total Physical Response, The Silent Way, Community Language Learning, Suggestopedia, and the Natural Approach

CALL began to develop in the 1960s with the first technology introduced into the classroom: the tape-recorder and film-strip projector [4]. The last 50 years of CALL can be divided into three stages: behaviourist CALL, communicative CALL and integrative CALL [5].

Stockwell [6] lists the various technologies that can be used in CALL as follows:

1. Interface design and authoring technologies.
2. Data management, access and portability.
3. Artificial intelligence and tutoring systems.
4. Automatic speech recognition technologies.
5. Natural language processing technologies.
6. Speech synthesis technologies.
7. Learner profiling systems.

8. Cognitive agent systems.
9. Mobile and communication technologies.

The listed technologies are explained in part V.

From the socio-behavioural aspect, Stroia [7] defines the advantages of using digital technology in the following respect:

1. Simultaneous addressing of several senses.

   Information can be perceived twice through two different channels (audio and visual).

2. Media offers a degree of variation and entertainment.

   This increases the motivation of the learner.

3. Virtual transposition to a natural environment.

   By using the new technology, the learner can virtually travel to the native environment of the foreign language, instead of physically travelling to a different country.

4. Communication with native speakers.

   Communication can be set up over the internet.

5. Manipulation of audio and video sequence.

   The audio and video recordings can be replayed and interrupted as many times as needed.

6. Updatability of digital materials.

   The ease of keeping materials up to date is especially notable for the online digital materials.

7. The authenticity of communication.

   The audio and video materials can have pieces of real communication.

8. Studying at learner's individual speed.

   The use of technology enables learners to study at their own pace, not having to follow the average speed of the whole group.

9. The availability of the means of learning.

   Learners can choose the time and place of study in case of a computer lesson.

10. Loss of insecurity.

    Learners do not feel shame in front of the computer when they make mistakes.

11. Interactive character of learning.

    Learners have to be fully active while doing their computer lesson and their minds cannot wander off.

12. Illustrating dynamic linguistic phenomena.

    Animations are used for the illustration of certain linguistic phenomena.

13. Immediate feedback.

    Computer provides immediate feedback for certain exercises, while for some there is still need for a human agent.

Using computers in language teaching and learning has prepared the basis for the development of new techniques and it has given teachers and learners a whole new perspective. The next part of the article gives real examples of using computers in an EFL classroom in the context of higher education and will support the advantages listed in this section.

III. GOOD PRACTICE IN EFL

The use of computers and technology in an EFL classroom at the College for Information Technologies in Zagreb will be explained in respect to four different language skills: reading, writing, listening and speaking. Different techniques are used in order to master these language skills, but in this article only the ones which use a certain technological aid will be explained. In this respect, different technologies (software and hardware) can be used. At this college, students use their own portable computers and have an uninterrupted access to the network.

A. Reading

The biggest breakthrough in mastering reading with respect to technology is definitely the invention of the internet. Updatability of information on the internet cannot even be compared to information in print. This is the reason why learners of language today can study and practice reading texts that deal with the most modern and up-to-date topics, as opposed to texts in coursebooks that were published a couple of months or years ago.

Using technology for mastering reading is used in two different ways according to the person who chooses the text. Online dictionaries are used in both of these situations.

The teacher chooses a text online, gives the URL address to the students and then they open it on their portable computers. The students are instructed to read the text and use an online dictionary in case of unfamiliar words. Discussion with the teacher, involving the whole class, follows after the reading. The teacher tries to elicit new words and writes them on the whiteboard. Comprehension questions are asked in the discussion.

In case of students being asked to choose a text online, they are assigned a certain subject area by the teacher. The students naturally choose different texts. The students are asked to read the most interesting paragraph to the class and to add a couple of sentences that best describe the topic of the article. Other students are incited to ask questions. In this discussion, which follows the reading, speaking skills are also mastered. One other example of combining language skills is when students are given a script to read before watching a video, when of course, reading and listening are combined.

The discussed techniques support the following advantages of Stroia [7], listed in part II of the article:

- Simultaneous addressing of several senses.

- A degree of variation and entertainment.
- Updatability of digital materials.
- Studying at individual speed.
- Availability of the means of learning.

B. Writing

Most of the writing is practiced by using the CD-ROM of the coursebook which offers interactive exercises. Students use the CD individually on their computers. The writing practice is divided into four areas: descriptions, academic, correspondence and articles. The exercises include cloze, matching and correct order. The CD offers the correction of answers.

Another way of practicing writing is after watching short videos, when the students are asked to summarize the main ideas of the video in a few sentences. The teacher checks the assignments after class.

In recent years, correspondence has mostly been carried out via e-mails. That is why practicing this kind of writing (formal or informal) is very important. Online exercises which provide practice in this area are used in the classroom.

The discussed techniques support the following advantages of Stroia [7], listed in the previous part of the article:

- Simultaneous addressing of several senses.
- Loss of insecurity.
- Interactive character of learning.
- Immediate feedback.

C. Listening

Listening to audioscripts from the coursebook was/is usually done by playing tapes/CDs on a portable player. The listening in this EFL classroom is done similarly, only the medium used is a portable or desktop computer. Playing audioscripts on computers enables better manipulation of the audio recording.

In addition, the internet offers a great selection of numerous audio clips, so the teacher can choose from presentations, videos, lectures, interviews, etc. which have not been artificially recorded as the ones on the course CD.

The mentioned techniques support the following advantages of Stroia [7], listed in the previous part of the article:

- Simultaneous addressing of several senses.
- A degree of variation and entertainment.
- Virtual transposition to a natural environment.
- Manipulation of audio and video sequence.
- Updatability of digital materials.
- Authenticity of communication.

D. Speaking

Within this language skill computers are used indirectly. The reason being that technologies like speech synthesis and artificial intelligence have not, yet, been used in this classroom. Students learn and practice their language skills using Microsoft Power Point software in order to prepare a presentation at home and then present it to the class.

During the lesson, students are asked to visit certain web pages and then report back to the class in smaller groups. This activity combines reading and speaking, while discussions after videos combine listening and speaking. Students are also encouraged to join webinars of their own interest, where they can practice listening and speaking in a native-speaker environment.

The mentioned techniques support the following advantages of Stroia [7], listed in the previous part of the article:

- Virtual transposition to a natural environment.
- Communication with native speakers.
- Authenticity of communication.
- Availability of the means of learning.

All the materials, URL addresses, etc. can be found on the official forum of the course, so that students can access them from home, especially in the case of being absent.

IV. SURVEY RESULTS

In order to reinforce the postulate of the benefits of using technology in an EFL classroom, a survey has been conducted among students of the College for Information Technologies in Zagreb. The survey conducted in 2010 at the same college showed that the students see considerable advantages in using technologically enhanced teaching and learning methods [8].

The same survey was given in 2013, where the number of valid respondents was 94, of which 8 were female. The questionnaire was composed of 10 statements. The respondents were asked to circle numbers from 1 to 5: 1 - I completely disagree, 2 - I mostly disagree, 3 - I cannot decide, 4 - I mostly agree, 5 - I completely agree, to show the degree of agreement with the following statements:

1. I am completely satisfied with my current English language textbook.
2. Extra materials are very important to me as an English learner.
3. Watching short movie clips in class helps me study the English language.
4. I sometimes need explanations when I am studying by myself.
5. I would like to have a book which would help me study by offering explanations even when I am not in class.
6. I miss having a dictionary as part of my English textbook.

7. I would like to have the possibility to make my own glossary and have it in a digital format.
8. Browsing the internet and using the online materials helps me in the study of the English language.
9. I enhance my knowledge of the English language by browsing through the web pages.
10. I am familiar with the term 'electronic book'.

Statement 2 was included in the questionnaire in order to show how students feel about using extra materials in general. 35.1 % of the students mostly agree and 30.9 % of the students completely agree with this statement, and only less than 10% students completely or mostly disagree (Fig.1).

Statement 3 shows the students' interest in audio-visual materials. 51.1 % of the students completely agree and 26.6% mostly agree that watching videos helps them in their study of the English language (Fig.2). The survey from 2010 gave similar results where 58% of the students completely agreed and 18 % of the students mostly agreed.

Statements 8 and 9 show the importance of the internet and various online materials in EFL teaching and learning. 57.4% and 61.7% of the students find the Web as a completely useful tool in English foreign language learning (Fig. 3 and 4). The percentages in the earlier study were quite similar, being respectively 53% and 68%.

The agreement of students to other statements in the questionnaire has not been individually analysed because of its lower significance to the reinforcement of the given postulate. The overall results are given in Figure 1.

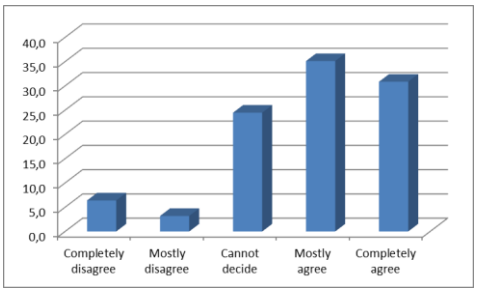


Figure 1. Statement 2

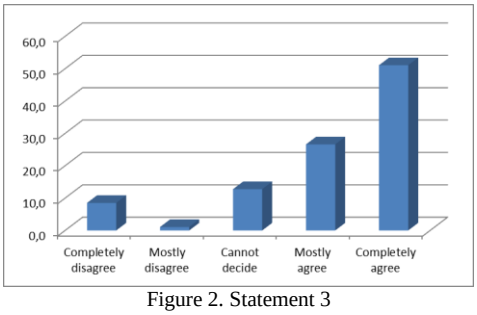


Figure 2. Statement 3

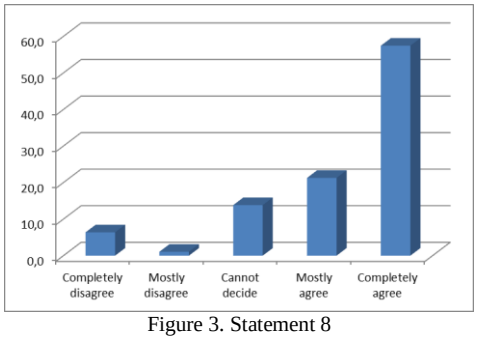


Figure 3. Statement 8

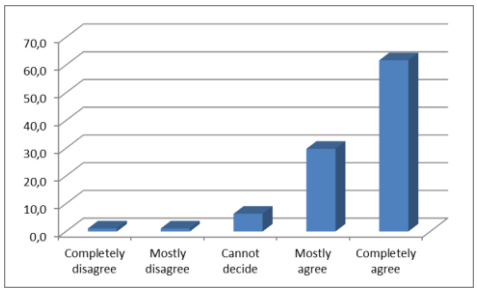


Figure 4. Statement 9

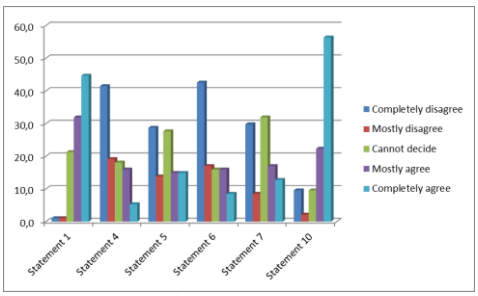


Figure 5. Results for Statements 1, 4, 5, 6, 7 and 10

V. TECHNOLOGY IN CALL

The explanation of the technology used in CALL which was listed in Part II is given in the following paragraphs.

1. Interface design and authoring technologies.

Authoring tools enable you to create multimedia software with less technical knowledge to master than with programming tools [9]. Teachers can thus use the right interface design and authoring software in computer assisted language learning.

2. Data management, access and portability.

The technologies used for data management refer to organizing, storing, securing and retrieving information. In language learning, this particularly refers to numerous online materials, the ease of access and the portability of information on different devices.

3. Artificial intelligence and tutoring systems.

The use of these technologies generates a new approach to language learning called the Intelligent Computer-Assisted Language Learning (ICALL). This

approach combines natural language processing and artificial intelligence in the form of intelligent tutoring systems [10].

4. Automatic speech recognition technologies.

These technologies involve recognizing speech and providing feedback. Special attention has to be paid to carefully designing learning activities so that the software would have a good recognition performance [11].

5. Natural language processing technologies.

NLP technologies are used in order to assess learner input and generate feedback according to grammatical and syntactical rules [12].

6. Speech synthesis technologies.

These technologies produce artificially synthetized human speech. They can be used in foreign language learning for transforming written text to speech. One example of speech synthesis can be found in online dictionaries where entries have an audio pronunciation.

7. Learner profiling systems.

There is a possibility of creating e-learning systems that are based on different learner profiles. With this approach, students are provided with learning materials which best suit their needs [13].

8. Cognitive agent systems.

An example of this technology is found in a study of a virtual agent used as a vocabulary trainer where via video stimuli learners engage in gesture-supported vocabulary training [14].

9. Mobile and communication technologies.

The rapidly changing world of ICT has introduced mobile and communication technologies which have brought about the world of social networking and microblogging. These kinds of technologies enable learners to stay in contact with the foreign language at all times, whether it is through receiving or producing information.

VI. CONCLUSION

It is unquestionable that technology enhancements bring about certain advantages. The same is true if you include technology in foreign language learning.

This article gives practical examples of using technology in an EFL classroom where numerous advantages of this kind of approach were listed. Using computers in the classroom adds a new dimension to the formerly used traditional approaches. Students are more exposed to the native language and use up-to-date materials which give them a feeling of actually being in the native environment of the foreign language.

The language of IT is English, which is why learning English as a foreign language in a computer-assisted environment has great benefits. At the College for Information Technologies students are aware of what technology has to offer in terms of foreign language learning. Other than help in the acquisition of language skills, technology can also enhance learner motivation.

In respect to the survey results and the methods described as good practice, FL teachers are encouraged to include computers in their everyday teaching methods, to blend the traditional and the electronic. However, teachers are not encouraged to use technology to make up for any lack of preparation, because they have to "put the same kind of thought into using technology in the service of the curriculum that they regularly do" [15].

REFERENCES

[1] L. Jin, M. Cortazzi, "Re-evaluating traditional approaches to second language teaching and learning", Handbook of Research in Second Language Teaching and Learning, Vol. 2, Ed. Eli Hinkel, Taylor & Francis, pp. 558-575, 2011.

[2] M. Azizinezhad, M. Hashemi, "A look at the status of computer assisted language learning and its applications," Procedia - Social and Behavioral Sciences 93, pp.121 – 124, 2013.

[3] J. Egbert, "CALL Essentials: Principles and Practice in CALL Classrooms," TESOL Inc.: Alexandria, 2005.

[4] A. Shafaei, "Computer Assisted Learning: A Helpful Approach in Learning English," Frontiers of Language and Teaching, Vol. 3, pp.108-115, 2012.

[5] P. Maftoon, H. Hamidi, S. N. Sarem, "The effects of CALL on Vocabulary Learning: A Case of Iranian Intermediate EFL Learners," BRAIN, Broad Research in Artificial Intelligence and Neuroscience, Vol. 3, Issue 4, Brain and Language, pp. 19-30, December 2012.

[6] G. Stockwell, "A review of technology choice for teaching language skills and areas in the CALL literature," ReCALL, 19 (2), pp. 105-120, 2007.

[7] M. Stroia, "Learning with Computers – A serious Challenge for the Didactic of Foreign Language Learning," Revista Academiei Fortelor Terestre, Social-behavioural sciences, Vol XVII, Nr.1 (65), pp. 38-42, 2012.

[8] D. Pešut, D. Živković, "E-book for IT experts learning English," MIPRO, 2011 Proceedings of the 34th International Convention, pp.1337-1342, May 2011.

[9] http://www.webopedia.com/TERM/A/authoring_tool.html

[10] T. Heift, "Intelligent Computer-Assisted Language Learning," The Encyclopedia of Applied Linguistics, November 2012.

[11] A. Neri, C. Cucchiarini, W. Strik, "Automatic speech recognition for second language learning: how and why it actually works", Proc. ICPhS, 2003. Available at http://www.scholar.google.hr

[12] M. Levy, G. Stockwell, "CALL Dimensions: Options and Issues in Computer-Assisted Language Learning," Routledge, 2013.

[13] P. Tzouveli, P. Mylonas, S. Kollias, "An intelligent e-learning system based on learner profiling and learning resources adaptation," Computers & Education, Vol. 51(1), pp. 224-238, August 2008.

[14] K. Bergmann, M. Macedonia. "A Virtual Agent as Vocabulary Trainer: Iconic Gestures Help to Improve Learners' Memory Performance," IVA 2013, LNAI 8108, pp. 139–148, 2013.

[15] R. J. Blake, "Brave New Digital Classroom," Second Edition. Georgetown University Press, 2013. Available at http://books.google.com/

Development of a computer system to support knowledge acquisition of basic logical forms using fairy tale "Alice in Wonderland"

Antonija Mihaljević Španjić*, Alen Jakupović*, Matea Tomić*
*Polytechnic of Rijeka, Business Department, Rijeka, Croatia
amspanjic@veleri.hr
alen.jakupovic@veleri.hr
mtomic@veleri.hr

Abstract - The knowledge acquisition of basic logical forms (term, statement, conclusion, and proof) is a prerequisite of the acquisition of any other knowledge and skills. Therefore, the knowledge acquisition of basic logical forms needs to be addressed in the earlier stages of education. In order to support the acquisition of knowledge and skills in the domain of logical forms, the authors of this paper propose the development of a computer based system for the presentation of thematic information, tasks and games based on the parts of "Alice in Wonderland" – fairy tale recognizable and popular due to Carroll's game with the basic forms of logical thoughts. The computer system would allow children and adults to gain the subject knowledge. A paper describes two initial phases of the aforementioned system - analysis and design. Analysis specifies in detail the basic system requirements, expressed in terms of learning outcomes. Design specifies in detail thematic information, tasks and games which will try to ensure achievement of learning outcomes specified in the analysis phase. The focus of design is in the use of a fairy tale Alice in Wonderland, and in the development of a computer based system.

I. INTRODUCTION

Logical forms indicate forms of human thought. The basic logical forms of thought are the term, statement, conclusion and proof. The term is encountered when there is a confrontation with an unknown object. In such a situation arises the need of understanding the meaning of that object and creating a clue about it. During a life and learning a person constantly acquires new concepts and expands his knowledge. In addition to directly perception of the object and acquisition of terms, it is possible to indirectly perform some conclusion on the basis of what is known about the object or event. When we want to confirm or refute a statement, this is made by use of proof, i.e. by application of rules of consistent logical reasoning. From the foregoing is evident that logical forms are in constant daily use and how much they are significant. Logical forms are subject to interest of logic as a scientific discipline that investigates shape or form and the ideal rules of thought. Logic and its subject logical forms are prerequisite for any human scientific or professional activity [6], [10], [11].

The intention of the author of this paper is to develop a computer system that would support the acquisition of knowledge and skills in the domain of logical forms. It would be available to a wide range of students, and would be tailored to the different age groups of students. The basic instructional method implemented in a computer system would be problem-based learning [1], [2], [5], [7]. The traditional method of teaching is subject-based learning. Figure 1 shows the basic difference between the two mentioned learning approaches [5].

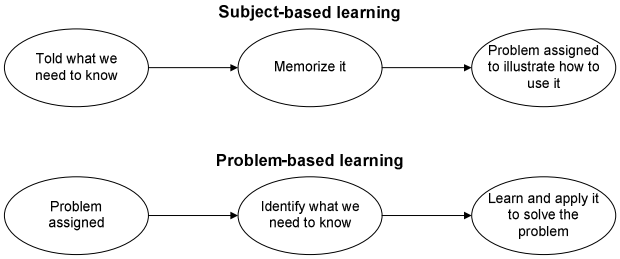


Figure 1. Difference between subject and problem based learning

In subject-based learning teacher provides content that is needed to learn. Students acquire the given content. Following the adoption, students are given problems to which they should apply the acquired knowledge.

In problem-based learning teacher first pose a problem to students. Guided by the problem, the students, together with a teacher, determine the content that is needed to resolve the problem. Then students adopt and apply identified content to a given problem. Problem-based learning is active, integrated, cumulative and connected [1].

Development of a computer system to support the acquisition of logical forms is based on the parts of "Alice in Wonderland" - fairy tale recognizable and popular due to Carroll's game with the basic forms of logical thoughts [4]. Also development of the system follows the ADDIE model, which consists of five phases: Analysis, Design, Development, Implementation and Evaluation [8], [9].

In addition to introduction, the paper consists of three chapters. In the second chapter - Analysis phase of system development to support the acquisition of logical forms - are shown the learning outcomes that students should

achieve, and for which the system will provide support. The third chapter - Design phase of system development to support the acquisition of logical forms – shows examples of the teaching content and the way they are presented to students by taking into account the problem-based learning method. The paper ends with a conclusion about the directions for further development of computer based system to support the acquisition of logical forms.

II. ANALYSIS PHASE OF SYSTEM DEVELOPMENT TO SUPPORT THE ACQUISITION OF LOGICAL FORMS

According to the ADDIE model of system development, analysis phase should determine the components that are needed in other phases of the system development. Basically we can say that the analysis should answer the question of what developed system must do, or which requirements should be satisfied by its work. Especially for the development of a system which should support the learning, analysis should provide answers to the following questions [8], [9]: Who is the learner? What is the instructional goal of the lesson? What are the delivery options? What instructional technology, if any, would enhance student learning? What will the students do to determine competency? What is the timeline? What are the online pedagogical considerations?

Computer based system to support the acquisition of logical forms should be used by learners of all ages, starting from elementary school age. In the first phase of system development focus will be on the students, because the authors of the paper participate in faculty teaching.

The learning outcomes whose achievement we want to be supported by the system are shown in Table 1. Stated are the learning outcomes for each logical form (term, statement, conclusion, and proof). In their definition has been applied Bloom's taxonomy of learning domains [3].

Computer based system to support the acquisition of logical forms should be available to the wide range of

TABLE I. THE LEARNING OUTCOMS FOR LOGICAL FORMS: TERM, STATEMENT, CONCLUSION AND PROOF

		LOGICAL FORMS			
		Term	Statement	Conclusion	Proof
BLOOM'S TAXONOMY OF LEARNING DOMAING	Knowledge	To recognize a word or sequence of words that signify the term in a given text or group of objects. To define the term.	To recognize simple statement and logical operators in a complex statement. Mark them by symbols (if they are written in some natural language).	To recognize simple logical elements in given logical form and classify them in the appropriate group – statement symbols and logical operators.	To recognize the logical forms of statements and conclusions in a given text and classify them in the appropriate group – the conclusion premises.
	Comprehension	To extract / identify higher and lower terms from the pair of terms or from a given text.	To appoint type of given statement and logical operators.	To identify the type of logical form - hypothetical syllogism, modus ponendo ponens, modus tolendo tollens, disjunctive syllogism, modus ponendo tollens, modus tolendo ponens.	To identify basic rules of conclusion in a given text using which one can prove a given statement starting from the premises.
	Application	To show given group of terms according to the range size.	To show simple statements and logical operators from which can be built complex logical statements.	To find logical statements and logical operators in a given text.	To mark by symbols all the logical forms – statements, conclusions and rules of conclusion by which it can be proved given statement starting from the premises.
	Analysis	To extract all terms from the given group of terms or text which are in certain logical relationship.	To extract all the simple logical statements and logical operators from the given set and by which one can build a certain type of complex statement. To extract all statements from the group of the statements or default text that are in some relation to logical square.	To extract all logical statements and logical operators from a set of logical statements, using which one can build a given form of conclusion.	To group logical forms of statements and conclusions, and the rules of equivalence that will be used in a proof.
	Synthesis	To find terms that are in corresponding relations with the given term and to build a table or devise a similar construction.	From the set of simple logical statements and logical operators to build a logical form of conclusion in a natural language. From the set of statements to select those statements that are in some relation to logical square.	To compile a logical form of conclusion by sequencing logical statements and link them by logical operators.	To build a proof using grouped logical forms, and basic rules for proving a given statement starting from premises.
	Evaluation	To reassess / show the relationship between given terms in Venn diagrams.	To check the validity of statement using a method of *reductio ad absurdum* and a boolean table.	To check the validity of conclusion using a method of *reductio ad absurdum* and a boolean table.	To build a new proof for the same statement based on the same premises, applying different rules and compare the existing proof with a newly built.

learners and that in the online and offline work mode. Availability should increase in the way that system can be used by desktop and laptop computers, but also by tablet computers and smart phones. It is assumed the minimum computer literacy of learners.

The system should support the learner-content interaction, i.e. without the instructor. This interaction will be supported with the instructional method of problem-based learning. To learners should be set a problem which they are trying to solve. In case they do not have enough knowledge for the solution of the problem, learners should be given the content that will allow them to adopt the necessary knowledge.

Through the established system learners should have the possibility to check the acquired knowledge. This should be realized in a way that they are given the problem which they are trying to solve – similar to the way in which the problem is set before acquiring knowledge. The difference is that now they do not have available content with the knowledge necessary to solve the problem.

The interaction of system and learners should be accomplished in several ways (for example by entering data, marking, positioning, choice, drag and drop, etc.).

III. DESIGN PHASE OF SYSTEM DEVELOPMENT TO SUPPORT THE ACQUISITION OF LOGICAL FORMS

Results of the analysis (the first phase of ADDIE model) enter the second phase of system development – the design phase. This phase needs to answer the question how the developed system would satisfy the requirements which are identified in the analysis phase. Especially for the learning supporting systems, design should provide answers to questions [8], [9]: which assignments student needs to perform in order to attain certain learning outcome specified in the analysis phase, which is targeted performance objectives of the student (what the student will do, under which conditions, and which the result is acceptable) and the manner in which will be tested the achievement of certain learning outcomes.

Not diminishing importance of responses to other questions, the authors of paper in this section provide an answer to the first question - description of the tasks that student needs to perform in order to reach the learning outcomes. This is to emphasize that system design wants to support problem-based learning. The logical form conclusion and its learning outcomes that fall in all learning domains (see Table 1) will be used as example (it includes all other logical forms). Descriptions have the following structure:

1. learning domain – learning outcome
2. description of the task which student needs to perform in order to achieve these learning outcome and its correct solution - the problem description
3. content that will be available to the student to solve the assigned task (problem) and reach a defined learning outcome.

In tasks for learning domain application, analysis, synthesis and evaluation, concrete examples of fairy tale "Alice in Wonderland" are used. In order to clearly present the learning outcomes, the corresponding description of the task, its solution and related content, the table view will be used. In analogy it can be created description also for other logical forms - term, statement and proof.

TABLE II. DESIGN FOR LOGICAL FORM "CONCLUSION" – LEARNING DOMAIN "KNOWLEDGE"

Learning domain: learning outcome
Knowledge: To recognize simple logical elements in given logical form and classify them in the appropriate group – statement symbols and logical operators.
Task – solution
From displayed logical form should be recognized logical elements and classify them in the appropriate group. (X ∧ Y) → ¬Z Statement symbols: ¬¬Z Logical operators: ¬(X ∧ Y) **Solution:** Statement symbols: X, Y, Z Logical operators: ∧, →, ¬
Content
Logical forms consist from statement symbols, and may contain logical operators. Statement symbols are used in the labelling of simple statements, and are used capital letters (e.g. A, B. .., X, Y, Z...). Logical operators are used for composing simple statements into more complex statements or for their negating. Logical operators are: conjunction ∧, disjunction ∨, implication →, equivalence ↔, and negation ¬.

TABLE III. DESIGN FOR LOGICAL FORM "CONCLUSION" – LEARNING DOMAIN "COMPREHENSION"

Learning domain: learning outcome
Comprehension: To identify the type of logical form - hypothetical syllogism, modus ponendo ponens, modus tolendo tollens, disjunctive syllogism, modus ponendo tollens, modus tolendo ponens.
Task – solution
Which type is the given logical form. (X ∧ Y) → ¬Z ¬¬Z ¬(X ∧ Y) **Solution:** Given logical form is type of modus tolendo tollens.
Content
There are six basic forms of conclusion that can be recognized according to the forms of premise and a conclusion.

Hypothetical syllogism	Modus ponendo ponens	Modus tolendo tollens
P → Q Q → R P → R	P → Q P Q	P → Q ¬Q ¬P
Disjunctive syllogism	**Modus ponendo tollens**	**Modus tolendo ponens**
P ∨ Q ili P ∨ Q ¬Q ¬P P Q	¬(P ∧ Q) P ¬Q	P ∨ Q ¬P Q

TABLE IV. DESIGN FOR LOGICAL FORM "CONCLUSION" – LEARNING DOMAIN "APPLICATION"

Learning domain: learning outcome
Application: To find logical statements and logical operators in a given text.
Task – solution
In displayed text it is necessary to find logical statements and logical operators. „Oh, you can't help that“, said Cat: „we're all mad here. I'm mad. You're mad.“ Alice didn't think that proved it at all; however, she went on: „And how do you know that you're mad?“ „To begin with “, said the Cat, „a dog's not mad. You grant that?“ „I suppose so “, said Alice. „Well, then“, the Cat went on, „You see a dog growls when it's angry, and wags its tail when it's pleased. Now I growl when I'm pleased, and wag my tail when I'm angry. Therefore I'm mad.“ „I call it purring, not growling“, said Alice. „Call it what you like“, said the Cat. **Solution:** Logical statements: A dog’s not mad.; A dog growls when it’s angry, and wags its tail when it’s pleased.; We're all mad here.; I'm mad.; You're mad.; I growl when I'm pleased, and wag my tail when I'm angry. Logical operators: therefore, then, and, not
Content
In logic the statement is a set of terms by which something is claimed or denied. Statement can be simple or composed. Statement „*A girl named Alice is the main character in a fairy tale written by Lewis Carroll - Alice in Wonderland. “* is an example of simple statement. Statement „ *A girl named Alice is the main character in a fairy tale written by Lewis Carroll - Alice in Wonderland,, and next to her in a fairy tale appear other interesting characters.“* is an example of composed statement. Composed statement consists of two simple statements related by some link that represents an appropriate logical operator. Displayed composed statement consist of two simple statements (*A girl named Alice is the main character in a fairy tale written by Lewis Carroll - Alice in Wonderland.* and *Next to Alice in a fairy tale appear other interesting characters.*) that are connected with „and“ which is a link for logical operator of conjunction. Linkers and appropriate logical operators are: 1. negation – linkers: no, nor... 2. conjunction. – linkers: and, but, than the, already, although, though, however... 3. disjunction – linkers: or, either... 4. implication – linkers: if … than, only if... 5. equivalence – linker: if and only if.

TABLE V. DESIGN FOR LOGICAL FORM "CONCLUSION" – LEARNING DOMAIN "ANALYSIS"

Learning domain: learning outcome
Analysis: To extract all logical statements and logical operators from a set of logical statements, using which one can build a given form of conclusion.
Task – solution
From displayed set of logical statements it is needed to extract those logical statements and logical operators by which one can build conclusion in form of modus tolendo tollens. A dog’s not mad. A dog growls when it’s angry, and wags its tail when it’s pleased. We're all mad here. I'm mad. You're mad. I growl when I'm pleased, and wag my tail when I'm angry. **Solution:** Logical statements: Dog's mad.; Dog growls when it's angry.; Wags its tail when it's pleased. Logical operators: Not = ¬; (Only if) Then = →; And = Λ
Content
It can be repeated content that is given in Tables 3 and 4.

TABLE VI. DESIGN FOR LOGICAL FORM "CONCLUSION" – LEARNING DOMAIN "SYNTHESIS"

Learning domain: learning outcome
Synthesis: To compile a logical form of conclusion by sequencing logical statements and link them by logical operators.
Task – solution
From given logical statements and logical operators should be compiled logical form of conclusion. Logical statements: Dog's mad.; Dog growls when it's angry.; Wags its tail when it's pleased. Logical operators: Not = ¬; (Only if) Then = →; And = Λ **Solution:** Dog’s not mad (only if) then (you see) a dog growls when it’s angry, and wags its tail when it’s pleased. (Is such that) a dog growls when it’s angry, and wags its tail when it’s pleased. Dog’s not mad.
Content
It can be repeated content that is given in Table 3.

TABLE VII. DESIGN FOR LOGICAL FORM "CONCLUSION" – LEARNING DOMAIN "EVALUATION"

Learning domain: learning outcome
Evaluation: To check the validity of conclusion using a method of *reductio ad absurdum* and a boolean table.
Task – solution
It is needed to check the validity of given conclusion using a method of *reductio ad absurdum* and a boolean table. (X Λ Y) → ¬Z ¬¬Z ¬(X Λ Y) **Solution:** Method of reductio ad absurdum (X Λ Y) → ¬Z ¬¬Z ¬(X Λ Y) ~~t t t t f t~~ t f t f t t t Conclusion is valid because the absurdity is in the first premise. Namely, implication is not true if the first element (implicator) is true, and the second is false. Method of boolean table X Y Z (X Λ Y) → ¬Z ¬¬Z ¬(X Λ Y) t t t t f f t f f t t t f t t t f t t t f t f t f t f f t f f f t t f t f f t t f t f t f f f t n f t t f t t f f t f t f t f f f f f f t t f t f Conclusion is valid, because there is no counterexample that would mean that from true premises follows the false conclusion.
Content
The validity of conclusion is one of the forms for preservation of truth. It can be checked by the method of *reductio ad absurdum* and by boolean tables. Both of these methods require translation of logical forms from natural language to the language of logic. Here we use a language of propositional logic. Language of logic has its own syntax and semantics. Syntax includes a list of symbols and formative rules, while semantics describes how formulas, occurred according to the rules of syntax, we can add the values. Building of boolean tables for composed logical forms starts with determining the values of simple statements. These values are recorded on the left side of table. Every statement has value of true or false. By increasing the number of simple statements increases the combination of values on the left side of table (for example, for two numbers the combination is 2 * 2 = 4, for three it is 2 * 2 * 2 = 8, or in general for n is 2^n). Values of composed logical form are set from simple elements to more complex. Method of reductio *ad absurdum* is a method in one line. Set of values go in reverse of set of values by boolean table – set of value takes place from values of main logical elements to those more simple. Both methods require knowledge of boolean tables of the main logical operations (negation, conjunction, disjunction, implication and equivalence).

IV. CONLUSION

This paper describes the first two phases (by ADDIE model) of the development of computer based system to support the acquisition of logical forms. The initial version of the analysis and part of the design is shown (for only one logical form of four of them). The system is based on problem-based learning and learner-content interaction.

The next phase in the development of the system is a further detailing of the design which should include the learning outcomes of all logical forms, definition of the ways of achieving verification of learning outcomes, and definition of levels of learning outcomes achievement which are considered acceptable.

Upon completion of the analysis and design phase follows the development phase. This phase should result in an initial version of the system with which learners will be able to be in online and offline interaction, and which will be widely available. For this purpose, the intention is to develop a system for the web environment (there is a possibility of interaction with other learners), but also as a standalone application for desktop, laptop and tablet computers, as well as smart phones.

In the implementation phase should be given an answer to the question of how the final version of the system will set up and how to make it available to learners. The intention is to set up a web version of the system on the server of the Polytechnic of Rijeka, which it already owns. On the same server will be set up a link through which the learner will be able to download the appropriate standalone application.

The evaluation phase should answer the question of how much the system successfully fulfils its purpose. To make this phase implemented, the intention is to put the system under evaluation of both, learners who have been interacted with the system, professors, as well as pedagogical experts.

REFERENCES

[1] B. B. Levin, "Introduction", in Energizing teacher education and professional development with problem-based learning, B. B. Levin, Ed. Alexandria, VA: Association for Supervision and Curriculum Development, 2001., pp. 1-7.

[2] C. E. Hmelo-Silver, "Problem-Based Learning: What and How Do Students Learn?", Educational Psychology Review, Vol. 16, No. 3, pp. 235-266, September 2004.

[3] D. R. Clark, "Bloom's Taxonomy of Learning Domains", http://www.nwlink.com/~donclark/hrd/bloom.html, 29.01.2014.

[4] J. J. Lecercle, Philosophy of Nonsense: The Intuitions of Victorian Nonsense Literature, London: Routledge, 1994.

[5] L. D. Spence, "Problem Based Learning: Lead to Learn, Learn to Lead.", http://n-imagecache.aldenhosting.com/~n1studyg//pblhandbook.pdf, 29.01.2014.

[6] L. S. Cauman, First-order Logic: An Introduction, Berlin: Walter De Gruyter, 1998.,

[7] M. Savin-Baden, A Practical Guide to Problem-based Learning Online. New York: Routledge, 2007.

[8] P. Jones and R. Davis, Instructional Design Methods Integrating Instructional Technology, in Instructional Design: Concepts, Methodologies, Tools, and Applications, M. Khosrow-Pour, Ed. New York: Information Science Reference, 2011., pp. 101-113.

[9] R. M. Branch, Instructional Design: The ADDIE Approach. New York: Springer, 2009.

[10] S. Kovač, Logika, Zagreb: Hrvatska sveučilišna naklada, 2006.

[11] S. Kovač and B. Žarnić, Logička pitanja i postupci, Zagreb: Kruzak, 2008.

Designing the Programming Assignment for a University Compiler Design Course

I. Budiselić*, D. Škvorc* and S. Srbljić*
* University of Zagreb, Faculty of Electrical Engineering and Computing, Zagreb, Croatia
{ivan.budiselic, dejan.skvorc, sinisa.srbljic}@fer.hr

Abstract - This paper analyzes the key properties of a programming assignment for a university compiler design course. After an overview of popular choices used at several universities in the world, we describe the assignment given to students of the *Programming Language Translation* undergraduate course at the University of Zagreb, Faculty of Electrical Engineering and Computing, and provide the rationale behind several design decisions through an analysis of its evolution in recent years. While most compiler programming assignments are based on tools like lexer and parser generators, our students actually build slightly simplified versions of these tools themselves. We discuss our experiences with this programming assignment over the last three years and qualitative differences in the results between this assignment and the tool based assignment we had been using before.

I. INTRODUCTION

Translation of programming languages and compiler technology is one of the topics that are ubiquitous in computer science undergraduate curricula. It is usually implemented as a single course in the curriculum that covers some portion of the *Programming Languages* knowledge area (KA) and a few topics from other KAs in the *Curriculum Guidelines for Undergraduate Degree Programs in Computer Science* published jointly by the ACM and IEEE [1].

However, while most of these courses cover a similar set of topics in lecture, there is no broad consensus on the design of the programming assignment that is an integral part of almost every compiler design course. In fact, this has been the case ever since this topic was included in the curriculum [2].

The common characteristics of programming assignments in most of the compiler design courses analyzed in [1] are that students build a complete compiler, and the success rate on the project significantly impacts the students' grades. Apart from this general commonality, the organization and scope of the compiler project diverge significantly.

Compiler courses are typically thought in the second half of the undergraduate program with prerequisites covering programming, basic algorithms and data structures, computer architecture and organization and sometimes introduction to the theory of computation. Since the compiler bridges the gap between the two abstractions of high-level programming languages and machine code, this order of courses in the curriculum ensures that students are familiarized with both of these abstractions prior to taking the compiler design course. Furthermore, a large part of a compiler is based on well understood and approachable formal models, and tools for building compilers are abundant. These factors put the compilers course in a somewhat unique position in the curriculum in that students can build a complete and complex system in their programming assignment. Consequently, the compiler project is frequently the largest and most challenging programming project students are involved in during their studies. A few papers mention that students talk about their compiler project experiences in job interviews [3,4], and in our own experience, we've found that a student's performance on this project provides valuable information for writing recommendation letters.

It is the scope of the project that allows the high level of diversity that is observed in practice. Specifically, the design of a compiler project is based on five somewhat competing goals, and different preference for these goals yield significantly different designs.

First, students should ideally get some insight into all stages of the compiler, and the project should reinforce the theoretical material presented in lecture [3,5,6,7]. Understanding the compiler is beneficial to students as it is one of the most frequently used tools in the software developer's toolbox.

Second, compiler technology has applications beyond translating higher-level programming languages into machine code, and students should ideally experience that in their project so that they can recognize opportunities for applying the technology in other domains of their professional work [6,7,8].

Third, the complexity of the project provides an opportunity or even the requirement to teach students software engineering best practices in software design, testing, the use of tools such as compilers and debuggers, scheduling work and other areas [3,4,6].

The fourth goal that is often encountered is teaching students teamwork and communication as the project is nearly universally done in groups [3,4,9]. This seems particularly important as industry feedback frequently suggests that students lack key teamwork skills [9].

Finally, the fifth goal that is often not explicitly mentioned is that creating a compiler can make students better programmers. Knowing how some language feature

is implemented in a compiler allows students to better understand the programming languages they actually use. For example, if the source language supports passing arguments by value and by reference, implementing this distinction in the compiler requires and reinforces complete understanding of the underlying concept.

In this paper, we discuss the challenge of balancing these goals through our experience with the compiler project in the *Programming Language Translation* (*PLT*) course at the University of Zagreb, Faculty of Electrical Engineering and Computing (UniZg FER).

The rest of the paper is organized in the following way. In Section II, we overview some related work on compiler design courses. In Section III, we overview the compiler project at seven leading computer science universities in the US and Europe. We describe the curriculum context of our *PLT* course in Section IV. In Section V, we discuss the organization of the *PLT* course and describe the evolution of the programming assignment from its initial to its current design. We conclude the paper in Section VI.

II. Related Work

Two useful classifications of compiler design courses and projects have been discussed in the literature. Schwartzbach [10] roughly divides compiler design courses into *front end heavy* and *back end heavy*. The former category focuses on formal languages and context free parsing, while the latter focuses on code generation, resource management and optimization for a realistic processor architecture. Semantic analysis and intermediate code generation, which are typically considered a part of the front end [11], are grouped into the *middle end* of the compiler, and both course categories cover these phases to some extent.

Under this categorization, a meaningful critique of both types of courses is that lexing and parsing are supported by easy-to-use tools, and that the compiler back end is typically reused as a component when building a compiler for a new language. Specifically, the bulk of the work in creating a new compiler is done in this middle end, i.e. in semantic analysis and intermediate code generation. Schwartzbach suggests that both front end heavy and back end heavy courses are somewhat constrained to using only simple languages in compiler projects. Instead, the paper presents a *middle end heavy* course and argues for projects that are based on large, feature-rich languages.

Waite [6] proposes three strategies for organizing the compiler design course and project that are mostly in line with the goals presented in the introduction. The *software project* strategy focuses on teaching software development techniques rather than compiler algorithms. To achieve this goal, compiler building tools are employed as much as possible. It is unclear which parts of the compiler should be thought in lecture and at what level of detail, but this strategy seems to stray away from a traditional compilers course in this sense.

Quite oppositely, the *application of theory* strategy focuses on theoretical models such as formal languages and attribute grammars. The goal of a course following this strategy is to teach students both the value and limitations of theory in computer science. The suggested programming assignments in this strategy are *tools based on theory*, which seems to imply that students should only build parts of a compiler.

Finally, the *support for communicating with a computer* strategy focuses on the applicability of compiler technology to many domains in software engineering. After completing such a course, students would be able to devise domain specific languages for various applications and use the compilers toolkit to implement these languages. Again, this strategy does not include a compiler project. Instead, students practice language design and use of tools.

III. Overview of the Compiler Project at Leading Computer Science Universities

In this section, we briefly overview the compiler design courses and accompanying programming assignments from some of the leading computer science universities in the US and Europe. The specifics of each course are shown in Table I.

We've estimated the percentage of lecture time devoted to lexical and syntax analysis (shown in the third column in Table I) from the courses' syllabuses available on their web sites. The three courses with no more than 15% of lecture devoted to these topics are either middle end or back end heavy by Schwartzbach's classification. They provide only a practical introduction to lexing and parsing with the goal of enabling students to either write or modify provided specifications for existing lexer and parser generators that are used mostly like black boxes. The remaining four courses with at least a quarter of lecture time devoted to the lexer and parser additionally cover the generator algorithms themselves at the expense of more advanced optimization approaches in the back end, but should not be classified as front end heavy.

All of the courses except the Columbia University course use fixed source and target languages in the compiler project. All of these source languages are object-oriented and typically syntactically and semantically resemble the language that is used in the introductory programming course in the curriculum. Similarly, the target architectures are typically those used in other parts of the curriculum as well.

Fixing the implementation language provides several potential benefits. First, it is easier to estimate the amount of code students will need to write to complete the assignment. Second, having a fixed implementation language makes it easier to provide starter code which can then be used to adjust the difficulty of the assignment without sacrificing the overall functionality of the compiler. Third, a specific lexer and parser generator can be used for that fixed implementation language and potentially covered in lecture or recitation. In fact, all seven courses use existing generators for these early phases of the compiler.

The project is usually done in small groups or even individually. Columbia uses the largest groups of five

TABLE I. KEY PROPERTIES OF COMPILER COURSES IN SOME OF THE WORLD'S LEADING COMPUTER SCIENCE UNIVERSITIES

University	***Course name***	***Lecture front end %****	***Source language(s)***	***Target language(s)/ architecture(s)***	***Implementation language(s)***	***Tools***	***Students per group***	***Project grade impact %***
MIT	6.035 Computer Language Engineering	10	Decaf	x86-64	Java	ANTLR	3-4	60
Stanford	CS 143 Compilers	30	COOL	MIPS	C++, Java	Lexer and parser generators	1-2	50
CMU	15-411 Compiler Design	10	L1-L4	x86-64	SML, Ocaml, Haskell, Java and others	Lexer and parser generators	1-2	70
Berkeley	CS 164 Programming Languages and Compilers	30	COOL	MIPS	Java	JLex, CUP	1-2	40
Columbia	COMS W4115 Programming Languages and Translators	25	Student designed	Student choice	OCaml	ocamllex, ocamlyacc	5	40
Oxford	Compilers	15	Oberon-like	Keiko/ARM	OCaml	ocamllex, ocamlyacc	N/A	N/A
ETH Zurich	Compiler Design I	30	JavaLi	x86 or similar	Java	JLex, CUP	2	66

* Approximate values

students with an emphasis on group organization [3]. On the other hand, groups of two students are most common, and students are often encouraged to solve the assignments with pair programming instead of splitting up the work.

Finally, the project accounts for between 40 and 70 percent of the students' grade in the class. To grade the project, most courses use a combination of objective testing where individual assignments are tested for correctness against a set of test inputs and either oral quizzes or extensive written documentation.

IV. THE CURRICULUM CONTEXT OF *PLT* AT UNIZG FER

Before discussing the design of the compiler project in our *PLT* course at UniZg FER, we will describe the position of the course in the computer science curriculum as it is an important factor in choosing the goals of the course and the design of the compiler project.

The *Programming Language Translation* course is taught in the fifth semester of the undergraduate curriculum. Students enrolled in the course previously complete an introductory programming course and a course on basic algorithms and data structures, a computer architecture course and a course covering the introduction to the theory of computation.

Both the introductory programming course and the algorithms and data structures course are taught using C and a small subset of C++ (basically C with classes). In the computer architecture course, students learn FRISC and are introduced to ARM. FRISC is a RISC processor architecture developed at Unizg FER specifically for teaching processor architecture. The *Introduction to Theoretical Computer Science* (*ITCS*) course extensively covers regular and context-free languages that are of particular interest for compiler construction.

Besides these enabling factors, there is one characteristic of the curriculum which makes the design of the compiler project more challenging. While students have an opportunity to learn several other programming languages through electives and skills, there is no single language that every *PLT* student is certainly familiar with except C. The costs and benefits of C make it a poor choice for the implementation language in a compiler course.

V. *PLT* COURSE ORGANIZATION AND GOALS

About a third of lecture time in *PLT* is devoted to lexical and syntax analysis. This provides students with an opportunity to apply what they've learned about automata theory in the previous *ITCS* course to two interesting problems which illustrates the general value of these models. Furthermore, understanding the algorithms behind lexer and parser generators makes it easier to avoid ambiguities when designing a language or to eliminate them if they arise.

The second third of the course covers semantic analysis and intermediate representations. Based on their programming experience, students generally have an intuitive understanding of the key topics in semantic analysis, such as declarations and type checking. By introducing attributes to the familiar model of a grammar, this intuitive understanding can be formalized in an approachable way. Similarly, attributed syntax trees can be easily understood through their relationship with parse trees that are also covered in *ITCS* and in the first part of *PLT*.

The final third of the course is dedicated to code generation and execution support, along with a limited introduction to both machine independent and machine dependent optimization. In this part, students draw from their experience writing FRISC and ARM programs.

The primary goal of the compiler project is to reinforce these concepts introduced in lecture with practical experience. In the remainder of this section, we describe the evolution of the compiler project in the last

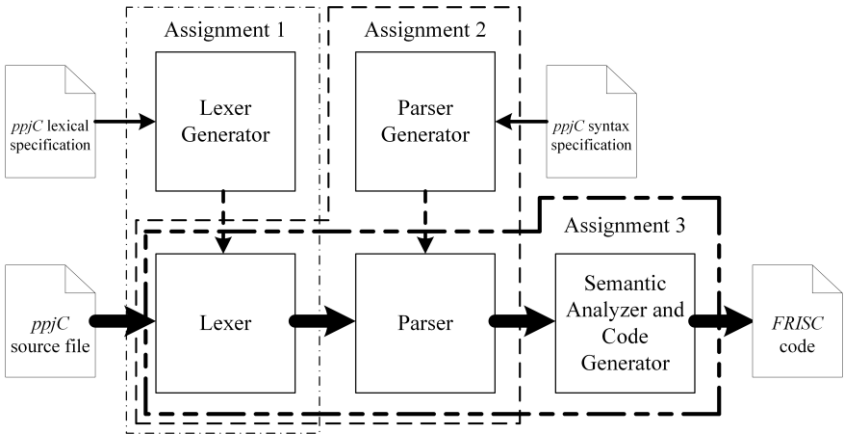


Figure 1. The design of the first iteration in the evolution of the compiler project in the *PLT* class at UniZg FER.

four years, from its initial organization to the one we are currently using.

A. Lessons Learned from an Ineffective Compiler Project Design

Up until 2010, the compiler project we used was fairly free-form, as students would design or choose their source, target, and implementation languages. This project design is intended to provide students with an opportunity for language design and familiarization with the compiler building tool chain, i.e. the second goal mentioned in the introduction is emphasized.

To prevent procrastination, we would set up two check points during the semester—one for lexical and syntax analysis and one for semantic analysis. At these check points student groups would demonstrate their current progress and group members would be quizzed on what they had done. At the end of the semester, groups would make their final code submission along with written documentation describing their source and target language, compiler architecture and key implementation ideas. Each group was also required to prepare a set of examples that showcase the capabilities of their compiler.

A teaching assistant would then read the documentation, run the provided examples, examine the compiler's source code, and come up with additional test cases. As a grading aid, we used an extensive list of possible language features, each of which was assigned some weighing factor corresponding to the difficulty of implementing it. This list included the most basic features like lexical analysis and support for variables and integral data to fairly advanced features like classes, inheritance and optimization.

After this preparation, the groups would come in for the final lab quiz focused on code generation, and the teaching assistant assigned to each group would go over the compiler's feature set with the group leader to make sure something wasn't overlooked.

The final scores were then scaled so that the most successful group would get 7.5% of the grade per member. The group leader was given the total group score and had to assign up to 7.5% to members, ideally depending only on their contribution to the project's success. The three lab quizzes contributed 2.5% of the grade each, so the project contributed a total of 15% to the grade.

While this project design allowed perhaps a few top students per generation to enjoy designing and implementing their own language, over several years of using it we noticed many significant problems.

The key problem was that at least half the groups didn't manage to finish the project. Difficulties usually arose during semantic analysis and code generation. This is understandable as lexical and syntax analysis were supported by *Lex* and *Yacc* equivalents, and appropriate definition files for many languages are available online. To get these stages working, students often simply modified existing generator input. This in turn made the start of the assignment very easy, giving students a false sense of the overall difficulty of the project and the time required to complete it.

Second, students were typically unsure about many design decisions in the early assignments as they didn't have a clear picture of what exactly they would be doing in later assignments. The most problematic point in this sense was the symbol table design which students often attempted to introduce in the lexer only to find later that their design was too limited.

Third, while students enjoyed the freedom to design their own source language, they often didn't choose an appropriate feature set. In particular, students often underestimated the complexity of some feature they wanted to include. Designing "the best" and most complex language became a goal for some groups, which often caused problems in completing the project.

Fourth, as the grading process was time consuming for teaching assistants and between 150 and 200 students take this class every year, we were forced to keep groups fairly large, usually between 10 and 12 students per group. Managing such a large group was understandably difficult for the group leaders, and this resulted in significantly different experiences for various members of the group. Usually, only the leader and a few other members would work on the compiler code, while others were assigned to creating test cases, writing the documentation or other supporting roles.

Finally, students often complained that the project is far too difficult for only 15% of the total grade. Furthermore, while we made every effort to keep project grading fair, this was a challenge as different teaching assistants graded different projects with different source

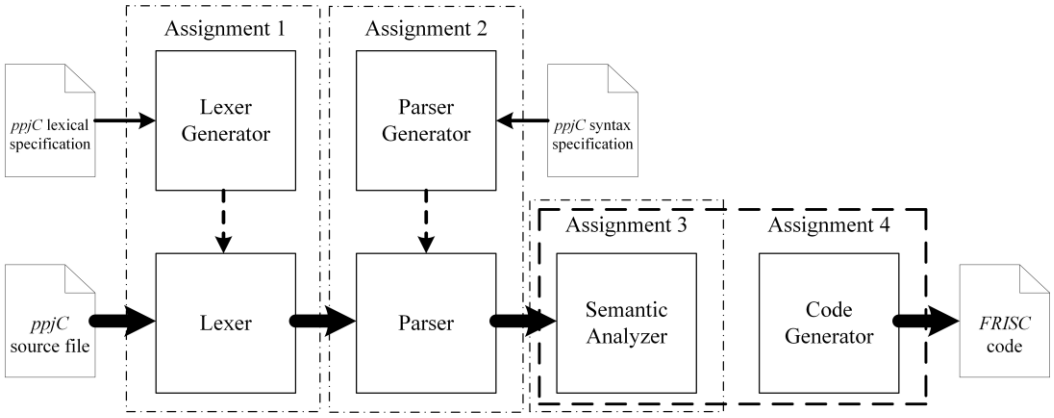


Figure 2. The design of the current compiler project in the *PLT* class at UniZg FER.

and target languages in an inherently subjective process. Specifically, it was difficult to correctly assign partial credit for features that weren't completely implemented, which was a common occurrence.

B. Evolution of the Compiler Project - the First Iteration

With the downsides of the old project in mind, in 2010 we set out to design a new project with the following ideas. First, we attributed most of the issues in the project to *underspecification*—students were given too many choices that they objectively weren't equipped to make. Most notably, we decided to specify both the source and the target language. We also included extensive advice on some key implementation details in the assignment document which was ultimately 60 pages long.

Second, we reduced the group size to nine students. While we were aware that the groups were still too large for effective organization, we had decided to make changes incrementally to avoid unexpected problems. For example, there was a concern that reducing the group size too much would result in some groups not having any strong programmers to lead the effort.

The design of this first iteration of the project is shown in Fig. 1. Unlike most other courses, we decided to have students implement their own lexer and parser generators instead of using existing ones. There are several motivating factors for this choice. First, all the algorithms required to write these generators are covered in lecture. Implementing these algorithms reinforces this knowledge. Second, the early assignments are made more challenging which is beneficial in that it doesn't portray a false picture of the complexity of the project as a whole, and students usually have more time to devote to projects early in the semester. Finally, building your own version of a tool forces you to understand it better, and hopefully teaches more than simply copying and editing existing lexical or syntax specifications.

With this in mind, the first assignment was to build a lexer generator that was a slightly simplified version of *Lex*. We provided students with lexical specifications for several example languages and for *ppjC* which was the source language we designed. We chose C as the basis for the language as that is the language that all of our students are exposed to early in the curriculum. The feature set of *ppjC* included a simplified integral and floating point type system, loops, branches, functions, single pointers, one dimensional arrays and structs. We intentionally kept the language fairly simple so that a majority of the groups would be able to complete the project.

The second assignment was to create a parser generator. We chose to use the *LR(1)*-parser as it allows the simplest and most intuitive grammars for specifying language syntax. Again, we simplified the syntax specification compared to a tool like *Yacc* so that the input is basically just the language grammar. As in the first assignment, we provided the syntax specification for several languages including *ppjC*. The students had to integrate the generated *ppjC* parser with the lexer they had generated in the first assignment.

Finally, the third assignment was to implement semantic analysis over the generated parse tree and generate *FRISC* code. We chose *FRISC* as the target processor as it is familiar to our students, which allows them to focus on the compiler itself. The semantics of *ppjC* were defined using natural language, mostly relying on the students' knowledge of C. Similar to the second assignment, students were again expected to integrate the new code with the results of the previous assignments.

As part of the third assignment instructions, we gave students a list of language features with weight factors, much like the one we had been using to grade the old project. The students were required to submit a similar list with the features they had implemented and test cases that exercise those features along with their documentation.

Overall, the changes we made felt like an improvement. While some students complained about the size of the assignment instruction text, nearly all groups managed to develop high quality solutions to the first two assignments. However, it was also clear that the third assignment needed significant additional improvement as many groups still stumbled in either semantic analysis or code generation. We identified two main reasons for this. First, compared to the first assignments, the third assignment was both underspecified and too big. Second, the fact that we required integration of all stages at every step made it difficult to recover from errors in earlier stages. For example, a small error in the parser generator made it nearly impossible to successfully finish the project.

C. The Current Design of the Compiler Project

In the next offering of the course, we made an effort to address these remaining problems. We split the third assignment into two separate assignments as shown in Fig. 2. Additionally, we completely removed the dependencies between all the assignments except the third and fourth. This means that the ability to complete an assignment is not negatively impacted by errors in earlier assignments. We achieved this by making the output of each assignment be the exact input for the next assignment. Specifically, the lexer outputs the token sequence of the input *ppjC* program which the parser transforms into a parse tree. The parse tree is then the input to both the semantic analyzer that outputs semantic errors and the code generator which is implemented as an extension to the semantic analyzer.

To further assist students in semantic analysis, we specified most of the semantics of *ppjC* in a formal way, using an attributed translation grammar, i.e. by adding attributes and computation rules to the *ppjC* grammar. The attribute values computation rules are designed for recursive descent on the parse tree, which is also covered in lecture. With this extension and additional implementation advice for the code generator, the assignment specification document grew to 80 pages, but we observed a significant improvement in the success rate of these later stages of the compiler which was our primary goal.

To address the grading problems, we introduced automatic evaluation of all assignments. Students submit their solutions to the web-based evaluator, which then runs several integration tests to check if the solution archive is properly formatted and correctly handles input and output. After the assignment deadline, all solutions are evaluated using 20-30 focused test cases that attempt to exercise various features of the particular stage of the compiler. The results of these tests contribute up to 15% of the student's grade. With the additional maximum 5% for each of the two assignment-oriented oral quizzes, we've increased the total grade impact of the project to 25%.

To reduce group management problems, we reduced the group size to six students. Additionally, to enable automatic evaluation, we constrained the implementation language choice to C, C++, C#, Python and Java. These are the languages the students are most likely to encounter in the curriculum depending on their course choices. Nearly all groups choose either Java or Python, with several groups choosing C# and C++ each year.

We also simplified *ppjC* by removing floating point numbers, pointers and structs. While this made the type system significantly simpler, it still offers a good insight into type checking and conversions.

With this design, we've achieved our primary goal of making sure most groups manage to implement at least some semantic analysis and code generation. In the last three years, the evaluation success rate was above 70% for the third assignment and above 50% for the fourth assignment, and only several groups fail to score any points on these assignments. The success rate for the first two assignments is consistently over 70%.

VI. Conlusion and Future Directions

In this paper, we've described the evolution and current design of the compiler project in our university compiler course. With a project with detailed specification of all the basic phases of the compiler, we've achieved significantly higher success rates in the later stages of the compiler than with the more free-form project we had been using in the past.

We've been getting positive feedback from students regularly, but many students still find the assignment too difficult. We are considering addressing this problem with designing two distinct assignment paths, where one would be a lower-credit version of the assignment focusing only on some parts of the compiler and language design. Additionally, we are also planning to further reduce group size which should minimize management problems, and more importantly, make the project experience similar for all students.

Acknowledgment

We would like to thank Ivan Žužak and Zvonimir Pavlić for their significant contribution to the improvement of the course and the compiler project in the last several years.

References

[1] The Joint Task Force on Computing Curricula, "Curriculum guidlines for undegraduate degree programs in computer science," ACM and the IEEE Computer Society, 2013.

[2] B. Appelbe, "Teaching compiler development," Proceedings of the 10th SIGCSE technical symposium on computer science education, vol. 11(1), pp. 23–27, 1979.

[3] A. V. Aho, "Teaching the compilers course," ACM SIGCSE Bulletin, vol. 40(4), pp. 6–8, 2008.

[4] W. G. Griswold, "Teaching software engineering in a compiler project course," Journal on educational resources in computing (JERIC), vol. 2(4), 2002.

[5] J. S. Mallozzi, "Thoughts on and tools for teaching compiler design," Journal of computer sciences in colleges, vol. 21(2), pp. 177–184, 2005.

[6] W. M. Waite, "The compiler course in today's curriculum: three strategies," Proceedings of the 37th SIGCSE technical symposium on computer science education, vol. 38(1), pp. 87–91, 2006.

[7] M. Mernik and V. Zumer, "An educational tool for teaching compiler construction," IEEE Transactions on Education, vol. 46(1), pp. 61–68, 2003.

[8] M. Ruckert, "Teaching compiler construction and language design: making the case for unusal compiler projects with PostScript as the target language," Proceedings of the 38th SIGCSE technical symposium on computer science education, vol. 39(1), pp. 435–439, 2007.

[9] W. M. Waite, M. H. Jackson, A. Diwan, and P. M. Leonardi, "Student culture vs group work in computer science," Proceedings of the 35th SIGCSE technical symposium on computer science education, vol. 36(1), pp. 12–16, 2004.

[10] M. I. Schwartzbach, "Design choices in a compiler course or how to make undergraduates love formal notation," Compiler construction: 17th international conference ETAPS 2008, pp. 1–15, 2008.

[11] A. V. Aho, M. S. Lam, R. Sethi, and J. D. Ullman, "Compilers: principles, techniques, and tools,", Pearson Education, 2006.

Micro Assessment SaaS Cloud Solution

A. Bundovski*, M. Gusev**, S. Ristov**
* AssecoSee, Skopje, Macedonia,
** University Ss Cyril and Methodius, Faculty of Computer Sciences and Engineering,
Rugjer Boshkovikj 39, Skopje, Macedonia
bundovski.aleksandar@gmail.com, marjan.gushev@finki.ukim.mk, sashko.ristov@finki.ukim.mk

Abstract - In this paper we present the architecture of a new Micro Assessment (MA) system for "micro knowledge and skills control" that allows the manager to get information and status on knowledge gained on behalf of the employees, on very short status cuts during the working hours, rather than scheduling special meetings with them in order to assess their gained knowledge. This kind of micro testing saves a lot of time, does not cause stress among employees and allows more time to be dedicated for learning.

Two studies were undertaken as a part of this paper: study on the "knowledge platform" market and study on the "need for individual and professional upgrades". Based on the findings it appears that both challenges can be answered by initiating Software as a Service solution for small and medium enterprises built upon "lifelong learning" concept enabled on a Cloud computing architecture. Tools for "lifelong learning" not only enhance individual's involvement in social issues, but they are also enabling them with competitive skills required at their workplace.

I. INTRODUCTION

The constant development of the information society is surely automating and shortening each business process in one company. Due to the recent economic crisis, most of the managers and employers are trying to reduce the costs and become more efficient, thus investing in new software solutions, which will also cut the expenses on company's payroll. Moreover, the efforts to upgrade and strengthen the whole value and supply chain are highly appreciable. To be successful means to enable an environment where the workforce should constantly be exposed to acquiring new knowledge on solutions and services that could be easily adjusted and utilized for a specific field and business within which their company operates.

This paper presents the architecture of a new product, which is defined as a micro assessment system that should be implemented as a software service. It is intended for micro assessment of skills and knowledge of employees in order to increase and improve the productivity of the company, execution and knowledge of users, and adding additional value to their professional development.

The MA system for "micro knowledge and skills control" also allows the manager to get information and status on knowledge gained on behalf of the employees, on very short status cuts during the working hours, rather than scheduling special meetings with them in order to assess their gained knowledge. This kind of micro testing saves a lot of time, does not cause stress among employees and allows more time to be dedicated for learning.

The learning environment provided by this product helps individuals to use business intelligence to make faster decisions and improve response to critical business issues, ultimately contributing to the overall performance of their organization. One of the major benefits is that this solution is based on the world trends in teaching called "lifelong learning". Lifelong learning [1] refers to a society in which learning possibilities exist for those who want to learn [2]. Learning is not restricted to the classroom and to formal learning inside learning institutions; it is an activity, which happens throughout life, at work, play and home. In the modern knowledge-intensive era, lifelong competence development has become a major challenge to our educational systems that have not changed their educational policies and pedagogical models to support life-long learning. There is an increasing demand for new approaches towards fostering lifelong learning perspectives [3]. The MA solution is just one more answer to the increasing demand for new approaches.

Our solution for small or medium enterprises (SMEs) is built upon ideas of Cloud computing and Software as a Service, and relies on the cloud scalability to ensure enough resources and enables single instance of the software servicing multiple customers, or multi-tenancy (virtually partition customers data and configuration).

The rest of the paper is organized as follows. Background and essential definitions are given in Section II. Details of the software design and implementation are presented in Section III. Section IV presents related work about existing solutions and designs and Section V discusses about features of the new architecture and implementation. Section VI is devoted to conclusions and future work.

II. BACKGROUND

This section gives a brief description of Micro teaching, learning and assessment concepts.

A. *Micro teaching*

Micro teaching firstly appeared in the 60-ies at the Stanford University, based on critical views of traditional lecturing in teacher education [5]. This teacher training technique employs real teaching situations for developing skills and helps teachers get deeper knowledge regarding the art of teaching [4]. Methods of micro teaching, which have been developed and improved in many ways throughout the last decades, are the baseline for micro learning on the level of acquirement of skills and usage of

teaching methods and also on the level of reflecting on the learning processes [5].

B. Micro learning

The basic premise of the first international conference for "Microlearning" [6] is that people can learn better, easier, more effectively in a more enjoyable manner if information is broken down into smaller units and if learning takes the shape of small steps [6]. Bruck [6] presents three aspects that characterize this shift to micro learning:

- Reduce the overflow and complexity of information and structures it into small and well linked sequences;
- Undertake this reduction from the perspective of the person who needs to cope with big amounts of new information; and
- Allow individual learners to choose themselves time, place and pace of learning.

C. Micro assessment

The next step in the learning process is the assessment. Ristov et al. [7] present four generations of e-Assessment and e-Learning systems. These systems start with monolithic assessment applications (first generation) upgraded to modular architecture solutions (second generation) evolving to a system based on SOA architecture (third generation). Consequently, the fourth generation is represented by the eco-systems based on the cloud computing.

Micro teaching and learning impose an appropriate method of assessment, including micro assessment of skills, knowledge and abilities of individuals. Therefore, the micro assessment method is covered through assessing smaller units of gained knowledge.

III. Description of the SaaS solution

In this section we describe the software design and implementation details of the MA SaaS solution for micro assessment.

A. Functional description

The main functionality of the software product is to inform and assess the employees in the companies of innovations and services the same company offers, as well as to measure their activity. In the same time the solution is to be developed as a cloud solution applying multi-tenancy and scalability architecture. In order to install this solution, the company-user needs to register in a system and download the installation package. Package installation is activated on each computer separately in order to enable the service to start running in the background.

1) Primary features:

- Background activity - Activates user interface (UI) when the user is inactive (Measurement of user inactivity starts when all of the following criteria are in place, in preset time frames: keyboard, mouse and chosen process(s) or application(s) inactivity. UI can be displayed as Question UI, displaying question with several answers, and/or Information UI, displaying information for the company.),
- Testing and Survey - There is an option to do internal testing or surveys of the entire company or in groups (by department, by location, by work position etc., with same or different tests.),
- Reporting - Measuring the activity of the user (reports) and results of testing (This feature is very attractive for managers and/or superiors because it allows them to assess the performance of their employees and their workload. For example, managers can measure employees' activity in terms of working or taking tests and learning, or even determine whether the employees are performing better in the morning or in the afternoon. In addition, different departments' performance can be measured and compared with this tool. Finally, the system will allow to review and download reports for all activities of the user, including performance, efficiency, duration, etc.)

2) Secondary features:

- Access - to learning materials through a Web browser,
- Presentation - delivery of learning materials in a specific format,
- Announcing - system is automatically activated when the user is announced on the domain of the company,
- Customization - allows semantic solution - adapt the system to learn the current knowledge of the employee,
- Manipulating with materials - adding new teaching materials, areas for study and testing, system allows changing the curricula, tracking versions of the curriculum and changes,
- Delay, prolongation of the specified time - if a user does not want to answer the question at this moment, the question should be able to delay for a specified time - can be delayed for a period of time pre-entered by the user,
- Multi Language support,
- Automatic recognition of user – by logging on the company domain, each user is automatically detected. User recognition does not depend on the device but on the user's domain credentials,
- Tenant Customization – the ability to customize some parts of the application, such as color of the UI or business rules.

B. Architecture design

Two main components of this product are presented in Fig. 1 as server (cloud) side and client side components.

The cloud side components include:

- **Micro Assessment Service (MAS) –** This is the brain of the solution. The main goal of this service is to distribute data relevant to users and to process their response, and evaluate whether the answer is true or false. This service processes all the messages and communicates with the Client side;

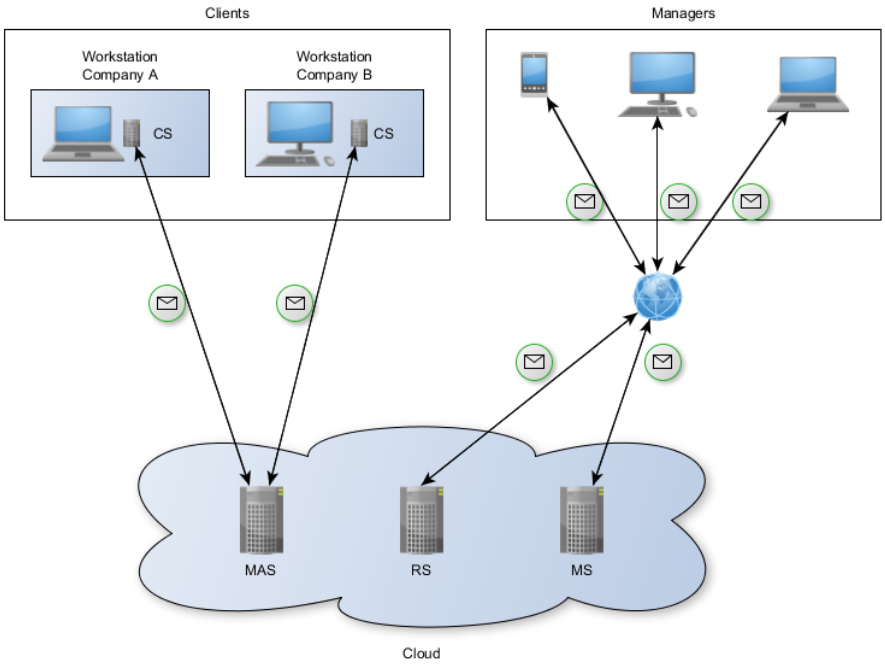


Figure 1. Architecture of the MA system realized as a cloud solution

- **Management Service (MS)** – A service used for users, courses, questions, exams and user authentication and authorization management. Most of the activities that are displayed on the clients display can be controlled by this service. The teaching material or the posted questions can be initiated and changed through this service,
- **Report Service (RS) –** The main function of this service is the preparation of reports from the answers and user activity. Access to these data is allowed only to authorized persons because of the sensitive data,
- **Database (DB) –** one database for all customers.

Client side component contains:

- **Client Service (CS)** – A service installed on every client computer, thin client installation. The main goal of this service is to monitor the activity of the user and display the data obtained from the MAS. When the user initiates transaction, this service sends a data request to MAS. After receiving the response from the MAS, data is displayed to the user. The answered question is sent back to the MAS.

There are two main user roles in MA system: managers and clients. The users in manager role manage the MA system functions: creation and administration of questions and answers in the system, testing, analysis of results, managing users and courses from his company, etc. The clients belong to a role that can take the exams and view their results.

C. Technical solution

There are several alternatives on how this technical solution can be developed to enable scalability and a multitenant pattern. The chosen technical solution is based on actions between the four major services mentioned above: CS, MAS, MS and RS.

The most utilized service of all will be the MAS service because it needs to be constantly active and available. This is due to the fact that all other client services are dependent on it, and the rest of the processes depend on the availability of this service. This service is actually the one that posts all the information and / or questions that are posted on behalf of the MS, and at the same time collects all the answers and reactions made on behalf of the client, that at a later stage are transferred to the MS and RS for further analysis, presented in Fig. 2.

On the other hand, MS and RS are secondary services that are needed periodically due to the fact that they only manage certain processes and activities. In addition, their actions are taken on request by the client (in this case management and HR team for example) and therefore the overload of these two services is not as big and as frequent as it is on the MAS service.

MS and RS are actually the enablers of the whole process to be started and finished, however the largest load of data absorption, distribution, display and collection is done by the MAS service. Consequently the MAS service will be the most overloaded and therefore needs to be scalable. In order to ensure this scalability of the system, we will install not one, but two (or more if needed) MAS services that will be enabled through the Network Load Balancing (NLB) technology. It is a clustering technology that is intended to provide high availability, reliability and scalability of mission-critical, TCP/IP-based services. These services would present a

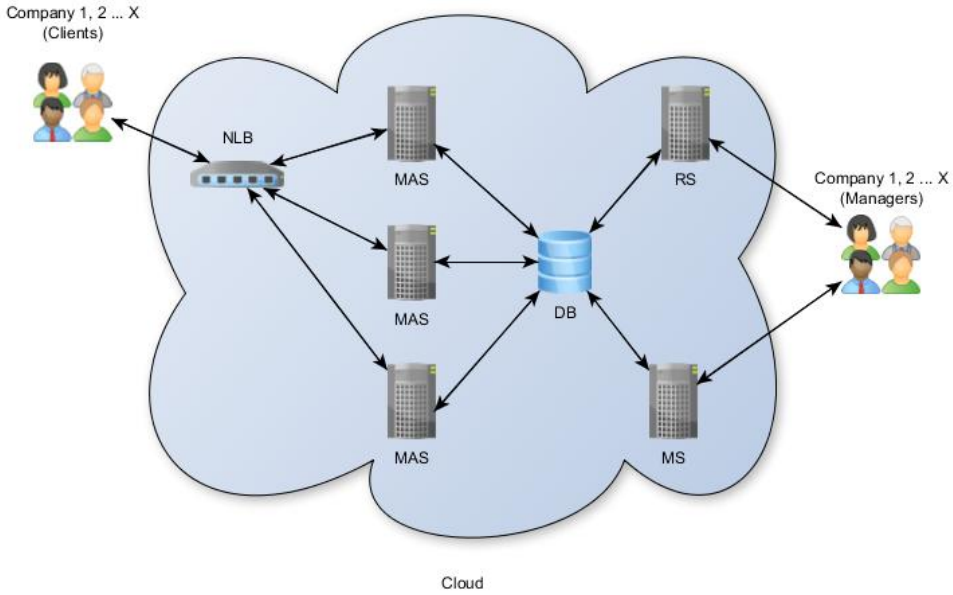


Figure 2. Micro Assessment system organization

"virtual service" address to the outside world, and when users attempt to connect, they would forward the connection to the most appropriate real service [13]. This assumes that the scalability will be guaranteed especially since the target clients of this Micro Assessment solution are SMEs, where the activity of all employees at given point of the day might be huge. As mentioned previously, our main goal is the MAS service to be always active and to be able to respond at any user inactivity.

In this paper we do not refer to security service and other related services. We concentrate on the definition of the core services required to realize the MA system.

Besides the scalability, our goal is to enable the MA solution on a multitenant architecture. In addition, given the fact that this is a cloud-based solution, we will have one joint database, one joint MS, RS and clustered MAS based on NLB technology. The proposed architecture setup will satisfy the multitenant pattern but will also cut on additional costs on different kinds of resources.

IV. Existing Systems

A. Analysis of existing solutions

According to our research of the "knowledge platform" market today, there are several systems for electronic learning, electronic testing and verification of knowledge; however those that meet the functionalities of Micro knowledge and skills assessment are SAP ® BusinessObjects ™ Knowledge Accelerator [9] and KnowledgePulse ® [11].

1) *SAP® BusinessObjects™ Knowledge Accelerator*

BusinessObjects Knowledge Accelerator is developed by SAP, a multinational consulting corporation. It is a strictly WEB-based learning, training and supporting tool, designed to help individuals' learn and apply business intelligence to their day-to-day activities [9], [10].

BusinessObjects Knowledge Accelerator provides end-users with relevant tutorials and reference materials to for self-service training and support of Business Objects products – when they want, at their desks, and at a pace that best suits them.

BusinessObjects Knowledge Accelerator helps individuals familiarize themselves with business intelligent processes, at their own pace and schedule, without having to leave the office. This helps organizations obtain maximum value from their Business Objects implementation as quickly and cost effectively as possible.

Drawbacks of the BusinessObjects solution are:

- Strictly WEB-based tool;
- Part of complex and expensive software;
- No interaction with the user after user inactivity; and
- No tests and survey.

2) *RSA KnowledgePulse®*

KnowledgePulse® is an advanced learning technology system developed by the Research Studios Austria. It consists of a client-server solution with a MicroStep Manager®, a content database, a user/course administration interface and a set of clients for PCs, notebook computers, handhelds and other devices [11], [12].

Knowledge Pulse as a software tool provides an opportunity to overcome a different material, such as foreign language, training for project management, training for product development or certificate required for promotion. The idea is that the absorption of knowledge to occur through small steps, which would avoid boring procedure of permanent and extensive learning, which is usually the result of eLearning systems.

For this solution one does not need a server infrastructure in their organization. KnowledgePulse® has a server infrastructure for an economical price. The server communicates with your PC or mobile phone via the

TABLE I. ANALYSIS OF MICRO ASSESSMENT SOLUTIONS.

Features	MA solution	Knowledge Accelerator	Knowledge Pulse®
tests and survey	✓	X	X
inexpensive	✓	X	X
no additional hardware expenses	✓	✓	X
interaction with the user after user inactivity	✓	X	✓
product of itself	✓	X	✓
displaying information	✓	X	X
displaying question	✓	X	✓
delay for the specified time	✓	X	X
measuring user activity	✓	X	X
measuring the results of testing	✓	✓	✓
access to materials via Internet	✓	✓	✓
easy to deploy	✓	✓	X
delivery of materials in a specific format	✓	✓	✓

Internet. When a computer or mobile phone is idle for a certain period, it sends small 'micro-lessons' in the form of flashcards with questions.

Drawbacks of the RSA KnowledgePulse solution are:

- No tests and survey;
- Expensive, server infrastructure need to buy from KnowledgePulse®;
- The server communicates with your PC or mobile phone via Internet;
- No delay after flashcards appear;
- On flashcards only displays questions, no displaying information for the company; and
- Algorithms and the method of calculating the data are performed on the client side. Thus, updates can be complicated, because it is necessary to deploy changes/updates to all client computers.

B. *Products comparison*

Table 1 presents the competitive features between proposed solution MA and the other two existing on the market, already analyzed and compared in [8]. MA solution satisfies all features, while the others not.

V. DISCUSSION

This section discusses advantages and benefits of the new MA solution and presents other solutions that have similar architecture as the new MA solution.

A. *Advantages of the new solution*

The implementation of a new Business Intelligence solution typically impacts a diverse range of user groups. This can make training and support overwhelming and costly. With this solution, one can affordably train and support a few hundred users at their own computers. In addition, it is extremely important provision of separate training available 24 hours a day. It is fully scalable and features extensive flexibility.

This solution speeds up user productivity, performance and knowledge, helping users add more value. This means faster decisions, better response to issues impacting critical business drivers and more consistent and accurate interpretation of data presented via reports and dashboards.

This solution will train users via job-specific skill development scenarios based on sample company data. In addition, it is customizable, hence it meets the requirements and relevancy of each user group, including varied training needs and tasks to be performed. Also it enables quick deployment of educational materials via the Internet. This means that training is delivered directly to the users' devices, giving them the flexibility they need to learn when and where it best suits them.

B. *Benefit*

There are three major benefits for the customers. This system allows measuring the activity and inactivity of employees and the managers can constantly have reports how active or inactive their employees are. Having a tool for educating your employees has never before been as practical, as this one. No more internal trainings will be needed, no more long staff meetings will be held, since all of the necessary information will be distributed by the tool. In addition to this, continuous education of employees will be enabled. Finally, testing the employees about the acquired knowledge will be on the tip of employers' fingers. Pop quizzes and ad hoc test can be introduced at any time. Also the users will be trained via job-specific skill development scenarios based on sample company data.

The companies and institutions that will have this solution in place will have state of the art mechanism, that none of their competitors have. They will be the new trend setters in measuring employees' performance. The value

added of this product is its multispectral purpose. It is not a necessary product, but it is a product that will give additional value and benefits both to the employer and the employee.

C. Solutions with similar funcionalities

Since e-Assessment cloud solution [7][7] has certain similar functionalities as the Micro Assessment solution, similar software architecture has been considered. In both cases major accent is put on the assessment activity and multi-client base that needs to follow the activities at same time. Therefore both scalability and a multi-tenant architecture are used in both solutions. Cloud is taken as a supporting platform since both solutions have a multi-client base.

VI. Conclusion and Future Work

Measuring employees' performance has always been one of the managers' biggest challenge. No more internal trainings will be needed, no more long staff meetings will be held. The new designed MA system is a product that can be beneficial from multispectral point of company's point of view. It educates employees, it measure performance and helps HR managers in their performance reviews and it tests employees how much they have learnt, i.e. how interested they are in company's prosperity, how motivated they are by their managers, how ambitious they are about getting promoted. Testing the employees about the acquired knowledge with pop-up quizzes and ad hoc test can be introduced at any time.

By enabling as many features as possible, one broadens the target groups in which the given solution might be applicable. Beneficiaries of this solution might be:

- SMEs (lawyer offices; bookkeeping offices; consultancy companies, i.e. computer based run businesses);
- Financial institutions (banks, insurance companies, pension funds); and
- Public administration (Government, Ministries, Governmental Institutions, Agencies of local self governance).

In this paper we presented a new model and architecture of a system for MA of skills and knowledge. The built MA system should be implemented as a software service that can run on cloud. The instructor can easily configure teaching materials and questions for knowledge assessment and skills to be able to join the existing systems for e-learning, etc.

The system is organized with two parts: the cloud side containing MAS, MS and RS services; and the client side with the CS service.

Finally, it is really a tool that might change the working and the educating culture and can bring a more realistic way of measuring employee performance.

In future we plan to design a pilot project to analyze scalability and multi-tenancy of the Micro Assessment cloud solution.

References

[1] D. N. Aspin, J. D. Chapman, "Lifelong learning: concepts and conceptions". *International Journal of lifelong education*, vol.19, no.1, 2000, pp.2-19.

[2] G. Fischer, "Lifelong Learning and its support with new media", *International Encyclopedia of Social and Behavioral Sciences, Discipline "Cognitive Psychology and Cognitive Science"*, Section Editor: W. Kintsch, Contribution No 41, 2001.

[3] R. Klamma, M.A. Chatti, E. Duval, H. Hummel, E. T. Hvannberg, M. Kravcik, E. Law, A. Naeve, P. Scott, "Social Software for Life-long Learning", *Educational Technology & Society*, vol.10, no.3, 2007, pp. 72-83.

[4] A. Remesh, "Microteaching, an efficient technique for learning effective teaching", *Journal of Research in Medical Sciences"*, vol.18, no.2, 2013.

[5] T. Hug, "Micro Learning and Narration, Exploring possibilities of utilization of narrations and storytelling for the designing of "micro units" and didactical micro-learning arrangements", Innsbruck, Paper presented at the *fourth Media in Transition conference*, MIT, Cambridge (MA), USA, May 6-8, 2005.

[6] P.A. Bruck, "Microlearning: Emerging Concepts, Practices and Technologies after e-Learning", in *Proceedings of Microlearning, Learning & Working in New Media*, Innsbruck university press, 2005, pp. 13 - 17.

[7] S. Ristov, M. Gusev, G. Armenski, K. Bozinoski, G. Velkoski, "Architecture and Organization of e-Assessment Cloud Solution", *IEEE Global Engineering Education Conference (EDUCON)*, 2013, pp. 736-743.

[8] A. Bundovski, M. Gusev, "A New Methodology to Analyze Micro Assessment Solutions", in *Web Procedings of ICT Innovations 2010*, pp. 315-329.

[9] SAP, SAP® BusinessObjects™ Knowledge Accelerator, Maximize Productivity with Effective Software Training, 2010.

[10] P. Olson, "Business Objects Announces BusinessObjects Knowledge Accelerator", San Jose, California, 2003.

[11] RSA, "KnowledgePulse, Small steps", 2013.

[12] P.A. Bruck, L. Motiwalla, F. Foerster, "Mobile Learning with Micro-content: A Framework and Evaluation", in *Proceedings of the 25th Bled eConference*, June 2012.

[13] K. Salchow, "Load Balancing 101: Nuts and Bolts", *White paper, F5 Networks, Inc*, 2007.

The role of IC technology in development and application of experimental methods and multivariate analysis

Mandi Orlić* and Marija Marinović **
*Polytechnic of Zagreb, Department of Civil Engineering, Zagreb, Croatia
**Department of Informatics, University of Rijeka, Rijeka, Croatia
morlic1@tvz.hr, marinm@inf.uniri.hr

Summary - Information and communications (IC) technology influences all human activities but also plays an important role in development of various scientific disciplines and other types of technologies thus enabling their large scale implementation. This paper presents the relations among IC technology, experimental methods and multivariate analysis. As the development of experimental techniques and multivariate analysis paved way to revolutionary discoveries referring to IC technology, this, as a feed back, enabled further development and implementation of the mentioned methods in the most diversified fields of human activities. The necessity of the integral approach, based on implementation of multivariate analysis and adequate experimental methods and respective IC technologies is illustrated in the example of solving the problem of preservation, protection and evaluation of cultural heritage objects. In the process, various experimental analytic techniques based on principles of physics and/or chemistry are used along with IC technologies. In order to obtain the best possible analysis and interpretation of series consisting of enormous number of data, aiming to filter only the most important information relevant for particular object, statistical approach is necessary where a prominent position belongs to multivariate analysis - a task that can only be accomplished by using adequate statistics software.

I. INTRODUCTION

Aspiration for speeding up and facilitating the solving of problems experienced in using various experimental methods and analysis of obtained results, as well as in setting the theoretical model independent of measurements, resulted in creation of various devices, which are predecessors of modern computers.

During several past decades, development of IC technology influenced all aspects of everyday life. Networked computers, e-mail, databases, Internet, World Wide Web (www), on-line electronic conferences, on-line search engines, new media, social networks and similar have become integral parts of everyday life and entire society has been facing the accelerated development and implementation of such technologies in daily life activities. New IC technologies, particularly crucial IC services like Internet, have considerably influenced the change and development of society, known as the information society. As of beginning of the 21st century, the image of society has been rapidly changing so that impact of new technologies and media influences the remodeling of total social conscience, including views on economy, science, education and politics. Development in technical, social and art fields has been based on new forms of computers and IC technology.

The magnitude of using Internet is best illustrated by statistical data, shown in Figure 1 and the fact that 2.7 billion persons or 39% of world population uses it [1]. Respective graph shows that from 1996 onward, the number of users has been growing exponentially. Yet, the development of IC technology increased the existing gap between developed and developing countries. Considerable increase in using IC tools in Croatia was noted after year 2000. In 2011, 70% of population was using computers [2], so it is fair to assume that today the number equals to 80% (Figure 2).

The goal of this paper is to explore how scientific notions are connected to experimental techniques and how multivariate analysis influenced the development of the most important IC technology tools (computers, Internet, lasers ...) and how IC technologies irreversibly enabled further upgrade and implementation of these methods in various exploration fields. On the basis of the performed researches, the paper shows the importance and possibilities for implementation of these technologies and methods in solving various problems connected to

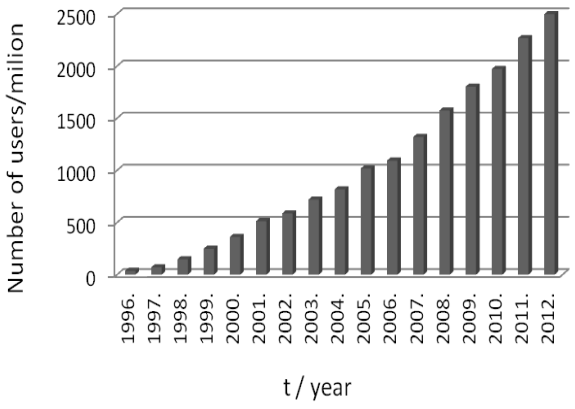


Figure 1. Increase of Internet users in the world from 1996 to 2012, according to World Stats.

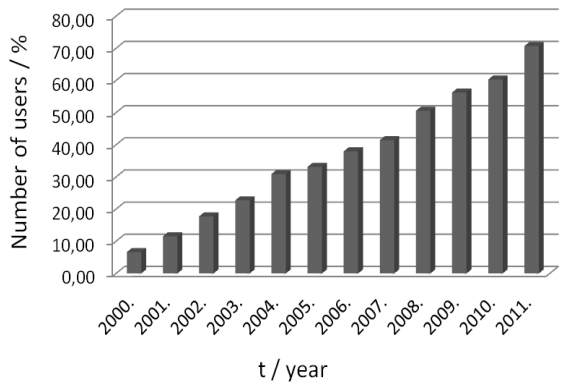


Figure 2. Presentation of increase in number of Internet users in Croatia from 2000 to 2011, according to Telecommunication Union

preservation and protection of cultural heritage objects.

After the introduction, interconnection of IC technology, experimental methods and multivariate analysis is presented as well as their role in cultural heritage. A separate chapter includes silver coin analysis, results of respective research and discussion of the results. This research aims to confirm the necessity of integral approach based on using multivariate analysis and adequate experimental methods and IC technology in acquiring relevant notions on cultural heritage object being explored.

II. CULTURAL HERITAGE AND TECHNOLOGY

Cultural heritage represents entire intellectual and material opus of individuals or groups in a given society. It is necessary to preserve the heritage, as it is of utmost importance for culture, history and identity of both individuals and entire people. Generally, cultural heritage includes material, movable and realty, and intangible heritage. Material heritage includes objects, while intangible heritage includes the totality of science, arts, philosophy, customs and similar.

Preservation and protection of cultural heritage, assumes identification, documentation, research, maintenance, protection, usage and promotion of its values and they play an important role in the process of knowing the cultural identity of given society and it is the proof of its existence and activity. In solving problems of preservation, protection and assessment of cultural heritage objects, it is necessary to emphasize the importance of interdisciplinary approach as, in the process, along with IC technologies, various experimental analytic techniques based on physical and/or chemical principles as used. In analyzing an enormous number of data obtained in experiments, of which some are to be eliminated while the others should be grouped and correlated, mathematical apparatus is indispensable, most often used in form of statistics and probability.

Implementation of digitalization techniques in the cultural heritage field enabled preservation and protection of original objects, which was done by making or printing publicly available protected 2D or 3D copies. Availability of digitalized materials through Internet infrastructure, i.e. through various network databases (virtual and digital libraries, archives and museums). Potential of 3D digitalization is used for a more authentic presentation of objects, for so called "virtual unification" (connection of various parts of an object located at distant locations), planning and carrying out restoration and conservation interventions by printing true replicas. 2D and 3D virtual reconstructions, made by means of IC technology, are most often used for popularization of cultural heritage, where digital reconstructions available on Internet and geolocation services enabled art and heritage to be even more available to users [3, 4].

The “experimental” method assumes process of observation of a phenomenon being examined under precisely defined conditions. These are most often laboratory conditions, enabling observation of phenomenon course and setting the very same conditions where, if necessary, the phenomenon may be reproduced [5]. Experimental methods, used for finding out quality and/or quantity element composition of material, belong to the group of spectroscopic methods. Application of such methods in exploring cultural objects enables to answer questions as to conditions of an object, degree of pollution, authenticity, chronology, origin or its material value. Before a cultural object becomes widely publicly available, it is necessary to analyze the object in detail in order to make decision on adequate preservation process. In exploring such objects, there are strict limitations in terms of selecting experimental methods adequate for implementation. Nondestructive, or minimally destructive, methods are selected. The reason is usually in object's value as it may be highly rare, has cultural and historical importance and variety of other circumstances, for example ownership or possession, which may require the nondestructive analysis. Besides, sampling and change of structural and aesthetic appearance of the object are to be avoided. If sampling is nevertheless necessary, a method requiring microscopic quantities of material required for analysis is selected. Complete information and answers in respect to some object are not possible to obtain using one single experimental method, rather, a combination of certain methods providing complementary results is necessary.

Non destructive or minimally destructive methods are particularly adequate for applying in cultural heritage field, for example X-ray fluorescence (XRF), Raman and micro-Raman spectroscopy, Laser-Induced Breakdown Spectroscopy (LIBS), thermography or multispectral imaging technology. Application of mentioned techniques produces emission spectra containing many complex data on the object. To obtain quality analysis and interpretation of strings containing large number of data and extract only the most important information relevant for given object, statistical approach is necessary, where multivariate analysis method holds a prominent positions. When analyzing such data types, most often used methods are principal component

analysis (PCA), cluster analysis and partial least squares (PLS) method. The main goal in implementing these methods is to reduce complexity of analyzed data set, particularly noted in high dimensionality and complexity of certain relations. These relations may be among some variables, but among measurements obtained for some samples and among some variables and some measurements. Therefore the PCA method or Cluster analysis is adequate for comparing several objects by some relationship, for example location where object was found and for obtaining information on their similarity or difference [6]. The partial least squares method is applied aiming to forecast some parameters on one or more analyzed objects. Most often, this refers to assessing quantitative element composition of sample on the basis of emission spectra obtained by adequate spectroscopic method.

Analysis of the ancient objects and identification of pigments in ancient fresco and modern paintings, as well as various stone and marble monuments, sculptures and ornaments and other glass or metal objects are good examples of applying nondestructive experimental methods in the world of tangible cultural heritage [7-12].

A. Experimental methods and IC technology

Set of experimental methods enabling to determine chemical composition and structure of a substance is called spectrometry. The beginnings of spectrometry are connected to scientific discoveries in early 17th century. The first scientific explanation of spectrum was the explanation of the rainbow by Marko Antonije de Dominis [13], the archbishop from Split. Complex nature of white light was discovered by Isaac Newton in 1666. Important discovery occurred in 1859, when Kirchhoff showed that all pure substances emit the unique spectrum [14, 15]. By the end of the 19th and in early 20th century various scientists were trying to explain the origin of spectrum phenomena. The problem of the universal explanation of spectrum for most elements was solved in 1925 by implementation of Heisenberg's uncertainty relation and Schröding wave equation [16, 17]. By the end of the 19th century, the existence of infrared, visible light and ultraviolet components of sunlight spectrum was known. Soon afterwards, radio waves, located far at the infrared side of electromagnetic spectrum, were discovered as well as X-rays and gamma rays, located far at the ultraviolet side of spectrum.

However, research of spectral phenomena of light, by the end of the 19th and in early 20th century, when optic spectrometry was at full blast, witnessed to the revolutionary invention of laser. This was preceded by series of researches in quantum mechanics, where most significant discoveries are connected to Max Planck and Albert Einstein.

Although as early as in 1870 Alexander Graham came to idea that visible light may be used as a communication medium, it was not possible to create a wave that would transport light from one point to another so his idea could not be realized until laser [1] was discovered in 1960. Multiple potentials of the laser were recognized so that production started in various laboratories all over the world.

In parallel with laser development, optic fibers were developed, featuring small losses and enabling light to be used as communication medium. Nowadays optic fibers build the backbone of modern telecommunication system. All forms of the modern communications, radio and TV signals, telephony, computer data and similar are based on signal carrier, i.e. wave with electromagnetic oscillation at certain frequency. Today, there is an enormous number of various lasers, which have been further developed, but new types have been discovered as well. Their implementation is enormous and they are applied, directly or indirectly, in almost all fields of human activity. Modern medicine could hardly function without using various types of lasers, particularly in surgery. In computer technology, lasers are used for storing and reading data on various media (CD, DVD), laser printers etc., while in military industry they are used for producing various weapons, although the original idea was to use lasers for development of spectroscopic methods. For past several years, implementation of lasers plays an important role in cultural heritage. Laser enable restoration and cleaning of various cultural (stone or metal) objects, including delicates works of art such as paintings or scripts [18].

Spectroscopy and its potentials of implementation have changed considerably, first by development of electronic computers that enabled complex calculations for large molecule spectra and offered increased precision for small molecules and then by using lasers.

Just like notions in physics led to the revolutionary discoveries in development of technology (lasers, computers, Internet,...), later they enabled further development of both theoretical and experimental notions. Over past several decades, spectroscopic methods have been applied in various fields of research, such as medicine, biology, ecology to researches in various cultural objects.

For quite a long time, computers are tools for acquiring new knowledges on nature and society. Notions on elementary processes and substance structure in the micro world are important for development of various science fields and therefore for development of new technologies in the macro world, so close to us all. Research in the micro world can happen in several ways. Experiment plays the important role there, where experiment generally consists of a device serving to make the microscopic particles suitable for research, then there is a device (most often laser) that shifts the particles to the selected energy conditions and finally, system that detects the particles or light photons. The function of computer is to control the experiment in terms of

[1] American scientist T.H. Maiman invented the first laser that emitted 694 nm wave length light; he induced laser emission by stimulated emission from ruby crystal excited by a flash lamp.

coordinating previously mentioned experimental entities. Data are collected, analyzed, saved and prepared for processing and presentation by means of computers. Yet, experiments are often too expensive as they may require high technology devices, so that there is a method of work where it is necessary to carefully design and precisely plan in advance. After that, basic processes should be recognized in the planned experiment, for which processes theoretical expressions in form of formulas mainly already exist and computer enables to create the computer simulated experiment, where basic measure values can be more easily changed as necessary than it could be possible in real experiment. Along with establishing relations between theoretical and experimental approach, IC plays an important role in data exchange process and possibilities for entire communication among scientists. Internet is used for basic communications, connecting and searching data bases, presentation and promotion of results in various science fields and similar. Web pages offer most diversified data, both old and the most recent scientific achievements, scientific magazines may be read, numerous books or instruments can be known or bought. The essence of Internet and web pages is in setting hyperlinks. Knowing a single one web address may provide the entrance and search of entire network.

B. Multivariate analysis and IC technology

Multivariate problems were addressed by numerous mathematicians such as Gauss, who, by the end of the 18th century set the basic methods of partial least squares or Galton, who, in the second half of 19th century, studied correlation matrix, defined correlation coefficient and linear regression [19]. His student, Karl Pearson, in 1901, trying to solve the data set dimensionality problem produced the interpretation of main components [20], while C. Spearman, explaining the human intelligence, formulated the basics of factorial analysis [21]. Till mid twentieth century, the largest part of the multivariate analysis was established.

The development of multivariate analysis was undoubtedly connected to the development of computer technologies and the first statistical programs. As early as in 1920', university and research laboratories in America used mechanically punched cards that served, among other things, to formulate complex statistical models such as variance analysis and linear regression [22, 23]. In mid 1960', the first statistical programs were developed, such as SAS (1966.), SPSS (1968.), Matlab (1970.), Minitab (1972.), etc. However, the revolution in implementation of the multivariate analysis started in 80' of the past century with development of personal computers, commercial multivariate program packages and particularly Internet that enabled availability of such programs to the "common" people. Till then, such programs, but computers as well due to their high prices, were used only in large scientific and research centers that could afford them. Today, it is rather difficult to count all programs used for various statistic processing, including the multivariate analysis. Prices for such programs are nowadays affordable to almost everybody. Many of these are offered at considerably lower prices for academic community and students. Great number of these is free, but such programs contain mainly only basic multivariate methods.

In order to understand completely the importance of Internet and IC technology in implementing multivariate methods for analysis the most different problems in various research fields, we analyzed scientific papers published on the Web of Science (WoS) database from 1955 to 2013. We searched the papers containing words *Multivariate Analysis*, *Principal Component Analysis* and/or *Partial Least Squares* in title, key word list or summary. Results are presented in Figure 3.

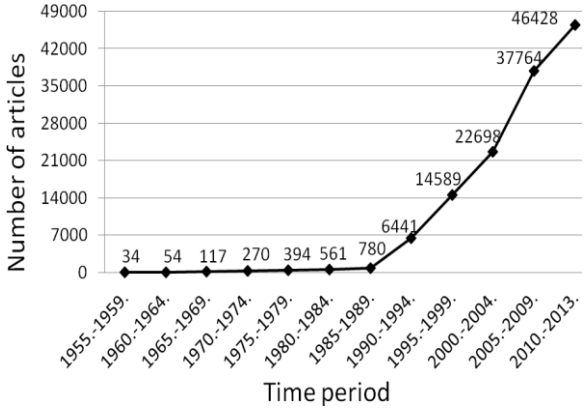


Figure 3. Number of works containing words *Multivariate Analysis*, *Principal Component Analysis* and/or *Partial Least Squares* in title, summary or key word list, published on the Web of Science (WoS) database from 1955 to 2013.

Based on these results, it can be seen that in early 1990' number of published papers grew considerably. The trend may be connected with the development of IC technology and particularly with availability of Internet. It can also be seen that multivariate analysis has been implemented on a large scale. In the same period, great number of papers were published in spectroscopy field as well, more precisely 2162 papers.

In order to comprehend as to how much multivariate analysis was used in combination with spectroscopic methods, we searched the database for papers containing words *Multivariate Analysis*, *Principal Component Analysis*, *Partial Least Squares* and/or *Spectroscopy* in title, summary of key words list. According to data shown in figure 4, a considerable increase in implementation of multivariate analysis in combination with spectroscopic methods started after year 2000, which matches the development of IC technologies.

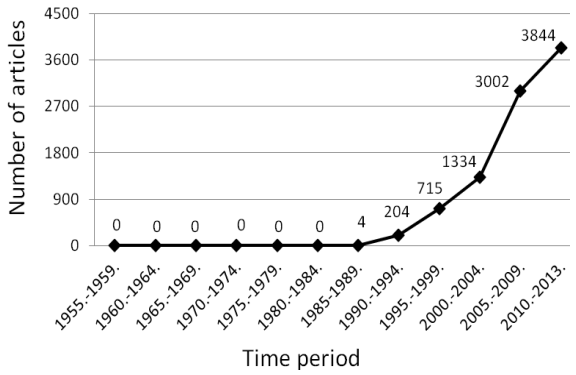


Figure 4. Number of works containing words *Multivariate Analysis, Principal Component Analysis, Partial Least Squares* and/or *Spectroscopy* in title, summary or key word list, published in the Web of Science (WoS) database from 1955 to 2013.

III. RESULTS OF THE SILVER COIN ANALYSIS

Nowadays, many nondestructive experimental techniques are used to determine the contents of elements in metal coins, but we used the Laser-Induced Breakdown Spectroscopy. The measurement was performed at *Laboratory for laser spectroscopy of cold plasma* at Institute for Physics in Zagreb. The goal of the experiment using the LIBS method was to make a shot of plasma spectrum generated by laser ablation of coin and determine chemical composition of the sample. Analyzed sample was a 10 crown coin, made in 1869 in Austro-Hungarian Empire. The silver concentration was 50%.

Used experimental setting consisted of Nd:YAG laser, optical system, sample (target), optic fiber, spectrometer system and computer. During measurement, gas argon was insufflated to target for spectral lines would be more emphasized and suitable for further analysis upon completion of laser ablation processing. When testing with laser ablation method, the sample is fixed on a stand, with it surface sloped at 45° to the laser beam direction. Measurement in one point lasted for 75 seconds, i.e. sample surface was shot by laser beams, each lasting one second, which resulted by 75 emission spectra. The first of them presents surface layer image, while the last one presents the bottom of the ablation made crater. The damage of sample resulting from ablation is micro sized and as such it is considered negligible.

Since the LIBS technique produces a large quantity of data, which are highly correlated, we used the principal component analysis technique for interpretation and reducing the dimensionality. The analysis was performed by means of the Unscrambler software exclusively intended for processing complex data sets by multivariate analysis.

Experiment data were processed using Principal Components Analysis (PCA) in order to select a few composition characteristic spectra out of 75 obtained by ablation at one point. Starting matrix **X**, of order 75 × 13508, of original data includes 75 rows, whose values represent spectra image after each consecutive second and 13508 columns with values that correspond to variables or wave lengths. After having run the PCA on the matrix, spectra were grouped in six groups. Objects (spectra) within one group are highly correlated in respect to observed variables (wave length). Therefore, it is sufficient to observe only one spectrum within each group for identifying elements. Figure 5 shows the results of the PCA, which explains 94% of variance with two main components. One spectrum was arbitrarily selected from each isolated group (16.3, 24.3, 33.3, 42.3, 66.3 and

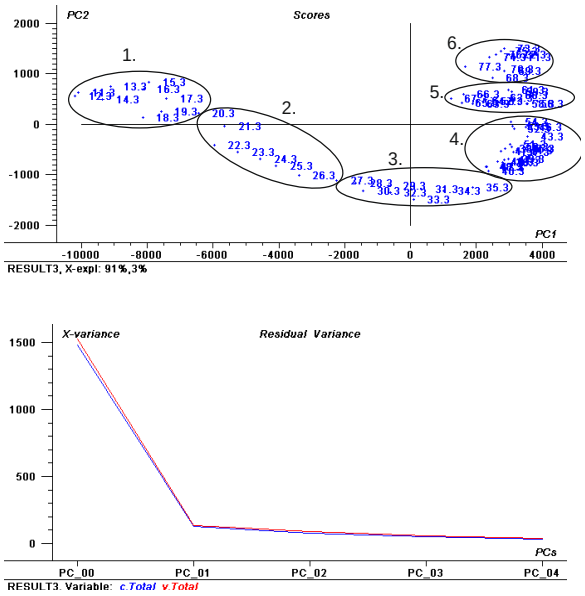


Figure 5. PCA of experiment data obtained by the LIBS method. The analysis was performed considering the first two main components, PC1 and PC2, where the first main component explains 91%, and the second 3% of variance in data.

73.3). Therefore, the element identification in sample was performed only from so selected emission spectra. Spectral lines read over entire range of wave lengths were compared to data in the NIST[2] database. Figure 6 shows the most intensive lines of identified elements from emission spectra, selected as described above, within wave length interval from 200 nm to 980 nm. Along with silver, in sample we identified presence of phosphorus, magnesium, copper, lead, calcium, sodium and chlorine from so selected emission spectra. Hydrogen and oxygen are from air since sample was not enclosed in vacuum chamber during measurement, while argon was purposefully insufflated to sample surface and therefore these three elements do not constitute its integral part.

We used the Origin software for processing the emission spectra and identification of spectral lines.

Since the LIBS method provides measurement by sample depth, we observed changes in intensity of detected lines. Namely, on the basis of changes in intensity of spectral lines of identified elements over time or by increase in depth from surface towards interior of sample, it is possible to make conclusion on presence of

[2] NIST = USA National Institute of Standards and Technology.

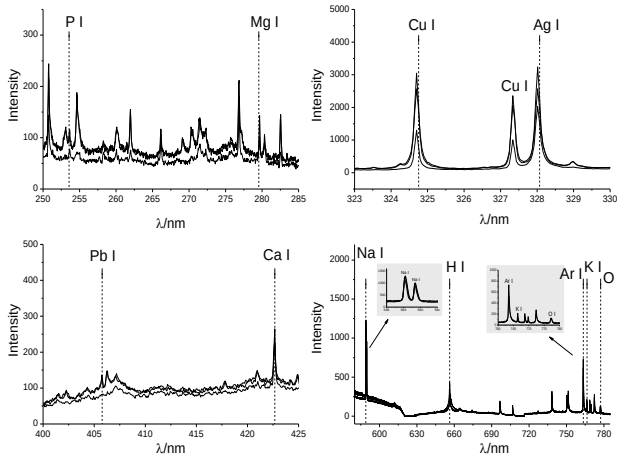


Figure 6. Identification of lines (elements) form LIBS spectrum of measured sample of wavelengths from 200 nm to 980 nm. Emitted radiation wave length is marked with λ, while Roman numerals following the elements represent their valency.

element in certain parts of the sample. Results of the observation are presented in figure 7.

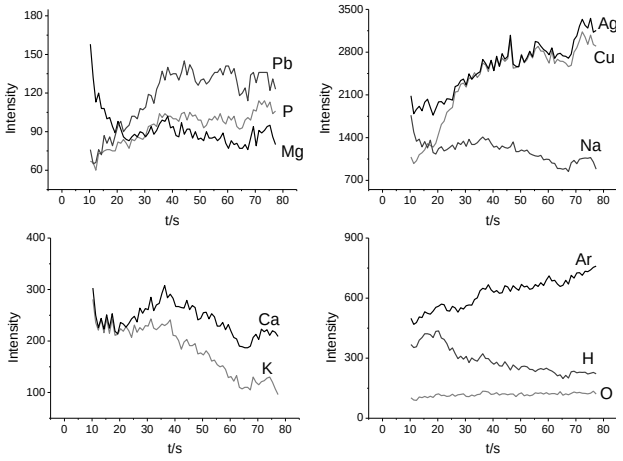


Figure 7. Intensity of most intensive lines for elements contained in sample as function of increase of depth, i.e., time. The Abscissa is time axis x, characterizing 75 different depths of craters generated by sample ablation at one spot. The ordinate represents change of line intensity from surface layers to the middle of sample.

IV. CONCLUSION

The researches performed point to the importance of the interdisciplinary scientific approach, based on the implementation of multivariate analysis and adequate experimental methods and respective IC technologies in solving problems in various activities. In this particular case, such an approach is indispensable for the process of preserving, estimate and protection of objects in the cultural heritage world. On the basis of the presented, we conclude that the development of experimental techniques and multivariate analysis resulted in important discoveries referring to IC technology, which in turn, irrevocably enabled further development and implementation of these methods. Accordingly, we expect that further development of IC technology will contribute to improvements of these methods, as well as to expand possibilities for their implementation in researching various problems in different research fields, including the cultural heritage.

REFERENCES

[1] www.itu.int/en/Pages/default.aspx

[2] www.internetworldstats.com

[3] H. Stančić, „Digitalizacija“, Zavod za informacijske studije, INTER-ing, Zagreb, 2009.

[4] H. Stančić, K. Zanier, „Heritage live: upravljanje baštinom uz pomoć informacijskih alata“, Univerzitetna založba Annales, Koper, 2012.

[5] R. Zelenika, “Metodologija i tehnologija izrade znanstvenog i stručnog djela”, Ekonomski fakultet u Rijeci, Rijeka, 2000.

[6] M. Orlić, M. Marinović, „Primjena metode glavnih komponenti u različitim područjima istraživanja“, Društvo i tehnologija 2012. – dr. Juraj Plenković, pp.189-198, 2012.

[7] N. Orlić, I. Jelovica, J. Dobrinić, C. Lofrumento, P.R. Salvi, “Analysis of ancient and medieval specimens using nondestructive spectroscopic techniques”, Nuclear Instruments and Methods in Physics Research A., vol. 580(1), pp. 739-742, 2007.

[8] I. Jelovica Badovinac, N. Orlić, C. Lofrumento, J. Dobrinić, M. Orlić, „ Spectral analysis of postage stamps and banknotes from the region of Rijeka in Croatia“, Nucl. Instrum. and Meth. in Phys. Res. A, vol. 619(1-3), pp. 487-490, 2010.

[9] N. Orlić, J. Dobrinić, P. Salvi, „Elementna analiza arheoloških objekata rendgenskom emisijskom spektroskopijom“, Radiation physics and chemistry (1993), vol. 51(717), pp. 4-6, 1998.

[10] M. Orlić, I. Jelovica Badovinac, N. Orlić, L. Mandić, C. Lofrumento, Cristiana, „Evaluation of elemental concentrations in environmental specimens using X-ray fluorescence and micro-Raman spectroscopy“, Nuclear Instruments and Methods in Physics Research. Section A. vol. 619(1/3); pp. 86-88, 2010.

[11] M. Orlić, J. Dobrinić, M. Marinović, „Multivarijatna analiza spektralnog sastava metalnog novca“, Proceedings of the 7th international numismatic congress in Croatia, INCC2013, Rijeka, 2013, pp. 164-172.

[12] M. Orlić, et al., „Primjena metode laserom inducirane plazme na analizu metalnih uzoraka“, Proceedings of the 7th international numismatic congress in Croatia, INCC2013, Rijeka, 2013, pp. 173-179.

[13] I. Supek, “Povijest fizike”, 2. izd., školska knjiga, Zagreb, 1990.

[14] R.A. Kyle, M.A. Shampo, “Gustav Robert Kirchhoff”, Jama, 1982, vol. 248(14), pp. 1739.

[15] M.A. Sutton, “The Early History of Spectroscopy”, Platinum Metals Rev., 1988, vol. 32(l), 28-30.

[16] J. Gribbin, “In search of Schrödingers cat: Quantum Physics and Reality”, Buntam Doubleday Dell Pub., 1985, pp. 47-73.

[17] J.D. Graybeal, G.S. Hurst, J.O. Stoner, “Spectroscopy”, Encyclopedia Britannica, 2009.

[18] R. Bordalo, “Laser Cleaning in Art Conservation. Part I: Laser Fundamentals”, e-conservation magazine, 2010, vol. 15, 46-53.

[19] M.A.K. Shiker, “Multivariate Statistical Analysis”, British Journal of Science, 2012, vol 6(1), 55-66.

[20] G.H. Dunteman, “Principal component analysis”, Sage Publications, Newbury park, 1989.

[21] H. Kane, C. Brand, “The importance of Spearman's g as a psychometric, social and educational construct”, The Occidental Quarterly, 2003, vol 3(1), 7-30.

[22] R.C. Milton, “Statistical Computing in the United States”, Environmental Health Perspectives, 1979, vol 32, 203-219.

[23] D.A. Grier, “The Origins of Statistical Computing” (www.amstat.org/about/statisticiansinhistory/index.cfm?fuseaction=papers)

The Views of Students and Teachers an Implementation af E-learning in Educational Process

Dragana Glušac*, Dragica Radosav*, Dijana Karuović*, Žolt Juhas*
* University of Novi Sad, Technical faculty "Mihajlo Pupin" Zrenjanin
glusacdragana.zr@gmail.com

Abstract –This paper presents a survey of students' opinions about the integration of e-learning in educational process. The results of how students are informed on distance learning and e-learning are presented. By analyzing the results, data can be obtained which clearly show what steps should be taken in order to implement a new methodology of education, e-learning methodology.

I. INTRODUCTION

E- learning is relatively a new product when we talk about new methods of learning. It represents a new type of education in our region which is not researched and implemented enough. Although there is a great number of people who want to improve their knowledge, this new type of education represents a potential risk of the unknown and therefore E- learning is being approached pretty carefully. This technology is now in developing phase and it will experience its maximum potential in the years to come. Nowadays, education in our country is performed in old, traditional way which is old fashioned and inappropriate for the 21st century and, as a result, Serbia is losing step with the world trends. The quality of education characteristic for our region in the past was recognized and respected but education should be developed and modernized today in accordance to the standards imposed by global trends.

Although there has been a lot of literature on e-learning so far they mostly include descriptions and technical information. We can rarely find empirical results given by pupils which explain effectiveness of this type of learning. It is interesting to obtain these data from primary school pupils because modernization of education has recently started in these regions and it is significant to pay atteention on pupils' informing from the beginning of their education. The facts about pupils' knowledge related to new technologies implemented in this type of education are of special significance, especially how much the pupils know about them and how helpful they are. It is important to emphasize that the very existence of new technologies in education is not sufficient but that the need for them exists. Only the symbiosis of these two factors can bring positive results in the long run. The survey of attitudes was carried out in Primary school "Jovan Mikić" in Subotica. Seven-form-pupils were interviewed by filling the questionnaire. By analysis of the results the assumptions are justified and these results will be presented in detail in this paper.

II. ACADEMIC FRAMEWORK OF E LEARNING

On-line education is a part of based on learning via Internet. It is wrong to understand E-learning only as digitalization of a classical text displayed on computer screen. Reading texts on the screen is not only bad for the eyes but it is monotonous too. Therefore, a crucial factor when creating on-line educational material is its extension and enrichment by multimedia content as well as implementation of tools for interaction enabled by Internet. For this reason it is necessary to adept the existing documents to cyber surroundings.

The first task within creation of such type of material for education is to identify the group which will use the material. This target group must be clear in relation to age, educational level, etc. Then, it is necessary to determine the aims of the survey. The next step is carrying out the survey which will conclude whether the aims correspond to those of the user's. In the case of e-learning there is an important rule that enables a user to identify him/herself with educational program. If it happens that a user is not interested in prepared material, he/she can leave the lecture by a simple click on the screen which cannot be done in classical education.

After realization of the first step it is possible to create educational material. Educational material should be divided into independent units – chapters according to the set aims. Within this task the application (ex. Trainersoft`s authoring tool) can help because it is easy to divide material in logical separate units by it. The more educational material is systematized the easier a user can use it. If it is possible, chapters should be limited in time not longer than 20 minutes. In E-learning this size of a chapter corresponds to one hour of classical lecture. Navigation within educational material is of high importance. Complicated educational material and lack of navigation possibilities result in frustration which leads to termination or withdrawal of users.

There are lots of new ideas which can be implemented or combined with those mentioned here within a creation phase. In the last ten years European Union has paid a lot attention to creation of digital society. Beside political and economic developing programs related to this issue, there has been a lot of researches in EU which resulted in the programs naturally related to researching this field. Nowadays, the effects of these programs and researches can be seen in the

society which is more digital than social. The result of this transiton is proved by the presence of digital education in institutions of higher education. Unfortunately, young experts are not given enough knowledge from this field and this unsatisfaction causes some changes at cultural, economic and political level. Because of challenges of digital society in various spheres of life in all political systems – for example, culture, education, health care, etc. – it is necessary to prepare and provide appropriate knowledge to students and experts in institutions of higher education therefore, if they have any dilemma or questions related to digital societies they will be able to recognize them and apply the achieved knowledge in solving them. It is recommended to apply 12 hints of Susan Moron-Garcie when educational material for E-learning is created:

1. Technology is used in education in order to make easier the access to quality educational surroundings.
2. In on-line education a user is in the centre and necessary activities are adjusted to aspects and opinions of users.
3. Let's try to create something new and not to copy what exists.
4. New technology brings new possibilities which should be implemented in education in order to increase interactivity and users' interest.
5. Positive, motivated and relaxed educational surroundings is essential in creation of educational material.
6. Slower tempo in getting used to new educational surroundings.
7. More frequent personal communication with users is necessary because of possible mistakes.
8. It is necessary to check knowledge and experience of users in the field of Information prior to group formation.
9. It is necessary to make time for testing and exercising in order to help users.
10. Personal ideas are not always the best solution. Users' ideas and suggestions are often good.
11. E-learning does not shorten the time. In the beginning more time is needed to prepare educational material but after that it is possible to save a lot of time.

Adaptation and improvement of existing methods for monitoring progress is good but it is necessary to monitor possibilities which new technology brings as well.

III RESEARCH METHODOLOGY

Based on these requirements - improvement of teaching methods, implementation of new technology, modernization of education – Primary school "Jovan Mikić" from Subotica was chosen. This school had shown interest for modernization and implementation of new technologies in order to achieve better success even before. School staff is consisted of a great number of young teachers who are not firmly linked to traditional way of education because they have experienced during their education modern methods. Pupils from the 7th form were chosen instead of eight-form-pupils because the latter had to prepare for secondary school preliminary exams.

For creating a questionnaire we needed a subject which most frequently deals with traditional way of education and is a result of modern technology as well. It can be concluded logically what subject belongs to the needed direction ; all this serves for comparing the research results which will be presented in detail. The results fulfill the expectation and for now it is enough to emphasize that implementation of new ideas in current school program is payable because in this way effectiveness of education can be improved.

The research results are based on two independent questionnaires presented in the Appendix. In order to increase pupils' attention it is necessary to limit the number of questions , minimum 10 and maximum 15. There are two kinds of questions in the questionnaire – classical questions which a pupil has to answer and questions in the form of a test where a pupil chooses one of the given answers. Questions in the form of a test can be categorized in two types. The first type has 2 offered answers (yes/no), and the second offers 5 answers. The second type was used in order to present results more precisely and differently. The majority of questions was expanded with one more possible answer (number 6), or the variation „Other" if neither of the offered answers expressed the attitude of the examinee. Both questionnaires contained identical questions adjusted to the subject.

The research results were analyzed by means of the program Stata 12. It is necessary to know certain notions for using this application.

Upon finishing the research the pupils were given CDs with video tutorial and installation of Adobe tools trial version for creation of a web page – Dreamweaver CS5. These tools were installed in IT study room in school. Video tutorial in which a creation of dynamic web page is presented was given as extra curriculum activity research. We waited for several days to give chance to greater number of pupils to watch video and then we visited the classes who took part in the research again.

The data obtained by this questionnaire are interesting especially those related to E-learning. The results were analyzed by means of the program Stata 12. The data were processed by „Fisher" and „Chi" trial. „Fisher" trial uses four cell (2x2) contingent tables for testing null hypothesis (independent elements). It is used for analysis of a small number of elements (according to the criterion for which the expected calculated limit value of any cell is smaller than 5). The trial does not use approximate value but it gives exact probability (p value) on the examined or even more extreme distribution beside the null hypothesis.

"Chi" trial is a type of analysis which serves for equation of expected and obtained results in which distribution of trial statistics in the case of null hypothesis

approximately describes „Chi“ square distribution. It is widely applied.

IV RESEARCH RESULTS AND DISSCUSION

During carrying out the survey all pupils were present. The questionnaire was filled in anonymously. The questionnaires were marked by coded signs in order to enable identification of pupils/respondents.

According to "Fisher type" the exact trial P value was higher than 0,005 (P<0,005) so the results obtained in this way were not valid. In other words, the research results are not representative and they serve only for presentation of opinions and attitudes of 7-form-pupils.

According to the survey, 54.06 % out of 76 pupils said that IT tools would help them in learning while 34 pupils thought contrary (*Table I*). Majority of pupils thought that the use of IT tools wasn't necessary but it can be useful. The Table presents precisely the number of pupils from two classes who like (17) and dislike Informatics (57).

The discussion was done for representative dimensions which can represent moderators.

		Would the use of IT tools help you to learn easier and better?		
		No	***Yes***	***Total***
Do you like Informatics classes?	***Yes***	32 56.14% 94.12%	25 43.86 62.50%	57 100% 77.02%
	No	2 11.76 5.88%	15 88.24 37.50%	17 100% 22.98%
	Total	34 45.94% 100%	40 54.06% 100%	74 100% 100%

Table I

Out of 76 pupils who participated in the research 30 of them would like a teacher to be present at classes physically (*Table 3*). From the Table can be seen in what extent the pupils need the presence of a teacher at classes.

		Do you think that a physical presence of a teacher at classes is necessary?		
		No	***Yes***	***Total***
	No	8 21.05% 50.00%	30 78.95% 50.00%	38 100% 50.00%
	Yes	8 21.05% 50.00%	30 78.95% 50.00%	38 100% 50.00%
	Total	16 21.05% 100%	60 78.95% 100%	76 100% 100%

Table 2

A half of the sample from the research thinks that a teacher's presence at classes is necessary (*Table 3*).

It was more difficult to obtain necessary data from teachers. Many of them didn't want to participate but still the idea was interesting to majority of them. Opinions of teachers didn't coincide with pupils' opinions which was expected.

More than a half of the pupils who participated in the research was not introduced to E-learning (*Table 7*). Only a few pupils would like to have this type of classes at this level of education. However, about 24% of pupils know or at least have an idea about e-learning.

Do yu know what is E-learning/distance learning?	***Answers***	***Answers in %***
No	45	59.21
I know what it is but I don't know what this type of class is like.	12	15.79
I know exactly what it means.	9	11.84
I know what it is and would like to attend such classes.	7	9.21
Other:	3	3.95

Table 3

V FURTHER DEVELOPING DIRECTIONS OF E-LEARNING

After the first part of the research both teachers and pupils who attended the lecture on possibilities of E-learning were interviewed. The students' results showed the following:

- ✓ 86% pupils think that video tutorial is useful and they would like to use it in the future.
- ✓ 36% of them would like to see a form of LMS within their education at Informatics classes.
- ✓ 100% pupils understand what E-learning is.
- ✓ 77% of them think that the use of E-learning would make their studying easier..

- ✓ 62% pupils think that there is no need for a teacher at Informatics classes if e-learning is used.

A high percentage of pupils was interested in further e-learning.

The teachers' results showed the following:

- ✓ 20% think that this method of learning is dangerous for pupils because they can find unreliable and unchecked subject matter as well as the people whose pedagogical background is problematical.
- ✓ 40% of them think that the implementation of this technology is more expensive than classical one so they don't want to use it.
- ✓ 30% teachers tried to use some of distance learning ways but they were not successful enough.
- ✓ 30% teachers think that the use of mobile phone in education can be appropriate but outside of school and definitely not at classes.
- ✓ all teachers share the same opinion according to which technical support for this type of education is not prepared yet.

VI CONCLUSION

In the last decade e-learning has experienced great expansion because of the progress at the field of ICT. Following the world trends and trying to improve quality of work in education a great number of educational institutions has included this method of learning in their regular programs. The results of this and similar researches show that there is a great interest and need for further expansion and improvement of these programs.

For their introduction it is necessary to have technical equipment and training as well as general education and knowledge of teachers because they are leaders of the whole process, therefore it is important to inform them about all advantages of new technologies and methods of learning.

The level of technical equipment was evaluated as satisfactory in the institution in which we carried out the research and there were also free solutions in the field of LMS which have been proved as successful in practice. We will emphasize two factors which, in great extent, contribute to recognition of e-learning as a crucial factor of developing strategy in this school :

- ✓ Education of a teacher because insufficient education and lack of knowledge related to implementation of new technologies can be a serious obstacle in relation to their introduction.,
- ✓ Appropriate motivation of teachers because resistance to changes is often a consequence of unreadiness for preparing a material for classes.

As for the sake of further researches and after this initial review of the state in this field, it is concluded that a detailed analysis of factors which influence implementation's success or failure of e-learning in this school should be the part of a plan. Afterwards, the model for implementation of e-learning will be developed on the basis of the obtained results.

Pupils' views about implementation of e-learning in educational process have confirmed our assumptions. The pupils would like to use any type of e-learning in their school but this aim can be fulfilled by appropriate informing on details of this type of learning. Teachers represent a key factor of success. Like we emphasized before, teachers should be motivated to make a step forward and accept new trends of learning and teaching in order to help their pupils enter the world of e-learning and make it the part of their education.

REFERENCES

[1] Forgó Sándor: Az új média és az elektronikus tanulás, Új Pedagógiai Szemle, 2008. 8–9. 91-97.

[2] [1] Hutter Otttó: E-LEARNING 2005, Műszaki Könyvkiadó, 2005.

[3] Kulcsár Zsolt: Az integratív e-Learning felé 2008. URL: http://mek.oszk.hu/06600/06695/06695.pdf

[4] Moron Garcia S., Willis L., 2009, Introductction to Pedagogic Research Toll Kit, ISBN: 978-I-904804-94-9.

[5] Bessenyei Istvan, - Stoffa, N., "Information Society Studies in Practice - a Networked Learning Case Study: Experiences of Teachers in NETIS project" (Istraživanje digitalnog društva u praksi - Primer mrežnog učenja: Iskustva predavača NETIS projekta), Halkidiki, 2008. Godine

[6] Kovacs Ilma, "Uj ut az oktatasban", BKEFKI Budimpešta, 1997. Godine

[7] Don McIntosh, "Learning Management System", Trimeritus eLearnig Solutions, 2007. Godine

[8] Pletl Szilveszter: Elektronikus ügyvitel a gyakorlatban, jegyzet 2007. VTŠ Subotica

[9] Martin Dougiamas, "Improving the effectivnes of tools for Internet based education", Curtin University of Technology, 2000. Godine

[10] Martin Dougiamas, "Developing tools to foster online educational dialogue", The University of Western Australia, 1999. Godine

[11] Juhas Ž., 2013., Master paper, 7th Grade students opinions of integrating e-learning in the educational process, Technical Faculty "Mihajlo Pupin", Zrenjanin

The Model of Remote Video Surveillance in Hierarchical Autonomous E-testing *WbeTS* System for Knowledge Testing

Damir Purković*, Eugen Ban **
* Faculty of Humanities and Social Sciences, Rijeka / Department of Polytechnic Education, Rijeka, Croatia
** Education and Teacher Training Agency, Branch Office Rijeka, Croatia
damir@ffri.hr
Eugen.Ban@azoo.hr

***Abstract* - The web-based system for online knowledge testing (*WbeTS*) is a hierarchical and modular autonomous system for remote knowledge assessment whose autonomy is based on remote video surveillance as one of the testing validation mechanisms. The main problem facing the system is the physical implementation of the video surveillance unit for remote monitoring of a student, as well as development of a decision making mechanism based on such monitoring. The monitoring would need to take place in conditions of widely available broadband Internet access and the use of a webcam. A possible solution is a model of the remote control described in this paper, which could solve this problem and make the system more reliable. The model is based on built-in elements and algorithms for visual recognition and motion tracking of the student, sound tracking, user (teacher) interface for assessing the validity of testing and of the system's kernel responsible for assessing the validity of the tests and reporting. The model is envisaged to be implemented accordingly to the wishes and needs of users, and can be integrated as a part of an integral e-learning environment. Although the model described is in principle technologically feasible, its practical realization requires professional refinement and experimental validation. Only by in-depth testing and experiences gained from the actual application of the system will be possible to identify all deficiencies and requirements that currently cannot be predicted.**

I. INTRODUCTION

Testing of the student knowledge is an integral part of every teaching and is as well of the e-learning environment. However, the implementation of such assessment of knowledge and the validity of the test results are often questionable. The reasons for this may vary: from the lack of skills and competencies of teachers to design and implement such tests, up to the so-called *methodological limitations* [1] of this kind of knowledge testing. Methodological limitations are evident in the inability to evaluate higher levels of achievement in the cognitive domain of achievements, according to the Bloom's taxonomy of educational objectives [2], [3]. The attempt to solve this limitation was to develop a concept of a hierarchical system for knowledge e-testing (*WbeTS*) [1], in which knowledge test should be implemented in stages, through established hierarchical blocks of various tasks, which difficulty corresponds to the level of achievement of these taxonomies. This hierarchical e-testing system is conceived as an autonomous system that enables less stressful and more casual and phased knowledge testing with a significantly higher level of validity in comparison to the results of the typical written or online tests. Phased accession enables the student to delay access to higher-level testing, having successfully solved the previous lower level, and the choice of final desired level of achievement. System at the same time, by using a subsystem for visual authentication and supervision of students, generally provides access to testing services from any location which greatly simplifies the technical preparation and organization of the testing performance. All this contributes to more casual and less stressful approach to knowledge e-testing, facilitating the organization and implementation of it, providing a more realistic insight into the actual level of achievement. However, the practical implementation of this system as a subsystem of an existing e-learning environment is facing a number of problems that can significantly impair quality and validity of the testing. One of the major problems needed to be successfully solved on the systemic level, on which the autonomy of the system is dependent, is the remote video surveillance of students.

II. PROBLEMS OF REMOTE SURVEILLANCE STUDENTS DURING E-TESTING

The validity of any knowledge testing system, including written and e-testing systems, is based amongst all on the content validity, i.e. on the choice of questions and tasks that measure what they are actually intended to. However, test results can vary depending on the subject's possibility to cheat. That is why it is necessary for any testing to be carried out for a limited time and in a controlled environment. This is validity that will further on be called *trivial validity*. Hierarchical autonomous knowledge e-testing system bases its autonomy on a remote student monitoring, which largely abolishes the possibility that the subject of testing cheats. Therefore, a remote surveillance system should provide an acceptable level of trivial validity with a minimum of necessary technical conditions. That minimum of technical requirements is relatively easy to determine and can be classified as a general requirement for the e-testing implementation, and the vast majority of students today can meet it. Such general requirements refer to the use of a standard personal computer (PC or laptop), or other categories of devices (notebooks or tablets) with the suitable Internet browser, a standard webcam with a microphone and

a broadband Internet connection. In order to avoid potential problems in video surveillance it is desirable to classify acceptable level of brightness and noise in the room. The monitoring subsystem should be carried out in terms of smooth network connection of the subject's PCs and the server. Therefore, server resources as well as the network throughput and resources of the client should comply with mentioned conditions, which is a standard practice today. However, the processing and transmission of video and image data can be a problem if one is working with high-resolution video signal or insufficiently optimized processing algorithms. Therefore, the overall video surveillance of the respondents should be based on a robust and well-optimized software motion and sound detection algorithms. Result will eventually be mathematically expressed as a numerical indicator of the trivial validation of the testing. The monitoring and detection process should be done in real time, while the filtering process and the data processing of the validity can take place with a delay.

III. MONITORING AND CLASSIFICATION OF MOVEMENT USING CAMERAS

In the contemporary visual surveillance systems different algorithms and techniques have been developed for the movement detection and classification. They are often tailored to specific applications in a particular area. Motion detection is one of the most active research topics in the field of computer vision [4] nowadays. Among traditional and widely known methods used to detect moving objects are: i) temporal differencing [5], [6], [7] ii) optical flow [8], and iii) background subtraction method [9], [10].

The temporal differencing method or the difference between correspondent pixels is, due to its relatively low level of programming complexity, not very demanding. This method comprises computer operations to determine the difference between two adjacent image frames in a sequence of a video clip images. The difference between extracted characteristic points of image frames is quantified by calculating Euclidean distance and variance. It was experimentally proved that the method of calculating the variance of the difference in picture frames is less sensitive to noise and can adapt to a variety of dynamic environments providing a good stability [4]. This method is not sensitive to luminance changes and can be adapted to different conditions of supervision and it is therefore suitable for the limited value detection and so-called harsh object movement.

The optical flow method is a key method for determining the speed and direction of a moving object. Optical flow represents an apparent speed distribution of light patterns [11] which may arise from the relative motion of objects and observers. The application of this method requires a short distance between two adjacent image frames, which makes multiple frame video more suitable for the application of this method, but at the same time it is much more complex to handle. This method uses the value of the points from the image, which results in moving field, and does not extracts characteristic image points as previous methods. Optical flow method does not require information about the previous scene, and can dynamically track the movements. Therefore, this method requires more computational resources due to long-time movement information processing and it can be greatly affected by even small changes in scene illumination. On the other hand, in the temporal differences method small variations in brightness can hardly affect the accuracy of detection, while the high brightness variations cause abrupt maximum value of the variance, which can usually be ignored [4]. In the practical application of this method for extraction of moving objects commonly is used Horn-Schunck algorithm [11] complemented with the modern techniques and algorithms to determine the contours of objects [8], [12].

Background subtraction method is, according to the manner in which it processes information about moving objects, similar to the functioning of the human eye [4] and it can be used to reduce the impact of changes in the moving scene by updating information about the background. HSV (*Hue, Saturation, Value*) space [13] is determined by three coordinates which define the tone (hue) of color, color saturation and brightness value to describe the color. In comparison to the RGB color space, HSV space is closer to the perception of the human eye and it is often used in image processing and in pattern recognition. Because of the frequent problem of reduced movement detection accuracy caused by variations in illumination, Stauffer and Grimson [14] proposed a process of updating the background by setting the background subtraction threshold. This procedure is based on the difference between the current frame and the one that represents the model of the background.

Background modeling is often performed with Kalman filtering approach [15] or by using statistical models [16]. In practical application, background updating process sets the subtraction threshold and then calculates the difference between the value of the key frame and background frame for each pixel. Pixels with a difference greater than the threshold value are considered to be a part of a moving object and their background value remains unchanged. Pixels with a difference of less than a threshold value are considered to be a consequence of changes in lighting and background frame pixels are replaced with the appropriate key frame pixels [4]. Since the object of observation is mobile, the background should be changed in real time – in practice this method is known as adaptive background subtraction technique [5]. It is difficult to adjust the background subtraction threshold using this method, which often leads to excessive operation upon pixel conversion from the RGB (Red, Green, Blue) to HSV space [13]. These operations can significantly slow down computation and implementation of monitoring in real time. Determination of the threshold for background subtraction, as a key part of this method, is often the subject of experimenting and empirical optimizing of the method.

In addition to these methods, which can be used individually, their particular advantages are more often combined and integrated recently in the different hybrid motion detection methods [4] , [17]. Hybrid method (Figure 1) can be designed in different ways and commonly uses a temporal differences method to extract the key frames for basic movement detection. After extraction of the key frame,

object in motion is being allocated using methods of optical flow and HSV background subtraction method. Video sequence will be processed with a most suitable method. Result of the processing is output data about moving objects, which can be additionally processed by a computer. Applying a hybrid method, here simplified, it is possible to accomplish optimum motion detection and characteristics of a moving object with minimal time and computational requirements needed to process video signals.

For the detection and classification of audio signals, in this case of speech, it is common to use VAD (*Voice Activity Detection*) techniques and algorithms [18], [19], [20], which are used for various applications, from audio conferencing and speech recognition, to speech and telephony coding [18]. Typical VAD algorithm consists of three stages: the first stage reduces the noise in the signal with certain methods, the second stage performs quantification of the input signal, and in the third stage the classification rule determines whether the signal is speech or not [18]. For the purposes of the subject, audio control of the minimum generic functionality of VAD algorithm is sufficient, based on which system is to extract speech from other sound signals.

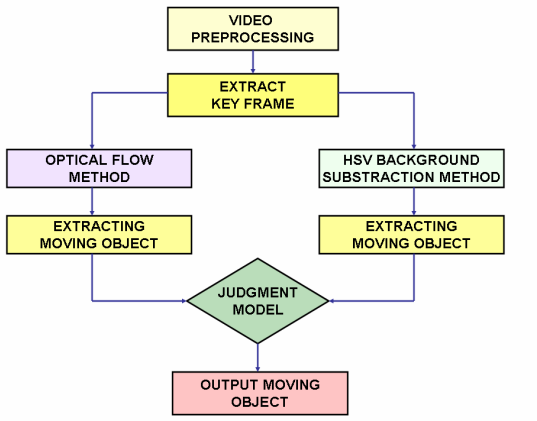


Figure 1. Flowchart of the hybrid method of visual tracking of a moving object [5]

IV. REMOTE RESPONDENTS SURVEILLANCE MODEL IN *WBETS* SYSTEM

Remote subject surveillance model (Figure 2) was developed by considering the requirements of it within a hierarchical e-testing system (*WbeTS*), current knowledge of the methods and algorithms for the detection and classification of movement and sound, as well as technical and technological limitations of hardware. The model is based on the use of a custom hybrid method of remote detection and basic movement characterization, and the use of a simple and robust VAD algorithm for sound detection. The model assumes that the trivial testing validity (V_T) can be described as a function of the so-called rough movement, or absence of subject from camera view (i), characteristic of fine head movements (g), lack of eye focus (o), and the detection of sound (speech control) (z). Given that the quantization of each controlled variable is more dependant of the duration than the number of detected phenomena, trivial validity can be described by a simple relation (1):

$$V_T = \sum_{1}^{Ni} t_i + \sum_{1}^{Ng} t_g + \sum_{1}^{No} t_o + \sum_{1}^{Nz} t_z \tag{1}$$

Where t is the duration of each occurrence that the system recognized as an anomaly and N is the number of occurrence of each anomaly. Accordingly, the trivial validity is actually an integer expression of the duration of all anomalies in seconds. This information has been proven to be the easiest to use in the analysis of video recordings of controlled online student knowledge testing.

Analysis of the video footage was conducted on a sample of 52 students of undergraduate and graduate studies at the Faculty of Humanities and Social Sciences in Rijeka, which were tested with 30-minute online tests prepared by the Department of Polytechnics. Testing and analysis were performed in order to classify movements of which trivial validity may depend, and the frequency of those movements in a completely valid test. Movements and anomalies which were entirely absent in a valid test were: rough movements, absence of subjects out of camera focus and sound anomalies related to the speech of subjects during the test. The position of the respondents head during the test has mainly been towards the video camera, with minor anomalies in all directions, which should be treated by the control algorithm as a tolerance orientation (an acceptable movement) of the head. During the valid test subjects blinked about 30 times a minute, each blink lasting less than a second, resulting in need for treating the blink as a movement and not detected as an anomaly by the tracking algorithm. According to this analysis, the valid test had an average of 23 movements of the head, which can be detected by the system as a change of the vector structure critical point set on the boundary lines of the head, which can be regarded as anomalies. These anomalies lasted between 0.7 and 3.4 s, which is an average of 177.45 seconds per test subject. These were anomalies that did not impair the validity of the tests, such as yawning, head scratching, stretching, etc. Therefore, duration of these movements could not cause copying or consultation amongst subjects, or similar activities that would disrupt trivial validity. When it comes to the eye focus, it is to be noted that the algorithm of the system, according to the analysis, should consider each eye position, where the lower third of the pupils are visible, valid eye focus. Thus, the analysis shows that the valid test has an average of only 3.7 absences of the eye focus per test, or altogether 15.2 seconds. From the analysis it can be concluded that the test, which is valid in terms of trivial validity, has up to 200 seconds of anomalies, or, $V_T \leq 200$ for a valid test.

Based on the analyzed movement which should be monitored by the system for compliance of trivial validation, above mentioned remote subject control model (Figure 2) has been designed and developed. The model, as an e-testing *WbeTS* subsystem, generally is consisted of user interface, which is an integral part of *WbeTS* system, and the

subsystem kernel, which performs video signal processing, video and audio signal filtering and validation. The user interface is designed for input of the desired parameters or requirements for monitoring facilities such as lighting, sound level, and the level of trivial validity. Projected video signal retrieved by the subsystem is, because of the computing resources optimization, limited to the usual webcam resolution of 320 x 240 pixels with 30 frames per second. Before retrieving the signal, the subsystem checks the desired conditions and, after they are met, retrieves the image that is "roughly" preprocessed, and converts it into a 2D binary image with the aim of reducing the computational complexity. At the beginning, the subject needs to choose the desired position, of which the system will warn him, in order to determine the reference relations between key points and the size of the observed scene (key frame). Rough processing includes identifying and monitoring key (characteristic) points, which represent the monitored virtual scene, on the subject's head. There is in total 9 boundary points of the subject's head and nose, and 8 points for the contour of each eye of the subject. After extraction of the key frame, system reduces a supervised background using the HSV background subtraction method, and tracks surfaces that bound the key points using optical flow method. Special attention is given to the monitoring of the respondent's eyes, or to be more exact, to the monitoring of the surface area of the eye and the pupil. For further algorithm elaboration it is important to emphasize that the monitoring of the points, which represent a key frame, should be performed with necessary movement tolerance for the entire area bounded by those points, and for the relation between these areas and the picture frame. Tolerance is necessary because the entire head of the respondent moves minimally during the test, while maintaining focus and mutual key point distances, which does not impair the validity of testing. Motion analysis found that overall differences in movement monitored area, and the points and their mutual ratio during a valid test do not exceed 12% of the initial reference value of those points. However, when making a practical implementation of the model, it is necessary to re-adjust and modify this value regarding experience. It should also be noted that the background subtraction algorithm should be performed under the condition that there is no other movements in the background of respondents, which means that the system may terminate testing if movement in the background appears, or treat the movement as a rough anomaly in testing, which may depend on specific implementation of the model. This isolated and reduced image with key points, areas that are monitored and seized background, is going through the proper filters in the further course of processing, which will then distinguish certain types of anomalies in accordance with their pre-defined characteristics. Therefore, the system detects every movement of the scene's key points, and determines whether the positions of the points, surfaces that they delimit, and their relations are within specified tolerance limits defined at the beginning of testing. Every absence of a key points, dodge and disproportion of mutual distances and/or surface areas which they delimit, are classified as anomalies and their duration is being determined. The lack of points is detected as an absence of the respondent or a rough movement, while is the disproportion of mutual distances and/or surfaces detected as an inappropriate movement of the head. At the same time surface areas of the eyes are being tracked as well, and each reductions below the predicted level is being detected and its duration is measured. If it lasts less than 1s, the movement is classified as a blink and the system does not record it, otherwise, the movement is classified as inappropriate eye focus and its duration time is marked of. The system also monitors and records sound environment during testing, which is classified as speech or other sounds based on the implemented VAD algorithm. When VAD algorithm classifies sound signal as speech, its duration is measured as sound anomaly in the testing. The duration of each of these anomalies is added to the previously recorded duration of equivalent anomalies, and their total time at the end adds up and turns into an integer amount of trivial validity V_T. Based on predefined trivial validity values, at the end of the testing the system decides whether the test is proper in terms of validity, and whether that information is entered into the corresponding field along with the test results of certain subject. Ultimately, this information is an integral part of the subject's test results and it is available via user interface that manages the e-testing system.

The described model is a generic algorithm structure for the remote surveillance implementation, which can be complemented by additional elements due to higher level of trivial validity. That said, the monitoring can begin with authentication of participants based on face recognition algorithm and the points taken as the parameters to be used during the surveillance. In addition, the system can, at the beginning of testing, or sequentially during the testing, retrieve images of the subject and store them in a database along with the test results. Stored images can be used when making an additional identity check, in the case of suspected subject swap during the test.

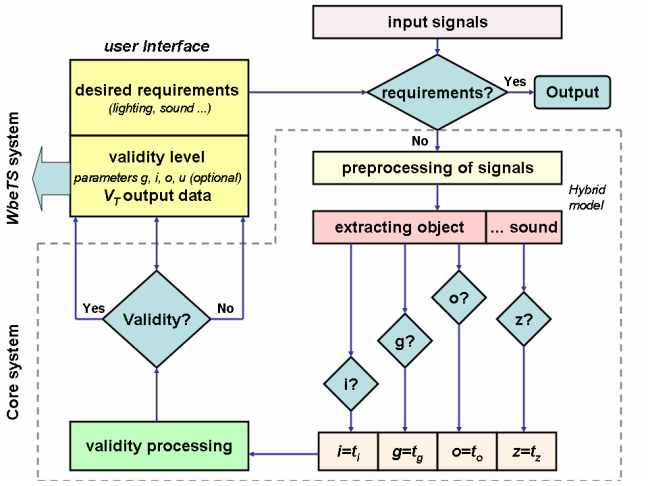


Figure 2. Simplified flowchart of the model of remote video surveillance in *WbeTS* system

V. POSSIBLE DIFFICULTIES IN THE IMPLEMENTATION OF THE MODEL

The model, here conceptually and partly technologically elaborated, needs to be developed and adjusted to the customer requirements in a specific e-learning environment, for experimental verification and improvement of the algorithm, the application, and of the model itself. There are possible problems in different tasks and levels of development and customization:

a) Problems in application development - the quality of the algorithms and procedures that will be developed to implement the model can greatly affect the applicability and relevance of application within which the model is implemented in the system. Video signals processing can use considerable computing resources, implying that the code should be well optimized and adapted to the standard computing resources used today, and the code should be tested in a real multitasking computing environment. During the application development a special attention should be paid to the eye focus monitoring algorithm, which is extremely important to ensure the quality of this kind of surveillance, but can be extremely challenging for software customization and optimization.

b) Problems of integration into the e-learning management system – as the model is a part of the WbeTS concept of e-testing system, and as such is designedto be integrated into any environment for e-learning, mostly into Moodle. The integration should follow developer's notes of the environment and its potential requirements and restrictions. Although it is not a limiting factor for the implementation, it is desirable that the model is developed in the form of a modular plug-in that can be easily integrated into the e-learning and e-testing system.

c) Problems in selecting the application's starting and executing method – as any network application is based on the communication between the client and the server, its performance depends on the quality and bandwidth of the interconnections. The application for remote control can be installed on the server, which means that the network signals travel to the server and are then processed there. Such a solution could, especially for a large number of simultaneous participants, quickly consume server resources and disable network communication. A much better solution is to run the application on the client side, which will at the end of the testing send the requested validity information and any additional information to the server.

d) Problems associated with the video surveillance requirements – as the subject of testing has the possibility to control the environment, there can be possible problems in meeting the testing requirements, which can be considered as a limitation from the subject's point of view. The environment in which the e-testing is conducted should be static and relatively quiet (home workspace, etc.), but can often be dynamic and noisy (public computer, waiting rooms, internet cafés, etc.). Furthermore, household lighting conditions can vary from poor home lighting to complete lack of it, or pretty dark environment, and so on. Because of these and similar problems, the implementation guidelines should clearly specify the required conditions regardless of the access location, and the application functioning requirements. Only in this way it will be possible to avoid many potential problems associated with the requirements for surveillance.

Although these groups of problems do not cover every possible situation, they are currently recognized and expected and they should be considered during the implementation. It is expected that future development and experimental verification of the remote control functioning in real life and/or in laboratory will discover additional problems and difficulties in the implementation of the model. Such development, soon to be realized, will also improve the model and the application, and will also discover new technological possibilities for realization of the desired goals.

VI. CONCLUSION

Modern knowledge e-testing systems, as a necessary and indispensable part of the e-learning system and environment, can greatly facilitate and accelerate the evaluation of the achievements of specific curriculum objectives. However, their trivial validity, because of which the teacher must supervise the process, often complicates the test implementation, especially when testing very large groups of students. Development of e-testing systems and concepts, and training of teachers on how to use them has largely counteracted all categories of validity. E-testing and online testing have become an integral part of contemporary mixed educational system. However, in order to satisfy trivial validity, it is still necessary to physically control the subjects. In this paper, a robust remote subject control model is presented, which is, on a theoretical level, trying to solve the problem of trivial validity. The model is based on the use of hybrid method for motion detection, VAD algorithm for sound classification, an algorithm for characterization and filtering of the detected movement, and the use of standard, affordable hardware and broadband internet. Although the model is designed to be a part of a hierarchical web-based system for knowledge e-testing (*WbeTS*), there are no obstacles for its implementation in other knowledge e-testing environments. For the model to become a part of the e-learning environment, it is necessary to develop a software module and integrate it in such environment, and distribute it to a wider audience. By implementation and testing in real conditions, it will be possible to further develop and improve the system and the model, and finally to gain users acceptance.

VII. REFERENCES

[1] D. Purković, Z. Kolumbić, A Concept of the Web-Based E-Testing system, MIPRO 2012, Biljanović, Petar (ed.), Rijeka: GRAFIK, 2012, pp.1514-1520 (ISBN: 978-953-233-069-4).

[2] L. W. Anderson, D. R. Krathwohl, A taxonomy for learning, teaching and assessing: A revision of Bloom's Taxonomy of educational objectives, New York: Longman, 2001.

[3] B.S. Bloom, Taxonomy of educational objectives: The classification of educational goals: Handbook I, Cognitive domain, New York;Toronto: Longmans, Green, 1956.

[4] Y. Du, X. Wang, S-T. Wang, S. Xue, Z. Qin, Hybrid moving object detection system based on key frame extraction, Industrial Electronics and Applications (ICIEA), 2011, 6th IEEE Conference on, pp 441-445, (ISBN: 978-1-4244-8754-7).

[5] R. Collins, A. Lipton, T. Kanade, H. Fujiyoshi, D. Duggins, Y. Tsin, D. Tolliver, N. Enomato, O. Hasegawa, A System for Video Surveillance and Monitoring: VSAM Final Report, Technical report CMU-RI-TR-00-12, Robotics Institute, Carnegie Mellon University, May 2000.

[6] D. Migliore, M. Matteucci, M. Naccari, A revaluation of frame difference in fast and robust motion detection, Proceedings of the 4th ACM international workshop on Video surveillance and sensor networks, pp.215-218, 2006.

[7] C. Zhan, X. Duan, S. Xu, Z. Song, M. Luo, An Improved Moving Object Detection Algorithm Based on Frame Difference and Edge Detection, Fourth International Conference on Image and Graphics, pp.519-523, 2007.

[8] T. Brox, A. Bruhn, N. Papenberg, J. Weickert, High Accuracy Optical Flow Estimation Based on a Theory for Warping, Proceedings of 8th European Conference on Computer Vision, Springer LNCS 3024, vol.4, 2004, pp.25-36.

[9] M. Piccardi, Background subtraction techniques: a review, IEEE International Conference on Systems, Man and Cybernetics, vol.4, pp.3099-3104, 2004.

[10] C. Wren, A. Azarbayejani, T. Darrel, A. Pentland, Pfinder: realtime tracking of the human body, Proceedings of IEEE Transactions on Pattern Analysis and Machine Intelligence, pp.780-785, 1997.

[11] B. Horn, B. Schunck, Determining optical flow, Artificial Intelligence, vol. 17, pp.185-203, 1981.

[12] H. Elbehiery, A. Hefnawy, M. Elewa, Surface Defects Detection for Ceramic Tiles Using Image Processing and Morphological Techniques, Proceeding of IEEE Transactions on Egnieering, Computing and Technology, vol.5, pp.158-162, 2005.

[13] http://en.wikipedia.org/wiki/HSL_and_HSV.

[14] C. Stauffer, W. Grimson, Adaptive background mixture models for real-time tracking, Proceeding of IEEE Computer Vision and Pattern Recognition, pp.246-252, 1999.

[15] C. Ridder, O. Munkelt, H. Kirchner,Adaptive Background Estimation and Foreground Detection using Kalman-Filtering, Proceedings of International Conference on recent Advances in Mechatronics, ICRAM'95, UNESCO Chair on Mechatronics, pp.193-199, 1995.

[16] O. Javed, K. Shafique, M. Shah, A Hierarchical Approach to Robust Background Subtraction using Color and Gradient Information, IEEE Workshop on Motion and Video Computing, Orlando, 5-6.12. 2002.

[17] P. Zhang, T-Y. Cao, T. Zhu, A Novel Hybrid Motion Detection Algorithm Based On Dynamic Thresholding Segmentation, Communication and Technology (ICT), 2010, 12th IEEE International Conference.

[18] http://en.wikipedia.org/wiki/Voice_activity_detection.

[19] M. H. Moattar, M. M. Homayounpour, A simple but efficient real time voice activity detection algorithm, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.176.6740&rep=rep1&type=pdf, 2009.

[20] J. Ramirez, J. M. Gorriz, J. C. Segura, Voice Activity Detection. Fundamentals and Speech Recognition System Robustness, Robust Speech Recognition and Understanding, M.Grimm i K.Kroschel (ed.), ISBN: 978-3-902613-08-0, 2007.

Web Application Development with Component Frameworks

V. Okanovic
Faculty of Electrical Engineering/Computer Science Department
Sarajevo, Bosnia and Herzegovina

Abstract – This paper describes development of dynamic web application using web frameworks based on web components. The paper considers several existing open source Java web component frameworks. Then, the paper provides basic features of the analyzed web frameworks and represents their main characteristics. To give an idea of which framework might use, the paper compares all analyzed web frameworks, and summarizes them and gives an overview of all their the most important characteristics.

I. INTRODUCTION

A web application framework is a software framework that is designed to support the development of dynamic web applications [1, 2, 3]. There are a lot of open source Java frameworks for web application development [4]. Depending on the type there are action based frameworks and component based frameworks.

To speed up web application development, action based frameworks mostly rely on external configuration files and less on Java code. Because of that they are not suitable for rapid development of web application.

The paradigm of component based frameworks is quite different. This type of frameworks creates a component abstraction in view tier. Thus, developer of web application defines a component in his application, and tells what the component does.

Regarding to numerous open source Java web frameworks, it becomes so difficult for developers to select a suitable framework for web applications development. Because of that, the purpose of this paper is to help web developers to easier choose the right framework for development of their web applications. The paper will analyze a few open source Java component based frameworks. Additionally, the paper will provide basic features of the analyzed frameworks as well as represent their most important characteristics. Also, in this paper the analyzed web frameworks will be compared and summarized.

II. OPEN SOURCE SOFTWARE

Open Source Software (OSS) is a software that can be freely used, changed and shared by anyone [5, 6]. OSS programming code is available on the Internet, so it can be used by people from all over the world. The software evolves when developers on the Internet can read and modify it freely. Many developers can work on open source software and they can modify and update it. The concept of open source software is good model for creating high quality software.

OSS is very important for student for several reasons. The programmers around the world are constantly developing many free tools and students can use them for their work. Students can learn from the best researchers and engineers in the world. They can open a software and see how programmers have organized it. By studying the source code written by the best developers, the students can learn to make better software. Thus open source software can help students to become better developers.

Also open source software is a good way for a young software developers to get an international recognition. Owners of existing open source projects are always looking for more collaborators. Developers can join to some open source project. By getting in the project, developers can create something useful as well as to develop their skills further.

For companies, organizations, governments and students in developing countries, OSS provides a way to prevent the widespread illegal copying of software, and an opportunity to raise the level of software development to international standards. The open source software has shifted proprietary software in many companies, governments and educational institutions. Besides being cheaper open source technology, executives in companies, governments and educational institutions have recognized that it offers better performance and reliability than proprietary solutions.

III. FRAMEWORK

In accordance with the free dictionary definition a framework is "a structure for supporting or enclosing something else, especially a skeletal support used as the basis for something being constructed" [7].

A framework can also be defined as a unitary entity formed of computer codes, scripts, classes, functions that offer a generic functionality and which is specialized or extended by a code generated by the user in order to obtain a specific functionality [8]. A framework provides architectural guidance by partitioning the design into abstract classes and defining their responsibilities and collaborations. A developer customizes the framework to a particular application by sub-classing and composing instances of framework classes.

Another defionition of a farmework is: a framework is typically an architectural pattern that provides an extensible template for applications within a domain [9]. By specifying a framework, developer specifes the skeleton of architecture, together with the slots, tabs, knobs, and dials that you expose to users who want to adapt that framework to their own context.

The most important benefits of frameworks are: modularity, reusability, extensibility and inversion of control. These major benefits provide what all developers want, and it is rapid application development, reuse of both code and design, reduced maintenance costs and ease of customization.

A. *Component based frameworks*

Unlike action based framework, philosophy of the component based frameworks is quite different. Component based frameworks create a component abstraction in the view tier. Event listeners capture user actions and map back to code in the controller tier. The major advantage of using components is modularity. Complex web pages can be composed easily using web components. Further, it is possible to reuse one or more web components on a few web pages of the same or another application. Moreover, component based frameworks make the collaboration between web designers and Java developers much easier.

A good framework can reduce the cost of developing an application. A Java component based frameworks are very popular type of framework. To give web developers an idea of which framework they might use, the following is analysis of basics features as well as representation of the most important characteristics of a few open source Java component based frameworks. For the sake of space, examples of web application development using these web frameworks will not be given or discussed in this paper. Also, specific examples and their detailed descriptions can be found in the cited references for these frameworks.

IV. TAPESTRY

Tapestry is open source component oriented web application framework written in Java [10, 11]. Using this framework developer can easily create robust and content rich web applications. An application developed with Tapestry is a collection of web pages. In Tapestry applications each web page is simple HTML page. The web pages are constructed using web components. That means each element of a web page is represented with some appropriate web component. Tapestry applications are built from HTML templates and specifications, and they can be run on any servlet container.

In web application development Tapestry implements Model View Controller (MVC) pattern [12, 13, 14]. In Tapestry applications developed using this design paradigm for separation, Model is consisted from domain object, View is represented with simple HTML files and Controller is implemented as XML page specification.

Tapestry framework is designed to support agile development of web application. In accordance with that, Tapestry takes care about many activities, such as dispatching user requests, monitoring the file system for changes of Java classes, managing server-side state, etc. In fact, Tapestry supports web developers to write less code. Thus web developers can use their time as much as possible for coding the application's functionality.

V. JAVASERVER FACES

JavaServer Faces (JSF) is Java web component application framework for creating user interfaces [15, 16, 17, 18]. JSF is part of the standard Java EE platform. Developed through the Java Community Process, JSF is a standardized display technology for building server-side user interfaces. JSF framework enables the reuse of web components. By using existing reusable web components JSF framework simplifies building of web pages and construction of user interfaces. There are many web components of JSF technology. Using them, this framework reduces the effort in building and maintaining web applications. Moreover, the framework provides processing component data and architecture for managing component state. In addition, JSF handles events and validates user input.

JSF application framework offers a clean separation between presentation logic and application logic. This framework implements Model View Controller pattern. JSF implementation of MVC design pattern makes applications much more manageable. Model manages data and logic of the application, View takes care about the user interface and Controller manipulates the user input. Regarding to team development, this approach of web application development allows that each member of the team can focus on individual part of work.

JSF framework uses JavaServer Pages for presentation logic and the standard tag libraries for representing the components of user interface. With JavaBeans components the framework defines functions of components on a web page. Beans configuration and page navigation rules are defined in configuration resource file of an application.

VI. WICKET

Wicket is a web component based framework [19, 20]. It is a web application framework for creating dynamic web pages by using web components. Wicket focuses on object-oriented design. By using this framework web pages can be built with reusable web components that can be extended by standard inheritance. A web component, such as a web page or a page section, is in the center of preoccupations in Wicket. This framework provides easy development of web applications. It uses only two technologies: Java and HTML. HTML pages are very cleanly separated from Java code. Thus, markup designer has complete freedom about how to design web pages. HTML designers can easely design and modify the pages by a visual HTML editor, like FrontPage, Dreamweaver, etc. Dynamic content processing and form handling is all handled in Java code. Each component in Wicket application is created in Java class and rendered in markup. The HTML file and Java class must have the same name. Also both of these files must exist in the same source structure.

Wicket provides a generic support for web applications. Web application concepts have corresponding classes in the Wicket framework and in the web application. For example, for the web application concept, Wicket has the *WebApplication* class. For a web page, Wicket has the *WebPage* class. Each web page in the web application has its own class that extends *WebPage*. Moreover, important feature of Wicket web applications is that they do not require XML configuration files. Instead XML configuration files Wicket uses a Java class for application settings.

VII. MODELIBRAWICKET

ModelibraWicket is open source Java component web framework and is part of Modelibra family [21, 22]. Besides ModelibraWicket, parts of Modelibra family are ModelibraModeler and Modelibra framework [21, 22].

ModelibraModeler is design and code generation tool. ModelibraModeler has been implemented to assist developers in domain model developing. The backbone of any software is a domain model. A domain model is a model of specific domain classes that describe the core data and their behavior.

Modelibra is a domain model framework [21, 22]. Its objective is to provide an easy to learn and easy to use framework. Its recommended use is in prototypes and small applications. It can be used in client and server applications. Modelibra has been designed to help developers of small applications in representing and using application domain models. Modelibra is useful for developing software around a domain model, for providing an easy installation for the users and for developing a web application to introduce software to the public audience. Modelibra uses XML files to save data of the current domain model. There is no need for any special software installation. However, Modelibra allows the use of both relational and object databases. The upgrade of an application from XML data files to a database does not require a single line of programming code to be changed. It is enough to change a domain configuration in XML and create a database and configure it to run the same application again.

ModelibraWicket makes a domain model alive as a web application, so that model data may be displayed as web pages and updated through forms. The model is metamorphosed into a web application with the help of its XML configuration. This web application can be considered as default application of the domain. It is important to realize that this web application is not a version that one would like to install as a web site. Its main purpose is to validate a domain model by designers and future users of the web application and consequently refine the domain model. In addition, ModelibraWicket has a collection of generic web components that may be easily reused in professional web applications to display or update entities. The generic web components may be easily reused in a specific web application. Their reuse consists of collecting data from a domain model by using Modelibra and by supplying the data as parameters to the web components. The generic web components from ModelibraWicket use Modelibra for a domain model and Wicket for application views.

VIII. COMPARISON OF WEB FRAMEWORKS

Today there are many open source Java web frameworks for development of web applications. Development of large number of web frameworks is caused by various reasons of web developers and user requests. Thus all web frameworks are very different and each of them has its benefits and limitations. Therefore, for web application developers it is very difficult to decide which framework to use for development of new project. To give web developers an idea of how a frameowork is suitable for their project, this section compares four above-described web frameworks and provides an overview of all their the most important characteristics.

The following are some factors that developers should consider when choosing a framework for creating a Java web application:

Maturity – determines how long the framework exists. Tapestry, JSF, Wicket and ModelibraWicket web frameworks exist for a number of years.

Community – refers to activity of community. Community of Tapestry, JSF and Wicket frameworks is active, but ModelibraWicket community is not active.

Documantation – determines the availability, quantity and quality of documentation. There are many books about Tapestry, JSF and Wicket frameworks. Also these three web frameworks offer good online documentation on their home pages. There are a few books and online manuals about ModelibraWicket framework, but this documentation is not large.

MVC structure – relates to how event calls are handled. All four frameworks (Tapestry, JSF, Wicket and ModelibraWicket) are component pull-based MVC frameworks.

Configuration – determines whether developer must create many XML, property or some other files to configure the framework. Only JSF is configuration framework, because in this framework there are several configuration files that must be configured.

Security – refers to implementation of security features for web application. Tapestry, JSF, Wicket and ModelibraWicket performs Java authentication and authorization security.

Localization – refers to internationalization, which means the ability to adapt web application to multiple languages. All four frameworks (Tapestry, JSF, Wicket and ModelibraWicket) support localization.

Ajax – relates to building of interactivity in web applications. Three of the four frameworks support automatic Ajax generation: Tapestry, Wicket and ModelibraWicket.

Page based – determines the ability of application development using web pages. All four frameworks (Tapestry, JSF, Wicket and ModelibraWicket) are a page based frameworks.

JSP based – relates to use of JSP pages to generate HTML. Only JSF is JPS based framework.

Template based – determines whether framework works with templates. Tapestry, JSF, Wicket and ModelibraWicket are a template based frameworks.

Inversion of Control (IoC) – refers to a mechanism used to look up internal or external services. Two of the four web frameworks directly support IoC: Tapestry and JSF.

Design and code generation – determines whether there is a possibility of design and code generation of web application. Only ModelibraWicket has this ability. For this purpose, ModelibraWicket uses ModelibraModeler, design and code generation tool.

An overview (of the existence or non-existence) of all these most important characteristics of the analyzed frameworks can be presented by a table (Table I).

TABLE I. OVERVIEW OF THE CHARACTERISTICS OF THE ANALYSED FRAMEWORKS

	Tapestry	**JSF**	**Wicket**	**Modelibra Wicket**
Maturity	+	+	+	+
Community	+	+	+	-
Documentation	+	+	+	-
MVC structure	+	+	+	+
Configuration	+	-	+	+
Security	+	+	+	+
Localization	+	+	+	+
Ajax	+	-	+	+
Page based	+	+	+	+
JSP based	-	+	+	+
Template based	+	+	+	+
IoC	+	+	-	-
Design and code generation	-	-	-	+

This table can be helpful for developers. Each developer can consider the table when selecting a component framework for creating a Java web application. Of course, many different criteria can be considered and developers can perform additional research if needed.

IX. CONLUSION

In this paper are discussed open source web component frameworks. The paper briefly describes four Java component frameworks: Tapestry, JavaServer Faces, Wicket and ModelibraWicket. These frameworks can be used for developing dynamic web application. Benefits and limitations of using these frameworks in development are indicated. The framework features are outlined, all four frameworks are compared and the overview of the most important characteristics of the analyzed frameworks is given.

Many people now wonder what to use for web application development. This paper can be used as a starting point for making a decision of selecting a framework for web application building. Thus, it can help web developers to increase their productivity. Also, this paper can be helpful for students. They can learn new concepts and simplifications, as well as advantages and limitations of all these four component frameworks. Thus, they can correctly apply their knowledge on real future projects as much as possible.

REFERENCES

[1] Web Application Framework (http://docs.oracle.com/cd/E19454-01/819-0726/s1afovew.html)

[2] T. C. Shan, Taxonomy of Java Web Application Frameworks, Internationl Conference on e-Business Engineering, Shanggai, China, 2006.

[3] Web Application Framework (http://docforge.com/wiki/Web_application_framework)

[4] Open Source Web Frameworks in Java (http://java-source.net/open-source/web-frameworks)

[5] Open Source Iniciative (opensource.org)

[6] Open source resources (http://opensource.com/resources/what-open-source)

[7] The Free Dictionary (www.thefreedictionary.com/framework)

[8] A. Sache, Choosing tThe Open-Source Framework for Web Application Development, Open Source Science Journal, 2010.

[9] G. Booh, J. Rumbaugh, I. Jacobson, The Unified Modeling Language User Guide, Addison-Wesley Professional, 1998.

[10] Tapestry (http://tapestry.apache.org/)

[11] H. L. Ship, Tapestry in Action, Manning Publications Co., 2004.

[12] E. Gamma, R. Helm, R. Johnson, J. Vlissides, Design Petterns: Element of Reusable Object-Oriented Software, Addison Vesley, 1994.

[13] E. Freeman, B. Bates, K. Sierra, E. Robbson, E. Freeman, Head First Design Patterns, O'Reilly Media, 2004.

[14] G. Seshadri, Understanding JavaServer Pages Model2 architecture – Exploring the MVC design pattern, JavaWorld, 1999. (http://www.javaworld.com/article/2076557/java-web-development/understanding-javaserver-pages-model-2-architecture.html)

[15] K. D. Mann, JavaServer Faces in Action, Manning Publications Co., 2004.

[16] H. Z. Chavez, JavaServer Faces: An Excellent Learning Tool, IEEE Distributed Systems Online, 2005.

[17] JavaServer Faces Technology (http://docs.oracle.com/javaee/5/tutorial/doc/bnaph.html)

[18] X. JunWu, Developing CRM System of Web Application Based on JavaServer Faces, The 2nd Internationl Workshop on Education Technology and Computer, Wuhan, China, 2010.

[19] Wicket (https://wicket.apache.org)

[20] M. Dashorts, E. Hillenius, Wicket in Action, Manning Publications Co., 2008.

[21] Modelibra (www.modelibra.org)

[22] V. Okanovic, Dz. Donko, T. Mateljan, Frameworks for Model-Driven Development of Web Application, The 9th WSEAS International Conference on Data Networks, Communication, Computers, Faro, Portugal, 2010.

Agile Management: A Teaching Model Based on SCRUM

Ž. Požgaj*, N. Vlahović* & V. Bosilj-Vukšić*
* Ekonomski fakultet Zagreb, Trg. J. F. Kennedyja 6, Zagreb, Croatia
zpozgaj@efzg.hr; nvlahovic@efzg.hr; vbosilj@efzg.hr

Abstract – Work presented in this paper is based on the notion of creating a teaching model based on the principles of agile management. Developed and described model is based on Scrum, one of the most widespread agile methods. Since Scrum is primarily a framework for managing software projects and software applications development, during the development of the model it was crucial to negotiate the differences between Scrum and the teaching process itself. Teaching model in this paper covers the whole teaching process from first lessons to the final exam. Model consists of three separate units: (1) presentation of theoretical knowledge; (2) practical knowledge and student projects; (3) exams and admission of grades. Each of these units represents a Scrum that results with particular „products". Basic characteristics of Scrum like incremental development, iterations, roles and artefacts are implemented in the teaching model using the same principles as in the original Scrum even though their content is adjusted to the specifics of the teaching process. Model is implemented in the Managerial Informatics master study course "Business Application Development" at the Faculty of Economics and Business University of Zagreb.

I. INTRODUCTION

One of the topics within the curriculum of the course Business Application Development (cro. Razvoj Poslovnih Aplikacija) is "Agile software development methods". In order to familiarise students with knowledge and skills within this topic more closely, we intended to organise the teaching process according to principles of agile management. The selected method was Scrum. Scrum represents a framework for software product development. Final goal of Scrum project is to create a product that complies with intended requirements. Teaching process aspires to reach the same goal as a Scrum project. Project, in this case, pertains to the organisation and conduct of the teaching process in current academic year according to requirements defined in the curriculum of the course, with the final goal – successfully complete teaching year. Analysis of teaching process in described manner has shown that teaching process can be compared and affiliated with any other process, even with software development process. Scrum is a typical software product development method. Characteristics of the method are described in detail in this paper through the presentation of the proposed teaching model. Teaching model is achieved through three separate Scrums; each Scrum containing three phases. Teaching process is developed in the same way as software product using all the stages of the software process life-cycle, where different roles are assigned to different actors, specific artefacts are used, respective activities and iterations. From presenting theoretical knowledge in the first Scrum, through presentation of practical knowledge and skills during the development of student projects in the second Scrum, until the finalisation of the teaching process with the student exams in the third Scrum. Students work on their projects in accordance to agile work organisation (including teamwork, pair-programming, presenting features as stories, weekly meetings, etc.)

Review of current literature indicates several attempts of implementation of Scrum in the teaching process. In most of the available cases Scrum is implemented only in the segment of the teaching process concerned with the organisation of student projects. As expected, courses where Scrum is used in this way usually cover the topic of software engineering [10]; [5] even though there are a few examples of using Scrum in courses dealing with other topics [4]; [1]; [6].

The rest of the paper is organised in five separate chapters. After the introduction, in chapter 2, Scrum methodology is introduced with descriptions of its basic characteristics. Chapter 3 is the central part of the paper where developed teaching model based on Scrum is described in detail. Chapter 4 indicates possible improvements of the organisation of the teaching process. Finally, Chapter 5 summarizes the key points in the conclusions.

II. SCRUM METHODOLOGY

The emergence of novel, so-called lightweight agile software development methods, evolved in the mid-1990s. The main characteristics of agile software development methods are iterative and incremental development, agility, flexibility in requirement definition, collaboration, self-organizing teams etc. Creators of these methods join the Agile Movement that publishes Agile Manifesto in 2001. Agile Manifesto details the philosophy and principles behind agile methods [8]; [9].

In 1995 Ken Schwaber and Jeff Sutherland present a study where they describe a new software development method and call it Scrum. The term scrum is derived from rugby, where a scrum restarts the game after the ball has gone out of play. Applying the interpretation of the term scrum to software projects, a software scrum gets the team

back together and gets everyone moving in the appropriate direction [1].

For Schwaber and Sutherland [7] Scrum is a process framework that has been used to manage complex product development. It is not a process or technique for building products; rather, it is a framework within which one can employ various processes and techniques. Speaking in terms of agile management, Cervone describes Scrum as an agile, lightweight process for managing and controlling software and product development predicated on a team-based approach [2]. It consists of three groups of phases: Pregame, Game (or Development) and Postgame (Fig. 1.).

Detailed description of activities and artefacts in each phase of Scrum is given in the rest of the paper.

According to 7th Annual State of Agile Development Survey for 2011. 54% software developers use Scrum. Scrum together with Scrum variants methodologies is used by 72% of software developers which makes Scrum the most popular agile methodologies being used [12].

III. TEACHING MODEL

Teaching is a process, teaching process is a project. Every project, including teaching projects, has its own respective goal (for teaching projects, according to the curriculum, the goal is to provide students with particular knowledge and enable them to master particular skills), its starting point and its ending point (during the semester). Complex projects, such as teaching project, require teamwork, mutual collaboration and cooperation. This can be seen through the relationship of teachers and students.

Following the ideas of agile approach to project development, Scrum methodology was used for the creation of the teaching model for the Business Application Development course.

Due to undeniable differences between teaching process and software process Scrum methodology had to be adjusted accordingly. First of all, every Scrum represents a complete wholesome processing unit that results with a particular product. This methodology is to be used to cover the complete teaching process of the course, from the introductory lectures until the final exam and admission of the grades. Three separate integral units were detected: (1) presentation of theoretical knowledge; (2) mastering practical skills through student projects; (3) exam and admission of grades. Each of these units represents a separate Scrum process and each of these deliver a specific outcome - product or several products. Consequently, teaching model for the Business Application Development course consists of three Scrums, or three Scrum units.

A. *About the course*

Course Business application development is an elective course in Managerial Informatics master study at the Faculty of Economics and Business University of Zagreb. This elective course consists of 20 hours of teaching (10 hours of lectures and 10 hours of practical work in computer lab). During the course students use software tools for UML diagrams, eDraw UML Diagram, and programming language Visual Basic, Visual Basic 2010 Express Edition. Workload of the course is expressed using European Credit Transfer and Accumulation System (ECTS) and the course is worth 4 ECTS credits. As we can see the number of course hours is relatively small and the number of ECTS credits is quite substantial, students are expected to work in teams outside the supervised part of the teaching process. Therefore students have additional time reserved for their independent teamwork in the computer lab, and they can choose from three different two-hour periods during the week. Teaching is provided by two lecturers, one for the theoretical topics and the other one for the practical part of the course. In the academic year 2013/2014, 33 students chose this elective course. They were allowed to organise themselves into working teams and, as a team, negotiate and define their project assignment topics.

B. *Roles, sprints and artefacts in the model*

Basic roles in the teaching model are Scrum Master and Scrum Team [7]. Teacher takes on the role of the Scrum Master. Scrum Master is the person responsible for the entire Scrum process and in this case Scrum Master is responsible for the realisation of the complete teaching

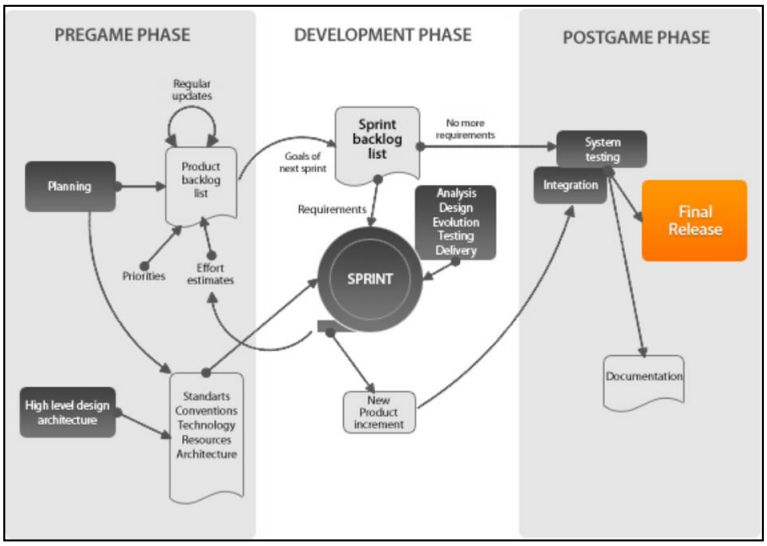


Figure 1. The Scrum phases [3]

process. This is why the role of Scrum Master includes the organisation of lectures and exercises, student consultations, advising students during project assignment topic selection, monitoring student progress and providing additional help during the development stages of student project assignments, grading of completed projects assignments, preparation and conduct of examinations, admission of grades, and so on. Role of Scrum Team is assigned to each student team. Students form teams by themselves and together as a team define project assignment topic. They delegate particular tasks among themselves during the development of the project, they communicate and consult with the Scrum Master, they observer the deadlines, write project documentation and present the project work when done. Role Product Owner is assigned to the team leader. Since student teams for the course Business Application Development consists of only two member, this role is not intended in this particular teaching model. In teaching processes where this role exists, Product Owner coordinates teamwork, takes care about the details pertaining to the project development progress and communicates with the Scrum Master. Product Owner represents the middleman between team members and Scrum Master. Soria et al. [10] recommend introduction of the additional role of Agile Coach in situations where students develop projects. According to [10] the Agile Coach encourages the teams throughout the Scrum process by cleaning the team's obstacles and emphasizing the use of tools. Designated Agile Coach represents the middleman between team members and Scrum Master and this role would be appointed to the lecturer that delivers the practical part of the course topics and exercises conducted in the computer.

Sprint represents a set of activities that are sequentially executed during each iteration of the development phase. In the teaching model, types of activities that will form each sprint are directly dependent on the Scrum unit the sprint cycle belongs to. For example, in the first Scrum unit activities within the sprints pertain to elaboration of theoretical knowledge about the topics of the course. In the second Scrum unit activities within the sprints follow the stages of the life cycle of the software product development.

Artefacts of the teaching model are specifically related by their content to the Scrum unit within which they are used. For example, in the first Scrum unit Product Backlog list, and later on Sprint Backlog list also, relate to requirements defined by the process of studying the theoretical knowledge about the topics of the course, while in the second Scrum unit both Product Backlog and Sprint Backlog lists relate to requirements defined by the project assignment topics i.e. particular software product being developed by the team. Project Topic Registration is also an artefact that is submitted by the student teams to the Scrum Master and it is only used in the Postgame phase of the first Scrum unit.

C. First Scrum – Theoretical introduction

Pregame phase contains activities that will enable students to participate in the upcoming Scrum units of the teaching model while successfully accomplishing and meeting the requirements that are continually requested from students during the teaching process. Activities defined for the first Scrum unit are providing students with the information about the course, presentation of theoretical knowledge about the topics of the course and preparation of students for the upcoming phases of the process. Key artefacts of this phase are Planning, Product Backlog list, Environment Architecture. Activities associated with the mentioned artefacts are performed by the Scrum Master. Artefact Planning involves creation of the teaching process agenda (possible innovations in the selection of the course topics in relation to the existing curriculum, determining scope of lectures for each topic by dates, planning the dates of midterm test and examination tests, defining methods of examination, determining primary and additional literature, determining dates for computer lab exercises, selection of software tools that will be used, and so on…) Product Backlog list contains features important for lecturer and students. So Product Backlog list contains each topic title, each weekly date for each lecture and corresponding topic that the lecturer will present, dates for the exam tests, instructions for the submission of student project assignment topics, deadlines for each activity, and similar information. Environment Architecture defines the required presentation equipment, lecture room features, computer equipment and corresponding software that is required during computer lab exercises and project development.

Game phase in the proposed teaching model spans over five weeks and in this period the teacher presents all of the theoretical knowledge topics as defined in the curriculum of the course. Product Backlog list is the basis for the creation of the Sprint Backlog list. Sprint Backlog contains a list of tasks to be completed during the Sprint. This is an ordered list of topics that the lecturer needs to present and they follow the stages of the software product life cycle. As topics are presented students gain required knowledge to be able to come up with project assignment topic for their own software product prototype during the second Scrum unit. List of basic and fundamental topics is enriched with topics such as software project management, software product quality metrics, and so on… List of tasks in Sprint Backlog that are to be completed by the students includes forming of project teams, completion of the Project Topic Submission forms, and similar activities.

Central part of this phase is dedicated to Sprints, i. e. incremental approach to completion of the tasks given in the Sprint Backlog list during separate iterations. Since the final number of hours of lectures for this course is ten, and there are two hours of lectures per week, model presumes five Sprints (5 x 2 hours). Each Sprint deals with one lecture. With the end of each Sprint, teaching process advances its progress in accomplishing the goals given in the Sprint Backlog list. Each lecture starts with the weekly „meeting“. The meeting agenda includes revision of topics presented on the previous weekly meeting, introducing students with the basic characteristics of the topic (or several topics) that will be presented during lectures and, finally, students get assignments for activities they should perform until the next weekly meeting (familiarise themselves with the content of a particular paper, retrieve particular information using Internet and similar…). Fifth

Sprint is different from the previous four since the location of the lecture is not in the lecture room but in the computer lab. Topic of this print is the modelling of business processes so the theoretical foundations are complemented with the drawing of particular UML diagrams using appropriate software tools.

After all of the items from the Sprint Backlog list are achieved, final phase of the Scrum unit begins – Postgame phase. For the lecturer this means that all of the required topics have been presented to the students. For the students that means that project teams are created, project topics determined and Project topic registration forms completed and submitted to the Scrum Master.

D. Second Scrum – Practical skills

Second Scrum unit of the teaching model is the most similar to typical Scrum that is originally defined as agile software development method. Since during second Scrum unit students develop their own software applications, Scrum is used as a framework for managing software applications development. Students develop their applications using Microsoft Visual Basic 2010 Express edition. Exercises are performed exclusively in the computer lab. Even though the primary part of this Scrum unit is dedicated to the development of software applications, basic information about programming in Visual Basic as well as the tutorial for the usage of the Integrated Development Environment is also covered. This is why in the Game phase during first three Sprints students are provided with necessary theoretical information after which exercises are performed using typical problem assignments that are similar to requirements in their project topics. Final two Sprints are dedicated to project work and the development of the software applications themselves.

In the beginning of the Pregame phase activity plan of the second Scrum unit is created. Product Backlog list is updated with the agenda of the practical part of the lectures. During planning Scrum Master is involved with defining the dynamics of presentation of programming basics by the topics and particular dates, defining practical examples of case studies related to specific lecture topics, defining required perquisite knowledge required for successful mastering of particular programing problems, defining deadlines for student projects and so on… Since in the previous Scrum unit 18 student teams submitted Project Topic Registrations, in the Product Backlog list only teams and their topics are entered in order to track basic features of student projects. This is why each Scrum Team compiles its own internal Product Backlog list. In these lists features of the software projects being developed are entered as well as corresponding requirements.

As mentioned earlier during the first Scrum unit, requirement for the particular software tools is defined as artefact in the Environment Architecture. Now, the testing of the installed computer resources is performed, and possible corrections are made if needed. When the final verification of the computer equipment is successful Game phase is ready to commence.

Game phase consists of five Sprints. After first three Sprints students were able to master the basics of programming and acquired required skills they are ready to start with the development of their projects. Final two weeks of the semester are allocated for the development of the projects, which corresponds to the last two Sprints of the teaching model. Scrum Teams have absolute autonomy over the organisation of their work on the projects, but within the limitations set by the rules of Scrum. This means that Scrum Teams compile their Sprint Backlog lists, define activities for each Sprint, define goals of each Sprint, delegate work tasks among team members, estimate timing and deadlines for particular Sprint activities, and so on… Since the amount of lecture hours in this period is reasonably small, Scrum Teams spend additional time and effort doing teamwork outside Sprint teaching model. This is why weekly meeting have pronounced importance during these two Sprints, since Scrum Teams report their progress from the previous week to Scrum Master. They report of what has been achieved in the previous period but also activities that they have planned for the next Sprint. This allows Scrum Master to follow the dynamics of the project development for each Scrum Team, giving them appropriate advice for the next Sprint, and so on… Each Scrum Team develops their project according to requirements given in the Sprint Backlog list, consequentially through the activities of each Sprint (Analysis, Design, Evolution, Testing, Delivery), iteratively using adequate number of Sprints and incrementally until all of the requirements are reached and the final version of the software product is developed.

Organisation of the work done during student projects is aligned with the recommendations of agile development of software products [9] and Extreme Programming (XP) [8]. Some of these recommendations include pair-programming and pair- process modelling using driver-navigator approach [11], writing stories (describing features and requirements) within the Product Backlog list and weekly meetings.

In the Postgame phase Scrum Teams continue their work by creating and compiling project documentation. Project documentation contains detailed description and disposition of the project assignment topic, UML diagrams (namely, decomposition diagram, use-case diagram at the system level, activity diagram and sequence diagram for a particular segment of the software product that was developed as a prototype), description of basic functionalities and characteristics of the software application, description of user interface, source code, course of testing activities, and final conclusions. Integral part of the project documentation that Scrum Teams attach to the documentation itself is the software product in both source code and executable form. All of these are submitted to the Scrum Master.

E. Third Scrum – Examination

Third Scrum represents the finale of the complete teaching process. By this Scrum everything that was set out by the curriculum of the Business Application Development course is completed. All of the topics given in the curriculum are presented, practical part of the course in the computer lab has been completed as well,

Scrum Teams have developed their software applications along with the appropriate project documentation and they have submitted the resulting work (project documentation and software applications) to Scrum Master. The last thing that remains in the teaching process to complete the course is the organisation and realisation of the examination. Third Scrum is concerned with examination on a particular date that is carried out as one complete unit. Final product of this Scrum for the Scrum Master is the conclusion of the exam date (project assignments are graded, exam is conducted, grades are registered) and for the students, grades are admitted in their indexes while the course Business Application Development has been successfully passed.

Pregame phase

In the Product Backlog list main properties for the examinations are registered: planned date of the examination, requirement that students that want to take the exam need to register for it, requirement that students need to have their index with valid entry noting that they have applied for the course in the beginning of the academic year. One of the most important artefacts that is entered in the Environment Architecture is the list of rooms that are reserved in the appropriate time when the examination is planned to take place. The method of publishing examination details is organized at the level of the Faculty and thus is not a part of the model.

Game phase

Sprint Backlog list contains the information about the exam date, including precise time and place where the exam will take place, as well as the list of students that have applied for the examination. Duties of the Scrum Master at this point are to compile exam questions and exam tests and prepare adequate number of test. Students are expected to take the test at the published date, time and place of the exam, while providing their indexes for inspection.

Since the model covers only one exam date, Game phase consists of only one Sprint cycle. Sprint lasts for a predetermined time period after which students submit the completed tests and after Scrum Master grades the written exams, students that have achieved minimum requirements and passed the test continue with the Postgame phase of the Scrum unit. Rest of the students are returned to the Pregame phase of this Scrum unit and are referred to the next exam date as it becomes available.

Postgame phase

In this phase for the students that have successfully passed the written exam Scrum Master calculates the final grade for the course where the final grade is structured as follows: 50% of the final grade is made from the grade of the written exam, 40% is contributed by the project assignment grade and the final 10% is contributed by the student activity during the course. The final grade is admitted in student index and registered in the exam list of the Faculty Information System.

IV. FURTHER IMPROVEMENTS IN COURSE ORGANISATION

The presented teaching model introduces a number of advantages and benefits for the students but it also faces a few challenges that need to be overcome and allow for improvements. Limitations of the model do not influence the structure of the teaching model but influence the suboptimal improvement of the teaching process quality. Some of the most prominent limitations are:

- Restrictive number of hours of lecture and exercises for the rich content of the course topics defined by the curriculum
- Limited background knowledge of the students that take this course that may require additional effort from the students to successfully finish their project assignments
- Relatively small number of students that creates challenges while forming student teams.

In order to improve the quality of the teaching process and resolve at least some of the limitations for the Business Application Development course an increase of the number of hours was proposed. From the current 20 hours additional 10 hours were suggested. This proposal was accepted by the Faculty bodies, so the next academic year there will be 10 hours of lectures reserved for the theoretical introduction to the course topics and 20 hours allocated for the practical part of the course that covers the development of student projects as well.

Next possible improvement pertains to the organisation of student project topics. The key idea here is that entire generation of students would develop the same complex software application and all of the Scrum Teams would make their contributions. They would still be independent self-organising and self-managing, but each Scrum Team would have 3 to 5 members. The project assignments would be given by the Scrum Master to the Scrum Team leaders. Scrum Team leaders would be assigned with the role of Product Owner according to the model described earlier in this paper, and each Product Owner would be assigned one segment of the software application that would be developed my their respective Scrum Team. The only restricting factor in this case would be the complexity of the student projects since there is limited background and prior knowledge of the students about course topics and programming. This limitation can be somewhat alleviated by the increased number of hours of practical work in the computer lab and Visual Basic.

V. CONCLUSION

Paper clearly shows how teaching process can be customised according to methodology initially planned for realization of completely different kind of processes. Scrum is chosen due to its specific advantages in comparison to other agile methodologies: (1) Scrum is presented as an integral part of the course in the topic „Agile software development methods“; (2) teaching process, as well as Scrum, is achieved through the incremental development: through presentation of theoretical knowledge, practical work and student project

as well as examination; (3) theoretical knowledge and practical skills are presented and thought iteratively: topic by topic; (4) student projects are organised and developed according to Scrum methodology: roles are assigned; required computer equipment is defined as a part of Environment Architecture; features of software products being developed are entered on the Product Backlog list; software requirements are defined according to Product Backlog list and the team member responsibilities are delegated through the Sprint Backlog list; software product is developed incrementally and iteratively through a number of Sprint cycles until final version is developed; (5) project documentation is created and compiled.

In this paper some of the limitations and possible improvements of the teaching model are given that could further raise the quality of the current teaching process.

LITERATURE

[1] M. Broussard: The J-School Scrum: Bringing Agile Development Into the Classroom, available at: http://www.pbs.org/mediashift/2014/01/the-j-school-scrum-bringing-agile-development-into-the-classroom/

[2] H. F. Cervone: Understanding agile project management methods using Scrum, available at: www.emeraldinsight.com/1065-075X.html

[3] Development process, ProfIT Labs Ltd – Custom software development company, available at: http://www.profit-labs.com/development-process/

[4] M.E.Grimheder: Can agile methods enhance mechatronics education? Experiences from basing a capstone course on SCRUM, available at: http://www.sciencedirect.com/science/article/pii/S0957415813000044

[5] V. Mahnic: A case study on Agile Estimating and Planning Using Scrum, available at: http://www.eejournal.ktu.lt/index.php/elt/article/viewFile/372/318

[6] R. Pope-Ruark, M.Eichel, S.Talbott, K.Thornton: Let's Scrum: How Scrum Methodology Encourage Students to View Themselves as Collaborators, available at: http://teachigandlearningtogether.blogs.brynmawr.edu/archived-issues/may-issues/lets-scrum

[7] K. Schwaber, J. Sutherland: The Definitive Guide to Scrum: The Rules of the Game, available at: https://www.scrum.org/scrumguide?gclid=CPS9gJ7XubwCFUT3wgodv1gAag

[8] Scrum available at: http://en.wikipedia.org/wiki/Scrum_(software_development)

[9] Scrum Alliance, Core Scrum – Values and Roles available at: http://www.scrumalliance.org/why-scrum/core-scrum-values-roles)

[10] A. Soria, M.R. Campo, G. Rodriguez: Improving Software Engineering Teaching by Introducing Agile Management,13th Argentine Symposium on Software Enginneering, ASSE 2012, available at: http://www.41jaiio.org.ar/sites/default/files/271_ASSE_2012.pdf

[11] R.L. Upchurch, L. Williams. "In Support of Student Pair Programming." ACM SIGCSE. ACM Publications, March 2001.

[12] VersionOne, 7th Anual Survey: 2011, The State of Agile Development, Full Data Report, available at: http://www.versionone.com/pdf/7th-Annual-State-of-Agile-Development-Survey.pdf

Ethical aspects and capabilities of social networks – students, teachers, and Facebook in elementary schools

Denis Vincek
Osnovna škola Ante Kovačića (Primary school)
Vladimira Nazora 1, HR-49250 Zlatar, Croatia
Phone: +385-49-466 832 +385-98-223 746 Fax: +385-49-466 370
E-mail: denis.vincek7@gmail.com

Abstract - While social network friendships between teachers and students are forbidden in Great Britain, they are allowed in Croatia. However, there are no recommendations on the student/teacher communication by any of the relevant institutions. A study conducted in the Ante Kovačić Elementary school in Zlatar, Croatia in June 2013 shows that more than two thirds of students have at least one teacher as a Facebook friend. Even though this social network is a convenient medium for teachers to make the learning faster and more interesting to students, this potential is seldom utilized. The study shows that both teachers and students think that their friendship and communication on Facebook cannot get them into trouble with the school board or the parents and that this does not undermine the teacher's authority. Until there is no regulation on student/teacher relations on Facebook in Croatia, every school can solve this through creation of its own Social Media Policy. In the absence of Media Education, the IT teachers and the school librarians are those who should inform the students about the critical use of Facebook and the media by teaching Information literacy.

I. INTRODUCTION

Mark Zuckerberg, a psychology student, who was also a very crafty programmer, started several services for online networking for his university friends at a very early age. These networking websites connected the students that took the same university courses (Coursematch), or were used to grade the physical attraction of the students from the university (Facemash). At the beginning of 2004, or to be more precise, on February 4th 2004, he started Thefacebook (at thefacebook.com), the social network for Harvard university students. By March, this social network had spread to Stanford, Columbia, and Yale and later to other universities across the USA and Canada but it was initially intended for university students only.

Mark Zuckerberg was soon joined by two of his university colleagues, Dustin Moskovitz and Chris Hughes. The network became profitable and Mark and Dustin left school to completely dedicate themselves to developing this social network that was named Facebook in the summer of 2005, buying the *facebook.com* for 200,000 dollars. But more money was needed to develop it and soon several million-dollar investments came, including the one from Microsoft [1].

At the time, Facebook was still a social network open exclusively to students. In order to catch up with and surpass MySpace, Friendster, Bebo and other social networks, Zuckerberg abandoned the idea of Facebook being exclusively for university students and on September 26th 2006, he opened the door to anyone older than 13 with an e-mail address who wanted to register [2].

The term friend very quickly got a completely new meaning and today Facebook has over 1.15 billion users [3] and there are over 1.6 million of Croatian citizens registered on this social network [4].

Communicating through social networks such as Facebook opens up a wide array of possibilities. If we examine Howard Rheingold [5] who claims that the power of virtual society and computer networks stems from four unique features of computer communication (it is a medium that a number of people can use to communicate; it allows people to connect in new ways; the power of technology will increase through the focus on digital forms of the media; cyberspace, the overall connection between people through computer and telecommunications has solid foundations to become the basic platform, the starting point for innovations that will enable the entrepreneurs to find new use for the new media), we can deduce that the power of Facebook is enormous as this network can be used for communication between more than one billion people connecting in new ways – communicating through notifications (messages, photographs, video and audio clips…), comments and likes.

However, one should also bear in mind that each form of communication mentioned can be visible to the entire community and this means a large number of people. When dealing with communication between teachers and other individuals, between students and other individuals, between two teachers, between two students, or a student and a teacher, it can quickly become available to all other members of the community, such as other students, teachers, homeroom teachers, school principals,

parents… Not only is this communication instantly visible, but is available for a longer period of time, perhaps even forever or until one of the participants deletes it.

This work will show the situation among the seventh and eighth grade students (since younger students are not allowed to have a Facebook account) and their elementary school teachers with regard to their level of knowledge of Facebook, and answer whether the students are Facebook friends with their teachers, how they established that friendship, what is the nature of their communication and whether it can cause problems in school or life in general. In addition, we shall see whether and how Facebook can be used for educational processes. This paper will also show an example of a country in which the communication between students and teachers through social networks is forbidden and explain what mechanisms are used to regulate that.

This paper aims to confirm or disprove the following hypotheses:

1. Some of the elementary school students and their teachers in Croatia are friends on Facebook.
2. Facebook is underutilized in the teacher/student(s) communication in class.
3. Both the teachers and the students do not think that Facebook communication could cause problems for them with the school board or parents.
4. Both the teachers and the students think that these friendships and Facebook communication help improve the teacher's authority.

II. GREAT BRITAN, AN EXAMPLE OF A COUNTRY IN WHICH THE TEACHER/STUDENT COMMUNICATION ON SOCIAL NETWORKS IS NOT ALLOWED

In Great Britain, schools have regulations policies that forbid communication between teachers and students on social networks. Not only that, these regulations often stipulate that Facebook, MySpace, Bebo, Twitter, Youtube, Internet forums [6] … must not be used during class and even the school employees are unable to access them through computers as the school's access to these websites is often blocked through filters.

Some of these regulations on using social networks are brief and concise and contain only three pages [6], while others contain up to 12 pages [7] with a detailed description for each potential situation.

Granby Primary School has the Personal Use of Social Media Sites Policy [7] that states that the employees are held responsible for the content published on their social networks and that they have to be aware that everything that they post is publicly available. This Personal Use of Social Media Sites Policy explicitly states that in order to prevent any form of conflict of interests, the employees are not allowed to send the friendship requests to students or to accept their requests. There are some exceptional circumstances when the student in question is the teacher's son or daughter, but not even then can the friendship be accepted immediately as the school principals have to be consulted on this matter.

The teachers may not post any kind of information on social networks that would reveal the identity of students, students must not be mentioned in the discussions, nor should the photos or video clips posted include students or their homes. The teachers are forbidden to post any kind of information that could cause the school to get a bad reputation. The Granby Primary School's Personal Use of Social Media Sites Policy forbids its employees to express their views or opinions on social networks as the stance or attitude of the school and the local authorities. Usage of any kind of offensive, discriminating, or in any way inappropriate language is forbidden as is disclosing any kind of confidential information.

The teachers should be aware that posting the photos on social networks that are visible to a large number of people can be used to establish where and with whom they work, and this could increase the risk of false accusations and threats. In addition, the social network sites may enable the identification of children or vulnerable adults which could even put the safety of the students at risk.

The policy regulations state that the teachers should be aware that students finding inappropriate photographs of their teachers posted on social networks could cause an embarrassing situation for the teacher and harm not only their professional reputation but the reputation of the entire school as well. This could also reveal the place of residence of the employees and this could have further implications on their individual safety. Because of this, the first and the basic guideline for the employees is to evaluate every information before publishing them on social networks and to set their privacy settings.

The Granby Primary employee groups, e.g. Facebook groups, are allowed, but the founders of the group are held responsible for all published content. Breaching any part of the guidelines results in disciplinary measures.

When it comes to any form of use of social networks in class, Social Media Policy and Use of Mobile Phones and Digital Photography Policy, enforced by the Worfield Endowed CE Primary School [6] states that the school principal has to be asked for permission before any kind of use.

Ambleside Primary School's Social Media Policy [8] is rigorous even when it comes to former students and states that the school employees are not only forbidden to be friends with the current students but with all former students as well until they turn 18.

III. CROATIA, AN EXAMPLE OF A COUNTRY IN WHICH THE TEACHER/STUDENT COMMUNICATION ON SOCIAL NETWORKS IS ALLOWED

CARNet has created a proposal for school safety policy named "Acceptable use of computers in school" [9], but in terms of using Facebook and other social networks with regard to student/teacher or teacher/teacher communication, there is no formal or informal

recommendation from the relevant institutions. Thus, everything is left at the liberty of schools, principals, or to be more precise, the teachers themselves (the only teacher that has dealt with this issue in more detail so far is Miljenko Hajdarović [10]), as one of the crucial factors in class. In other words, everything that is not forbidden is allowed. However, one should bear two things in mind regarding this communication.

A. *The students and the teachers are always a part of the didactic triangle "teacher – content – students"*

First of all, the teachers must never forget their role (in class). They are one of the three factors that make up the so-called didactic triangle [11]. The teacher is an expert that organizes an efficient educational process and the class itself by teaching the students. The teachers should always (indirectly) let the students know that they, like the second factor of the class, are acquiring education through systematic teaching and individual learning. The third factor of the didactic triangle is the teaching or educational content, and this is what might connect the teachers and the students in communication through social networks. This exact type of communication as the special factor of education that is vital for participating in social networks emphasizes the need for information technology which turns the didactic triangle into a didactic quadrangle

B. *Many-to-many and the rules of conduct on Facebook*

Secondly, unlike the old, classic media with expensive technology that included the so-called "one-to-many" communication, Facebook allows you to have a meaningful "many-to-many" communication [5]. This is possible as you do not need expensive technology to become a user of this social network. This fact implies two things.

Facebook is easily available to all users, even primary school students. Most students have personal computers or laptops that have fast Internet connections (ADSL) with flat rate access. A lot of them have an ADSL device that enables them to have wireless Internet connection with a satisfactory signal throughout their entire home and this allows them to be constantly connected to Facebook using either laptops or smart phones that can be purchased at very reasonable prices.

The following aspect is even more important. The classic media, e.g. the newspapers, radio, and television, are centralized. The new media, including Facebook as a social network, generally do not have their own journalists, producers, directors, or editors, that is, persons that would control which notification is appropriate, polite etc. And it is among children that this could cause a problem as they are at an age at which they could, intentionally or unintentionally, cross the line either as a creator or as a receiver of inappropriate messages.

Therefore, it would be good to adhere to general rules of behaviour while online, that is, to behave according to the code of desirable conduct on the Internet, the so-called netiquette [12] that originally stems from Usenet and has now spread onto the entire electronic communication [13] as is the case with Facebook. This social network has whole array of statements, standards, and rules [14] that one should read before creating a personal profile and that one is supposed to adhere to, and particular attention should be paid to setting the level of protection in privacy settings.

Despite of everything, one must always bear in mind that Facebook (much like the rest of the Internet) contains content not appropriate for children (impolite or offensive). The USA tried to solve this problem in 1996 through legal means as a part of the telecommunications reform when the American Congress passed the Communications Decency Act [5], but the Supreme Court overruled this act as unconstitutional, claiming that Internet communication cannot be limited to children only.

IV. METHODOLOGY

The method for collecting data chosen for this study was the survey method since it is the most convenient method for descriptive and causal studies and also because the costs are minimal. The "distribute and collect" survey had 12 questions mostly closed-type questions that were formulated relatively simple and created so as to raise interest of the surveyed students as one should bear in mind that the participants in hand were children. The surveys were filled out in June 2013 by the students from seventh and eighth grades and Ante Kovačić Elementary School in Zlatar was chosen as a sample for the study.

A total of 44 out of 47 seventh-grade students (94 %) filled out the questionnaire as did 64 out of 65 eighth-grade students (98 %). This means that the data regarding the use and familiarity with this social network was collected from 108 out of 112 students (96 %) older than 13 that can be legally registered on Facebook.

The survey questionnaire that the teachers and school associates (school counsellor, school psychologist, librarian) filled out contained 14 questions and was structured similarly to the one given to the students and 31 out of 35 school employees filled it out (89 %).

V. RESULTS AND DISCUSSION

A. *Students and Facebook*

Out of 108 students that took part in the survey, 102 of them have an active Facebook account. As much as 98 per cent (43) of seventh-grade students and 92 per cent (59) of eighth-grade students are Facebook users.

The question "How would you rate your knowledge and usage of Facebook?" had four possible answers and no students chose the answer "none at all". While the eighth-grade students rated their knowledge of Facebook with high marks (36 students chose "excellent", 23 students chose "good"), four seventh-graders said that their knowledge of Facebook is "poor" while a similar portion of students chose "excellent" (19) and "good" (20).

The question "Do you know how to set the privacy settings?" yielded similar answers. All eighth-graders

chose "YES" while only two seventh-graders answered "NO".

The explanation for better familiarity and skills of eighth-grade students lies in the fact that their age allows them to have more experience with social networks as does the fact that the students cover the topic of privacy settings in school library and homeroom classes that deal with information literacy.

As much as 70 per cent (30) of seventh-grade students and 71 per cent (42) of eighth-graders have at least one teacher or other school employees (counsellor, psychologist, librarian) as a Facebook friend. Of those students, 70 per cent of them communicate with their teachers through Facebook (likes, comments, notifications).

The survey showed that students are not afraid that their communication with their teachers could cause any problems with the school board or parents. Some 77 per cent of seventh-grade and 76 per cent of eighth-grade students are not afraid of that scenario (Figure 1).

All friendships between the students covered by this survey and their respective teachers were established solely on the request of students.

When asked "Do teachers lose some of their authority with virtual friendships with students?", 69 students (67,6 %) answered "NO", 17 students (16,7 %) claimed "YES", while 16 students (15,7 %) think that teachers actually gain authority through this (Figure 2). The percentages are similar for students from both grades, and for students who are friends and who are not friends with their teachers. The only exception is that one portion of eight-graders that are friends with their teachers (29 %) believe that the friendship increases the authority of the teacher.

This is a very important piece of information as we carry our teachers in our memory as the living reminder of the time when we were formed as persons. The authority is the first trait that makes good teachers as it gives them credibility and sets them as the firm anchorage for the students [15]. The eighth-graders see this as a sign of maturity from their teachers.

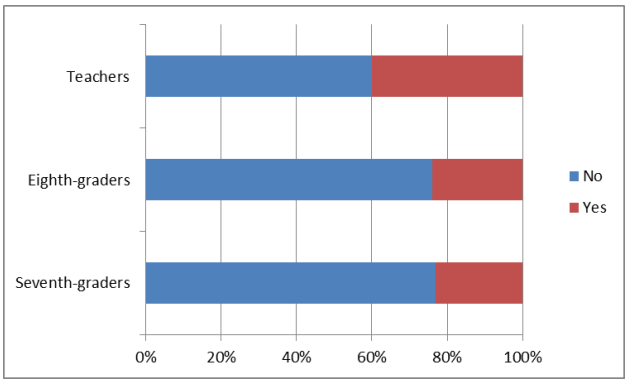


Figure 1. Do you think teacher/student relations could cause problems for teachers/students with the school board or the parents?

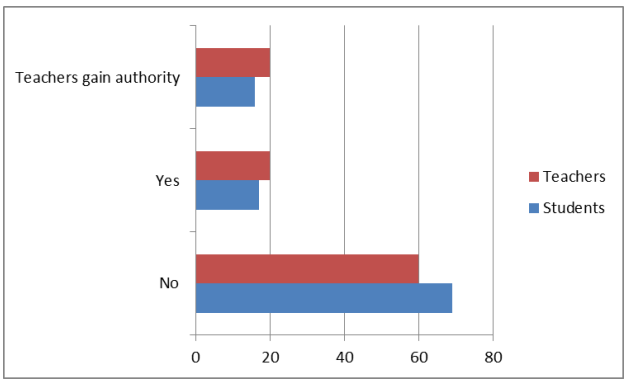


Figure 2. Do teachers lose some of their authority with virtual friendships with students?

When it comes to using Facebook for school needs, the survey included the question "Have you ever used Facebook for education?" The percentage of students who have used it is 65 per cent for seventh-graders and 81 per cent for eighth-graders.

When asked about their school use for Facebook, the students provided the following answers: "I ask my colleagues about my homework, I ask for help regarding homework, I check if I did my homework right, I practice for my tests, I ask other students what I should focus on when studying, we have a class group and we are talking about which high school to choose, Facebook can only be good, I broaden my horizons, we talk about creating posters and preparing presentations, we discuss on how to improve the relations in class, I ask about the things I don't understand, we talk about the way the school is organized when the school has a project day the day after, I ask about what we've covered in school if I was sick, I ask my teachers to explain what I don't understand, I practice my foreign language by talking to my friends from other countries."

B. Teachers and Facebook

The survey was filled out by 31 out of 35 teachers and school associates (school counsellor, psychologist, and librarian) in Ante Kovačić Elementary School, that is, 89 per cent of employees. Only 12 of them have an active Facebook account which means that only 39 % of them are users of this social network. When looking at their age, we may see that 67 per cent of school employees younger than 40 have a Facebook account, while only 21 per cent of employees older than 40 have a Facebook account.

They also had to rate their level of familiarity with Facebook. Half of them rated it as "excellent", 42 per cent as "good" and only one of them rated it as "poor". The teachers younger than 40 rated their skills with Facebook higher than their older colleagues. All school employees said that they know how to set the privacy settings.

Only 42 per cent of school employees that have their profiles on Facebook are Facebook friends with their students regardless of their own age. They established their friendship exclusively when requested by their students and all teachers that are Facebook friends with

their students use it to communicate with them (likes, notifications, and comments).

The question "Do you think communication with your students could cause problems for you with the school board or the parents?" yielded conflicting answers from the teachers. As much as 40 per cent believe there is such a risk (Figure 1). However, 60 per cent of school employees think that they do not lose any authority through this communication, 20 per cent state that they gain some authority and the remaining 20 per cent said that they lose some of their authority (Figure 2). The answers to this question are encouraging if we consider role the teachers have in terms of accepting desirable modes of conduct by their students.

The teachers stated that they use Facebook in communication with their students to answer to their questions regarding school content, homework, tests (only when students are not sure about something) or to answer any questions students may have regarding school-related activities and the organization of classes.

However, no answers were given regarding the teacher/parent(s) or teacher/teacher(s) communication in this survey. Facebook could be in such a way that the teachers send the parents various notifications regarding school activities, organization of classes due when special events are planned in school, the students' obligations, field trips and related issues. Facebook could also be used to connect all teachers within a school, teachers for certain groups of classes or teachers of same classes from neighbouring schools to exchange the preparations for class, discussing classes, and sharing publications or Internet content used, forwarding information regarding professional seminars and exchanging ideas, experiences, advice and suggesting solutions to various problems. All three types of communication (teacher/student(s), teacher/parent(s), teacher(s)/teacher(s)) is best done within closed groups.

When asked "How would you react if you see inappropriate content on your student's profile", the teachers said they would: "talk with the student, warn the student about the inappropriate behaviour, notify the school associates (school counsellor, school psychologist, or school librarian), notify the parents if I were his/her homeroom teacher, indirectly warn the student – so that everyone in class can hear".

VI. CONCLUSION

The chalk and blackboard are slowly becoming a thing of the past, overhead projectors are also not seen in classes very often and mass media are not bad at all if the teachers know how to use them properly.

The study shows that more than two thirds of elementary school students have at least one of their teachers as a Facebook friend. Considering the fact that almost all students have Facebook accounts, this social network represents a convenient medium through which the teachers could make learning fun but also make the process of understanding the teaching content faster and easier: from posting useful and additional materials regarding the lesson covered in class, reminding the students about their homework and other school obligations, discussing the topics from homeroom classes to working on class, school, or even inter-school/international projects or notifying the students about the results of their exams – of course, all through closed groups or class pages. However, the teachers very seldom utilize the potential of this social network, although with some rare exceptions [16, 17]. This is also reflected in the fact that less than one half of teachers that have Facebook accounts have students as their Facebook friends even though there is no obstacle against that in Croatia.

After all, the study shows that both the teachers and the students do not think that their mutual communication through Facebook could get them into trouble with the school board or the parents. The percentage of students who think this way is 77 %, while 60 % teachers feel the same way, and it would be good for some rules of conduct in teacher/student relationship to be established. Adults are accustomed to act according to certain predetermined patterns, and written recommendations or rules would create the space for the teachers to feel safer in their communication with students on Facebook and/or would increase the number of teachers that have their students as Facebook friends. More teachers in this communication would pave the way towards new ideas for using Facebook in class. Creativity can only happen where there is enough freedom within certain boundaries.

When it comes to authority, the study shows that Facebook friendship and communication between teachers and their students does not diminish the authority of the teacher. This is particularly important as the job of a teacher is public and they have to pay attention to their general and professional competence and their authority and general personal profile at all times. The authority is based on reciprocity and mutual respect and love and this is the crucial aspect of every humane [18] relationship, especially in education. In this respect, the teachers have to be careful to avoid any superficial, confusing, and especially inappropriate communication. The teachers have to be aware that their every post (texts, photos, video clips) can instantly become visible to their students. However, one could set the "custom" privacy settings so that the students cannot see any of the posts made by the teacher in question. The alternative is to create two profiles: the official one (through which the teachers communicate with their students) and the private one. However, this is not in accordance with the Statement of Rights and Responsibilities [19] that stipulates that you must not create more than one personal account. When it comes to communication through private messages, the teachers should be as formal as possible and not erase the messages or conversation from their Inbox.

There is no doubt that social networks represent an important means of communication, but the fact remains that the legislature is not keeping up with this, or to put things mildly, it has not done enough. It is hard to believe that we might get any kind of regulations or recommendations for teacher/students Facebook interaction in near future here in Croatia, considering that the media were mentioned only six times in the National

Youth Action Program from 2002, and only three times in the National Youth Action Program from 2009 to 2013, while the draft of the Strategy for Education, Science and Technology from 2013 does not even mention the media, media literacy, or the media culture [20] even though the media will be the school of the future [21] and even though the 2007/65/EC and 2010/13/EU directives require that EU member countries include media literacy and Media Education into their systems of education [22]. As long the situation is like this, it is up teachers to choose whether or not to befriend their students on Facebook.

Until the relevant institutions create regulations or recommendations for teacher/students Facebook relations, a possible solution for each school is to bring their own social media policy which would clarify that friendships between teachers and students are allowed and that it is up to the teachers to choose whether or not to befriend their students on Facebook, which means the teachers are those who accept and not those who send friend requests. The teachers can show their proactivity through closed groups or class pages in which friendship with students is not a condition for communication.

Since the teachers know almost all the students in their schools and know which class they attend (if they are uncertain, they can use the E-Matrix to check [23]), the social media policy should also contain a statement which clearly indicates that the teachers are not allowed to accept friend requests from the students under the age of 13, accept such students into groups or create class pages for students under the stated age.

The policy should also regulate the procedure in case of inappropriate behaviour from one of the students, which can be a warning, either public or via private Inbox message, with the request that the student in question immediately deletes the inappropriate content, or the notification for the homeroom teacher, school counsellor or school psychologist. Inappropriate communication can stem from the fact that the students of today are media literate, but only in the technological sense and not in the critical sense [20].

In the absence of Media Education, the teachers without adequate training are forced to teach children about the media. This void could be filled by IT teachers or the school librarians who could talk about the critical use of Facebook as a part of their Information literacy programme.

The school social media policies should open up the possibility of using social networks in classes through class pages or closed groups.

In addition, the social media policy should be used to emphasize that the teachers, due to their role in the intellectual, moral, and social development of the children, are held responsible for the content published on their social networks and should conduct themselves on Facebook the way they behave in class with students, or in their life outside of school, which primarily means avoiding discussing school and work and mentioning the students in their posts on Facebook.

REFERENCES

[1] D. Vincek, Fenomen društvenih mreža : utjecaj Facebooka na kretanje broja posjetitelja internetskih foruma : diplomski rad [na Odsjeku za informacijske i komunikacijske znanosti Filozofskog fakulteta Sveučilišta u Zagrebu]. Zagreb: D. Vincek, 2011.

[2] D. Zeevi, "The Ultimate History of Facebook." Internet: http://socialmediatoday.com/daniel-zeevi/1251026/ultimate-history-facebook-infographic, Feb. 2, 2013 [Sep. 29, 2013].

[3] D. Cohen, "Facebook has 1.15B monthly active users, but how many are invalid?" Internet: http://allfacebook.com/form-10q-invalid-accounts_b122538, Jul. 26, 2013 [Sep. 29, 2013].

[4] "Gdje ima najviše korisnika Facebooka u regiji." Internet: http://metro-portal.hr/gdje-ima-najvise-korisnika-facebooka-u-regiji/76533, Feb. 15, 2013 [Sep. 29, 2013].

[5] N. Prelog, Novi mediji : skripta [za studente novinarstva na Fakultetu političkih znanosti]. Zagreb: N. Prelog, [2000?]

[6] "Social media policy and use of mobile phones and digital photography policy." Internet: http://www.worfield.shropshire.sch.uk/uploads/files/social_media_policy_october_2012%5B1%5D.pdf, Oct, 2012 [Sep. 30, 2013].

[7] "Personal use of social media sites." Internet: http://www.granby.leicester.sch.uk/PDF/documents/POLICIES/STAFF/Personal%20Use%20of%20Social%20Media%20Sites%20policy.pdf, Aug, 2012 [Sep. 30, 2013].

[8] "Social media policy." Internet: http://www.amblesideprimaryschool.co.uk/SiteData/Root/File/Policies/social%20media%20policy%20latest.pdf, [Oct. 2, 2013].

[9] "Prihvatljivo korištenje računala u školi." Internet: https://sec.carnet.hr/assets/Uploads/Prihvatljivo-koristenje-racunala-u-skoli.pdf, Jan. 29, 2013 [Oct. 6, 2013].

[10] M. Hajdarović, "Preporuke nastavnicima za odgovorno korištenje Facebooka." Internet: http://hajdarovic.com/wp-content/uploads/2013/08/preporuke_facebook.jpg, Oct, 2013 [Oct. 7, 2013].

[11] V. Poljak, Didaktika. Zagreb: Školska knjiga, 1985.

[12] S. Hambridge, "Netiquette guidelines." Internet: http://www.ietf.org/rfc/rfc1855.txt, Oct, 1995 [Sep. 29, 2013].

[13] D. Vincek, Internetski forum kao medij : seminarski rad [na Odsjeku za informacijske i komunikacijske znanosti Filozofskog fakulteta Sveučilišta u Zagrebu]. Zagreb: D. Vincek, 2010.

[14] "Uvjeti i pravila korištenja usluge Facebook." Internet: https://www.facebook.com/policies/, [Sep. 29, 2013].

[15] V. Andreoli, Pismo učitelju. Zagreb: Školska knjiga, 2006.

[16] D. Drenjanac, C. M. Dumančić, "Facebook u obrazovanju," in MIPRO 2013 proceedings [DVD]. Rijeka: MIPRO, 2013.

[17] I. Kustura, "Smiju li biti prijatelji : učenici i učitelji sve češće 'frendovi' na Facebooku," in Večernji list, Sep. 2, 2013, p. 10.

[18] M. A. Peuc, "Kriza društva – kriza škole – kriza odgoja : učiteljski autoritet danas i sutra," in Školske novine, br. 27, Sep. 17, 2013, p. 11

[19] "Izjava o pravima i odgovornostima," Internet: https://www.facebook.com/legal/terms, Dec. 11, 2012 [Oct. 7, 2013].

[20] I. Kanižaj, "Izloženost, povjerenje i manipulacije: mediji u svakodnevici djece i mladih," presented at 4. okrugli stol za školske knjižnice, Sisak, Oct. 3, 2013.

[21] Z. Miliša, "Posljedice tabloidizacije medija na svijest i ponašanje djece i mladih," presented at 4. okrugli stol za školske knjižnice, Sisak, Oct. 3, 2013.

[22] I. Kanižaj, "Zašto u Strategiji nema medijskog odgoja," in Školske novine, br. 28, Sep. 24, 2013, p. 8.

[23] "Lista učenika – Osnovna škola Ante Kovalića," Internet: https://matice.mzos.hr/Ucenici/ucenici.aspx, [Feb. 14, 2014]

Computer Classroom Operating System Security (Windows)

Mario Zovkić*, Tedo Vrbanec**
*Primary school Barilović, Barilović & Primary school Katarine Zrinski Krnjak, Krnjak, Croatia
**Faculty of Teacher Education, University of Zagreb, Zagreb, Croatia
*mario.zovkic1@skole.hr, **tedo.vrbanec@gmail.com

Abstract - Computer security, school security policy and school computers usability should be important for all school computers and network resources users. But instead of implementing security policy, users often unintentionally compromise computer functionality and security out of ignorance or negligence. Since Windows is the default operating system in the majority of Croatian schools, students often have to use computers in the role of administrator. Bearing in mind the multitude of new technological solutions and continual changes which happen in this industry, this paper tries to provide an overview of software solutions for IT administrators of school computers, which are also applicable in libraries, colleges, universities and other public places which use computers with Windows OS.

Keywords: Classroom, Security, Windows, Policy.

I. INTRODUCTION

Windows is the *de facto* standard operating system for desktop computers in Croatian schools. It is in IT teacher's best interest to keep the maintainance of computers as simple as possible and to require as little interventions as necesarry during classes. Because of that, and due to ignorance or negligence, the configurations in computer classrooms are mostly intended for the users in the administrator role, regardless of who is logged in: a teacher or a student. When students are logged in as administrators, they have the freedom to do whatever they want. As they are curious by nature, they like to explore on their own to see what they can change on the computer or what they can access and make all sorts of mischief: change the desktop backgrounds and screen savers, delete files from the computer, install programs and games and use the installations to add different toolbars in the web browser. For children who have an advanced knowledge of computers, possible limitations and restrictions represent a kind of challenge. In this context, we will show some guidelines for increasing the operational security in computer classrooms where Windows OS is installed, but these guidelines are partially applicable to other operating systems as well.

II. OPERATIONAL AND PROGRAM SECURITY

"Operational security includes two aspects of computer security; the first represents the ways of increasing the awareness of possible actions among the potential victims, while the second represents the ways in which computer criminals can be prevented from committing these acts". [1] It should be made clear to the employees that security is threatened and that everyone shares the risk and the responsibility. "The only way in which operational security can function completely is by implementing operational security in the programs of physical insurance as well as the personnel security and communications within an organization. In fact, the operational security is primarily used as an aid in making those security programs as productive as possible." [1]. This could be argued. Namely, the authors of this paper believes that operational security, apart from the education of users, is the most important element of the security system of an organization, as well as of a person as an individual or as a part of an organizational structure. However, it is perfectly obvious that an organization can never stop searching for vulnerable spots, breaches of security and methods of prevention, and can never stop discovering successful breach attempts, all this along with continual training of employees and users. [1]

"There are two types of systems: open and closed ones. If you use open source software you can analyze and inspect the code of security mechanisms and practically be sure that it will do what it is designed to do and nothing more. That means that you can check it for existence of trapdoors and backdoors, and in the end be sure that it will not send any data to manufacturer's servers (it is really not important what it sends, it is important whether it sends the data without your permission – this is how your privacy gets violated). If you use closed source software, the only thing that can help are patches and service packs, but you will not have a guarantee of the data leek (except for the manufacturer's promise)."[2] A recent discovery in the computer industry is the "cloud computing", which is [3] "a technology that keeps up data and its application by using the internet and central remote servers". According to Harauz and Kaufman [4] "we have the ability to utilize scalable, distributed computing environments within the confines of the Internet." Regarding data, i.e. data security, the safety of these systems is questionable: "So let's pause to consider the alter ego of the cloud as panacea: the cloud as a platform for malice. How bad could it get? [5]" If we look at the system administrator's side, it is an excellent solution, because the other side takes care of the security, integrity and maintenance of the computer. There are global players, and in Croatia a local ISP, "Croatian Telecom", started offering their Cloud computing service at the end of May 2013, where the user is given the right to use their computer. The user is shown a desktop with the applications that the ISP offers and the user immediately has a functional computer without the necessity to maintain it or worry about its security. [6]

One of the fundamental postulates of operational security is to create a user account that doesn't have administrative authorities (limited user). In Unix-like operating systems that is the only possibility, and completely normal, but in Windows OS it is an option that makes the daily use very complicated. Very often, the users change the settings of the operating system without even being aware of it. Any need for intervention or

repairing from the administrator, usually an IT teacher, is a waste of valuable time during classes or leads to unpaid working hours of the IT teacher in the role of system administrator. Considering the installed OS – Windows, the computers in Croatian schools often have inadequate hardware capacity. Apart from that, it is a fact that Windows OS works more and more slowly as more time passes from the day of the installation (problem with Windows Registry). Altogether, computers are slow, and become slower with time, which increases the amount of time that needs to be spent on them. In the modern world, which is turbulent and accelerated, it is important to minimize the amount of time necessary for computer maintenance, i.e. to strive towards minimizing that time by automatization. There are on average 19 computers in Croatian computer classrooms [7]. Based on that, we can approximately calculate the necessary time for their minimal maintenance, providing that there are no unforeseen situations. It is necessary to update the OS monthly, which takes at least ten minutes, depending on the speed of the computer, speed of the administrator and the internet connection. It equals a total of four classes a month, if we don't count the updating of user programs, control programs and additional programs; such as Java, Flash, Shockwave, programs for pdf files, video and audio files, etc., and sometimes the updating of the latest versions of antivirus programs. Web browsers are the favorite targets of attack on computers. Therefore it is important that they, as well as their add-ons and plugins, are independently updated. Web browsers such as Firefox, Google Chrome and Opera can be independently updated, but they require administrative authorities. Internet Explorer usually requires an intervention from the administrator during the installation of a newer version.

Servers should be equipped with smart UPS in case of an unexpected electricity loss and voltage stabilization, in order to keep the public services alive, and in order to protect the hardware from damage with controlled shutdown, in case the power supply loss lasts longer. It is desirable to protect the BIOS of personal computers with a password, because although the password is easy to cancel, it is necessary to open the PC case, which slows down and discourages the potential attacker. Safety copies of the data which is important for the school should be kept on a dislocated medium. The likelihood of a flood in schools is minimal, but the likelihood of a fire is greater. Permanent copies should be made at least once a year at the end of every school year. Safety copies of the data should be made depending on the importance and the amount of the generated information, and for critical data, safety copies should be made automatically, in real time.

III. Access to Operational Security

There are two basic types of computer classrooms in Croatian schools: thin client with or without multiseat technology and PC/laptop based classrooms. In a thin client approach, one or more powerful servers provide the central intelligence and control center for the security system. The end-user terminals display the graphical interface and process input from the mouse and keyboard. Most of the intensive processing happens on the server(s). It is also possible to start up the OS of the computer through the network using the Open Thin Client OS (http://openthinclient.org/).

Multiseat systems resemble thin-client systems with client interfaces (mouse, keyboard and monitor) attached directly to the computer that does the processing. That computer runs a program or operating system capable of managing multiple user sessions at the same time. Userful Multiplier and Windows Multipoint Server 2012 are two programs that fall into this category. PCs are computers that only one user can work on at a certain point – the one that is logged in to the operating system at that point. Locking the computers (which we will explain below) has its advantages:

a) there is less likelihood that a computer will pick up a virus or some other malware,
b) the administration of the computers is easier, (which leaves more time for other things),
c) there is less disk fragmentation due to constant installation/uninstallation of applications,
d) there are fewer (non-standard) applications which represent a safety hazard.

We will suggest several ways in which a satisfactory level of OS computer security can be achieved.

A. Group Policies for Various User Groups

These group policies reinforced with passwords make a good start. For example, administrators have all the rights (installation/uninstallation of programs; adjusting, adding and removing hardware), while regular users have the right to start an internet browser, office applications, etc. This approach has its limitations, because it is necessary to create several user groups and more access levels (for example, student, teacher, administrator), and these groups need to be administrated. Nevertheless, using group policies is much more efficient if the computers are cloned, or even better, if a domain group policy is used so it can be easily controlled by the domain controller.

B. Lockdown Software

This type of program enables the administrator to determine which devices users can access, which programs they can start, which folders they can see and access (for example, hiding the control panel and certain programs). These programs are not demanding, i.e., they don't require a large amount of knowledge or a centralized server. Depending on the computer classroom configuration and the program options, the administrator can control all these setting from one central computer. Disk protection programs that automatically delete all changes on the disk upon the next computer start up, user log out, or some other setting of the program, also fall into this group. Using this approach, a lot of people will think, "Why bother using antivirus protection?" But bearing in mind the increasing ingenuity of malware programmers and the fact that viruses keep getting more and more complicated and hard to discover, installing an AV solution for a greater computer safety is still recommended. Also, by updating, you don't only install patches, but also newer versions of programs (for example, Internet Explorer), as well as various improvements. Although the installation of an antivirus solution decreases the total available performance of the computer (which is sometimes extremely important), it is recommended to have an active and updated antivirus software. Each computer should be adjusted so that it

automatically downloads virus definitions. These are some software solutions for locking the settings: Fortres Grand Fortres 101, Deepfreze, Faraonics WINselect, KioWare, Rollback RX, Faraonics DeepFreeze and Centaurion Technologies DriveShield.

Deskman program [8] has one button to lock down workstations. It has the kiosk mode which tightly secures desktops and locks computers and it also has the ability to set up custom restrictions. It can freeze an application to create a safe environment and stop unwanted applications.

Deep Freeze [9] allows you to restore a PC on every restart. It has a virtual drive, an on/off protection switch so if administrators need to make any updates on a deepfreeze workstation, they have to turn off the deep freeze protection which requires a system restart.

When compared to above programs, Drive Vaccine [10] does not need to be turned off in order for any updates or software installations to be made, and protects the Master Boot record. Below we present a table with an overview of possibilities of particular programs:

TABLE 1. COMPARISON OF COMPUTER LOCK DOWN AND SYSTEM RESTORE PROGRAMS

Type of program	Name	License	Limitations / requirements	Publisher(URL)
System restore or disk protection	Deep Freeze	Commercial	Linux version only supports the SuSE Linux Enterprise Desktop, 10% free hard drive space	Faronics(1)
	Fortres Grand Clean Slate	Commercial	There is no version for Windows 8 and Linux	Fortres Grand(2)
	RollBack Rx	Commercial	There is no Linux version	Horizon DataSys(3)
	Drive Vaccine	Commercial	There is no Linux version	Horizon DataSys(4)
	Returnil System Safe 2011	Free for non-commercial, personal home use	Does not work on Windows 8, does not have Linux version	Returnil Virtual System(5)
Lockdown PC	Toolwiz Time Freeze	Freeware	Does not support Linux	Toolwiz(6)
	HDGuard	Commercial	Does not support Linux	RDT – Global(7)
	Faraonics WINselect	Commercial	Does not support Linux	Faraonics(8)
	SiteKiosk 8	Commercial	Does not support Linux	Provisio(9)
Application Launch Restriction	Windows 7 UAC	Commercial	Requires Windows 7	Microsoft

Table legend
(1) http://www.faronics.com/en-uk/products/deep-freeze/standard/
(2) http://www.fortresgrand.com/products/cls/cls.htm
(3) http://www.horizondatasys.com/products_and_solutions.aspx?ProductId=1
(4) http://www.drivevaccine.com/
(5) http://www.returnilvirtualsystem.com/returnil-system-safe
(6) http://www.toolwiz.com/products/toolwiz-time-freeze/
(7) http://www.hdguard.com/
(8) http://www.faronics.com/en-uk/products/winselect/
(9) http://www.sitekiosk.com/SiteKiosk/Default.aspx

If a school has a "server-client" logical network configuration, the administrators can control permissions from one computer. In case of a peer-to-peer network, the best solution is to clone the adjusted computer. When you clone the Windows OS (with the exception of the sysprep option for Windows 7), it is almost always necessary to have the same or quite similar (chipset) configurations of the cloned computers because of HAL, which is considered to be the driver for the motherboard and allows instructions from higher level computer languages to communicate with lower level components, but prevents direct access to the hardware. [11]

C. VHD (Virtual Hard Disk)

A virtual hard disk is a file that encapsulates a hard disk image. [12] It may contain what is found on a physical HDD, such as disk partitions and a file system, which in turn can contain files and folders. It is typically used as the hard disk of a virtual machine (VMware, Oracle VM VirtualBox, Windows Virtual PC, Microsoft Hyper-V Server). VHD contains the disk image which is loaded into the memory of the computer and all changes can later reflect on the OS itself (or not, depending on the options), so a layer is created between the computer resources and the program itself and there is no danger of compromising the security and integrity of the OS which is installed on the physical disk. It is recommended to use the native Windows 7 boot because the virtualization program still puts a burden on the computer.

D. Live CD/DVD

A live CD or a live DVD is a completely bootable computer operating system which boots from a CD, DVD or a USB and runs in the computer's memory. Live USB flash drives have the ability to write changes back to their bootable medium.

IV. WINDOWS SPECIFIC ISSUES

Windows XP is an incredibly resilient OS and is used by the majority of teachers in schools, but for which the support expired on April 8, 2014. Usually, the computers have a version of Windows already pre-installed on them, so the installation of a newer version would either be a license violation, or the right to an OS upgrade would need to be bought. **This is where the responsible government department comes in and mostly makes (pays for) an agreement with Microsoft that allows "a free" OS upgrade for the computers that already have a license, which is renewed every 3 years.** However, the Windows OS upgrade often does not end adequately, and it is necessary to reinstall the entire OS and all the required program support. But using outdated computers [7] that aren't capable of **running** newer OS versions eliminates any possibility for change within the Windows domain, which is notoriously demanding regarding hardware. Also, OS upgrade is a rather time-consuming job for an IT teacher, even if it is just for one computer, so it's normal that they are usually not thrilled about doing it. Alternatively, a version of Linux OS could be installed, but that brings many other problems (mostly subjective, due to user habits, and to a smaller extent objective), that could all be summed up as lack of time and/or desire to change.

As a result of the agreement that was signed between the Department of Science, Education and Sports and the Microsoft Company in June 2011, about licensing Microsoft software for primary and secondary schools, as well as for higher education institutions and public science institutes - „Microsoft Campus and School/Academic Agreement", all primary and secondary schools encompassed by the agreement have acquired the right to use licensed Microsoft software that can be downloaded from the web page: https://msdc.skole.hr/. Also, in November 2013, the Minister of Science, Education and Sports Željko Jovanović, Ph.D. and the Chief Executive

Officer of Microsoft, Steve Ballmer, signed the "Education Transformation Agreement" (ETA) with the aim to improve the Croatian education system [13] which, among other things, includes free use of Office 365 system for about 550 thousand students, college students, teachers and professors. There is also an interesting possibility for the IT teachers who have started using Windows 7 (but available and limited only for the Enterprise and Ultimate versions [14]): creating VHD disks and starting the computers from those disks. VHD is a format in which the disk is virtual. It contains everything that is on the hard disk of the computer (partitions and files) and it is used as the physical disk of a virtual computer. It enables the computer to set it up and to boot from the OS that is stored on it. This technically demanding principle is made simpler for teachers by selfless enthusiasts: Mark Minasi and his "Steadier State" [15], and the Finnish programmer Sami Laiho and his tool "Wioski" [16]. Wioski is a "self-healing kiosk-computer running Windows 7/8". It is a "wrapper that surrounds your own WIM-file of Windows 7/8 OS installing it on a VHD-file and using a differencing disk for writing incremental data." [16] Steadier State "freezes the computer as an image and creates a snapshot of the current activity as a virtual C:\. Restarting gives the option to roll back or keep the new snapshot." [15]

Windows 8 shows great functionality in the form of "refresh" and "reset" functions. [17] "Refresh" preserves user settings, user data, and applications bought through the Windows store. Everything else is removed and restored to defaults. Reset purges all applications and data, and reinstalls the operating system reverting it to a brand new fresh installation which requires reentering a license key and performing initial setup when completed. Programs that check whether a particular program has a newer version (such as Secunia PSI) can be helpful with computer maintenance. Another interesting possibility is using one program to start the installation of many others, such as InstallPad (http://installpad.philisoft.com/), where we first need to enter the URL addresses from which the instalation will be downloaded, and then the program downloads and installs them on the computer on its own. It is even better to visit Ninite.com (http://ninite.com/) and choose all the programs that need to be installed from the web page, and then they are downloaded and installed through one installer on their own.

Legal aspect of Windows licenses

When it comes to creating images, bear in mind that Microsoft Volume Licensing programs are not a source of full license for the Windows operating system. These programs offer only upgrade licenses for the Windows desktop PC. [18] Volume License media can be used to deploy software only to licensed desktop PCs. A customer cannot acquire an initial or “full” Windows desktop PC license through any Microsoft Volume Licensing program. The Windows desktop PC operating system upgrade license is for upgrades only. [18]

Reimaging is permitted if the copies made from the Volume Licensing media are identical to the originally licensed product. Windows Enterprise is not available outside the Volume Licensing programs and, therefore, is not eligible for reimaging. [19] Customers can use these copies from Microsoft media only if they are the same product and version, contain the same components, and are in the same language. [19] Reimaging using Volume Licensing full version media requires that customers have licensed the Windows desktop operating system either preinstalled through the OEM or as an FPP retail product. [19] A licensed version of Windows 7 OS is also required for the Windows upgrade license through Volume Licensing, where "customers must first acquire a full license for a qualifying operating system preinstalled by an OEM". [18] Bear in mind that one cannot actually upgrade from Windows XP to Windows 7. [20]

Windows Deployment

The most popular open source deployment solutions nowadays are Clonezilla and FOG. Clonezillla has an archaic text-based interface and it is not recommended for novice users while FOG provides a web-based interface accessible from any computer on the network. It also has the ability to deploy images from mobile devices. [21] If an administrator has to deploy images across the network, setting up a dedicated cloning server is required. "For Clonezilla SE, this means setting up a Diskless Remote Boot in Linux (DRBL) server by installing it on top of a Debian or CentOS server." [21] Another option (for small networks) is to use the DRBL live CD to convert the machine into a temporary server. If all of the machines for imaging are not on the network, Clonezilla is a must, because FOG can clone only from an imaging server. FOG, by default, requires hosts to be registered (recording a unique MAC address) before imaging and before deployment (it can be overridden with an enabled Capone plug-in [22]). Clonezilla doesn’t require registering a host.

When it comes to installation, FOG "installation script creates the components required for an imaging server on top of a standard LAMP server." [21] The "FOG server is designed to integrate with existing network infrastructure such as DHCP and DNS servers, although it can also set these up on its own." [21] If an existing DHCP server is used, PXE traffic must be forwarded to the FOG server and the firewall must be configured to pass traffic to the server. The FOG service has additional features and can "schedule tasks, such as deploying images, installing and managing printers, tracking access to cloned machines, powering up a computer remotely using Wake-On-LAN, and installing applications remotely via a feature called snap-ins." [21] "With snap-ins you can directly push the executable installer for a simple application to a Windows computer." [21] Snap-ins can also be used for pushing software updates to client computers. FOG's menu has additional features and "offers options to wipe a disk, restore deleted files, scan a disk for bad blocks, and run a virus scan using ClamAV." [21]

If the person in charge of computer maintenance decides to use the Windows server, they have the technology called WDS (Windows Deployment Service) at their disposal. WDS is included as a Server Role in all 32-bit and 64-bit versions of Windows Server 2008. Deployment must be performed with the "sysprep tool" which is included in every instance of Windows 7 and Windows 8. The machine must be sysprepped (all system-specific information from an installed Windows image is removed; including the computer security identifier (SID)) and then the image can be deployed using WDS. There is no limit to the number of times that the Sysprep command can run on a computer, but the clock for Windows Product Activation begins its countdown the first time Windows starts. Sysprep/generalize command can be used to reset the Windows Product Activation a maximum of three

times. If the Sysprep command needs to be run multiple times on a single computer you have to use the SkipRearm setting in the Microsoft-Windows-Security-Licensing-SPP component to postpone resetting the activation clock. [23] There is probably no need to have outside experts perform the deployment in schools (as far as students are concerned), because according to article 12, paragraph 11 [24], students' documents are only kept until the end of each school year. Below we present a table with some solutions for Windows deployment.

TABLE 2. POSSIBLE SOLUTIONS FOR WINDOWS DEPLOYMENT

Name[URL]	License	Specific requirements
Smart Deploy[1]	Commercial	It has a platform pack (PC drivers for different vendors)
Fog project[2]	Free	**A**pache, **M**ySQL, **P**HP but they are automatically downloaded by the software
Clonezilla[3]	Free	-
System Center 2012 R2 Configuration Manager[4]	Commercial	Active Directory schema must be extended, SQL Server is required, and multiple site server roles have to be configured.
Symantec Altiris Deployment Solution[5]	Commercial	-

Table Legend
[1] http://www.smartdeploy.com/
[2] http://www.fogproject.org/
[3] http://clonezilla.org/
[4] http://www.microsoft.com/en-us/server-cloud/products/system-center-2012-r2-configuration-manager/default.aspx
[5] http://www.symantec.com/deployment-solution

V. RECOMMENDATIONS FOR THE IMPROVEMENT OF SCHOOL COMPUTERS SECURITY

Administrators should lock the rooms where computers are stored, and also set up authorized video surveillance. If, despite the fact that all deadlines have expired, security policy still doesn't exist, it should be imposed on schools by the responsible government department. It must be introduced to all school employees who should be given a test to make sure they are familiarized with the security policy, under the threat of having their salary reduced until they have successfully passed it. After that, all other users such as outside collaborators and students should be familiarized with the policy. Users of public computers should sign an agreement stating they will use the computers in conformity with the security policy. This could be achieved by having a simple application appear when the computer starts, with the following message: "By using this computer you accept the terms of use that are in accordance with the institution's regulations and the ISP." The application could be more complex so that before using the computer, the user must answer correctly three randomly picked questions about the security policy of the institution and the ISP. Apart from the digital message on the computers, detailed regulations should be displayed in the classroom so that everyone could read them. Regulations should include the ways the equipment and resources can be used, as well as the proper behavior. A few guidelines that should be listed:

1) Who has the right to use the computers? (for example, the age limit for using the computers, chat rooms and social networks as well as a list of web pages allowed for persons under 18)
2) Strictly forbidden actions, i.e., a ban against disrupting the physical integrity (against slamming, hitting and breaking the computers, equipment and the cables), as well as a ban against moving the equipment (mouse, keyboard, printer…)
3) For what purposes the computers can be used and a ban against sharing files and illegal distribution of copyright protected material.
4) A ban against changing the settings of the computer's OS (downloading illegal programs and malware, using BitTorrents, visiting and downloading pedophile and porn material), the possibility or impossibility to print, save files to various media, run external programs or download programs.
5) Disciplinary actions for those who don't abide by the previously stated rules (the possibility of charging for damage, reporting to authorities, banning against further access and use of the computers and equipment). Public computer labs should "protect themselves from possible lawsuits by educating their staff, clients, and students about copyright law." "Users need to be educated that installing unlicensed software is a prohibited activity." [25]

As more and more devices (PDAs, mobile phones, tablets) are connected to the internet, communication security becomes more important. It is necessary to design and configure a network according to the rules of the trade – it is not a layman's job and it should be given over to an outside company or CARNet, in case the IT teacher is not competent enough. Within the domain of operational security, if Windows are used, then XP should be abandoned and upgraded at least to Windows 7 and also a functional solution should be implemented for locking the settings and/or restoring the computer to a desired state. We recommend the VHD technology and its possibilities, since it is available within Windows, but a live CD/USB of a Linux version is not a restrictive solution since the school curriculum doesn't mandate the platform/software, but only the topics that have to be covered. If it is financially possible, the cloud, i.e. a virtual desktop is an even simpler solution. If the computers are too weak/old and the Windows 7 or newer versions cannot be installed on them, then the options of replacing the computers with newer ones or installing a newer version of Linux distribution should be considered, or the PC computers should be converted into clients with the acquisition of a strong Windows server. Concerning the Windows deployment, it is necessary to keep in mind the Windows licenses, and in practice we recommend a version of Windows server and Windows deployment service, considering that their use in schools is free of charge, or perhaps open source programs such as FOG and/or Clonezilla. "Clonezilla SE makes the most sense for networks with clients that don't require much administration after being imaged, such as those in Internet cafes, libraries, and school labs." [21] If the machines aren't connected to a network, Clonezilla must be used. Taking care of computer security also means creating a set of rules for using the computer equipment, and taking care of all aspects of security must never stop.

VI. CONCLUSION

Nowadays, the concern for security on the internet, on the computer, on social networks and in everyday life is

becoming more and more important. Since more and more devices are being connected to networks, and technology and mobile devices are becoming more available and have become a constituent part of everyday life, we should think hard about introducing the topic of IT security and computer security into our curriculums. In the primary school practices, there is very little regard for computer security. In the authors' opinion, this will not change until a very serious incident occurs in an institution; until then, very little attention and time will be spent on it. It appears that it all comes down to the IT teacher, i.e., to how much will and knowledge he/she can invest in the students' education and the computer classroom. Investing time in some solutions for locking the settings, deleting the changes made on the computer, live distribution, or a cloud solution are some viable options for every person, but it all greatly depends on financial capacity, responsibility and concern for the confidentiality of information, and on personal attitude, knowledge and skills. Regardless of the path, there are two crucial choices: a) whether to use Windows or not and b) whether to lock the settings or give complete freedom (all users as administrators)? Things are well as long as the responsible government department is making agreements with Microsoft and gives Windows products for teachers to use "free of charge", but what can be done once there is no more money or once the politics instruct all state institutions to switch to open code for some reason? In that case, the only legal and correct solution is to switch to an "alternative", free of charge and much safer OS (for example Linux). This way every line of the code would be under control and we would be sure that a certain part of the code is working the way it was meant to. Concerning the question of whether or not to lock the settings or give complete freedom to users, the authors believe that it is more productive to give almost complete freedom to users, i.e., to use some sort of technology for deleting the changes, i.e., for restoring all the settings, even if that means that computers should often be cloned. Further research is necessary in the field of computer security, such as: what is the level of computer security in Croatian institutions, and how much do the IT teachers in schools (especially the ones who play the role of system administrators) know about computer security and maintenance which include broad topics such as security, hardware, software, networks and application of IT technology in class, HCI and creating digital educational material. Finally, how much do the students, whom the schools and teachers are here for, know about IT and computer security and do the perceived shortcomings in the system significantly influence the knowledge they acquire through informal channels?

REFERENCES

[1] Miroslav Bača, Uvod u računalnu sigurnost. Narodne novine, 2004.

[2] Dragan Pleskonjić, Nemanja Maček, Borislav Đorđević, Marko Carić, "'Security of Computer Systems and Networks' Book Preview," ComSIS, vol. 4, no. 1, 2007.

[3] Priyanka Arora, Arun Singh, Himanshu Tyagi, "Analysis of performance by using security algorithm on cloud network," international conference on Emerging trends in engineering and management (ICETM2012), pp. 23–24, 2012.

[4] John Harauz, Lori M. Kaufman, Bruce Potter, "Data Security in the World of Cloud Computing," IEEE Security and Privacy, vol. 7, no. Issue 4, pp. 61–64, Aug. 2009.

[5] Blumenthal, M.S., "Hide and Seek in the Cloud," IEEE Security Privacy Magazine, vol. 8, no. 2, pp. 57–58, 2010.

[6] Hrvatski telekom, "T-com cloud computer," 2013. [Online]. Available: https://www.hrvatskitelekom.hr/poslovni/ict/cloud/racunalo. [Accessed: 01-Jul-2013].

[7] K. Pavlina, A. Pongrac, B. Latas, "Hardware equipment of computer classrooms in Croatian elementary schools," MIPRO, 2012 Proceedings of the 35th International Convention, pp. 1325 – 1327, 25.05 2012.

[8] Anfibia, "Deskman," 2013. [Online]. Available: http://www.anfibia-soft.com/products/deskman/. [Accessed: 30-Jun-2013].

[9] Faraonics, "Deep Freeze," 2013. [Online]. Available: http://www.faronics.com/products/deep-freeze/. [Accessed: 30-Jun-2013].

[10] Horizon Data Sys, "Drive Vaccine," 2013. [Online]. Available: http://www.drivevaccine.com/. [Accessed: 30-Jun-2013].

[11] Wikipedia, "Hardware abstraction," 18-Oct-2013. [Online]. Available: http://en.wikipedia.org/wiki/Hardware_abstraction. [Accessed: 28-Jun-2013].

[12] Liang Yang, Anthony F. Voellm, "Hyper-V Virtual Hard Disk (VHD) Performance White Paper," A Microsoft White Paper, Mar. 2010.

[13] The Government of the Republic of Croatia, "The Government President Milanović and the Minister with the Chief Executive Officer of Microsoft," November 2013. [Online]. Available: http://public.mzos.hr/Default.aspx?art=12804. [Accessed: 23-Dec-2013].

[14] "VHD limitation," technet.microsoft.com, 22-Oct-2009. [Online]. Available: http://technet.microsoft.com/en-us/library/dd799282%28WS.10%29.aspx#BKMK_limitations.

[15] Mark Minasi, "Stadier State," 2012. [Online]. Available: http://www.steadierstate.com/.

[16] Sami Laiho, "Wioski," 2013. [Online]. Available: http://www.wioski.com/.

[17] Microsoft, "windows refresh, restore," 2013. [Online]. Available: http://windows.microsoft.com/en-US/windows-8/restore-refresh-reset-pc. [Accessed: 19-Feb-2013].

[18] Microsoft, "Operating System License Requirements: Initial Operating System and Transfer of License." 2013.

[19] Microsoft, "Reimaging Rights," 2013. [Online]. Available: http://download.microsoft.com/download/3/D/4/3D42BDC2-6725-4B29-B75A-A5B04179958B/Reimaging.pdf. [Accessed: 13-Mar-2013].

[20] Mirosoft technet, "Windows 7 Upgrade Paths," 17-Jun-2009. [Online]. Available: http://technet.microsoft.com/library/dd772579.aspx. [Accessed: 13-Mar-2013].

[21] Mayank Sharma, "Clonezilla vs. FOG: The clone wars," 18-Mar-2013. [Online]. Available: http://www.openlogic.com/wazi/bid/275172/Clonezilla-vs-FOG-The-clone-wars. [Accessed: 01-Aug-2013].

[22] FOG, "Fog project Wiki," 02-Oct-2011. [Online]. Available: http://fogproject.org/wiki/index.php?title=Plugins:_Capone. [Accessed: 05-Aug-2013].

[23] Microsoft technet, "How Sysprerp Works," 07-Aug-2010. [Online]. Available: http://technet.microsoft.com/en-us/library/dd744512%28WS.10%29.aspx. [Accessed: 13-Mar-2013].

[24] MZOŠ, "Pravilnik o načinima i postupcima i elementima vrednovanja učenika u osnovnoj i srednjoj školi," 2010. [Online]. Available: http://dokumenti.ncvvo.hr/Dokumenti/pravilnik_vrjednovanje_os_ss.pdf. [Accessed: 30-May-2013].

[25] Hilary Naylor, "Securing Your Computers for a Public Computing Environment," 10-Jul-2007.

Elementary school students using Facebook: A case study of Croatian elementary school ban Josip Jelačić

Martina Draženović
Ban Josip Jelačić elementary school, Zagreb, Croatia
martina.drazenovic@gmail.com

Abstract - After the survey was conducted two years ago in this particular school, I determined certain difficulties regarding usage of Facebook among our students. During the last two years I addressed the problem and begun giving social networking lectures to students, their parents and teachers, later I have incorporated social network workshops in curriculum and especially emphasized the importance of Netiquette. This paper will deal with the status of Facebook in Croatian elementary school ban Josip Jelačić, it will point out the problems that students had and still have, and it will show what has changed during this short period of time.

I. INTRODUCTION

The usage of Facebook can start from thirteen years of age, but it has made clear by the students themselves that it can start earlier than that. During my time in this school, many students younger than thirteen approached me with some kind of problem which happened to them while they were using Facebook. Since they do not fulfill required age limit, I instructed them to tell their parents and, if required, to inform the police. I held presentations and conducted workshops with students, their parents and colleagues from my school. It is very important to emphasize that during these workshops we never logged in Facebook nor visited the Facebook page. Although challenged with this situation, we managed to execute these workshops successfully.

A. Facebook in school

According to "Ordinance on the treatment of teaching staff in schools in undertaking measures of preserving student rights and report each of these rights violation to authorities" [1] the school is required:

a) To inform students and parents about the rules of safe use of modern technologies, particularly mobile phones and the Internet,

b) To inform students and parents about the consequences of inappropriate communication on social networks (insulting, embarrassing, public lynching, photos, videos, etc.) and on the conduct of school facilities concerning information about violence in the electronic media in accordance with the Protocol of procedure in cases of violence among children and youth. [2]

This means that even though Facebook cannot be used until thirteen years of age, the school is required to introduce students and their parents to social networking.

B. Present needs

The students need to be prepared to use social networks before they turn thirteen. Numerous cases of abusing via Facebook prove this thesis. Teachers should be able to help their students (delete student's Facebook profile or to help them secure it better). For the present time, this is impossible because Facebook states that no person younger than the age of thirteen can create an account (and it is considered that no one under this age has Facebook account). There aren't any laws in Croatia which would help teachers bypass this situation if in deed their students were younger than thirteen and would want to erase their account or secure it better. The only thing the teachers can do (if the student is below 13 years of age) is to teach about Facebook via presentation or conversation.

II. THE RESEARCH

A. Past research

During 2012, we had an outbreak of Facebook incidents in our school. We weren't familiar with the law regarding using Internet in school (later we found out that there is no such law which would apply to school) and since we haven't had any incidents before, we had to learn how to deal with this situation. This situation led us to reexamine our school policies and act accordingly. As an informatics teacher, I felt obligated to react and try to lower the risks. Therefore, I researched the problem and provided guidelines in a manner of lectures and workshops to students, their parent and fellow colleagues. From these guidelines, my school modified its policy introducing restrictions concerning virtual behavior and cellphone usage on school territory. The research has been conducted in 2012 amongst students who take informatics classes and are older than 13 years of age. Total of 97 students participated in the research. According to the research, 73% of surveyed students were using Facebook which indicates the popularity of Facebook in our school as well as that all 100% of the students were familiar with Facebook. 69% stated that they use Facebook more than an hour a day which also indicates the importance of Facebook in their lives. 39% of surveyed students stated that they had received threatening messages via Facebook and 12% stated that they do not know every person they added as a friend on Facebook. 21% of them stated that they didn't share their account information with their

parents and 77% have computer in their own room which indicates the lack of adult supervision while online. [3]

B. Specific problems

Since 2012 we had encountered some specific problems which concerned our students. Students experienced hacked Facebook accounts and cyber-bulling (insults, threats, groups and pages created strictly to offend and mock them). Later, they realized that some of their behavior while online can endanger them; leaving an open profile, sharing their detailed personal information, taking photos of themselves and their family, liking various groups and pages which promote violence and sharing their critical thoughts about the community they live in. Also, I discovered frustration with collecting "likes" on Facebook. Many of my students said that it is very important to them to collect as many "likes" as possible which would indicate the popularity of the individual in school. Some of them admitted that they added strangers as their friends on Facebook because the more friends they had on Facebook the more they feel popular in "real life".

C. Workshops

Since our informatics textbooks don't cover social networking [5], I was forced to improvise in a way that I incorporated social network workshops in curriculum. I divided workshops into 5 parts:

1. "Introduction to social networking" – here we covered general information about social networking (variety of social networks, their popularity in the world and user agreements).
2. "Benefits of social networking" – here we covered positive sides of social networking, how to use social networks to help their community, reach to other people and how to use them for future business possibilities.
3. "Dangers of using social networks" – we discussed various hurtful situations which might happen as a direct result of incautious behavior while online. This is the part where we spent most of our time discovering and identifying various risks and how to avoid them. Students gave their personal input on the negative experiences they had while using Facebook and how they dealt with it.
4. "Laws" – we discussed laws which would apply while an individual incautiously uses Internet. Laws are divided into two groups, criminal [9] and misdemeanor [10]. We identified the connection between certain behavior and the gravity of penalty. This part also included student's understanding of law threw their parent's eyes since their parents are held accountable for their child's behavior until they turn 14 years of age.
5. "Netiquette" – we discussed Netiquette and practiced proper virtual communication. [11]

After finishing these 5 parts we decided to raise awareness towards using social networks and distribute this information to our community. The students built web-page about dangers of Facebook, they made posters and presentations. We published the link to their web-page and presentations on school's official website and displayed posters around school property. Parents also had a role in raising awareness. After initial parent-teacher conference, during which they were introduced to social networking and potential dangers of using them, they demanded to stay updated, they organized themselves (if something happens they would all know about it and try to reduce the damage) and they moved computers from their children's rooms where they would be able to supervise their child while online. Parents commended the initiative and admitted that they had poor knowledge of social networking as well as what might happen to their children and to them. I introduced my colleagues into social networking and while some of them had already been using Facebook to distribute school assignments and news updates to students, the others had poor knowledge about social networks. My colleagues welcomed the initiative and collaboration amongst students, their parents and themselves. [4]

D. The students

Total of 120 students voluntarily completed anonymous survey. They had been given 26 questions regarding Facebook. The goal of this research is to see how the students are contemplating Facebook now in comparison from two years ago. Some of the students are the same, but most of them left the school since 2012. I have highlighted some of the questions which are the same in both researches:

1. Are you a Facebook user?

2. Are you familiar with Facebook?

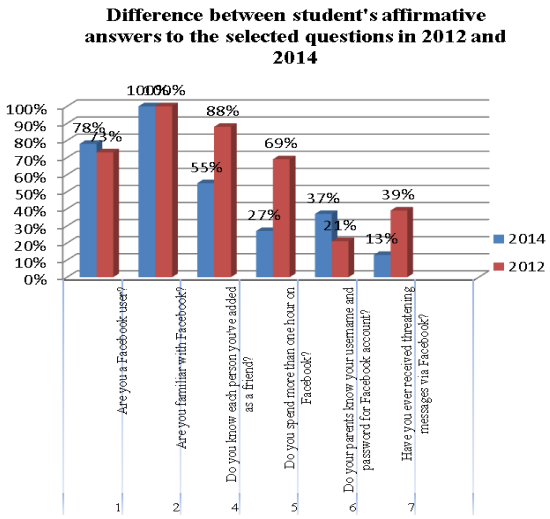
3. Where is your computer located in the house?

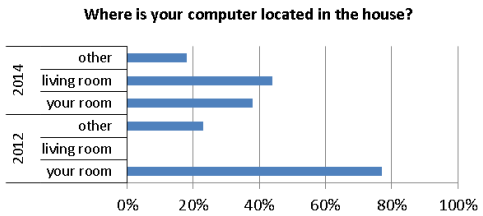
4. Do you know each person you've added as a friend?

5. Do you spend more than one hour on Facebook?

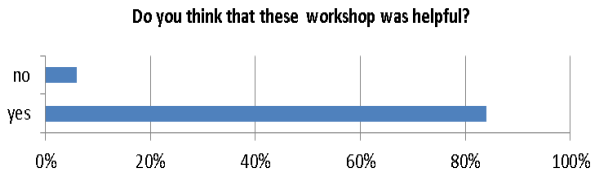
6. Do your parents know your username and password for Facebook account?

7. Have you ever received threatening messages via Facebook?





Students who were surveyed in 2014 were asked additional questions regarding workshops. The question „Do you think that the workshops were helpful?“ was crucial for the workshops because it tells me whether I should reexamine these workshops, change them or simply cancel them.



E. The parents

The parents were given surveys to voluntary fill out after parent – teacher conferences. They were asked to anonymously answer several questions regarding dangers of using social networking and workshops. Total of 33 parents were surveyed. 32 out of 33 parents think that workshops were very useful and that I should continue conducting them. 13 out of 33 parents are using Facebook.

F. The results

The results regarding student's surveys:

1. 5% more students are Facebook users.

2. All of the students are familiar with Facebook.

3. In 2012, 69% of students spent more than an hour on Facebook while in 2014 that number decreased to 27% which indicates the fall of Facebook popularity in our school.

4. In 2014, 16% more students shared their account information with their parents.

5. In 2012, 39% students received threatening messages whilst now that number is 13%.

In just two years a serious differences can be seen with students using Facebook. According to this research, popularity of Facebook is starting to decline, students spend less time on Facebook and receiving threatening messages had also declined. In 2012 almost 80% of students had computer in their room, whilst now that number is below 40%. 84% of the students think that social network workshops I conducted helped them. Parents asked me to give presentations more often to them and to their children, to keep emphasizing the dangers of using social networking and to keep updating them.

III. CONLUSION

This research indicates that Facebook isn't as popular among students as it was two years ago, it indicates that students now have more supervision at home while online and it shows decrement in negligence while using Facebook.

REFERENCES

[1] Ordinance on the treatment of teaching staff in schools in undertaking measures of preserving student rights and report of each of these rights violation to authorities, "Narodne novine", broj 87/2008., 86/2009., 92/2010., 105/2010., 90/2011., 16/2012.. 86/2012. i 94/2013. , Zagreb, October 2013

[2] Ordinance on the treatment of teaching staff in schools in undertaking measures of preserving student rights and report of each of these rights violation to authorities, "Narodne novine", broj 87/2008., 86/2009., 92/2010., 105/2010., 90/2011., 16/2012.. 86/2012. i 94/2013. , Zagreb, October 2013, article 20

[3] MIPRO (The Role of Facebook in Elementary School Education), Published in: MIPRO, 2012 Proceedings of the 35th International Convention, 21-25 May 2012, Page(s): 1398 – 1400, Print ISBN: 978-1-4673-2577-6, INSPEC Accession Number: 12865552

[4] MIPRO (Incorporating social network workshops in elementary school curriculum), Published in: Information & Communication Technology Electronics & Microelectronics (MIPRO), 2013 36th International Convention on 20-24 May 2013, Page(s): 622 – 624, Print ISBN:978-953-233-076-2, INSPEC Accession Number: 13768286

[5] Informatika 8 : udžbenik iz informatike za 8. razred s multimedijskim CD-om / Fredi Glavan, Alfa, 2011.

[6] Informatika 7 : udžbenik iz informatike za 7. razred s multimedijskim CD-om / Ida Srdić, Goran Kadić, Branko Hrpka, Alfa, 2008.

[7] Informatika 6 : udžbenik iz informatike za 6. razred s multimedijskim CD-om / Ida Srdić, Branko Hrpka, Goran Kadić, Alfa, 2011.

[8] Informatika 5 : udžbenik iz informatike za 5. razred s multimedijskim CD-om / Ida Srdić, Alfa, 2011.

[9] http://www.zakon.hr/z/98/Kazneni-zakon

[10] http://www.zakon.hr/z/52/Prekr%C5%A1ajni-zakon

[11] http://en.wikipedia.org/wiki/Netiquette

Evaluation Study and Results of Intelligent Pedagogical Agent-led Learning Scenarios in a Virtual World

M. Soliman*, and C. Guetl*,**
* Graz University of Technology/IICM, Graz, Austria
** Curtin University/School of Information Systems, Perth, Australia
muhamed.soliman@gmail.com, christian.guetl@iicm.tugraz.at

Abstract - The use of intelligent pedagogical agents (IPAs) has shown to provide positive learning results and new learning possibilities. The IPA has specific importance in immersive learning environments to enable 24/7 availability, offer a learning companion opportunity, and increase learner engagement. In that regard, a proof of concept prototype implementation of an intelligent pedagogical agent (IPA) in the Open Wonderland virtual world environment was created. This paper reports a qualitative evaluation study and experiment performed by a team of six experts in relevant areas of expertise. Those areas include cognitive science, computer science, e-education, and virtual worlds. The experiment studied key prototype components in relation to four learning scenarios with distributed control between the learner avatar and the pedagogical agent to answer questions relevant to their effect on learning attributes such as motivation, engagement, and the learning experience. Given the qualitative nature of the experiment, the paper also analyzes and reports results relevant to expert input of how the prototype can better contribute to future pedagogical agent realizations and the impact on learning enhancement.

I. Introduction and Background

In attempts to discover various methods for new technology enhanced learning, immersive environments provide interesting properties for learning support. Those properties include 3D visualization, collaborative learning, and more to add to convenience, remote access, and cost effectiveness, [1]. A particular implementation of immersive environments is virtual worlds which has Open Wonderland [2][3] as an example. Virtual world environments are accessible through Internet browsers while representing each user as an avatar. The user avatar can thus roam into the 3D space and perform available learning activities alone or in collaboration with others who are online.

An important challenge in virtual worlds is learning support. Given the convenience of the time and location, the immersive environment can be occupied with many users. It can be infeasible to achieve 24/7 human learning support through avatars. Furthermore, a learner who cannot find peers to collaborate with or be in an inhabited world can find it uninteresting to proceed.

The use of intelligent pedagogical agents (IPAs) has shown to provide positive learning results and new learning possibilities. [4] shows increase in learner perception and performance through the motivational role of the *persona* effect of animated pedagogical agents. IPAs are life-like characters that can be thought of as points of interaction between the learner and the immersive environment to provide learning support, [5]. There has been research in the area of pedagogical agents from different dimensions, [6][7]. The motivation for incorporating IPAs in an immersive environment target interesting properties of 24/7 learning support, enable learning peers, and offer intelligent learning support.

The goals of the IPA are pedagogical oriented to increase learner engagement and interactivity in the virtual world and to provide effective learning services while monitoring learner progress. To discover feasible interaction possibilities between the learner and the IPA in learning scenarios, in the learning context, a simulation was achieved to show interaction patterns [8]. Furthermore, adding intelligence abilities to the IPA was also studied in [9].

Realizing the IPA has been reported in [10] providing a prototype of an IPA in the Open Wonderland virtual world to offer different types of services. In order to achieve it, a Non-Player Character (NPC), a multi-modal communication module, and supporting elements were developed and amended with Bot Chat module. The purpose of the multi-modal communication module is to enable interaction with the IPA. Furthermore, a voice chat was added through voice synthesis tool, [11]. The developed prototype in reference to the performed simulation provided further details of interaction in the implementation of Open Wonderland virtual world. This formed different learning scenarios, in each of which the pedagogical agent provides specific learning services with objectives. Fig. 1 shows a visual 3D scene for a pedagogical agent and a learner avatar in corresponding interaction in relation to an experiment object in the Open Wonderland virtual world, [2].

In order to assess the prototype and evaluate its impact on the learning experience in the virtual world, an evaluation experiment is required. There were two approaches to obtain the evaluation: through user evaluation or through expert evaluation. The second approach, which is selected, is in relation to the method of performing a qualitative experiment to obtain experts opinion on various aspects.

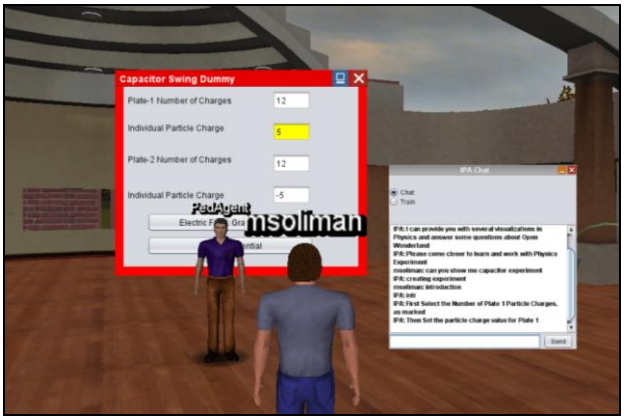


Figure 1. IPA interacts with a learner avatar on an experiment object in Open Wonderland. A text chat tool is used for communication.

II. EXPERIMENT SETUP

A. Research Questions and Design

The following hypotheses are formed to be checked in the experiment:

H1: The IPA, in the provided setting, provides appropriate mechanisms for increasing learner engagement in the virtual world.

H2: The IPA, in the provided setting and through awareness of the learning object, furnishes important guided instruction as an approach for enhancing the learning experience.

H3: The IPA contributes positively to learner motivation through guidance and task-completion support.

H4: The IPA provides the learner with an appropriate feedback mechanism that is positive to the learning experience.

H5: The pedagogical agent positively contributes to the learning experience in the virtual world setting.

In addition, the study has the following objectives:

O1: Evaluating and testing hypotheses H1:H5.

O2: Evaluating appropriateness and alternatives of the design elements such the text chat tool.

O3: Evaluating how can the prototype implementation potentially support the IPA concept.

B. Apparatus

To test hypotheses and objectives, the prototype setting is used through four learning scenarios in Open Wonderland. Participants refer to experiencing the following scenarios before filling a questionnaire.

- **IPA Proximity:** Captures learner intention to start learning activity as well as IPA's attempt to engage the learner through two proximity zones.
- **Using Multi-modal communication:** Refers to Learner to IPA interaction in either text or voice. Answers to questions are given through an AIML module, [10][12].
- **Tutoring:** The IPA tutors a sample simulation experiment.
- **Activity Feedback:** The IPA observes and provides feedback to the learner in relation to the learning object visual module.

C. Participants

Given the qualitative nature of the experiment, the study incorporated six experts in relevant fields. All of the participants are working at least one year in research on related topics having experiences in the following: computer science, cognitive science, e-education, and virtual worlds. Table I shows their areas of specialties and their distribution. Some of the experts are in research of more than one field. The age of the participants are in the range of 24-44 with (Mean = 30.3 & SD = 7.2). Three (50%) of the participants are females and three (50%) are males.

TABLE I. EXPERTS AREAS OF SPECIALTY AND RELEVANCE TO THE STUDY.

Area, relevance, and number of participants (%) in each area.		
Area	***Relevance***	***Number (%)***
Cognitive Science	Thought processes of the learner, exposure to artificial methods for e-education.	3 (50%)
E-education	Experience with tools and methods for electronic instruction.	4 (67%)
Computer Science	Facilitating methods of IPA foundations.	5 (83%)
Virtual Worlds	Main immersive environment the pedagogical agent operates in.	5 (83%)

D. Procedure

The procedure of the study includes four main steps for each participant, as follows:

1. Filling pre-questionnaire by the participant
2. Introduction to the system
3. Participant performs four IPA supported scenarios
4. Filling post-questionnaire

E. Pre and Post Questionnaires

The purpose of the pre-questionnaire is to obtain demographic data and prior expert knowledge in the relevant areas as discussed in Section C. The contents of the post-questionnaire are for the hypotheses and objectives testing. The majority of the questionnaire questions are Likert based, except questions 2, 3, and 21. The post-questionnaire also obtains qualitative opinions through open questions. Table II shows Likert and multiple choice questions of the post-questionnaire.

TABLE II. POST-QUESTIONNAIRE QUESTIONS.

Q	***Likert Question*** *(1-Strongly agree, 2-Agree, 3-Neutral, 4-Disagree, 5-Disagree)*
1	**I think using the question answering tool supports the learning activity.**
4	**I felt that IPA gestures are helpful to the learning experience.**
5	**I thought that the voice capability of the pedagogical agent is important.**
6	**I felt that the pedagogical agent has some sort of intelligence.**
7	**I find the method of detecting avatar proximity is good to determine learner intention to run the activity.**
8	**The proximity scenario helps to increase learner attention to the learning activity in the virtual world.**
9	**In tutoring and activity feedback scenarios, the pedagogical agent increases learner attention to the learning task.**
10	**I felt that the IPA encourages me to interact more with it.**
11	**I felt the IPA encourages the learner user to interact more with the environment (learning object).**
12	**I think the tutoring scenario by the IPA contributes positively to the learning experience.**
13	**I think the observing and providing activity feedback scenario is important to the learning experience.**
14	**I think that the IPA support to learning by doing improves the learning experience in this virtual world setting.**
15	**Overall, the IPA contributes tactics that positively support the learning activity in a virtual world**
16	**I felt that the IPA furnished guided instruction with the learning task, as an approach for enhancing the experience.**
17	**Through the task-completion support, the IPA helps improve learner motivation.**
18	**I learned something new about pedagogical agents.**
19	**I think the IPA is an essential learning component to an immersive learning environment.**
20	**The virtual world can become more interesting to the learner being more inhabited with pedagogical agents.**
	Multiple Choice Questions
2	**Which tool is preferable for activating pedagogical agent functions and obtaining its responses?** *(Choices: Text Chat tool, Traditional 2D menu, Other)*
3	**Having both the text and voice chat options are important to the user.** *(Choices: Yes, No)*
21	**I view that the virtual world learner user will prefer to learn with a similar pedagogical agent***….(Choices: None of the time, Some of the time, Most of the time, All of the time)*

III. RESULTS

A. Pre-Questionnaire Results

Participants answered the open question "*Which application types you find a virtual world useful for*" with different application types of *collaboration, simulation and visualization, language learning, scientific research,* and *games*. Their views on the advantages of virtual worlds for e-education included: *flexible location, visualization in 3D, spatial information, immersion, freedom, collaboration, simulation of history, physics, learning together for physically apart students, and meeting native speakers for language learners*. In response to stating disadvantages of using a virtual world for e-education, the following are included: *The demanding requirements of good Internet connection and graphics, usability issues, being hard for new users, user interface challenges, not always up to date, can seem to be complicated for users with low IT skills.*

Only two (33%) participants indicated prior knowledge about pedagogical agents with "Agree" response, three (50%) are neutral, and one participant (16.7%) doesn't have pedagogical agent knowledge (disagree). One participant (16.7%) agrees to have prior experience with pedagogical agents in virtual worlds, one is neutral (16.7%), three (50%) disagree, and one (16.7%) strongly disagrees. Participants indicated expectations on the IPA as: *to provide guidance and hints, explanations and repetitions, being intuitive, easy to use, and flexible.* They see advantages of IPA in the virtual world to: *help in the learning efforts, help to perform a task, having more knowledge, provide easy instructions, support learning at any place and any time.* Their indicated views on disadvantages included: *distraction of the task, and of possessing less in personality.*

B. Post-Questionnaire Results

The purpose of the post-questionnaire is to test for hypotheses and objectives. The majority of choice questions are in Likert scale [11][13] (1-Strongly agree, 2-Agree, 3-Neutral, 4-Disagree, 5-Strongly disagree), except for three questions. The post-questionnaire also includes open questions to obtain qualitative opinions relevant to hypotheses and prototype design elements and how to improve them. Table II shows choice questions in the post-questionnaire. While there is a disagreement, in specific areas of research, for calculating the mean of Likert items, the reason is due to assumptions on equal distances among the items. Thus, further measures are added including the Standard Deviation (SD), the Mode (most repeating response), and the frequency for each response, [14]. Table III summarizes the responses to each Likert question giving the various measures used.

IV. ANALYSIS OF RESULTS AND HYPOTHESES EVALUATION

Each hypothesis is tested based on different related questions from the questionnaire. While Table II shows the questions, Table III shows the results and calculations for individual questions. In addition, open questions provide qualitative input and discussed below.

A. ***H1:*** *The IPA, in the provided setting, provides appropriate mechanisms for increasing learner engagement in the virtual world.*

This hypothesis is tested through questions 8-11. On the overall, participants agree on questions relevant to this hypothesis. A combined score of (M = 1.95, SD = 0.75, Mode = 2) is obtained assuming equal weights. On the overall, 83.5% of the participants agree, 4% disagree, and 12.5% are neutral.

Open question responses such as "*You feel personally involved when a chat starts", "Very nice approach to make learners curious", and "You feel personally involved when a chat starts"* indicate support to the hypothesis in relation to specific elements of the prototype.

TABLE III. LIKERT QUESTIONS: RESPONSES AND BASIC CALCULATIONS

Q	*Mean*	*SD*	*Mode*	*No. of Participants for each Response (1-Strongly agree, 2-Agree, 3-Neutral, 4-Disagree, 5-Disagree)*				
				(1)	(2)	(3)	(4)	(5)
1	2.17	1.17	1	2	2	1	1	0
4	3.00	1.10	3	0	2	3	0	1
5	2.33	1.21	1	2	1	2	1	0
6	3.33	1.21	2	0	2	1	2	1
7	1.33	0.52	1	4	2	0	0	0
8	1.50	0.55	2	3	3	0	0	0
9	2.00	0.63	2	1	4	1	0	0
10	2.33	1.03	2	1	3	1	1	0
11	2.00	0.63	2	1	4	1	0	0
12	1.83	1.17	1	3	2	0	1	0
13	1.50	0.84	1	4	1	1	0	0
14	1.67	1.21	1	4	1	0	1	0
15	2.17	0.98	3	2	1	3	0	0
16	2.33	1.03	2	1	3	1	1	0
17	2.00	0.63	2	1	4	1	0	0
18	1.67	0.52	2	2	4	0	0	0
19	2.33	0.82	3	1	2	3	0	0
20	1.50	0.55	2	3	3	0	0	0

B. ***H2:*** *The IPA, in the provided setting and through awareness of the learning object, furnishes important guided instruction as an approach for enhancing the learning experience.*

This hypothesis is tested through questions 12, 13, 14, and 16. On the overall, participants agree on questions relevant to this hypothesis. A combined score of (M = 1.83, SD = 1.05, Mode = 1) is obtained assuming equal weights. On the overall, 50% strongly agree, 29% agree, 8.3% are neutral, and 12.5% disagree in combining with equal weights. In Agree/Disagree calculation, 79% agree on the overall. Accordingly, it is approximated to support this hypothesis with *Agree*.

Participants indicated support to this hypothesis with *responses "Hints/Visual Hints (yellow marks) where was the focus"* are liked*, "The IPA telling me about right and wrong inputs."* Participants stressed the importance of providing more detailed and specific feedback by the IPA as an improvement suggestion.

C. ***H3:*** *The IPA contributes positively to learner motivation through guidance and task-completion support.*

This hypothesis is checked through questions 8, 17, and 20. On the overall, participants agree on questions of this hypothesis with combined score of (M = 1.67, SD = 0.59, Mode = 2). On the overall, 100% of the participants agree, if excluding one neutral opinion. Accordingly, this hypothesis is supported with *Agree*.

Participants indicated strong support to this hypothesis with several responses as a good approach such as *"Fits very well", providing guidance, It is not forcing the learner", "By motivating and by giving hints & how to solve the simulation",* and more.

D. ***H4:*** *The IPA provides the learner with an appropriate feedback mechanism that is positive to the learning experience.*

This hypothesis is checked through questions 9, 13, and 16. On the overall participants agree on questions relevant to this hypothesis. A combined score of (M = 1.9, SD = 0.87, Mode = 2). On the overall 77.78% of the participants agree, 5.56% disagree, and 16.67% are neutral. Accordingly, it is approximated to support this hypothesis with *Agree*.

Participants indicated support to the hypothesis through open questions such as "*Feedback helps during experiments", "Wrong answers (actions) gets corrected immediately".* They also provided area of improvement in this aspect which is to "*provide more specific feedback of what went wrong, and how to perform it correctly".*

E. ***H5:*** *The pedagogical agent positively contributes to the learning experience in the virtual world setting.*

This hypothesis is checked through questions 12-17 to give an overall result of (M = 1.91, SD = 0.97, Mode = 1) with equal weights. On the overall 41.67% of the participants strongly agree, 33.33% agree, 8.33% disagree, and 16.67% are neutral. Accordingly, it is approximated to support this hypothesis with *Agree*.

Responses to open questions in relation to particular elements also support this general hypothesis. The hypothesis is also related to Q21, in which 5 participants (83%) chose some of the time while one participant (17%) indicated most of the time choice to conclude its positive contribution to learning experiences in the virtual world, but not in all cases.

F. *Measuring Prototype Tools Appropriateness to the Learning Activity*

Questions 1-5 are relevant to tools appropriateness to the learning activity; see Table II for questions and Table III for results. Experts in general agreed to the importance of the text chat tool and preference over a traditional 2D menu to interact with the pedagogical agent, and activate its functions. Less importance is then given to the IPA gestures. Furthermore, suggestions are provided to enhance the text chat tool through auto-completion features and action recommendation.

G. *How the Prototype Supports the IPA Concept*

Questions 18-21 are relevant to generic statements about the prototype in relation to the IPA concept in general; please see Table III for calculations for each question. In general, the prototype has provided input to have something new about pedagogical agents with agree (M = 1.67, SD = 0.52, Mode = 2), see Fig. 2. Participants either

agree or strongly agree to support IPA influence on giving users the feel of the virtual world being more inhabited (M = 1.5, SD = .55, Mode = 2). Fig. 3 shows the resulting calculation of this question of 3 (50%) participants strongly agree and 3 (50%) agree.

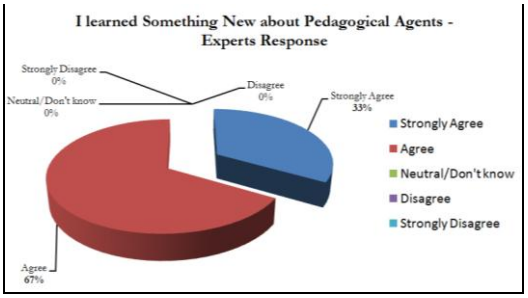


Figure 2. Results of questions Q18 which is related to the IPA concept.

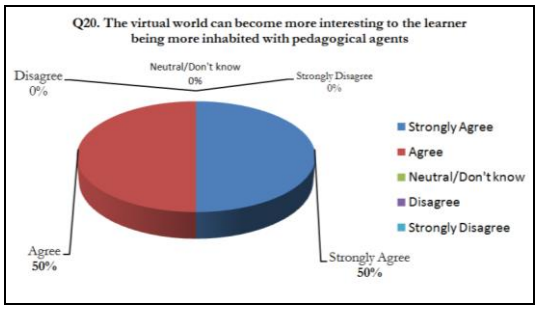


Figure 3. Results of questions Q18 which is related to the IPA concept.

H. Further Study Conclusions and Outlook

The study can thus shed light into most important aspects of the learning scenarios and design elements of the prototype. One of the key scenarios is the proximity scenario, as the IPA attempts to reason about learner intentions for performing learning activity and hence motivates to get engaged accordingly. The relevant questions (Q7, Q8) had results of (M = 1.33, SD = 0.52, Mode = 1) and (M = 1.5, SD = 0.55, Mode = 2) and with 100% agreement, see Tables II, III.

Regarding prototype elements importance, experts indicated the importance of the text chat module as a key component for communication with the IPA and triggering actions. They also suggested elements of enhancement with adding auto-complete feature and action recommendations by the pedagogical agent.

V. CONLUSION

The paper described an experiment setup and results to evaluate intelligent pedagogical agent learning scenarios in a virtual world. The purpose is to evaluate relevant hypotheses to pedagogical agent influence on the learning experience in a virtual world. The study had a qualitative nature, and was conducted by six expert participations from cognitive science, computer science, e-education, and virtual worlds. A pre-questionnaire and post-questionnaire are filled by each participant. The questions were in relation to four learning scenarios by the pedagogical agent in the Open Wonderland virtual world environment. Most questions are Likert questions and open ended questions to capture general opinions. On the overall, support was obtained on the relevant hypotheses (H1-H5), as discussed.

On the overall, **t**he pedagogical agent positively contributed to the learning experience in the virtual world setting. The results showed particular support to learner engagement addition by the pedagogical agent and the importance of the proximity learning scenario. Through the open questions, the study suggested key prototype elements including a text chat module and suggested further improvements to the learning scenarios.

ACKNOWLEDGMENT

Appreciation goes to anonymous expert participants from TU-Graz for contributing their time and effort to provide valuable contribution to this study.

REFERENCES

[1] Immersive Education, http://immersiveeducation.org/, Online, Accessed March 15, 2014.

[2] OpenWonderland, http://openwonderland.org/, Online, Accessed March. 15, 2014.

[3] J. Kaplan and N. Yankelovich, "Open Wonderland: an extensible virtual world architecture," IEEE Internet Computing, 2011, 38-45.

[4] J. Lester, S. Converse, S. Kahler, T. Barlow, B. Stone, and R. Bhogal, "The Persona Effect: Affective Impact of Animated Pedagogical Agents," In Proceedings of CHI '97 (pp. 359-366), Atlanta, March 1997, Miami, Florida, USA.

[5] M. Soliman and C. Guetl, "Intelligent pedagogical agents in immersive virtual learning environments: a review," Proceedings of 33rd International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO, May 24-28 2010, Opatija, Croatia.

[6] B. Weusijana, R. Kumar, and C. Rose, "MultiTalker: Building Conversational Agents in Second Life using Basilica, " Second Life, Education Community Convention, Purple Strand: Educational Tools and Products, 2008, Tampa, FL.

[7] W. Johnson, J. Rickel, R. Stiles, and A. Munro, A., "Integrating Pedagogical Agents into Virtual Environments," Teleoperators and Virtual Environments. Vol. 7(6), pp. 523-546. 1998.

[8] M. Soliman and C. Guetl, "Simulating interactive learning scenarios with intelligent pedagogical agents in a virtual world through BDI-based agents. International Journal of Engineering Pedagogy (iJEP), vol. 3, no. 2., pp. 41-47, April, 2013.

[9] M. Soliman and C. Guetl, "Evaluation of intelligent agent frameworks for human learning in virtual worlds," International Journal of Engineering Pedagogy (iJEP), vol. 1, no. 3, pp. 45-48 Oct. 2011.

[10] M. Soliman and C. Guetl, "Implementing intelligent pedagogical agents in virtual worlds: tutoring natural science experiments in OpenWonderland," The IEEE Global Education Conference, EDUCON 2013, March 12-15, Berlin, Germany.

[11] Mary Text To Speech, http://mary.dfki.de/, Online, Accessed March 15, 2014.

[12] AIML: Artificial Intelligence Markup Language, http://www.alicebot.org/aiml.html, Online, Accessed March 15, 2014.

[13] R. Likert, "A technique for the measurement of attitudes. Archives of Psychology," vol. 22, no. 140, pp. 1-55, 1932.

[14] H. N. Boone, Jr. and D. A. Boone, "Analyzing Likert data," Journal of Extension, vol. 50, no. 2, 2012.

LeCTo: a rich lecture capture solution

P. Pale, J. Petrović and B. Jeren
University of Zagreb, Faculty of Electrical Engineering and Computing
Department of Electronic Systems and Information Processing, Zagreb, Croatia
juraj.petrovic@fer.hr

Abstract - Lecture capture is a term referring both to learning materials based on recordings of live lectures as well as software and hardware support systems enabling lecture recording and reproduction. As lecture captures are today becoming more and more valued learning materials, there is a strong need for quality yet simple to use lecture capture systems supporting their creation. In this paper we describe a software solution for rich lecture capture recording and reproduction. LeCTo, an application developed to maximize lecture capture learning potentials and simplify lecture capturing process is presented through its functionality as well as problems and solutions encountered during its implementation.

I. INTRODUCTION

Lecture capture learning materials are today used not only in distant learning [1], but also as live lecture supplements [2]–[4]. Like most digital learning materials, lecture captures offer learning advantages like scalability and self-pacing, they can include multimedia rich materials to account for the recommendations of today dominant learning theories [5], [6], and finally they can provide the learner with the freedom to choose when and for how long will he learn.

Today there are various commercial and free lecture capture systems available focusing on different aspects of lecture capturing process: some focus on simplifying the recording process [7], [8], or automatizing camera control during recording [9], [10], some attempt to improve lecture reproduction properties by allowing the consumer to focus and manage only target region of interest in the lecture video [11] or by improving synchronization mechanisms [12], while some systems focus on hosting and indexing of lecture capture materials [13].

In this paper, we describe LeCTo (Lecture Capture Tool) focusing on several identified aspects of lecture capture systems. In particular, LeCTo is designed to:

- simplify lecture capturing process,
- enable creation of multimedia rich lecture capture materials, so called rich lecture captures, and
- offer a simple solution to their online management and reproduction.

II. A RICH LECTURE CAPTURE

The idea preceding the implementation of LeCTo was that lecture captures are not necessarily just recordings of events and learning materials from a live lecture, but rather they could be digital learning materials based on the lecture recording and enriched using additional materials to fully utilize the potential of digital learning environments. A rich lecture capture can contain the following types of contents:

- One or more video files that may be recordings of the lecturer, the desktop of his computer, his smartboard writings, a live experiment or demonstration or any other video material. A rich lecture capture can contain any number of videos, although it is usually two: the lecturer and his desktop.
- Additional lecture contents, in particular: links to additional online learning materials, online learning material that will be presented not as a link, but as the very content it refers to, typically graphical material or online applications, a quiz, speaker notes, lecture transcript or related textual materials and frequently asked questions or FAQ that students pose during the lecture.

Starting from the definition of a rich lecture capture, LeCTo was designed and implemented to support recording and reproduction of such multimedia learning materials. LeCTo has two components: LeCTo Recorder, a desktop application for recording lecture video materials, LeCTo Player, an online application for reproduction of rich lecture captures, and finally LeCTo PowerPoint plug-in, to help creating additional lecture contents.

III. LECTO RECORDER

LeCTo Recorder is an application enabling recording of one or several videos related to the lecture using any video camera device attached to the computer. Recording parameters that can be set include recorded file location, source camera for each video stream, video parameters (bitrate, framerate, and video scaling factor for each video stream), and audio parameters (source microphone and audio bitrate). Configuration settings can be saved, exported or imported, and several presets are also available.

Once the recording configuration has been defined, the lecture can be recorded using *record* and *stop* buttons of the application interface. Recordings are stored as separate files. Flash video (*.flv*) file format was chosen as supported video format since it is one of the most common file formats for delivering online videos today.

IV. LECTO POWERPOINT PLUG-IN

LeCTo PowerPoint plug-in was designed and implemented to provide support for creating the additional content to the lecture (Additional Lecture Content, ALC). It was designed as a Microsoft PowerPoint plug-in, as it is one of the most commonly used programs for creating lecture slideshows. Plug-in and contains of six buttons offered in the add-ins tab on the PowerPoint ribbon.

Four of those six buttons correspond to additional lecture content types: links, web contents, quizzes, and FAQ, while the fifth button allows the user to add textual indices or slide tags, enabling a more efficient navigation over the lecture. Each of those materials is edited in a small pop-up window which appears upon clicking the corresponding button. The fifth type of additional lecture content is speaker notes, can be edited directly in the space provided for them originally in PowerPoint. Every presentation slide can have all five types of ALC associated to it, meaning that every slide can have its own links, tags, online content, quiz, FAQ and speaker notes.

The sixth LeCTo button in the add-ins tab has a function of exporting all of the edited additional lecture contents into textual XML files with *.alc* extension which can later be used in LeCTo Player. Each content type is stored as a separate file.

Another important function of LeCTo PowerPoint plug-in is recording of slide transition times. Once installed, this add-in will track slide change times for every PowerPoint presentation performed in the full-screen mode. When the presentation is finished, a dialog box is opened asking for the user to choose whether the performed presentation was part of a lecture recorded by LeCTo Recorder. If so, a XML file containing all slide transition timestamps with *.cce* (context change event) extension is created. Otherwise, the transition timestamps will not be saved. Transition timestamps are accompanied with slide titles and slide ID's generated by PowerPoint.

V. PLAYER SETUP FILE

Once all lecture materials have been created (*.flv* video files, *.alc* files and *.cce* file), they need to be uploaded to the web to enable their online reproduction. LeCTo is designed so that lecture materials can be stored at virtually any online location, or even at several different locations. This means that institutions with available web space can utilize their own storage locations, while private users can store their files using any of the numerous free online file storing services.

The exact URL location of every lecture file can be specified to the player one by one. More convenient way is to define them in the player setup file, an XML file with *.psu* extension, which can be generated using LeCTo Recorder. The user only needs to define URL's for every file he wants to include in the lecture in a simple input form. The *.psu* file is generated automatically upon filling in the corresponding form in LeCTo Recorder.

VI. LECTO PLAYER

LeCTo Player is a browser-based application designed to enable reproduction of rich lecture captures. In order to reproduce a rich lecture capture, the user should upload the corresponding *.psu* file defining locations of online lecture materials to the LeCTo application, but he can also manually add URL's of all additional lecture *.alc* resources he wants to use. Alternatively, a LeCTo lecture can also be started directly if the link to the *.psu* file is added to LeCTo Player's URL. There is no limit on number of files included in the lecture: any number of video files as well as any number of *.alc* files can be included in the lecture, although providing the learner with too many learning materials will probably result in cognitive overload [14]. A practical limitation is also space available on the screen.

In the LeCTo Player's main interface each of the lecture materials, represented by a corresponding file, is displayed in a separate window so it can be easily manipulated. All LeCTo windows are opened within one, same web browser window. All *.alc* files, i.e. additional lecture contents, are synchronized with the video files and display only the contents associated to the current context being defined either by time or by presentation slide ID of the lecture.

Lecture navigation in LeCTo is supported in three ways: through chapter or slide index, video slider as well as through permuted index. Permuted index is an index of all words appearing in any additional lecture contents presented with two words appearing before and two after the target word to help to determine the context of its appearance in the text. Each word in the index is a link to the target chapter or presentation slide.

VII. IMPLEMENTATION

A. LeCTo Recorder

LeCTo Recorder was implemented in Java in order to make it a platform-independent application. Surprisingly, Java has a relatively weak multimedia support, so several multimedia support libraries were taken into consideration before exact implementation.

- Xuggler is a library enabling efficient manipulation with media streams directly from Java, effectively hiding the unneeded low-level manipulation aspects, yet it had to be rejected since it did not provide camera support.
- JavaCV is a Java interface to OpenCV, a library of programming functions for real time computer vision. Although it could have technically been utilized, it had been rejected since it provides a large variety of functions not needed and not originally intended for the purposes LeCTo Recorder would be used for.
- JMF or Java Media Framework was perceived as a potentially useful Java multimedia support for LeCTo Recorder, yet it turned out its last stable release was released in 2003 and will probably soon become deprecated. Two similar libraries with the same potentials and problems including lack of updates during the last few years are LTI-CIVIL and FMJ (Freedom for Media in Java).

- Libvlc is a C++ library for supporting HTTP streaming, video, desktop and audio recording, and was chosen for LeCTo Recorder implementation since it provided necessary support for desired LeCTo functionalities. Libvlc is included in the installation of VLC Player, a popular free and open source cross-platform media player that is regularly improved through new releases.

1) Vlcj library

Since Libvlc is written in C++ programming language, it cannot be utilized directly from Java code. This problem of accessing and using Libvlc library from Java was solved by utilizing vlcj, a library enabling Java interface to Libvlc and its methods. vlcj is an open source library using Java Native Access library to access native code (in this case C++) directly from Java. To achieve this, Java Native Access library *jna.jar* and *platform.jar*, enabling easier native code access depending on the operating system, and the location of the Libvlc need to be added to the project execution path. Libvlc is located in the VLC Player installation folder and has a different extension depending on the operating system*: .dll* on Microsoft Windows, *.so* on Linux, and *.dylib* on Mac OS X.

2) Camera detection

One of the necessary parameters to utilize vlcj for camera stream recording is the camera name or ID. The detection and recognition of cameras connected to a computer has turned out to be a platform-specific problem. On Windows platform, Direct Show Java library (DSJ) can be used to identify cameras currently connected to the computer. On Linux platform, camera device detection is based on reading the device names set as video0, video1 and so on from the /dev/ directory. A more elegant solution left for the future improvements includes support camera detection using Video For Linux For Java library shortly called v4l4j. Both DSJ and v4l4j libraries include many other functions which were not needed in this implementation of LeCTo Recorder. In Mac OS X, only the predefined system camera is offered to be used, since an older version of VLC Player not including all required functions was still recommended in the time of implementing LeCTo Recorder. Future improvements for Mac OS X include utilizing VLC Player 2.x Libvlc libraries which enable camera detection support. Libraries currently utilized for LeCTo Recorder are schematically represented in the Figure 4.

3) LeCTo PowerPoint plug-in

LeCTo PowerPoint plug-in was adapted from Pyramidia plug-in application, a part of one of LeCTo's predecessors [15], [16]. LeCTo PowerPoint plug-in was designed as a Microsoft Office extension written in C# using .NET framework. All the additional lecture content edited in the presentation is stored directly in the PowerPoint presentation making its storing and transporting as simple as possible. In the very presentation, all types of additional lecture contents are stored as invisible textboxes on the corresponding presentation slide.

Once they are created, the additional lecture contents can be exported from PowerPoint presentation to *.alc* files. Each *.alc* file uses the following XML scheme:

- type attribute, appearing at the beginning of the file defines to which of the five ALC types the file corresponds;
- <window> element including child elements with metadata on size and location of the corresponding ALC window in LeCTo Player, the name of its author and a list of its contributors;
- <frame> element including child elements describing ALC content (exact child elements depend on the type attribute defined at the beginning of the file) and display time information (either as exact time interval during which the defined content should be displayed or presentation slide ID to which it should be associated to).

The *.cce* file, automatically generated after a presentation in PowerPoint, also uses an XML scheme with <frame> elements, usually each referring to a presentation slide or chapter. Each <frame> elements contains the following child elements:

- <contextID>, referring to the unique integer number automatically assigned by PowerPoint to each presentation slide,
- <title>, defining the title of the slide or presentation chapter, and
- <occurrence>, defining the start and end time or just duration of the slide or chapter.

Since *.alc* and *.cce* files are stored as text, they can be edited or even created even outside PowerPoint in any textual editor.

B. Player setup file

The player setup file defines all that is needed to watch a LeCTo lecture: it defines lecture materials and their URL-s (*.flv* video files, *.cce* time events file, and *.alc* supplemental files). Since it is also a textual file, the *.psu* file defining lecture resources and their locations can easily later be re-edited. The *.psu* file uses the following XML scheme:

- <video> elements including child elements describing display properties and URL of each lecture video material,
- <cce> element defining the .cce file URL, and
- <alc> elements including child elements describing window size, position, and URL of each lecture ALC material.

Just like *.alc* and *.cce* files, *.psu* file is also stored as text, so it can be edited late if any of the contents changes its location.

1) LeCTo Player

LeCTo Player was implemented in javascript and PHP programing languages. The basic idea behind the

implemented structure is that as much processing and implementation logic should be performed on a viewer's computer rather than on the server. From the implementation perspective, LeCTo Player can be described in two segments: manipulating input files and lecture reproduction.

7.4.1. Manipulating input files

To start a lecture, one must first define LeCTo input files comprising a lecture capture on the LeCTo Player initial page, which should include at least one *.psu* file defining online locations of lecture video and optionally additional *.alc* files. The *.psu* file, as well as optional *.alc* files, can be uploaded through viewer's computer or can be defined by typing in their URL's. The file upload form has been supported through uploadify, a jQuery plugin enabling multiple file upload functionality.

Once they are defined, the input files should be transferred to a viewer's computer where they will be parsed to extract required information, minimizing server's processing load. However, due to the javascript's same origin policy those input files cannot be transferred directly. Same origin policy means that, due to possible safety issues, an application running in a web browser cannot access files at a different domain then the one it belongs to. This rule does apply to web browsers, but not to web servers enabling a simple solution of first using the server to download all the defined input files to a predefined temporary server location, and finally transferring files from the server to the viewer's computer stored as XML DOM objects. XML DOM defines a standard for accessing and manipulating XML documents, in this case *.psu*, *.cce* and *.alc* files.

When all input files have been transferred to a viewer's computer, it is time to initialize the lecture. The LeCTo user interface has been designed using jQuery user interface library enabling support for elements like buttons, menus and dialogues, and their manipulation. Lecture initialization is performed through opening of video dialogues (windows), one navigational dialogue and dynamically generated dialogues for all defined .alc files.

7.4.2. Lecture reproduction

To reproduce lecture video file, LeCTo Player utilizes jwplayer, an embedded platform independent video player. It was chosen since it enables easy web page integration, is HTML 5 compatible and has a javascript API. It is also free for usage on non-commercial sites. jwplayer supports two types of video transport: using HTTP protocol and using RTMP (Real Time Messaging Protocol). If the remote server where video files are stored supports RTMP protocol, videos can be streamed so the user can access even the parts of the lecture that have not been downloaded yet, which is not the case for regular HTTP transfer. Alternatively, video streaming can also be supported through HTTP pseudostreaming protocol. During the lecture reproduction, the video files are synchronized so that if any of the play, pause or slider controls is used, a method performing the same action on all other video streams is performed.

Synchronization of additional lecture contents with lecture videos is performed using a simple javascript mechanism which checks lecture time every 1.5 seconds during the lecture replay. After the lecture time has been checked, the same mechanism checks weather any of the ALC dialogues should be synchronized and performs the synchronization if necessary.

VIII. CONCLUSION

The applications described in this article, LeCTo Player, Recorder and PowerPoint plug-in, present a full solution to lecture capturing, enrichment, and reproduction. Each of those components is designed to be as simple as possible, yet offers authors of digital learning materials based on lecture captures a wide range of possibilities and platform independence. LeCTo Player, Recorder and PowerPoint plug-in can be accessed and used under the Creative Commons license at www.lecto.org.

ACKNOWLEDGMENT

The work described in this paper was conducted under the research project "Complex System Modeling" (036-0362214-1987), supported by the Ministry of Science, Education and Sports of the Republic of Croatia.

REFERENCES

[1] T. Liu and J. R. Kender, "Lecture videos for e-learning: current research and challenges," in IEEE Sixth International Symposium on Multimedia Software Engineering, 2004. Proceedings, 2004, pp. 574 – 578.

[2] J. Day, J. Foley, R. Groeneweg, and C. van der Mast, "Enhancing the classroom learning experience with web lectures," in Proceedings of the 2005 conference on Towards Sustainable and Scalable Educational Innovations Informed by the Learning Sciences: Sharing Good Practices of Research, Experimentation and Innovation, Amsterdam, The Netherlands, 2005, pp. 642–645.

[3] J. McElroy and Y. Blount, "You, me and ilecture," in Proceedings of the 23rd Annual Conference of the Australasian Society for Computers in Learning in Tertiary Education: Who's Learning? Whose Technology?, Sydney, Australia, 2006, pp. 549–558.

[4] K. Lovell and M. A. Plantegenest, "Student utilization of digital versions of classroom lectures," J. Int. Assoc. Med. Sci. Educ., vol. 19, no. 1, pp. 20–25, 2011.

[5] R. E. Mayer, The Cambridge handbook of multimedia learning. Cambridge University Press, 2005.

[6] J. L. Plass, R. Moreno, and R. Brünken, Cognitive load theory. Cambridge University Press, 2010.

[7] T. Chiueh and P. Deng, "Lectern: a digital desk system for video-free course-lecture capturing and playback," in 2000 IEEE International Conference on Multimedia and Expo, 2000. ICME 2000, 2000, vol. 1, pp. 3 –6 vol.1.

[8] P. E. Dickson, W. R. Adrion, and A. R. Hanson, "Automatic capture and presentation creation from multimedia lectures," in Frontiers in Education Conference, 2008. FIE 2008. 38th Annual, 2008, pp. T2A–14 –T2A–19.

[9] C. Zhang, Y. Rui, L. He, and M. Wallick, "Hybrid speaker tracking in an automated lecture room," in IEEE International Conference on Multimedia and Expo, 2005. ICME 2005, 2005, p. 4 pp.

[10] B. Wulff and R. Rolf, "OpenTrack - Automated camera control for lecture recordings," in 2011 IEEE International Symposium on Multimedia (ISM), 2011, pp. 549 –552.

[11] A. Mavlankar, P. Agrawal, D. Pang, S. Halawa, N.-M. Cheung, and B. Girod, "An interactive region-of-interest video streaming system for online lecture viewing," in Packet Video Workshop (PV), 2010 18th International, 2010, pp. 64 –71.

[12] G. Schroth, N.-M. Cheung, E. Steinbach, and B. Girod, "Synchronization of presentation slides and lecture videos using

bit rate sequences," in 2011 18th IEEE International Conference on Image Processing (ICIP), 2011, pp. 925 –928.
[13] M. Wald and Y. Li, "Synote: Important enhancements to learning with recorded lectures," in Proceedings of the 12th IEEE International Conference on Advanced Learning Technologies, ICALT 2012, 2012, pp. 521–525.
[14] R. Mayer and R. Moreno, "Nine Ways to Reduce Cognitive Load in Multimedia Learning," Educ. Psychol., vol. 38, pp. 43–52, Mar. 2003.
[15] I. Miletic, P. Pale, H. Pandzic, H. Bogunovic, and B. Jeren, "The structure of the Pyramidia e-learning tool - the programmer's point of view," in 14th International Workshop on Systems, Signals and Image Processing, 2007 and 6th EURASIP Conference focused on Speech and Image Processing, Multimedia Communications and Services, 2007, pp. 180 –183.
[16] P. Pale, I. Miletic, Z. Kostanjcar, H. Pandzic, and B. Jeren, "Development of Pyramidia: an integrative e-learning multimedia tool," J. Electr. Eng. Educ., vol. 48, no. 3, pp. 264–279, 2011.

Using Data Mining for Learning Path Recommendation and Visualization in an Intelligent Tutoring System

I. Jugo, B. Kovačić and V. Slavuj
University of Rijeka, Department of Informatics, Rijeka, Croatia
{ijugo; bkovacic; vslavuj} @ inf.uniri.hr

Abstract – Since the creation of Web-based learning systems researchers have tried to make them adaptive to various characteristics of learners in order to increase the quality of learning. The application of data mining (DM) methods on learning system logs is often used as a basis of intelligent tutoring systems (ITS) that have the ability for automatic adaptation of some aspect of the learning process. Such a system was developed at our institution in the previous years and has been used in a number of courses. To improve our system we added an integration application to create a continuous feedback loop with a DM tool. Our goals, from the student's perspective, are to improve the quality of knowledge acquired by students as well as shorten the learning time by offering recommendations based on mined patterns of learning paths through the knowledge domain. From the perspective of teachers, our goal is to create a rich data visualization system (in the form of a Web application using standard Web technologies to create visualizations) to give them new insight into students' behaviours. In this paper we present the structure of our system and the research design that will be used to verify its results.

I. INTRODUCTION

The main problem with e-learning is lack of feedback concerning students' work. The two most commonly used e-learning systems in the EU are Moodle and Blackboard, which, however, in their basic installation version do not offer any tools that could be used for tracking the work and behaviour of each student. Such information would be of great help to teachers and would enable the adaptivity of learning dynamics, make communication easier, etc. In addition, it would help teachers in identifying patterns in student's learning paths through the learning domain as well as in identifying possible problematic learning units.

Adaptive e-learning systems aim to improve student's results by taking a particular aspect of the system and adapting it to the identified characteristics of the student [1]. Certain systems adapt the way they present the learning content based on the learning style of the student, others adapt content adaptivity based on the assessed level of knowledge, yet others adapt the difficulty of test questions, etc.

An intelligent tutoring system for e-learning based on dialogue was developed at our institution [2]. The system guides the learner through the learning domain by alternating between sequences of learning about particular concepts from the domain and sequences of knowledge testing, in order to offer support and skip the concepts that have been learned previously. The system was further enriched by adding the module for statistical analysis of student activity [3], which aims at providing feedback to both the student and the teacher. Feedback to the student includes an overview of his/her progress, more specifically, the number of visits for each knowledge unit (concept), the number of questions answered and tests taken, and the success rate of her answers. Conversely, the teacher receives detailed information on each student's progress and knowledge, but also an overview of the entire domain, which allows her to take appropriate action concerning any problematic concepts in the domain. In order to assess the impact of the implemented changes on system users and their attitudes towards using the system, we have undertaken an evaluation which revealed the possibilities for further development of the system [4].

Educational Data Mining is an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students, and the settings in which they learn [5]. In the last overview paper of the state-of-the-art in EDM [6] a new set of important research objectives was set and among them were: a) EDM tools have to be designed to be easy to use for educators or non-expert users in DM, b) the DM tool has to be integrated into the e-learning environment as one more traditional authoring tool.

In this paper we put forth our model of an adaptive ITS paired with an integrated EDM tool designed for educators and non-experts in DM. Our research aims to develop a system that offers students a better learning environment, teachers a better supervision over the teaching process, and DM experts the possibility to continue their work in a specialized tool and easily add new analyses into the system.

The paper is organized into five sections. Section 2 describes the theoretical foundations of this research and situates this work in the area. Section 3 presents the architecture of our system and points out its most important contributions. In Section 4, we describe our research plan that will be used to verify the expected results of the application of the system at our Department. Finally, we conclude the paper with an emphasis on future work.

II. RELATED WORK

Educational data mining (EDM) is a relatively new appellation for an area of research in which classic DM methods (originating from the areas of machine learning and artificial intelligence), along with some new ones, are used on data collected from e-learning systems in order to clarify the problems related to e-learning. The results of these methods are often used as a basis for automatic adaptation of e-learning systems to the needs, expectations and habits of their users. Although the term EDM appeared around 2005, some authors started using DM methods to analyse or predict the behaviour of ITS users much earlier [7].

Today, EDM is a fast growing area of research with the number of papers published per year rising exponentially. Authors applied existing DM methods and developed new ones specifically for EDM to research a very wide number of topics regarding e-learning, user success prediction, user clustering, user modelling, intelligent recommendation systems, intelligent examination systems, etc. There are a small number of overview papers of the field, e.g., [8] (published in 2007), [9] (published in 2009), while the last one was published in 2010 and cited over 300 references [6].

Our research deals with and connects two main areas: a) learning path pattern analysis as a basis of creating an intelligent recommendation system, and b) visualization of those patterns, particularly data visualisation in the Web browser using standard Web technologies (to avoid the need of using plugins such as Java or Flash).

Many authors have developed applications for DM-based data visualizations, but rarely using standard Web technologies. In [10] authors developed a Flex/Flash based application in the field of Bioinformatics, while in [11] authors developed a Java/Matlab based application that accepts data file uploads and returns results from a small set of DM algorithms. In [12] authors developed a student forum activity visualization tool using the Scalable Vector Graphics (SVG) web standard. A very recent paper [13] describes a rich data visualization system for visualizing student activities in a ITS that was developed as a standalone Java application that will be used as a reference point for our visualization that will differ in that it will be developed as part of our Web-based Integrator application using standard Web technologies (SVG, Canvas) and emphasize on real-time updates that will allow the teacher to watch users navigating through the knowledge domain and indicate their successfulness on related exams. In [14] authors created a Java application that used DM algorithms (namely, Hypertext Probabilistic Grammar – HPG) to discover patterns in navigation trails (paths) through the ITS content and create static visualizations using directed graphs and text labels to point out the direction and frequency of mined paths. A similar visualization approach was used in [15] where interactions with a motivationally aware tutoring system called M-Ecolab and its previous version without that capability (Ecolab 2) where mined to discover patterns in behaviour in connection with motivational feedback. In [16] authors used DM algorithms (namely, association rules and classification trees) to find rules in students usage of the ITS and visualize those rules to teachers in a tool called Author Assistant.

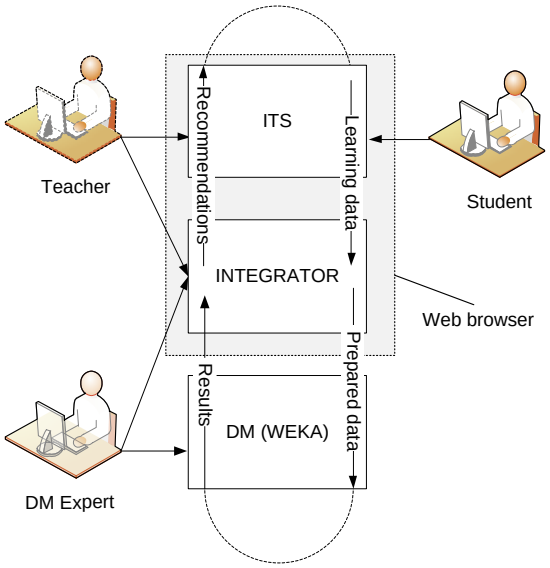


Figure 1. Learning system integrated with a DM tool [17]

Our research aims to combine these two areas of research by creating an intelligent tutoring system that will offer recommendations to students to both strengthen their knowledge acquisition process and shorten learning time. It also aims to offer relevant insight to teachers by providing real-time visualizations of learning activity patterns with additional information about students' results in testing their knowledge. All this information will be based on an integration module that will collect data from the ITS, send it to a DM tool, retrieve results and feed it into modules for offering suggestions and data visualization. To the best of our knowledge this type of system was not developed previously.

III. SYSTEM ARCHITECTURE

The main idea behind this integrated system is to allow non-experts to best use the potential of DM in e-learning, as well as to improve the existing e-learning systems by introducing adaptivity through recommendations.

The utilization of DM in e-learning systems or everyday classrooms is far from widespread use. We presume that by integrating DM tools with e-learning systems (and the educational process in general) we can help broaden the application of DM in education significantly. This integration should be seamless to the end user. In this way the teachers could stay in the familiar environment of the Web browser and the e-learning system they regularly use, and start using new DM powered features that will bring novel, helpful insights to their work.

The overall process of DM defines a set of steps that have to be followed to apply it to the target system and gain useful insight. Although some steps of this process are very technical, we feel that it can be implemented in such a way that (once connected to the ITS) it is mostly invisible to the teacher, and that it can be done in a familiar environment of a Web browser. In the case of

EDM being applied to an ITS, the result of this system would be a continuous feedback loop from the ITS (providing usage/learning data) to the Web interface (providing various prepared EDM procedures), to a DM tool (performing DM algorithms on prepared set of features). The feedback then goes back to the Web interface (providing insight via rich data visualizations) and finally, the user implements the suggested changes into the ITS. This process is depicted in Figure 1 [17].

We propose a system consisting of four basic modules. Block diagram of the system can be seen in Figure 2. First, the Learning Data module is used to make the connection to the ITS and define data sources needed for various EDM actions. The sources consist of attributes from multiple tables. This module monitors data sources and periodically pulls new data from the ITS. This module is used by the DM expert. For our system, the primary interest lays in knowledge domain structure analysis, so data about paths through the knowledge structure created by the students is needed. We also analyse data concerning access times, learning session duration and frequency, learning materials usage, quiz and test results, etc. Once data sources are defined, this module does not need further attention from the DM expert.

The central part of the Integrator Web application is the Actions Repository module. It is envisioned as an extensible set of DM-based operations of interest to the teachers (analysis toolbox). Every action offers a different analysis concerning the learning process, for example, prediction of student success in a course. Actions define the necessary data preparation routines, feature extraction, discretization, etc., needed for the DM algorithm to be used. Actions are run in predefined data intervals (week, month, and semester) or in an interval defined by the teacher. This part of the system is primarily used by teachers. Concerning our system, one of the first Actions we will develop will be pattern discovery in student's paths through the knowledge structure.

The DM expert is able to connect to the system, analyse the ways teachers use the actions, and modify or create new Actions. As the system will be open source, other developers will be able to develop and install new actions. Each Action employs the DM tool in the background, concealed from the end-user. The DM tool is contacted via API calls, by the Input/Output module that receives data from the Action selected by the user, and creates a set of arguments to make a correctly formatted call to the Weka API. It also handles the output file of the DM tool by preparing it to be used by the Data Visualization module, which also has the role of the user interface of the application. So, this module consists of a data visualization area, the Actions toolbar and a data manipulation (intervals, filter) area. The key contribution is that the user interface is not DM-oriented, i.e. it does not mimic user interfaces of common DM tools like Weka, but is instead user-oriented and optimized to bring real value in the form of insight into the learning progress of the students. Consequently, it optimizes the overall efficacy of the educational process.

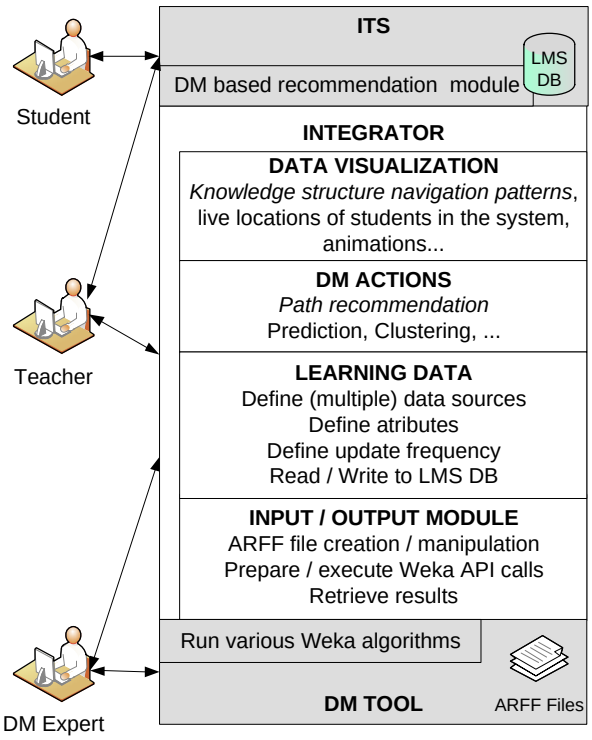


Figure 2. Block diagram of the system

Teachers can use the Data Visualization module, run various Actions and use the provided insight to modify the ITS in two Web browser tabs, without the need to learn about or even see DM tools, algorithms, database queries or any other technical details. The role of a DM expert is minimized after the initial connections with the ITS are defined in the Learning Data module. Although the system could be fully integrated into our ITS, we believe that creating a separate system with powerful connection abilities will enable other LMSs and educational data providers to use our system for their analysis and new applications.

The conceptual design of our visualization system is presented in Figure 3. Because the knowledge structure (domain) can contain hundreds of units (concepts) the objects representing them need to be kept small in order to display the whole structure on one screen. Concept names and student names will be revealed by tooltips when the teacher positions the mouse pointer over the objects. As the visualisation will be refreshed every few seconds teacher will be able to see students moving from concept to concept.

The number of students at each concept is displayed in a circle. The thickness of the lines and the arrow orientations display the common paths though the knowledge domain. Red lines depict returns indicating a possible problem with teaching materials or the structure of the domain. Dashed lines represent the original structure of the domain as created by the teacher. The

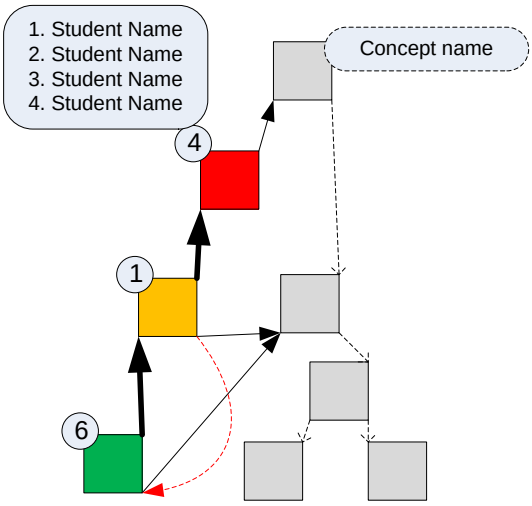


Figure 3. Visualization concept design

colour of the concept depicts the average student success on exams about that particular concept. Grey objects represent "no-data". This part of the system is currently under development.

IV. RESEARCH DESIGN

The development of this system is a part of an extensive research into the ways of improving the quality of acquired knowledge of students and shortening learning time. In addition, it will enable teachers to easily use DM on learning data and to get a visual insight into student activities. The research phases are described below.

A. *System development and implementation*

Basic functionalities for students are: learning (through lessons or individual knowledge units), examination (can be used for pre-testing, practice and grading) and collaboration (each lesson and knowledge unit have forums where students can post/read questions for the teacher or other students).

From a technical perspective this is a MVC Web application based on the Phalcon PHP and JQuery frameworks with a responsive user interface that can be used on smartphone, tablet and PC monitor screens without any loss of functionality.

New versions will be available at http://ditus.uniri.hr as the development progresses.

B. *Acquiring and transforming data*

To perform any kind of data analysis, we need to capture raw data from the application. There are two main approaches to this: using Web server log data or creating a custom data format and a logging service. Both approaches have their pros and cons: Web logs are "free", i.e., they are gathered without our intervention, but contain a very large amount of unwanted/unusable entries that need to be removed before data is pre-processed. Using a custom log format and a logging service requires development time, but grants a clean set of entries that can already be prepared for DM, which saves the time needed for data preparation/cleaning and, partly, pre-processing. We plan to choose the latter approach: adapt the well-known and widely used PSLC DataShop Tutoring Message Format [18] to our needs and develop a logging service as an integral part of the ITS and DM integrator application. Once the integrator application is connected to the ITS database, it can transform the relational data to our log format which can then be used for DM.

C. *Data mining for learning path pattern discovery and recommendation generation*

There is a large number of well-known pattern discovery algorithms used today. To discover patterns in learning paths of ITS users, some authors employed the already existing algorithms, while others developed new ones. Algorithm selection depends on the overall goal of the DM process. When trying to discover daily learning patterns, Hung and Zhang in [19] used sequential association rules. In [20], the author developed an algorithm for detecting "to-be-enhanced" learning paths and used fuzzy sets to make learning recommendations more understandable to end-users. Before using mined learning path patterns for making recommendations, we need to connect them to final student results in the course or some other source of data that will help in identifying which patterns to use. In [21], authors used sequential pattern mining to extract most frequent learning patterns and then applied distance-based clustering to group students according to their learning performance. In [22], authors proposed a cross-level learning suggestion process using FP-tree and FP-growth algorithms.

In order to identify the patterns that will be the base of our path recommendation process, we will grade mined patterns according to the following two criteria: the student's overall course success and the total time they needed to cover the course material (number of patterns, number of items in a pattern, total duration of a pattern). After this, we will have a set of preferred patterns to be used in the recommendation module of the ITS.

All the DM procedures in our system will be performed according to the CRISP-DM model [23].

D. *System evaluation*

Before we begin the final experimental phase, both the visualizations (for the teacher part of the system) and the overall user satisfaction with the system (relatively small sample of student population – testers) will be measured using the approach described in [4].

Evaluating data visualizations is a complex problem due to the fact that it cannot be measured objectively. We will implement the three step data visualization evaluation approach used in [13] which was based on [24]

E. *Final experiment and results verification*

The final stage of the research will consist of an experiment that will verify whether the initial goals of

the research, from the standpoints of both students and teachers, have been met. For the previous version of the system, two knowledge domains (Operating Systems 1 and 2, consisting of approximately 80 knowledge units per domain, and approximately 200 questions per domain) have already been developed and used over 2 semesters with groups of up to 100 students. Two more domains will be developed (Web development 1 and 2) for the purposes of the experiment. Students will be divided into three groups: a control group (will not use the system), the first test group (will use the ITS without recommendations) and the second test group (will use the ITS with DM-based recommendations). The final results will be statistically analysed and used to either confirm or reject our hypothesis that this system will both improve the results of the learning process and shorten learning time required to cover the course materials.

Even though it is still being developed, at the time of writing of this paper, the system had a responsive user interface, basic learning functionalities and communication with the DM tool. We are still developing "live" data visualizations in the browser using standard Web technologies without the need for any additional plug-ins.

V. CONCLUSION

Educational data mining has developed to a point where its findings can and should be made easily available to those that can benefit from it the most, namely teachers at various levels of the educational system. As e-learning, e-classrooms, computers and tablets are becoming more common, so should the widespread application of EDM. One of the main obstacles preventing the widespread use of DM in learning is the need for expert knowledge. We aim to overcome the obstacle by creating an easy-to-use Web-based system, integrated with an ITS used on a daily basis, which offers an extendable set of Actions providing useful information. Teachers do not have to know that these Actions run DM algorithms on ITS data in the background; they just receive useful information at the click of a button. This system, paired with rich data visualization, can bring a new dimension to the e-learning experience. Students will benefit from the suggestions given by the ITS, which aims at time saving and learning "the right way" also powered by DM algorithms. The main contribution of the system will be the continuous integration of the two systems, which will enable the overall system and its algorithms to be employed on daily updated data, and provide the teachers and students with better results as well as useful trend analysis.

REFERENCES

[1] P. Brusilovsky, Adaptive Hypermedia, User Modeling and User-Adapted Interaction, vol. 11, pp. 87-110, 2001.

[2] B. Kovačić, I. Jugo, Applying a Distance Learning System Based on Dialogue in e-commerce, Proc. of 32nd Int. Conf. on information and communication technology, electronics and microelectornics, MIPRO, pp. 36-39, 2009.

[3] B. Kovačić, I. Jugo, V. Slavuj, Improvement of system for distance learning based on dialogue by appliance of statistical analysis, of 35th Int. Conf. on information and communication technology, electronics and microelectornics, MIPRO, pp. 1794-1797, 2012.

[4] V. Slavuj, B. Kovačić, I. Jugo, Assessing user satisfaction with a system for e- learning based on dialogue, Proc. of Document numerique et societe DOCSOC, pp. 221-236, 2013.

[5] International Educational Data Mining Society, http://www.educationaldatamining.org, accessed: 10.01.2014

[6] C. Romero and S. Ventura, Educational data mining: A review of the state-of-the-art, Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, vol. 40(6), pp. 601-618, 2010.

[7] O. Zaïane, J. Luo, Web usage mining for a better web-based learning environment. In Proceedings of Conference on Advanced Technology for Education. Banff, Alberta, pp. 60–64, 2001.

[8] C. Romero, Ventura S. Educational Data Mining: a Survey from 1995 to 2005, Expert Systems with Applications, Vol. 1, No. 33, pp. 135-146, 2007.

[9] R.S.J.D. Baker and K. Yacef, The state of educational data mining in 2009: A review and future visions, Journal of Educational Data Mining, vol. 1(1), pp. 3-17, 2009.

[10] J. Koelling, D. Langenkaemper, S. Abouna, et al., WHIDE-a web tool for visual data mining colocation patterns in multivariate bioimages. Bioinformatics, Vol. 28, No. 8, pp. 1143-1150, 2012.

[11] M. Zorrilla, D. García-Saiz, A service oriented architecture to provide data mining services for non-expert data miners, Decision Support Systems, Vol. 55, No. 1, pp.399-411, 2013.

[12] S. Jyothi, C. McAvinia, J. Keating, A visualisation tool to aid exploration of students' interactions in asynchronous online communication, Computers & Education, Vol. 58, No. 1, pp. 30–42, 2012.

[13] M.W. Johnson, M. Eagle, M. Barnes, InVis: An Interactive Visualization Tool for Exploring Interaction Networks, Proc. of Int. Conf. on EDM, pp. 82-89., 2013.

[14] C. Romero, S. Gutiérrez, M. Freire, and S. Ventura, Mining and visualizing visited trails in web-based educational systems, Proceedings of the 1st International Conference EDM 2008, pp.182-186, 2008.

[15] G. Rebolledo-Mendez, B. du Boulay, R. Luckin, E.I. Benitez-Guerrero, Mining Data from interactions with a motivational-aware tutoring system using data visualization, Journal of EDM, Vol. 5, Issue. 1, pp. 72-103, 2013.

[16] C. Vialardi, J. Bravo, A. Ortigosa, Improving AEH Coursed through Log Analysis, Journal of Universal Computer Science, Vol. 14, No. 17, pp. 2777-2798, 2008.

[17] I. Jugo, B. Kovačić, V. Slavuj, A proposal for a web based educational data mining and visualization system, Proc. of the 5th Int. Conf. on Information Technologies and Information Society IT IS, pp. 59-64, 2013.

[18] Guide to the Tutor Message format (v4) - PSLC DataShop, pslcdatashop.org/dtd/guide/, accessed: 10.01.2014.

[19] J. Hung, K.Zhang, Revealing Online Bahaviors and Activity Patterns and Making Predictions with Data Mining Techniques in Online Teaching, Journal of Online Learning and Teaching, Vol. 4, No. 4, pp. 426-437, 2008.

[20] G. Hwang, A conceptual map model for developing intelligent tutoring systems, Computers and Education, Vol. 40, pp. 217-235, 2003.

[21] J. Su, S. Tseng, W. Wang, J. Weng, J.T.D. Yang, W. Tsai, Learning Portfolio Analysis and Mining for SCORM Compliant Environment, Educational Technology & Society, Vol. 9, No. 1, pp. 262-275, 2006.

[22] Y. Huang, J. Chen, S. Cheng, A Method of Cross-level Frequent Pattern Mining for Web-based Instruction, Educational Technology & Society, Vol. 10, No. 3, pp. 305-319, 2007.

[23] C. Shearer, The CRISP-DM Model: The New Blueprint for Data Mining, Journal of Data Warehousing, Vol. 5, No. 4, pp. 13-22, 2000.

[24] C. Plaisant, The challenge of information visualization evaluation, Proc. of the working conference on Advanced visual interfaces ACM, pp. 109-116, 2004.

Security analysis of wireless network access following 802.11 standard in educational institutions of the Republic of Croatia

Aleksandar Skendzic*, Bozidar Kovacic **

*University of Applied Science, Gospic, Croatia

** University of Rijeka/Department of Informatics, Omladinska 14, Rijeka, Croatia

**Phone: (385) 51-345-064 Fax: (385) 51- 345 207 *E-mail: askendzic@velegs-nikolatesla.hr, **bkovacic@inf.uniri.hr,

Abstract: Security of wireless networks is lower than the security of wired networks and it remains he greatest security risk of a local network . This is primarily due to the communication medium (air), which is accessible in a specific radius of action as well as the inadequate use wireless network protective mechanisms. Besides, administrators of wireless networks usually respond to damage caused by a system security breach only after it had already happened. The aim of this research is an evaluation of the safety parameters that affect the security of wireless networks. For this purpose a questionnaire was designed to test the expertise in educational institutions in the Republic of Croatia: CARNet system engineers will evaluate the parameters that affect the security of local WLAN. The collected data show individual evaluation of wireless network security parameters that affect the WIFI security. The results will be used for recommended safety guidelines.

Keywords: *wifi, security policy, educational institutuions, CARNet*

I. INTRODUCTION

The development of information-communication technology greatly contributes to the quality improvement in teaching at secondary school and university levels. The rise of computer networks and internet has created new opportunities for communication between teachers and students. Global developments and empirical reseearch indicate increased use of wireless access to local networks at the primary school levels too. Croatia is lagging behind global trends on primary and secondary school level, while the introduction of wireless access at university education level keeps up with the European average. The reasons for this, has been stated, are financial problems, lacking interest in the leading bodies of educational institutions and problems of network infrastructure maintenance. In order to find good a solution for the use of wireless networks including adequate security level, a sound development strategy is needed, departing from the analysis of actual possiblities and needs.This shoud be the basis of their impelementation. Accordingly it is necessary to offer pupils and students supported acess of wireless network from their personal computers.

II. WIRELESS NETWORK STANDARD IEEE 802.11

The emergence of wireless network standard IEEE 802.111 has expanded the possibilities of local networks based on IEEE 802.3 ethernet standard. The possiblity of wireless connectivity to local networks (LAN) was added. Bearing in mind the use of radio waves in the context of medium carriers of widely spread signal, this signal is used to connect clients and make their communication possible from access points (AP). Standard IEEE 802.11 (11 meaning work group that developed it) was published in 1997, and prescribed the following specifications as shown in table 1 [11].

TABLE 1. SOME SPECIFICATIONS OF IEEE 802.11 ETHERNET STANDARD

Reach within buildings	depending on the number and type of obstacles
Work frequency	from 2.4 to 5 GHz
Bandwith	0.7 Mbit/s and up
Data transfer	2 Mbit/s and up[2]
Reach	Up to 75 meters (in open air) depends on the standard (a,b,n,g,..)

III. SECURITY BREACH OF WIRELESS NETWORKS

Network security can be endangered by events which compromise availability, credibility, integrity and confidentiality of saved or transferred information and/or service provided by such networks or systems.

[1] International non-profit organisation IEEE (*Institute of Electrical and Electronics Engineers*).

[2] Standard IEEE 802.11 ac has transfer of 7 Gbit/s.

Vulnerability of wireless networks is due either to security weakness or to system shortcomings. The attacker can take advantage of system vulnerability in order to compromise its security with the aim of unauthorized access. Regular control of system vulnerability can eliminate threats to a system's security.

The principal security problems are:

- Program vulnerabilities,
- Malicious code,
- Social engineering.

Potencial sources of attack on wireless networks can be classified as follows:

- Internal attacks,
- External attacks.

Internal attacks may come from the network system itself. Users of a network system (operators, operatives authorized to access the system) fall under direct jurisdiction of system's administrator. Valid users can harm the network system and wireless network itself by accidental errors.

External attacks come from without the system and such threats are beyond the system administrator's control. There are many possible external threats, ranging from criminal organizations, commercial companies bent on collection of confidential data, investigative agencies in search of information, to simple enthousiasts trying to prove their ability to interfere with wirelesss network security. It is important to point out that wireless networks are often potential attackers' targets due to their characteristic capacity of receiveing radio waves.

Security policy defines planning and describes goals or security procedures. Security policy can be defined as the declaration of aims reached by particular procedures. Conditions under which security policy can have an effect are general – they are not specific. The implementation of security policy is a draft of security within a specific system [1].

Peltier [1] argues that an information security program should be a part of a security protection program of an organization's entire property. A security program is not made to satisfy the need for security or revision requirements, but as a process providing the management instruments necessary for management of responsibility. An organization's management is responsible for securing necessary protection control of an organizations/ínstitution's property where the security policy is implemented.

- In order to make use of optimal methods of risk management by means of security policy, it is necessary to:
- Know the number of users of a wireless network and their identification,
- Define network servers and the services that they provide,
- Define authorized persons who may introduce wireless network equipment ,
- Have a clear definition of the user rights level and procedures of gaining and losing user rights,
- Determine and secure physical access to wireless access points,
- Define security properties of wireless access points,
- Define wireless network type and program support (type and model),
- Define the procedure in case of loss of a network device,
- Define the procedure in case of a security incident,
- Define the procedure for use of encryption and management of keys,
- Define the frequency and range of the estimated security risks.

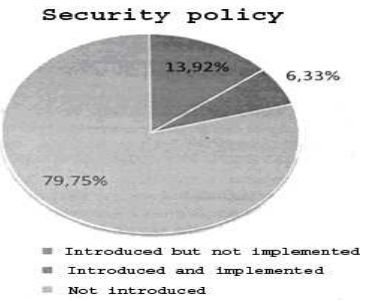
IV. SECURITY POLICY IN EDUCATIONAL INSTITUTIONS OF THE REPUBLIC OF CROATIA

Security policy in educational institutions as well as in other business organizations involves every single computer user. Final users are responsible for administering their own passwords in computer systems. On the other hand system administrators are in charge of supervision, interventions and recovery of security lapses. Crucial responsiblity in carrying out security policy lies with every individual (employee of the institution). Security policy cannot predict every possible breach of security integrity, but it can secure a solution to every problem. Bearing in mind levels of responsibility, a user on the lowest level imperils the security of the entire network if he/she permits compromise of his/her own computer. Employees of educational institutions (both teaching and other personnel) often do not take care of data privacy and the manner of use of network resources. Quite often, due to negligence, computer systems are in jeopardy. Research results show that employees themselves are a high security risk in most instutions.

Institutions in Croatia react to damage done to the security system only after it had already happened. Radojević [3] draws attention to the fact that security in educational institutions is subject to gross deviations: on the one hand he sees it as a major problem when it is discovered, to oblivion when the problem is solved – till the next such incident.

The CARNet survey from 2009 showed that only 45% of the surveyed institutions declare to have proscribed security protection. The survey forms were filled out by CARNet coordinators, i. e. persons who represent the insitution's interests in the CARNet community. The results show that the CARNet coordinators are not experts of computers or of informatic profile, rather they are professionals in the field that is of primary importance for the institution in which they work.

The following graph shows the results from the survey held in November 2010 among system engineers. 20% of the institutions have a defined security policy and only 6% of them do indeed carry it out.



Graph 1. The results of CARNet questionnaire about using security policies in the educational institutions [3].

V. A PROPOSAL OF A GENERAL WIRELESS NETWORK PROTECTION CONCEPT AND RESEARCH METODOLOGY

In the construction of a security system it is necessary to follow certain steps that can also be validated in practice. The funcionality of a security system must be periodically checked or (incidentally) due to the awareness of increased risk of attack on wireless network.

The definition of security level entails a creation of a security system based on optimal choice of parameters that affect it. Efficient functionality of a system requires determining acts that internally connect all elements of the security system. These acts steered as well.

It is also necessary to define security policy and security strategy. Determining security strategy includes the defnition of the security concept that is being constructed. Depending on the security concept of educational institutions, the introduction of the chosen system will be entrusted to system engineers within the institution, or to an external company. While defining the security system, internal resources for construction and execution of necessary protection measures will be considered, as well as the cost of the utilization of external services. According to the graph 2. 100% percent of security level is not possible.

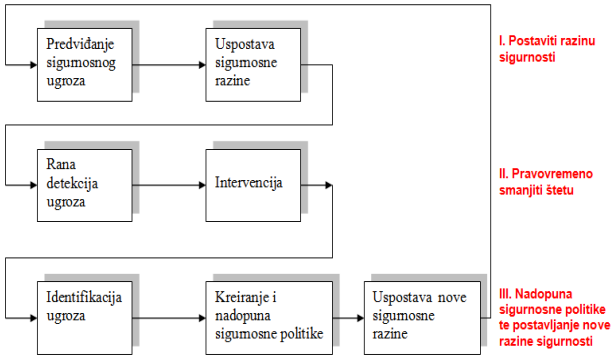
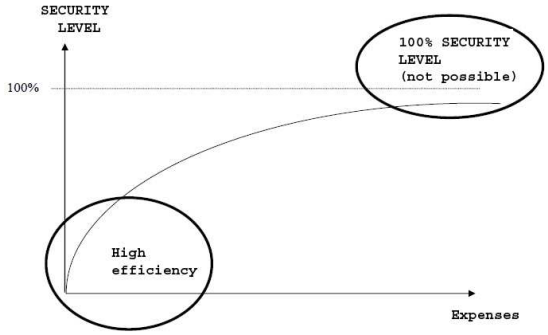


Figure 1. A proposal of a general wireless network protection concept and additional measures of security policy . Author's own source.



Graph 2. Graphic presentation of possible security levels [10].

In order to investigate and evaluate security parameters that influence wireless network security, a questionaire was composed to test the evaluation of exprets in educational institutions of parameters affecting local wireless network security. A random sample of respondents (**N=59**), system engineers employed in the educational institutions – CARNet members (35 faculties and 24 high schools). Questionnaire was distributed online over the CARNet mailing list *sistemci.hr* at September 2013. Total number CARNet member institutions is 1651 and due to technical constraints[3] and the unavailability was not possible to contact all member institutions. One of the reasons is that the member institutions (primary and secondary schools) have no system engineers (like college) than resource administrators who do not necessarily know the professional field of network security. For the evaluation was the scale of 5 levels of intensity: 1 = not at all important parameter in safety , 2 = partly important , 3 = parameter is important for safety , 4 = parameter very affects the safety , 5 = parameter completely affects the security . The study analyzed the level of security of wireless local area networks, which are based on the calculated average score of selected parameters: the use of encryption , type (model) of wireless network equipment , network coverage , the use of firewall , the

[3] Respondents are those members whose system engineers are subscribed to the mailing list *sistemci.hr*

number wifi users , the existence of security policies at the institution , used network services and the use of a VPN .

The questionnaire consisted of ten questions:

1. Does a local wireless network exist in the institution?
2. Does a security policy exist in the institution?
3. Evaluation of the wireles network in the institution.
4. Evaluation of importance of wireles security parameters, such as :

   a)Encryption algoritms/certification,
   b)Type (model) of network equipment,
   c)Spatial wi-fi network coverage,
   d)Use of *firewal*l,
   e)Number of wireles network users,
   f)Stricly defined security policy,
   g)Simplicity of authenication and authorization,
   h) Types of network serves/services,
   i) Use of VPN;

5. Existence and use of encryption algoritms/certificates in the existig local network;
6. Existence and use of RADIUS protocol and LDAP;
7. Evaluation of efficiency of RADIUS protocol in the sercurity user authenitcation;
8. and 9. Occurence and number og security incidents in the local wireles network;
10. Time needed for detection of security incidents in the local wireless network.

Here are the following answers of the questions from questionnaire.

A. Question nr. 1. Does your institution have a wireless network?

88% of the respondents answered positively to the question on implemented local wireless network in the institution where they are employed, and 12% responded negatively. One could expect that the majority of institutions do have a wireless network due to broad avalilability of wireless technology, acceptable price of equipment and rather simple implementation of computer network in the institutions. The institutions have also recognized economic acceptability of wireless network implementation of the 802.11 standard in comparison to the more expensive cabeling by means of the ethernet 802.3 standard. Wireless connenction possiblities have also increased mobility of users in the institution. This is of great importance when it concerns the use of new teaching and learning systems and of new applicable technologies like e-diary or Tablet that can be used in the education process.

B. Question nr. 2. Does your institution have a defined security policy?

To the question concerning the existence of security policy in the institution 53% of the respondents answered negatively and 47% answered postively. One of possible reasons for omitting security policy as an important factor of security of information systems in educational institutions is the estimate of the importance of information in need of protection. There is a widespread belief that in such institutions information is not essential and thus needs no protection. That is the the main reason for the lacking security policy in educational institutions. The Law on information security suggested the need for security policy under which the security of wireless network access also definitely falls [4]. At primary school level security policy practically does not exist, while academic subjects have differing records when it concerns defining and implementing security measures. The main reason for lacking security policy in institutions of higher learning is the same as in primary schools. An example to the contrary are the systems existing primarily in academic institutions like ISVU, Prehrana, IFIS, NISpVU and SmartX. These systems are applied only in academic institutions, thus not on primary and secondary school levels. In order to implement efficient security policy, system enigineers in academic institutions should have the support from the governing bodies since the adequate protection system requires financial resurces. Besides, investment in security policy is justfied by the very nature of information they are supposed to protect. In conclusion: the existence of security policy in educational institutions is of paramount importance for the security of their information system and thus of wireless network access.

C. Question nr. 3. How do you judge the security level in your institution?

The estimated wireless network security level shows that 47% of the respondents finds the security level satisfactory. Institutions do care about the wilreless network security level, but there is still room enough for raising its quality. System engineers in academic institutions again need the support of governing bodies to finance the expense of more efficient protection, even more so since the resources must come from each institution individually. 5% of the respondents claim not to use any security mechanism for wireless network access protection. It is possible that the institutions possess network equipment with wireless network access, but they do not use it (for example ADSL routers). It is not very probable that the possiblity of wireless access exists, though unprotected by certain protection mechanisms. Additional 5% of the respondents claim to use exceptionally high level of security, which implies that thay have a more substantial backing from the institutiom's governing bodies/management. 22% of the respondents declare that their institution maintains a satisfactory security level and 20% of them report a high level of securty.

D. Question nr. 4a. Evaluate the importance of cited parameters in planning wireless network security.

Encryption algoritms and certification have emerged in answer to question 4 as one of the most important factors in wireless network security. 54% of the respondents consider encryption and certification to be the most important local wireless network security mechanisms and only 2% believe that these are not important for seccurity access to wireless network. On the other hand the respondents gave highest marks to encryption and certification as the most important security mechanisms. Three types of algoritm for secure communication were represented acccording to the following standards: a) Wired Equivalent Privacy - WEP, b) Wi-Fi Protected Access - WPA, c) Wi-Fi Protected Access II – WPA2 [11]. One can guess that the most frequent encryption algoritm in use is WPA2, bearing in mind that this one meets the highest security standards in comparison to both WPA and WEP (see chapter II). Encryption algoritms are the most frequently used security mechanisms in wireless network access protection and as such are implemented in program support of wireless networks devices (routers, access points, wireless clients). Due to the problem of compatibility of wireless network equipment it is necessary to gain additional information on applicability of encryption algoritm to every single network communication device.

E. Question nr. 4b. Do you use of firewall?

Answering the question on the use of firewall 61% of the respondents consider firewall to be of paramount importance for the wireless network security. Firewall got the highest mark. Firewall has the role of filtering the wireless network traffic. In the context of wireless network according to the IEEE 802.11 standard, firewall can be used as an additional security mechanism. It is placed between the local wired network and network devices that make possible wireless connection to a local network (in most cases access points or routers). Protective walls are fitted in such a manner that they require a rigorous authenication and authorization of network access and as such are a very efficient mechanism in the network protection. The respondents have apparently recognized the importance of firewall for the secure access to wireless networks. It is also important to stress that the use of firewall is only one additional mechanism in the security of both wired and wireless networks, although it only provides protection on the network borders, having no influence on the internal security.

F. Question nr. 4c. What is the number of network users?

41% of the respondents find that the number of network users is extremely important for the security planning. The number of wireless network users is proportionate to the security level. If the number of users is high, the number of possible security dangers grows too. A small number of users does not require a more complex approach to the wireless network security. The reason for this is the possibility of detailed monitoring of every single user at the moment when he/she uses the wireless network. 27% and 20% of the respondents consider the number of wireless network users as important and very important in planning the network security and only 5% of the do not consider it important.

G. Question nr. 4d. The type (model) of network equipment.

32% of the respondents think that the type (model) of network equipment is important in the optimal planning of wireless network security. Considering that the market offers a variety of active wireless network equipment, one can expect variety in their technical characteristics and possiblities of use of recommended security mechanisms. One can find in the market equipment marked „WIFI certified“, issued by the non-profit manufacturers' organization Wi-Fi Alliance, whose goal is stimulation of mutual compatibility and promotion of their products. Wi-Fi Alliance, with a network of independent test laboratories carries out tests in order to safeguard mutual compatibility of active wireless equipment and to render support to secure communication. The cited certificate [5] guarantees interoperability, security, simple installation and reliability in operation. 29% of the respondents evaluated the type (model) of wireless equipment as exceptionally important in the planning of security, and only 2% find the type and model of wireless equipment not important in the planning of wireless network security.

H. Question nr. 4e. Spatial coverage by wi-fi network.

Spatial coverage was considered as exceptionally important in the planning of network security by 57% of the respondents. 0% of them consider the spatial coverage as not important. 17% and 19% of the respondents consider it as being important and very important respectively. Bearing in mind the covered space and the limited range of the wireless network signals emitted by the active wireless equipment (access point, wireless routers) and the phenomenon of “secondary petal?” effect the higher degree of spatial coverage by the wireless network represents also a higher security risk. Therefore it is extremely important to know, at the installation of active wireless equipment, what will be the spatial coverage and thus spatial accessibility of the wireless network.

I. Question nr. 4f. Strictly defined security policy.

The existence of security policy in the institution (organization) was recognized by 42% of the respondents as exceptionally important for the security of the information system in the institution as well as in the case of of local wireless network security. 24% and 27% of the respondents consider the security policy as important and very important respectively, and 0% of them think that a strictly defined security policy is not an important security factor in the planning of the wireless network

security. As has been mentioned in Chapter IV of this thesis, CARNet has elaborated a proposal of security policy for schools titled Acceptable use of computers in schools and another document titled Security policy of information systems for the CARNet members. These documents treat the security policy in general terms, and focus on the user rights and responsibilites. These documents carry also instructions realting to the use and wireless network security (elaboration of the documentation of the computer network in the institution, record of the assigned IP addresses range to specific parts of the computer system, fine tuning of the wireless network before operation, the importance of encryption of the network traffic, implementation of the encryption and authenitaction method, etc. The task of the security policy inculdes also monitoring changes in properties of the newtwork equipment in order to respond swiftly and efficiently to security incidents. Bearing in mind the presented research findings one can conclude that the respondents are well informed with respect to the importance of the security policy in the institution and on the influence of security policy in the planning od security of the local wireless network. In conclusion one can say that security policy should be planned often, should be implemented and supplemented by new rules and guidelines (recommendations) in order to make it as efficient as possible in preventing security risks and damage to confidetial information in the system.

J. *Question nr. 4g. Simplicity of authentication and authorization.*

53 % of the respondents consider the simplicity of the user authentication and authorization to be of the paramount importance for the wireless network security, 27% of them find it very important and 12% important. Seen from the user side it is necessary to make simple authorization possible. Complex systems where the authentication and connection to wireless network is complicated for the user (like hand written SSID and IP address and network domains) should be avoided, except in cases when security standards are set very high. Educational institutions should choose a wireless security model that provides high security level, but should also simplify the user authenication and access of wireless services. It is interesting to note that 0% of the respondents consider simplicity of authentication/authorization process unimportant for a certain security level. User authentication should as a rule take place in such a way that the user can find the corresponding wireless network SSID and authorize himself by introducing access data (e. g. from AAI@edu.hr system). System EDUROAM also provides a simple access provided the user makes use of the client application on the client computer. podešene na klijentskom računalu (eng. Supplicant).

K. *Question nr. 5. Do you use RADIUS protocol and LDAP in the local (wireless) network?*

The presence of the RADIUS protocol together with LDAP service has been reported in 39% of the institutions and 61% of them do not use RADIUS+LDAP service. It is important to stress that academic institutions (polytechnics, universities and scientific institutes) do use RADIUS protocol te LDAP imenik. The reason for this is that they possess their own infrastructure with CARNet backing, as the key institution supporting member institutions. Elementary and secondary education does not use ADIUS+LDAP in theiw own systems. Electronic user identities in the AAI@edu.hr system are based on the existence of the RADIUS+LDAP systems. From there the users access other systems like ISVU or specific network services (FTP). Security level is reached by the use of the network protocol itself, LDAP service and firewall with the aim of limiting network traffic svrhu ograničenja u mrežnom prometu and permitting the use of specific network services. This system is used in the access of wireless network whereby the security is not based on the use of encryption algoritms, but on the certificates (WPA+TKIP)and client applications on the users' computers korisničkim računalima (e.g. secureW2 client application with in advance defined profile "Eduroam" [6]). The stress lies on the user authenitication process which is carried out by EAP[4]-TTLS protocol. According to the CARNet CERT document CCERT–PUBDOC–2008–01–216 [7], EAP protocol is used on the data network layer OSI (eng. Open System Interconnection) of the referring model. Implementations of the protocol meant for networks with routers and for switched networks are available. In case of wireless networks the impelementation takes place at access points and switches. The advantage of this authentication protocol is the possibility of expanding it ? by selection of the authentication method and simple addition of new methods without having to change the entire implementation of the protocol.

L. *Question nr.6. Which encryption /certification algoritm do you use in the wireless network protection?*

39% of the respondents declared that they use WPA2+PSK certification in order to protect the network from unauthorized access. 12% of them use WPA encryption algoritm, 8% WEP, 15% do not use any encryption algoritm/certification, 7% use WPA and the rest is used by 5%. Although the majority of the respondents recognized the WPA2 certification more efficient in preventng security breaches than WPA/WEP

[4] Extensible Authentication Protocol EAP (Extensible Authentication Protocol Eng.) enables authentication using a supported method of forty, and is commonly used in wireless computer networks.

algoritms, the vulnerability of WPA2 certification is reflected in the use of WPA2+PSK[5] [9].

M. Question nr.7. Evaluate the effciency of the RADIUS protocol and LDAP in the security user authentication.

The efficiency of the RADIUS protocol and LDAP imeničkog service in the security of user authentication according to IEEE 802.3 and IEEE 802.11 standard got from 39 % of the respondents the highest mark. 30% and 27% of them value it as good and very good.5% of the respondents gave it a negative mark. RADIUS protocol and LDAP imenički service are among the main components in the AAI@edu.hr. System. LDAP imenik contains data on electronic identities, RADIUS server is primarily meant for authentication aimed at security in network access (wired and wireless access as well as AOSI web service for data access and maintanence of the LDAP imenika. It is important to stress that the wireless network access due to the need for a higher security level necessarily requires the use of encryption algoritm that protects it.

N. Question nr.8. Have you recorded a security incident in your institution that happened by way of wireless network?

19% of the repondents report at least one security incident and 81% of them claim that no such incidents have happened. Recording security incidents can be related to the question number 3., where 47% of the respondents declared that their institution has a satisfactory level of access protection to wireless network system. That brings us to the conclusion that the institutions take care about the security of wireless network and that they are not exposed to simple security perils that can damage the system (e.g. unprotected wifi network signal, the use of WEP encryption, etc.).

O. Question nr.9. how long did it take you to discover the incident?

The answers to the last question can draw attention to the current state of affairs in the educational institutions. 42% (5) of the respondents report that they needed several days to discover the incident that had occurred. 33% (4) of them discoered the incident after several hours . It is important to note that question number 10. Is consequentially connected to the question that preceeds it on the record of security incidents. 19% (11) of the repsondents have recorded security incidents. Bearing in mind the survey results on can conclude that the security perils are noticed mainly on time. If the institutions could employ more qualified system engineers and/or make a better division in the IT department of the institution, the result would be even better. It is a fact that the system engineers in the educational institutions often deal with a wide scale of IT tasks and thee remains less time to cope with monitoring the system security. Reaction to a signalled problem ususally occurs when one of the services stops functionioning, or does not function any more while possible „interception“ of security perils requires constant monitoring of the system and more work from the system engineers. Anyway, system engineers in the educational institutions do take care of the security system and they also repair the damage to it on a short notice. 0% of the respondents declared that they needed longer than a month's time to discover a security incident.

[5] Pre Shared Key – shared secret key with one of the modes WPA2 certification in the security of wireless networks. The same is not proved reliable because it is possible to bypass through the available software tools such as Aircrack.

VII. CONCLUSION

The research results presented in this paper refer to the field of security of wireless networks with the emphasis on the educational institution in the Republic of Croatia. The reasons for choosing the relation of wireless network security and the educational institutions are the security lapses and perils that occur in practical use of the network and in the analysis of security parameters. A very important role of network administrators, that are not present in all educational institutions, is also stressed in the present paper. Besides, the knowledge of the network security is not equally spread in all educational institutions. And since security levels range from complete absence of security to protected networks, the present paper aims at presenting the current state of affairs in the matters of local wireless network security in educational institutions in order to sketch possible future guidelines for the optimal security model. Some average quantitative values of security parameters and assigned weighting factors in the research are shown in table 2 and table 3.

TABLE 2. AVERAGE QUANTITATIVE VALUES OF SECURITY PARAMETERS (RESEARCH)

Security parameter	Average score (research)
1. Encription	**4,03**
2. Type of network hardware (equipment)	**3,74**
3. Wifi coverage	**4,15**
4. Using of firewall	**4,24**
5. Number of wlan users	**3,77**
6. Defined security policy	**3,93**
7. Network services	**3,81**
8. Using of VPN	**3,41**
9. Using RADIUS + LDAP	**3,66**
Average score (total)	**3,86**

TABLE 3. ASSIGNED WEIGHTING FACTORS (WEIGHTS) SECURITY PARAMETERS ACCORDING TO THE RESULTS OF RESEARCH

Security parameter	Ponders (value)	Ponders %
1. Encription	0,15	15,00%
2. Type of network hardware (equipment)	0,08	8,00%
3. Wifi coverage	0,18	18,00%
4. Using of firewall	0,21	21,00%
5. Number of wlan users	0,09	9,00%
6. Defined security policy	0,13	13,00%
7. Network services	0,1	10,00%
8. Using of VPN	0,02	2,00%
9. Using RADIUS + LDAP	0,04	4,00%
Total	**1,00**	**100 %**

According to research presented weights show a high correlation (strong relationship defined within ± 0.70 to ± 0.90) with the results of the study (r = 0.78). Two parameters (VPNs, RADIUS + LDAP) have low weights (2% and 4%) and are almost negligible in the final evaluation, because their ratings are almost no influence on the final result. Table 4 shows the values of the security parameter in the overall security that can display numerical values of 1 - Insufficient security level to 5 - very high level of security.

TABLE 4. SECURITY LEVELS (VALUES) OF RESEARCH

Security level	Value (MIN)	Average (AVG)	Value (MAX)	
1	0	20	29	
2	20	40	49	
3	50	60	69	
4	70	80	89	(r)
5	90	100	99	0,779604014

The present paper researched and analysed the theoretical properties of wireless network security, the existing pieces of legislation relating to security policy in the educational institutions, the well-known types of security perils, the survey of possible security parameters, the specific cases of security lapses and the statistical review, based on a sample, of the current wireless network security status in the educational institutions. The presentations of the theoretical wireless network security properties includes also the study and analysis of the literature dealing with the subject matter and general properties of secuity as well.

VIII. REFERENCES

[1] Barman, S., Writing Information Security Policies. New Riders Publishing, 2002. pp. 3 and 4.

[2] Peltier, R., Thomas, Information Security Policies and Procedures. Auerbach Publications, 2004.

[3] Radojević, Branko, Problematika provođenja sigurnosne politike u visokoškolskim ustanovama RH. MIPRO 2011, zbornik radova, Opatija, 2011. pp.196.

[4] Zakon o informacijskoj sigurnosti, Narodne novine, Zagreb, Hrvatska, http://narodne novine.nn.hr/clanci/sluzbeni/298919.html (22.2.2013.)

[5] WiFi Alliance, http://www.wi-fi.org/wi-fi-certified%E2%84%A2-products (21.08.2013)

[6] STUDOM,http://studom.srce.hr/fileadmin/StuDOM/template/docs/StuDOM_2009_Microsoft_XP_bezicno.pdf (23.08.2013)

[7] CIS,http://www.cis.hr/www.edicija/LinkedDocuments/CCERT-PUBDOC-2008-01-216.pdf (23.08.2013)

[8] CARNet,http://www.cert.hr/sites/default/files/NCERT-PUBDOC-2010-12-001_0.pdf (23.08.2013.)

[9] http://blog.g0tmi1k.com/2010/02/video-cracking-wifi-wpawpa2-aircrack-ng.html (11.2.2013.)

[10] Hadjina, N., Zaštita i sigurnost informacijskih sustava (nastavni materijali), FER, Zagreb, 2009.

[11] Tanenbaum, A. S., Wetherall, D. J.: Computer networks 5. edition, SAD, 2011., pp. 300.

Use of iPads in foreign language classes

K. Starkl Crnković
OŠ Vežica, Rijeka, Croatia
klarastnkv@gmail.com

Summary

The implementation of iPads in the process of teaching and learning brings new opportunities for teachers and learners in and beyond the classroom. This work deals with the experience of the very first five months of the 1:1 model of implementation of iPads in foreign language classes with 11 and 12-year-olds in a primary school in Rijeka, Croatia.

I. INTRODUCTION

By implementing iPads in the teaching process of a group of eight graders, Primary School Vežica started last year a unique project in Croatia. Being the results highly encouraging, the 1:1 implementation model was extended to fifth and sixth graders at the beginning of this school year. Teachers of all subjects have adopted iPads to support their teaching and have been training students to use their own iPads to personalise their learning.

The way iPads are used in and beyond the classroom depends, to some extent, on the subject taught. This work will deal with the way iPads have been integrated in foreign language classes with 11 and 12-year-olds throughout the period of the very first five months of their implementation.

II. GETTING USED TO iPads

Adaptation to the use of the device ran smoothly because most of the students had been using smart phones for a few years and was already well-acquainted with touch screen technology. Furthermore, navigating the touch-based software is a very intuitive process, so none of the students encountered major problems.

During the first three-week-period of familiarisation with the device's main options, such as installing and using an app, an e-book or dictionaries, using writing tools like *Pages* and using the school's virtual learning environment (*Edmodo*), teachers discussed with the students how to do with their iPad what they used to do before with pens and paper, desktop PCs and books as well as how they can learn in ways that were never possible before.

III. iPads AT WORK

Being English he first foreign language for all students in the above mentioned school, teachers of English can easily take advantage of the enormous number of apps available in English. But teachers of other languages such as German, Italian and French (taught to smaller groups of students as second foreign languages in the same school) have successfully implemented the use of iPads in their classrooms being the number of educational apps available for the study of these languages substantial (although somewhere lacking for that particular age group). Besides the greater amount of existing apps for the study of English (as a first language or foreign language), teachers of English can choose from a wider range of educational apps in their target language designed for other subjects in order to integrate topics and material from other subjects into English in a cross curricular learning vision.

With more than 500,000 apps available for use the question of how you find those that best suit your students' educational needs and programme sometimes arises. Collaboration among teachers is in this case essential. In the case described the use of iPads has encouraged greater sharing of resources and ideas among teachers. Teachers can also refer to web sites which provide lists of apps for particular subjects and needs, categorise most used apps according to Bloom's taxonomy and

give useful advice on how to use some apps effectively.[1]

However, we have been able to try out the use of iPad in every phase of the learning process from motivation to revision and assessment for all of the four skills: listening, reading, speaking and writing without completely replacing traditional methods of teaching or becoming a paperless classroom.

Being the iPads fully compatible with online teaching and learning platforms such as *Quizlet*, we made a great use of them in the foreign language classroom, instructing at the same time our students to make the best use of them outside the classroom.

Tools such as audio and video recorders are highly useful in foreign language classes. It is easy to imagine how the iPad with an incorporated camera and microphone makes the difference. Not only the way learning takes place, but also the way homework is completed changes. Students can be encouraged to take part to a live debate for their evening homework and marked on their level of engagement and responses or can be given the task to use the language in a particular exercise, record their voice and send it to the teacher. The latter activity can be a great opportunity for the less self-confident students, who usually do not do much talking in class, to build their confidence gradually.

IV. COMMUNICATION, COLLABORATION AND SHARING

A major effort was put into getting students used to new ways of communication, collaboration and sharing.

The regular, ongoing contact between teacher and student made possible by the device, gives each student individually feedback and information as to which skills or knowledge they should approach at each stage and allows discovery of student's interests, strengths and areas for improvement.

Teachers can in that way help their students learn how to personalise and channel their learning. They can become facilitators for personal learning rather than instructors of facts. The device enables students to approach and process new language content by themselves, in pairs or small groups extremely well. For instance, while reading a new text in the target language, students have at hand a number of options by which they can look up the meaning of a new word or expression, hear it pronounced (and replay it as many times it is necessary), find the translation in their mother tongue, see the word used in different contexts, find images, videos, links and so one. The result is a more integrated learning experience which makes students used to problem solving and handling complex tasks and which is more engaging and because of that, more memorable for students.

In such a learning environment students are expected to put much more effort into the learning process than they used to do before. They are also expected to take responsibility for their learning. Our fifth-grade-students have been instructed to create, use and reuse their material in different ways in order to study, reinforce or revise lexical units or grammar points. They decide for which words or expressions they need to practise spelling or pronunciation and they insert them in apps such as *Dragon Dictation* or *Spelling Notebook* which they can easily access as many times they feel they need to. Most importantly, each time they access and practise the material, they get immediate feedback about their progress. They are able to check and keep track of their progress and became aware of their own learning. They are in that way taught to plan their learning and become more organized.

Not only communication between teachers and students has been enhanced, peer communication has also improved. The fact that students can easily share materials of any kind (texts, audio, photos, videos, links) gives great opportunity for collaboration to take place. After certain content has been learnt, it can be put in a similar or different context with the employment of students' creativity. *Puppet Pals* is an app which has proved to be very effective in the language learning process because it allows students to create an animated story and make the characters talk by recording their voices. In fact it makes

[1] See for example
iPads for Learning
http://www.ipadsforeducation.vic.edu.au/education-apps;
Bloomin' Apps, http://www.schrockguide.net/bloomin-apps.html
103 Interesting Ways to use an iPad in the Classroom,
https://docs.google.com/presentation/d/1UJnqyK52X26az5zN5kr1ISTNPwbdUbxfG6M-9WMW4Do/present?slide=id.i21

them practice the target language in a fun and relaxed context. If the aim is practising writing, than a suitable app could be *Strip Designer*, an app for creating comics.

The last two apps mentioned are the ones that seem to be the most appealing to our foreign language students and that have been rated as very effective by our teachers. However they are, like the other apps mentioned in this work, just a few drops in the ocean of learning possibilities offered by the use of iPads in and beyond the classroom.

V. CONCLUSION

In the described case of implementation of iPads in the teaching and learning process both teachers and students have experienced positive results such as continuous access to engaging educational materials; different ways to help students learn new concepts, express their creativity and personalize their learning; easy ways for learners to connect, collaborate and share and, in more general terms, education moving from a content to a problem solving paradigm.

Both teachers and students have become aware of the enormous potential of the use of iPads, a potential that still has to be explored and put into practice firstly by those who have in mind the pedagogies the device supports: teachers and educators.

Tablet PC & smartphone uses in education (TabletTours)

Krunoslav Bedi
School of Civil Engineering, Art and Craft Čakovec
Športska 1, HR-40000 Čakovec, Croatia
krunoslav.bedi@ck.t-com.hr

Abstract - This paper presents the development and design of TabletTours application, high-school students' extracurricular project activity. TabletTours application is designed and implemented as a virtual tablet-based travel guide.

The additional purposes of this paper are to examine the manner in which students use new technologies (tablet PC, smartphone) and to showcase an example of a successful implementation of acquired ICT knowledge and skills in a project-oriented classroom.

Keywords: tablet PC, smartphone, multimedia computer interaction

* In 2013, the application "TabletTours" was presented and won the 3rd Prize at Info kup – Primošten 2013. The trailer of the game can be accessed via – http://www.youtube.com/watch?v=NGKlsL30wlM.

I. INTRODUCTION

In this day and age children are immersed into the digital world. To overlook and / or ignore this fact would be a fatal mistake. The teaching staff have to be given a fair chance to effectively capacitate the children for the world in which they live and will continue to live in. A change in the teaching process itself is necessary. Two approaches to enable a wider population to benefit from ICT revolutions are: to enhance level of ICT literacy (including basic, functional technology and computer education), and to design appropriate IT tools around the capabilities of users [1].

Digital technologies, primarily tablet PCs, smartphones or laptop computers, are currently imposing a need for a new pedagogical paradigm. The new learning environment calls for educational and technological innovations. The involvement of educational experts, professors and teachers is crucial to ensure that new multimedia technology serves the educational process benevolently. There is a need for further investigation of student attitudes and teachers views about the use of tablet PCs [2].

Tablet PC is the most popular computer of our time which is a portable personal computer typically smaller than a notebook computer but larger than a smart phone, and it is easy to transfer the data by means of its internet and memory device [3].

Smart Phone: It is a kind of communication device that has been designed by adding the features of PDA. Due to the fact that smart phones have mobile operating system and many applications, they are very common devices used actively in all areas for different purposes [3].

The need for information and the increasing degree of educational autonomy require a proactive approach hence curiosity and exploration in the digital world, which are the natural drivers of learning, enable an even faster retrieval of desired information. The integration of mobile technologies into the educational system goes beyond its sole availability [4].

Social networks are not only a new, increasingly important social dimension but also a gateway to improving speaking and writing skills.

In a traditional educational environment only the teachers, possibly parents, could have read student assignments. However, it is now necessary to take into account that these assignments will be available and visible to a wider audience. New technologies are also suitable for students with learning disabilities. The use of tablet PCs becomes a tool to give greater attention to this type of student with school difficulties [5].

Characteristics and personality traits of each child are improved and made flexible by digital technologies. A growing number of scientific disciplines indicate the importance of interactive learning approaches which are made available through digital contents and applications. In simple terms, multimedia-digital contents become a relevant source of information displayed in an attractive way with the possibility of direct communication and student interaction. Such digital contents open many audio and visual channels for students to choose from according to their respective personal preferences. This situation motivates exploring the potential of education-aimed netbooks and tablet PCs with a view to determining which of these devices would be the most appropriate for the classroom [6].

For example, a cleverly designed digital textbook becomes an important source of the latest information presented in a visually appealing way with the possibility of direct communication and interaction between the student and the "environment". Namely, the digital textbook allows simultaneous audio and visual stimuli. If a child-student shows more interest, available links and websites provide the possibility for further knowledge deepening in any area of interest. Smartphones are no longer cutting-edge communication gadgets, but are now necessities in peoples' lives. For smartphone users, their phone is the first thing they look at in the morning, and the last thing they look at before going to sleep [7].

The concept of a new upcoming school system changes the role of the teacher / educator significantly. It is obvious that the teacher is no longer the primary source of information. The focus shifts towards giving meaning and sense to new information in a wider context which would allow the children to understand these complex processes better. The teacher would also have a significant role in helping students engaging in the process of abstracting and differentiating important and reliable information from the less important one. Due to the fact that unlimited access to large databases imposes an entirely new task for the children, the professional help of teachers is extremely important in this crucial cognitive process. This reality has prompted considerable interest in finding fruitful ways of integrating mobile technologies, such as wireless laptops and tablet computers, into educational settings [6].

II. THE APPLICATION OF DIGITAL TEACHING IN THE SCHOOL SYSTEM

To educate and optimally prepare the children for the digital world is not only one of the fundamental tasks of the contemporary school system but a responsibility of the society as a whole. The quick and recent upsurge of mobile IT devices is a relevant development for education [8].

Is it really so or is it yet another craving for profit? Large firms always have strong marketing strategies and their expensive computers offer a number of features regardless of the fact that many of these software packages and options are rarely, if ever, used. We have failed to develop a proper system of setting professional educational criteria and assessing the quality of the offered equipment and educational materials. It is therefore not difficult to assume that the educational website "market" could become a commercial jungle very soon if an effective way of establishing order and adherence to generally accepted criteria is not introduced.

The slow and complex process of preparing textbooks for print contributes to delays in the teaching material modernization. A tablet PC, which every schoolchild could have eventually, would provide a systematical syllabus update offering children the latest knowledge and scientific discoveries tailored to fit their age. Should we compare the prices of a lower price range laptop/tablet to a set of class textbooks, workbooks and school paraphernalia, the investment would certainly pay itself off within a period of two or three school years. We all agree that children are a country's greatest wealth and therefore the quality of their education determines the overall development of a society. With the growth and convergence of mobile and social computing technology, mobile media are accorded new social meaning manifested in different aspects of our everyday life [9].

III. DIGITAL TEACHING EXPERIENCES

Many countries are turning to digital textbooks with the ambitious goal of fully digitalizing textbooks for primary, secondary and tertiary educational institutions by the year 2015. However, there are initiatives which include combining digital with paperback textbooks. Mobile devices such as personal digital assistant (PDA), mobile phone, and Tablet PC are nowadays more user friendly and user convenient than before. They are coming with major improvement in memory storage, interactivity features and high data transfer speed. Mobile devices are neutral to teaching and learning theories; they can be used with traditional learning theories such as behaviorism and new learning theories such as constructivism [10].

The development of new communication technologies, especially collecting and processing of required or new information, is likely to change the curricular context in the near future, emphasizing the individual approach and the technical knowledge necessary for using all available sources of information and knowledge. The proliferation of mobile technology provides a myriad of opportunities to support learning and performance both inside and outside the classroom. With mobile technology, the learning environment can go with the student to the field site, to the laboratory and beyond. There is an opportunity to leverage mobile technology to better support students not only in the classroom, but also as students navigate to the context of their learning. Mobile technology opens the door for a new kind of learning and performance support in the field, providing anytime and anywhere access to information, processes, and communication [11].

Today's digital communication technology, if used wisely, could most certainly serve as a useful eduactional platform for developing quality school programmes and curricula. Mobile media devices (MMDs), consisting of smartphones and tablet computer s (e.g., iPad, Surface RT), not only strengthen our social bonds but also provide us a

completely new way to interact with information [12].

A very interesting developmental project "One Laptop per Child (OLPC)" was launched in the year 2002. Its aim was to open new doors for children in developing countries and give them a chance to learn and access information easily. If we consider the fact that in today's world there are still almost two billion children without adequate access to education, which is readily available to their peers in developed countries, each step taken in this direction is not solely a humanitarian gesture. It is a firm foundation for strong and long-term development and offers assistance to developing countries. The project of a technical device which would enable easier learning thus modernizing the educational process certainly transcends the Internet Age.

The schools are to use the iPads for replacing physical book and submitting coursework, while there will be no longer be a place for the more traditional pens, pencils, blackboards or even classrooms, which are replaced by 'subject rooms' [13]. The BYOD (Bring Your Own Device) model is becoming an increasingly attractive and inexpensive option. That is, students bring their own devices. The cost of the devices, their maintenance and connectivity plans are usually shouldered by the students. In a world increasingly dominated by mobile devices BYOD projects can be implemented quickly and cheaply in areas where most students have such devices. However, BYOD has serious limitations if it fails to accommodate students who do not have mobile hardware or creates situations where some students have more sophisticated devices than others [14].

IV. „TABLETTOURS" APPLICATION PROJECT

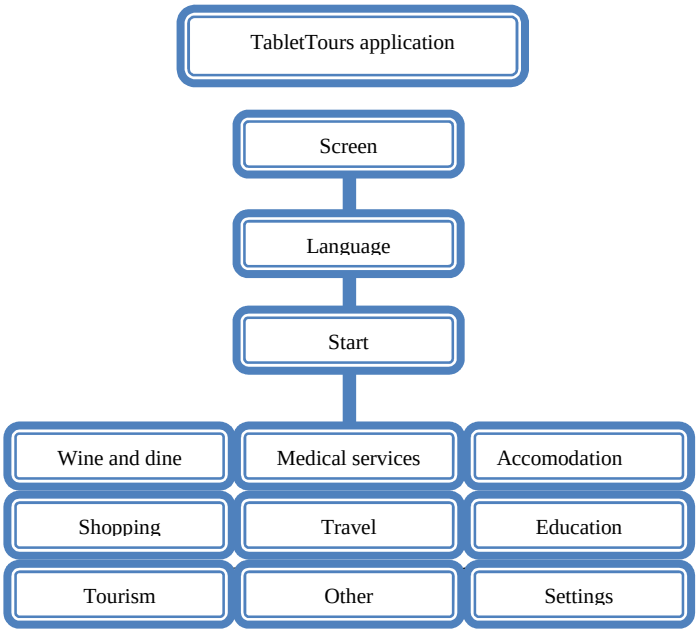
Two fourth-grade students started creating the entitled „Tablet Tours" application under the mentorship of the author of this paper. A wide range of software normally used in the four-year course was utilized. In a four-year web-designer/media technician course the entire necessary software infrastructure has been provided by the Ministry of Education and Sports of the Republic of Croatia. The work itself was team-organized since teamwork had proven to be far more efficient than individual work, more precisely, one of the students doing the programming and the other one the visual designing. The students worked on the application during their optional subject/activity time with the key role of the coordinator assigned to the mentor, i.e. the optional subject teacher, who is also the author of this paper. In their teamwork, the students chose a specific name, "DreamLand"", which can be seen from the Youtube attached trailer. The assigned name has given the teamwork a whole new dimension and a touch of "professionalism" as well as additional motivation for the students. A Facebook profile (a group) promoting the application was created for the purpose of the State IT Competition – Infokup 2013. Primošten.

TabletTours project, derived from project-based learning, was developed in co-operation with the local Tourist Board. Prior to the development of the application itself usability and marketing-driven approaches had been contemplated on so that the application could be launched in practice.

The application is designed not only as a virtual guide but as a specific guide for tourists as well. In fact, travel guides offered by TabletTours application can be used as a helpful tool in a particular field of work or sightseeing. Tourists / customers can use the application in accordance with their particular interests.

The application idea comes from the observations and experiences of students during school or travel (for example, a school trip).

Application scheme:



Processor	1GHZ+
Memory	256MB
Resolution	Automatic
Operating systems	Android v2+, iOS, Windows 7, 8

Teamwork

Developmental background

Software used in the making of „Tablet Tours“ and assigned documentation: Adobe Flash CS5.5 Professional, Adobe Premiere Pro CS5.5, Adobe Photoshop CS5.5, Adobe After Effects CS5.5, Microsoft Word 2010, te Camtasia Studio 8.

Graphic design

The graphic designer had to develop a functional, dynamic, understandable and interactive user interface. His additional jobs were designing a promotional leaflet as well as the project's website. The graphic designer used the following software: Adobe Photoshop CS5.5 i Adobe Flash CS5.5 Professional.

Programming

The task of the programmer was coordinating all the graphic design concepts with the software solutions. In the making of „TabletTours“ application the programmer used ActionScript as the basic software system which ensured the application would work properly. The programmer showed the respective skills in the use of Adobe Flash CS5.5 Professional with AdobeAir technology (for Android, for iOS, Windows).

Video design

The task of the video designer was coordinating the multimedia and content requests with functionality and the educative quality. He made this possible by using the following software: Adobe After Effects CS5, Adobe Photoshop CS5.5, Adobe Premiere CS5.5.

Audio design

The application's audio design consisted of producing trailer music and adjusting sound to „listening“ option.

Application control

The tablet has a touchscreen enabling touch control of the appplication. However, if the tablet supports a dock keyboard, a mouse/keyboard can be used. The application is adjusted for Android, iOS and Windows 8 operating systems.

Application screenshots

TABLE 1. PANELS/BACKGROUNDS FROM TABLETTOURS APPLICATION



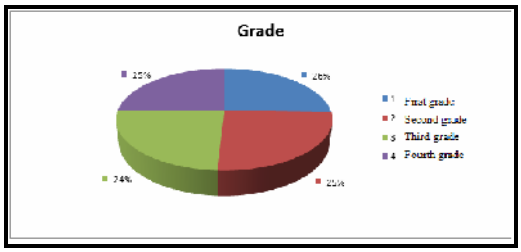
V. RESEARCH METHODS

In order to obtain relevant data, a research involving a sample of 311 students (first, second, third and fourth grade respectively) was conducted. All of the students involved attend the following vocational programmes in School of Civil Engineering, Art and Craft Čakovec: media technician, web-designer, interior architecture designer and ceramics designer. The research was conducted by survey. The survey sample included both sexes, 140 male and 171 female students. The students were given the possibility to withdraw at all times, to participate or decline participation, therefore one can assume that the obtained data is relevant.

The aim of the survey was to get feedback on:

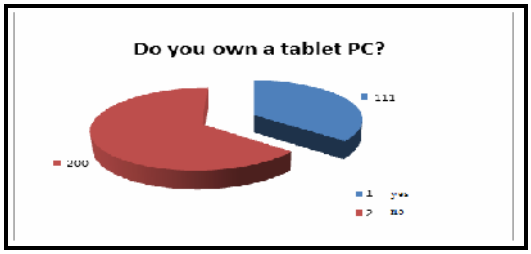
- whether today's students have a tablet PC or a smartphone
- whether today's students use or will use a tablet PC or a smartphone
- whether students use a tablet PC or a smartphone to follow lectures
- the extent in which they justify the use of a tablet PC or a smartphone
- the reasons for using new technologies instead of "paper" handouts
- any indication of future use
- whether the results indicate a change in digital content acceptance

VI. RESEARCH RESULTS



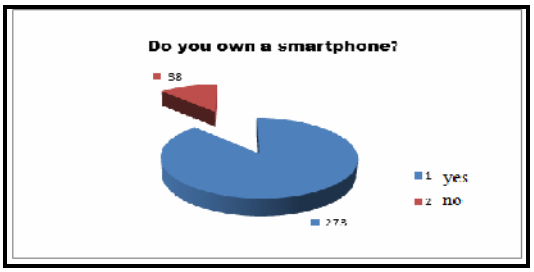
Graph 1. The percentage of students participating in the survey

An equal number of first-, second-, third- and fourth-grade students participated in the survey, as follows; 26 percent of first-grade students, 25 percent of second graders, 24 percent of third graders and 25 percent of fourth-grade students.



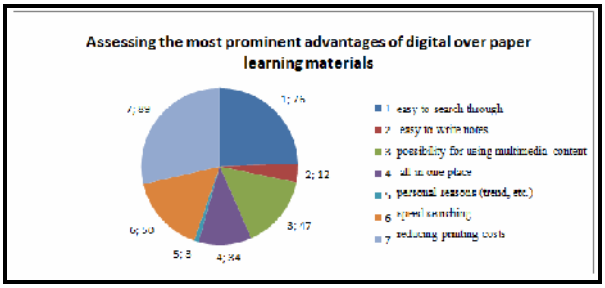
Graph 2. Owning a tablet PC

The majority of the students included in the survey do not own a tablet PC.



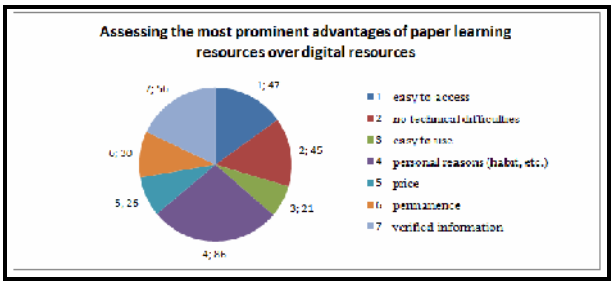
Graph 3. Do you own a smartphone?

The given chart illustrates the fact that the majority of the participants own a smartphone, which could be explained in the context of popularity of smartphones among younger population.



Graph 4. Assessing the most prominent advantages of digital resources over paper learning resources

The students consider speed searching to be the most prominent advantage of digital over paper learning resources (89 out of 311 students). The students have also sincerely admitted the importance of personal reasons (50 out of 311).



Graph 5.Assessing the most prominent advantages of paper learning resources over digital resources

When assessing the advantages of paper learning resources over digital ones, the largest number of responses falls in the "personal reasons" category, i.e., habit. It is obvious that the students still consider paper-based learning material to be indispensable. The evident uniformity of the remaining answers indicates that there is no undue preference.

VII. CONCLUSION

Tablet PCs offer a new perspective to the children by enabling a simpler and friendlier way of learning, gathering new information, switching to interactive and / or multimedia world. Here and now learning has the ability to engage learners because of its authentic learning and context based applications [11].

We all agree that an abrupt change is not good for the educational system. New technologies, in spite of their many advantages, should not replace paper textbooks completely.

Therefore, the common conclusion is that a combination of digital and paper textbooks is the most acceptable one, that is, tablet PCs should be made available as classroom tools. It is not possible to replace all paper materials with digital ones without affecting learning efficiency. We should keep in mind that there are advantages and limitations in the use of new technologies in lesson planning. Although tablets are increasingly popular in education, there are still obstacles and barriers that may limit the use of them in the educational landscape and more specifically in teaching [15]. However, there is only one possible answer to the question „will new technologies change the teaching process“ – a resounding yes.

REFERENCES

[1] P. Kanthawongs and P. Kanthawongs, Perception of Primary School Students, Parents and Teachers toward the Use of Computers, the Internet and Social Networking sites Original Research Article *Procedia - Social and Behavioral Sciences, Volume 88, 10 October 2013, Pages 282-290*

[2] H. Dündar and M. Akçayır, Implementing tablet PCs in schools: Students' attitudes and opinions Original Research Article, *Computers in Human Behavior, Volume 32, March 2014, Pages 40-46*

[3] I. Goksu and B. Atici, Need for Mobile Learning: Technologies and Opportunities Original Research Article *Procedia - Social and Behavioral Sciences, Volume 103, 26 November 2013, Pages 685-69*

[4] D. Ifenthaler and V. Schweinbenz, The acceptance of Tablet-PCs in classroom instruction: The teachers' perspectives Original Research Article *Computers in Human Behavior, Volume 29, Issue 3, May 2013, Pages 525-534*

[5] F. Ferrer, E. Belvís and J. Pàmies, Tablet Original Research Article *Computers & Education, Volume 56, Issue 1, January 2011, Pages 280-288* Ferran Ferrer, Esther Belvís, Jordi Pàmies

[6] C. Alvarez, C. Brown and M. Nussbaum, Comparative study of netbooks and tablet PCs for fostering face-to-face collaborative learning Original Research Article *Computers in Human Behavior, Volume 27, Issue 2, March 2011, Pages 834-844*

[7] Y. Lee, C. Chang, Y. Lin and Z. Cheng, The dark side of smartphone usage: Psychological traits, compulsive behavior and technostress Original Research Article, *Computers in Human Behavior, Volume 31, February 2014, Pages 373-383*

[8] M. B.W. Kobus, P. Rietveld and J. N. van Ommeren, Ownership versus on-campus use of mobile IT devices by university students Original Research Article *Computers & Education, Volume 68, October 2013, Pages 29-41*

[9] C. Lai, An integrated approach to untangling mediated connectedness with online and mobile media Original Research Article *Computers in Human Behavior, Volume 31, February 2014, Pages 20-26*

[10] D- S. Al Hamdani, Mobile Learning Research Article *Procedia - Social and Behavioral Sciences, Volume 103, 26 November 2013, Pages 665-674*

[11] F. Martin and J. Ertzberger, Here and now mobile learning: An experimental study on the use of mobile technology Original Research Article *Computers & Education, Volume 68, October 2013, Pages 76-85*

[12] B. Zhong, From smartphones to iPad: Power users' disposition toward mobile media devices Original Research Article *Computers in Human Behavior, Volume 29, Issue 4, July 2013, Pages 1742-1748*

[13] http://tabtimes.com/news/education/2013/08/21/dutch-ipad-schools-open-their-doors-first-time visited: 15.12. 2013.

[14] http://michaelrice.weebly.com/tablets-in-education.html visited: 15.12. 2013.

[15] D. Audi and R. Gouia-Zarrad, A New Dimension to Teaching Mathematics Using iPads Original Research Article *Procedia - Social and Behavioral Sciences, Volume 103, 26 November 2013, Pages 51-54*

Informatika kao izborni predmet ili izvannastavna aktivnost u razrednoj nastavi

Jelena Šurić, mag. prim. educ.*, Tea Pavičić, mag. prim. educ.**, dr. sc. Mario Dumančić, doc.***

* Osnovna škola „Braća Radić“, Koprivnica, Croatia, suric.jelena@gmail.com
** Prva osnovna škola, Varaždin, Croatia, tea232pavi@hotmail.com
*** Učiteljski fakultet, Sveučilište u Zagrebu, Croatia, mario.dumancic@ufzg.hr

Sažetak - **Glavni cilj ovog rada i istraživanja bio je sistematizirati i prikazati važnost uvođenja informatike kao izbornog predmeta ili izvannastavne aktivnosti u područje razredne nastave. Provjeren je i odnos djece u dobi od 7 do 10 godine prema digitalnoj tehnologiji, njihova empatija i znanje. Istraživanje je provedeno na prigodnom uzorku od 400 učenika razredne nastave u pet osnovnih škola u Varaždinu i Koprivnici. Svaki je učenik ispunio anketni upitnik Računalo i ja. Rezultati su pokazali da gotovo svi učenici posjeduju i koriste računalo i Internet. Većina ispitanika dio svog slobodnog vremena provodi koristeći digitalnu tehnologiju isključivo za zabavu i suzbijanje dosade. Rezultati konkretnih informatičkih pitanja pokazuju nezavidno znanje učenika o osnovama digitalne tehnologije, no njihova zainteresiranost za stjecanjem novih, informatičkih znanja stvara temelje za stručnim uvođenjem učenika razredne nastave u svijet informatike.**

I. UVOD

Glavni cilj nastave informatike podrazumijeva osposobljavanje učenika za korištenje osobnog računala. To korištenje osobnog računala obuhvaća znanja o dijelovima računala i njihovim ulogama te o programima za upravljanje i sigurnom radu na računalu. Zašto bismo ova znanja i sposobnosti ograničili samo na učenike viših razreda osnovne škole? Zašto učenicima nižih razreda osnovne škole, koji također žele, znaju, mogu i uče vrlo brzo, uskraćujemo tu priliku?

U Nastavnom planu i programu Republike Hrvatske ne postoji dio koji bi obuhvatio informatiku od 1. do 4. razreda osnovne škole. Ovaj složeni zadatak izrade nastavnog plana i programa još uvijek nije napravljen. Izrada programa informatike mora obuhvatiti niz specifičnosti poput starosne dobi učenika, njihov psiho-fizički razvoj, mogućnosti ICT tehnologije, specifičnosti metodičkog i didaktičkog pristupa itd. Na primjer, uzimajući u obzir dob, znamo da se sposobnosti rješavanja problema kod djece razlikuju ovisno o dobnim skupinama, kao i ono što su oni u stanju razumjeti. Osim toga, ne smijemo uzeti u obzir samo globalni koncept (Nastavni plan i program, udžbenike za različite razrede, scenarije predavanja i terminologiju koja se koristi u zadatcima za učenike), nego i pedagoški materijal i odgovarajući obrazovni softver.

Bischof i Sabitzer u svome članku *Computer Science in Primary Schools – Not Possible, But Necessary?!* neostvarivost naslovom postavljene teze vide u problematici planiranja i programiranja nastave informatike i određivanja njezinog glavnog cilja [3]. Prema njima, najvažniji didaktički principi, iza samog sadržaja, koji nužno mora biti prilagođen spoznajnim i razvojnim psihosocijalnim karakteristikama djeteta, moraju podrazumijevati istraživačko učenje i timski rad. Autori smatraju da bi glavni cilj provođenja nastave informatike u razrednoj nastavi između ostaloga trebao biti: povećanje, odnosno stvaranje interesa za korištenje ICT, motiviranje djevojčica u korištenju tehnologije, proširivanje stava prema ICT, usmjeravanje prema pravilnom načinu korištenja računala i pokazati učiteljima da podučavanje područja ICT nije toliko teško kao što mnogi od njih pretpostavljaju [3].

U članku Laurie Patsalides, *Why and How to Use Computers in Classrom,* pronalazimo neke prednosti pravilne integracije računala u nastavni proces gdje bi učenici stjecali vještine obrade teksta radeći zajedno na projektima i razvijali međusobnu suradnju između učenika te između učenika i učitelja. Rad na računalu zahtjeva više razine razmišljanja, potiče učenje putem pokušaja i pogrešaka te razvija odgovornost, samostalnost i osjećaj postignuća [11]. U Velikoj Britaniji istraživanja su pokazala izrazito poboljšanje obrazovnog učinka nakon ulaganja u ICT u osnovnim britanskim školama [13]. Učinak ICT-a promatran je u trima područjima: engleskom jeziku, matematici i znanosti na kraju osnovnog obrazovanja. Utvrđena je pozitivna korelacija između financiranja u ICT po učeniku i postignućima u engleskom jeziku.

Ron Canuel i Thierry Karsenti 2009. godine pokreću istraživanje u jednoj kanadskoj osnovnoj školi. Istraživanje je pokrenuto s ciljem stjecanja što boljeg razumijevanja prednosti i izazova u korištenju prijenosnih računala u osnovnim i srednjim školama. U svrhu istraživanja podijeljeno je oko 5 600 prijenosnih računala učenicima od 3. do 11. razreda. Rezultati istraživanja prikazuju mnoštvo prednosti korištenja prijenosnih računala u nastavi grupiranih u četiri glavne kategorije: olakšavanje i unaprjeđenje školskih zadataka,

unaprijeđeni psihosocijalni faktori – motivacija, autonomija, interakcija (između učenika, učitelja i roditelja) i pažnja, olakšan pristup informacijama, učenje je prilagođeno pojedincu, stjecanje i razvoj vještina i otvorenost prema svijetu i prilika za budućnost. Spomenute prednosti jasno pokazuju da korištenje prijenosnog računala ili intenzivno korištenje informacijsko komunikacijske tehnologije ima veliki utjecaj na učenički uspjeh te potencijalno i na njihov budući akademski uspjeh, društveni i poslovni život [7].

Proučavajući neke europske obrazovne kurikulume i razine školovanja uočeno je da se vrijeme uvođenja informatike u školovanje djeteta značajno razlikuje. Njemački, Švicarski i Austrijski obrazovni kurikulumi informatičko obrazovanje najčešće započinju oko 13.-14. godine života. Francuski obrazovni sustav visoko je centraliziran i podijeljen u tri faze: primarno obrazovanje (od 6. do 11. godine), sekundarno obrazovanje te više obrazovanje (diplomski studij, magisterij i doktorat). Primarno obrazovanje traje 5 godina i podijeljeno je u dvije glavne razine učenja. Informatika (Technology) kao predmet javlja se u drugoj razini primarnog obrazovanja koju pohađaju djeca od 8. do 11. godine života i zauzima prosječno 2,30 sati tjedno. S nastavom informatike nastavlja se i na sljedećoj obrazovnoj razini (prva razina sekundarnog obrazovanja) samo s povećanim brojem tjednih sati.

Ne želimo reći da korištenje računala u nastavi i učenju ima samo pozitivne utjecaje, ima i negativnih, no njih ne smatramo prijetećima ukoliko naglasak u cjelokupnom radu stavimo na pozitivnu stranu, a ono negativno osvijestimo i pokušamo korigirati. Uzmemo li u obzir nužnost i neophodnost primjene informatike u radu i životu djeteta ne preostaje nam drugo nego fokusirati se na osmišljavanje načina, metoda i programa rada u nižim razredima osnovne škole.

„(...) spomenuli smo da je pred vratima učionice uvijek gužva, ali još su nezaboravniji komentari učenika koje čujemo kad idu kući: Bilo je super! Baš je super na infici! Šteta da je već gotovo! Najčešće rad ne prekida školsko zvono, kao što je uobičajeno, već gašenje računala od učitelja – inače neki ne bi još dugo poslije zvona napustili učionicu! Nije li to najbolji dokaz da većina učenika računalo shvaća igračkom, a da njegove velike odgojno – izobrazbene uloge nisu niti svjesni.“ [8]

II. ISKUSTVA UČITELJA INFORMATIKE NIŽIH RAZREDA

Uvođenje informatičke kulture kao i samog predmeta informatike u škole odvijao se vrlo sporo. Nakon niza pokušaja i pokusa, 1985. godine informatika prodire u cjelokupan odgojno-obrazovni sustav, uključujući i vrtiće. Taj je način uvođenja informatike bio vrlo kritiziran i slabo prihvaćen od strane društva i samih djelatnika odgojno-obrazovnih ustanova. Radni materijal bio je neprilagođen i neprimjeren za pedagošku uporabu te nerijetko nepotpun, a izobrazba učitelja bila je nedovoljna i najčešće se temeljila na kratkim tečajevima koji su bili organizirani od strane škole. Kao glavni uvjet da bi informatika prodrla u škole, navodi se postojanje nastavničke ekipe koja bi bila informatički obrazovana i dovoljno motivirana za rad [9].

Tko danas može opisati mjesto koje će informatika zauzimati u obrazovanju 2060. godine, kada u mirovinu odu nastavnici koji su počeli raditi 2010. godine? No, Lesourne ističe činjenicu da prelazimo iz društva materije i energije u društvo znakova. Živimo u svijetu gdje će svaki pojedinac morati znati mnoštvo jezika te gdje će njegova uspješnost i njegovo snalaženje u društvu ovisiti velikim djelom o sposobnostima da dekodira još prije nekoliko godina nepoznate znakove. Sudeći prema navedenim činjenicama, informacijska će tehnologija morati imati svoje mjesto i u početnom obrazovanju osnovnoškolaca [9].

U Republici Hrvatskoj, informatika kao predmet u razrednoj nastavi, provodi se danas samo u nekim školama. Kako bismo saznali nešto više o samoj realizaciji i učinkovitosti predmeta informatike, ispitali smo nekoliko učitelja koji predaju informatiku u nižim razredima na području Republike Hrvatske.

Zanimalo nas je tko provodi nastavu informatike, kolika je zainteresiranost učenika za spomenuti predmet te koje su, po njima, prednosti ili eventualni nedostaci nastave informatike u nižim razredima.

Saznali smo da se informatika kao predmet u nižim razredima provodi isključivo kao izborni za koji učenici ne dobivaju opisne ni brojčane ocjene. Nastavni kadar koji provodi nastavu informatike vrlo je raznolik. Većinom su to učitelji razredne nastave koji imaju završenu informatiku kao pojačani predmet ili učitelji razredne nastave s položenim informatičkim tečajem, dok su rijetko na tim mjestima zaposleni diplomirani učitelji informatike. Posebna edukacija za provođenje ovog predmeta uglavnom nije potrebna. Učitelji obrađuju teme po HNOSU i uglavnom nisu zadovoljni planom i programom koji je tamo ponuđen. Smatraju da je programiranje u nižim razredima suvišno i da je to bespotrebno opterećenje učenika u toj dobi. Neki su pak nezadovoljni i kažu da su prepušteni sami sebi, što se tiče plana i programa. Ne rade po udžbenicima jer je to izborni predmet i kao takav ne smije zahtijevati od roditelja dodatno financijsko opterećenje, što bi kupovanje istih podrazumijevalo.

Zainteresiranost učenika za spomenuti nastavni predmet već od prvog razreda je maksimalna. Gotovo 100% učenika pohađa informatiku kao izborni predmet u školama koje to omogućuju. Učenici dolaze s bogatim predznanjem. Gotovo svi učenici znaju upaliti i ugasiti računalo te pokrenuti neku igru. Malo je onih kojima je to prvi susret s računalom, no prema učiteljima, to je najbolji način da učenici pravilno nauče koristiti računalo.

Poučeni svojim osobnim iskustvom i radom, učitelji informatike smatraju da uvođenje informatike kao predmeta u niže razrede ima znatne prednosti i da je više nego potrebno. Kao glavne prednosti navode da ranije učenje informatike sprječava usvajanje pogrešnih navika vezanih uz korištenje i rad za računalom koje je kasnije jako teško, gotovo nemoguće, ukloniti. Umjesto besmislenog surfanja i igranja igrica, učenici kvalitetno primjenjuju i nadograđuju svoje znanje. Samostalno i kontrolirano istražuju i služe se Internetom, prezentiraju svoje uratke te svakim novim otkrićem stječu dodatnu motivaciju pokazujući veliki interes za daljnji rad. Također, razvijaju logičko mišljenje i prihvaćaju računalo kao jedan od „alata“ koji im može pomoći kod rješavanja svakodnevnih zadataka. Naposljetku, predznanje spomenutih učenika daleko je veće u višim razredima od učenika koji se tek tada upoznaju s informatikom kao predmetom.

Osim spomenutih prednosti, dobili smo uvid i u neke od nedostataka koje za sobom nosi uvođenje informatike kao predmeta u niže razrede osnovne škole. Većina učitelja kao glavni razlog navodi neadekvatnu stručnu spremu učitelja koji provode informatiku te smatra da bi bilo potrebno organizirati više informatičkih tečajeva. Veliki problem predstavlja i slaba opremljenost škola informatičkom opremom, što onemogućuje provođenje nastave. Kod učenika pak postoji opasnost od razvijanja ovisnosti o računalima, ona im postaju jedina zabava i preokupacija u slobodno vrijeme zbog čega može doći do raznih senzornih oštećenja i zdravstvenih problema (pretilost, tromost, umor).

III. METODOLOGIJA ISTRAŽIVANJA

Budući da je cilj našeg istraživanja bio ispitati, sistematizirati i prikazati važnost uvođenja informatike kao izvannastavne aktivnosti ili izbornog predmeta u razrednu nastavu provjerili smo kakav je odnos učenika razredne nastave prema informatici, računalu i internetu. Na osnovu toga, postavili smo cilj istraživanja: Prikupiti informacije o korištenju i poznavanju računala i interneta učenika razredne nastave, saznati njihovo mišljenje o nastavi informatike, provjeriti na koje se sve načine učenici služe računalom te kako se pritom osjećaju.

Istraživanjem smo željeli provjeriti odnos učenika razredne nastave prema informatici i digitalnoj tehnologiji stoga polazimo od sljedećih hipoteza: Većina učenika posjeduje računalo i koristi ga više od 1 sata dnevno. Učenici većinu svog slobodnog vremena provode koristeći računalo no to ne podrazumijeva visoku razinu njihova informatičkog znanja. Učenici već u samim početcima korištenja računala uviđaju važnost informacijske pismenosti.

IV. UZORAK

Istraživanje je provedeno školske godine 2012./2013. na uzorku od 400 učenika razredne nastave, od toga 190 djevojčica i 210 dječaka. Istraživanje je provedeno u pet osnovnih škola: u Varaždinu (207 učenika – 51,75% ispitanika) i u jednoj koprivničkoj osnovnoj školi (193 učenika – 48,25% ispitanika). Uzorak je prigodan.

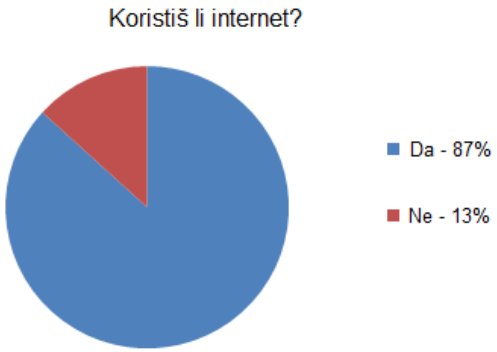
TABLICA I. PRIKAZ UZORKA - BROJ UČENIKA PREMA SPOLU (M – MUŠKO, Ž - ŽENSKO) I RAZREDU (R)

	1.r.	**2.r.**	**3.r.**	**4.r.**	**UZORAK**
m	58	34	59	59	210
ž	56	46	43	45	190
Ukupno	114	80	102	104	400
%	28,5	20	25,5	26	100

Odnos učenika prema računalu i internetu te njihov stav prema nastavi informatike ispitan je anketnim upitnikom *Računalo i ja.* Upitnik se sastoji od 16 pitanja. Pitanja su većinom zatvorenog tipa, od kojih neka pitanja nude mogućnost samostalnog davanja odgovora. Zadatak učenika bio je odabrati i zaokružiti jedan od ponuđenih odgovora ili dopisati svoj odgovor. Ispitivanje su proveli autori ovoga rada u dostupnim osnovnim školama. Nakon što su učenicima objašnjeni svrha i cilj ovog anonimnog istraživanja i način njihove provedbe pročitana je uputa o načinu ispunjavanja anketnog upitnika i krenulo se u samu provedbu koja je trajala 15ak minuta.

V. PRIKAZ REZULTATA

Sukladno našim očekivanjima, prije provedenog istraživanja, analizom anketnih upitnika utvrđeno je da 98% učenika posjeduje i koristi računalo, a čak 87% njih služi se internetom.

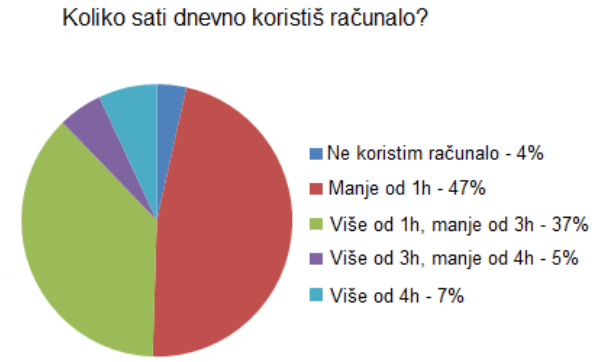


Graf 1. Prikaz učenikovih odgovora na pitanje o posjedovanja računala (N=400)

Dublja analiza odgovora na pitanje o zadovoljstvu učenika prilikom korištenja računala pokazala je da se većina učenika (62,5%) osjeća sretno, kod 37% njih korištenje računala ne izaziva nikakve posebne emocije (osjećaju se isto kao i kad rade nešto

drugo), dok se samo dvoje učenika (0,5%) prilikom korištenja računala osjeća tužno. Učenici su upitani znaju li njihovi roditelji uvijek što oni rade na računalu. 87,8% ispitanika odgovorilo je da su njihovi roditelji upućeni u njihove aktivnosti na računalu dok navodno manji broj roditelja (10%) ne zna čime se njihova djeca bave kad koriste računalo.

Pri analiziranju odgovora na pitanje koliko sati dnevno koriste računalo dobili smo s jedne strane zabrinjavajući, ali također i poticajni podatak. Smatramo da je dobiveni postotak zabrinjavajući obzirom na dob ispitanika, njihove razloge i informiranost o tehnologiji koju koriste, što ćemo detaljnije i prikazati daljnjom analizom našeg istraživanja. Međutim, ovaj visoki postotak prikazuje njihovu visoku motiviranost i interes za korištenje računala, što bi se svakako trebalo nadopuniti, iskoristiti i usmjeriti k ostvarenju zajedničkog društvenog i osobnog cilja, a to je vođenje mlađih naraštaja k informatičkom društvu znanja.



Graf 2. Prikaz učeničkih odgovora na pitanje o vremenu koje provode za računalom (N = 400)

Analiza odgovora na tvrdnjama *Kada najčešće koristiš računalo* (Tablica II.) pokazala je da najviše učenika (63,3%) računalo koristi tek kad izvrši sve svoje školske obveze. Veliki broj učenika (17,5%) naveo je da računalo koristi kad im roditelji dopuste ili kad imaju slobodnog vremena. Važno je istaknuti kako 11,7% učenika navodi da računalo koristi kad god želi.

TABLICA II. DESKRIPTIVNA STATISTIKA POJEDINIH TVRDNJI O VREMENU KORIŠTENJA RAČUNALA

Tvrdnja	Dječaci		Djevojčice		Ukupno	
	N	%	N	%	N	%
Uopće ne koristim računalo	4	1,9	6	3,2	10	2,5
Kad napišem svu zadaću	137	65,2	116	61,1	253	63,3
Kad sam sam/a kod kuće	12	5,7	8	4,2	20	5
Kad god hoću	23	11	24	12,6	47	11,7
Neki drugi odgovor	34	16,2	36	18,9	70	17,5

Učenici su također ispitani o razlozima zbog kojih najčešće koriste računalo (Tablica III). Uspoređujući odgovore djevojčica i dječaka možemo zaključiti da većina djevojčica (45,3%) koristi računalo samo kad im je dosadno, dok najveći broj dječaka (39%) koristi računalo zato što to voli. Zanimljiv je podatak da veći broj djevojčica (14,2%) koristi računalo za izvršenje nekih zadataka, nego dječaci (9%). Iako manje vole računala od dječaka, djevojčice u većoj mjeri uviđaju prednosti i mogućnosti računala.

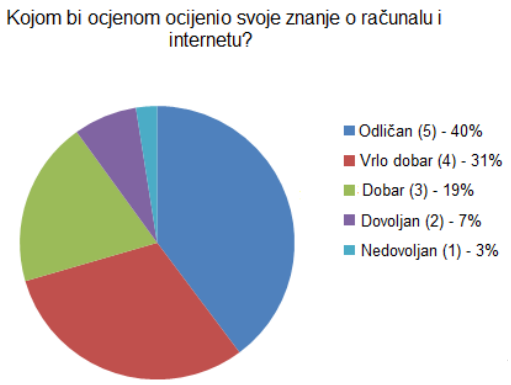
TABLICA III. DESKRIPTIVNA STATISTIKA POJEDINIH TVRDNJI O RAZLOZIMA KORIŠTENJA RAČUNALA

Tvrdnja	Dječaci		Djevojčice		Ukupno	
	N	%	N	%	N	%
Samo kad mi je dosadno	76	36,2	86	45,3	162	40,5
Zato jer to volim	82	39	61	32,1	143	35,8
Zato jer svi moji prijatelji koriste računalo	10	4,8	4	2,1	14	3,5
Kad imam neki zadatak	19	9	27	14,2	46	11,5
Nešto drugo	23	11	12	6,3	35	8,8

Većina učenika, njih 83%, kaže kako su čuli da na Internetu postoji mnoštvo opasnosti koje vrebaju na njih, dok preostalih 17% nije upoznato s tom činjenicom. Činjenicu o postojanju ružnih i/ili neugodnih stvari koje se događaju na Internetu, potvrdilo je 16,7% učenika, dok na naše zadovoljstvo, ostali učenici koji koriste Internet (75,8%) nisu imali neugodnih i/ili ružnih iskustava.

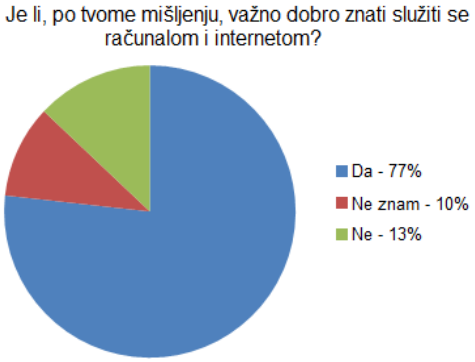
Vođeni pretpostavkom da učenici nižih razreda osnovne škole provode veliki dio svog slobodnog vremena koristeći računalo, kratkim pitanjima željeli smo provjeriti njihova znanja o osnovnim dijelovima računala i računalnim programima. Rezultati su pokazali da 87,25% učenika zna da je monitor dio računala na kojem vidimo sve što radimo, dok njih 12,5% to ne zna. Pitanje *Za što nam služi program Word* donijelo nam je sljedeće rezultate: 50,7% učenika točno je odgovorilo na postavljeno pitanje, njih 9% misli da Word služi za gledanje filmova, 7,1% učenika misli da se Wordom pretražuje Internet, dok 33,2% odabire odgovor *Ne znam* na postavljeno pitanje. Nešto bolji rezultati dobiveni su teorijskom pitanju o programu Bojanje. Gotovo svi učenici (93,5%) znaju da je Bojanje program koji služi za crtanje. Teško je sa sigurnošću reći, je li tako visoki postotak pokazatelj njihova znanja ili rezultat sugestije samog naslova programa.

Veliki broj učenika ne zna što je monitor, Word i Bojanje, a ipak je većina njih je pokazala nekritičan stav prema svome znanju.

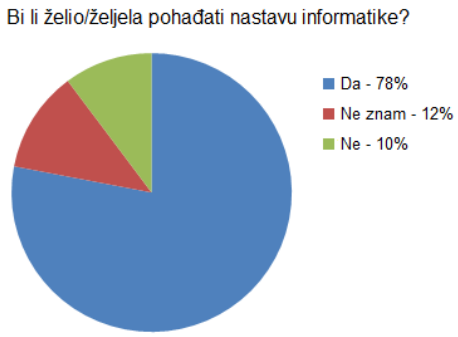


Graf 3. Prikaz samovrednovanja učenika (N = 400)

Na kraju našeg istraživanja zanimalo nas je jesu li najmlađi naraštaji svjesni činjenice da je računalna tehnologija neizostavni aspekt razvoja današnjeg društva i temelj budućeg razvoja. Graf 4. pokazuje da učenici uviđaju važnost tehnologije, a iz Grafa 5. vidljiv je njihov izrazito visok interes za digitalnom tehnologijom.



Graf 4. Prikaz odgovora o važnosti korištenja računala i Interneta te učenikovih preferencije prema informatici kao nastavnom predmetu (N = 400)



Graf 5. Prikaz odgovora učeničkih preferencije prema informatici kao nastavnom predmetu (N = 400)

VI. ZAKLJUČAK

Dvije neosporive činjenice su da živimo u društvu koje nužno treba informatički sposobne i pismene pojedince te da su djeca od svoje najranije životne dobi u dotaciju s digitalnom tehnologijom. Ovim radom željeli smo prikazati važnost, ali i mogućnost vođenja djece niže životne dobi do informatički sposobnih i pismenih pojedinaca koji su potrebni našem društvu.

Cilja našeg istraživanja bio je prikupiti informacije o korištenju i poznavanju računala i Interneta učenika razredne nastave te saznati njihovo mišljenje o nastavi informatike. Zanimalo nas je i na koje se sve načine učenici služe računalom i kako se pritom osjećaju.

Rezultati istraživanja sugeriraju zaključak da gotovo svi ispitanici koriste računala što ih ispunjava i čini sretnima (62,5%). Računala najčešće koriste u svrhu zabave dok je edukativna uloga računala slabije zastupljena. Iako su pokazali nezavidno znanje o osnovama računala, ipak smatraju da je ono izrazito visoko, samim time što provode veliki dio svog slobodnog vremena koristeći računalo.

Iz cjelokupnog rada vidljivo je da našu poruku i ideju o važnosti uvođenja informatike kao izbornog predmeta ili izvannastavne aktivnosti u razrednu nastavu podupiru, ne samo anketirani učitelji, već gotovo i 80% naših ispitanika. Spoznaja da mlađi članovi naše društvene zajednice uviđaju važnost i potrebu informacijske pismenosti, ali da ujedno imaju i visoku razinu motiviranosti za istu, trebala bi biti veliki poticaj vođama našeg odgojno – obrazovnog sustava, ali i društvu u cjelini, da ih krenemo voditi od najranije dobi prema željenom cilju.

LITERATURA

[1] G. Baert, M. Galton, P. Honeth, J-M. Sivirine, M. Thurler, "iNovaCije u osnovnom obrazovanju," Zagreb: Školske novine, 2002.

[2] Lj. Bakić – Tomić, M. Dumančić," Odabrana poglavlja iz metodike nastave informatike," Sveučilišna skripta, Zagreb: Učiteljski fakultet, 2012.

[3] E. Bischof, B. Sabitzer,"Computer Science in Primary Schools – Not Possible, But Necessary?!, „Computer Science, Vol 7013, str. 94-105, 2011.

[4] Education in France [online]. Wikipedia.org. Dostupno na: http://en.wikipedia.org/wiki/Education_in_France [12. listopada 2013.]

[5] Fakultet organizacije i informatike Varaždin (2011)Rang liste [online] Dostupno na: http://www.foi.unizg.hr/sz/obavijesti/rang_liste [20 travnja 2013.]

[6] French Primary School – Ecole Primaire [online]. French Property.com. Dostupno na: http://en.wikipedia.org/wiki/Education_in_France [12. listopada 2013.]

[7] T. Karsenti, S. Collin,"Benefits and challenge of using laptops in primary and secondary school:An investigation at the Eastern Townships School Bord," Université de Montréal, 2011. Dostupno

na:http://www.etsb.qc.ca/media/etsb_pdf.file/enCA/synth%20seAnglais.pdf [30. studenog 2013.]

[8] I. Kniewald,“Računalo umjesto igračke,“ Umjetnost i dijete, Vol. XXVII, str.239-243, 1995.

[9] J. Lesourne,“Obrazovanje & društvo,“ Zagreb: Educa, 1993.

[10] Ministarstvo znanosti, obrazovanja i sporta. Nastavni plan i program za osnovnu školu. Zagreb: Ministarstvo znanosti, obrazovanja i sporta, 2006.

[11] L. Patsalides,“ Why and How to Use Computers in the Classroom“ [online] Bright Hub Education, 2013. Dostupno na: http://www.brighthubeducation.com/elementary-school activities/5638-benefits-of-computer-time-in-the-classroom/ [27. studenog 2013.]

[12] A. Semeov, A.J. Knierzinger, K. Martcheva, S. Roesvik, E. Schmidt,“Informatics for primary education,“ Moscow: IITE, 2000. [online] Dostupno na: http://unesdoc.unesco.org/images/0013/001303/130330eo.pdf [15. travnja 2013.]

[13] O. Silva, S. Machin, S. McNally,“New Technology in Schools: Is There a Payoff?,“ The Institute for the Study of Labor (IZA) [online], DP No. 2234, 2006. Dostupno na: http://ftp.iza.org/dp2234.pdf [26. studenog 2013.]

[14] A. Vučetić Škrbić,“E-škola: Informatika postaje nezaobilazna u školi,“ [online] Slobodna Dalmacija, 2009. Dostupno na: http://www.slobodnadalmacija.hr/Hrvatska/tabid/66/articleType/ArticleView/articleId/72362/Default.aspx [20. travnja 2013.]

Primjer projekta u nastavi informatike – što je lijepo ljudskom oku

Vesna Skočir
X.gimnazija "Ivan Supek", Zagreb, Republika Hrvatska
vesna.skocir@skole.hr

Sažetak - Projektna nastava ima velike potencijalne mogućnosti. Jedna, osobito važna, je da se kroz takav oblik nastave učenici mogu upoznati s istraživačkim elementima. U radu je predstavljen projekt grupe učenika trećega razreda prirodoslovno - matematičke gimnazije koji su ponovili eksperiment Gustava Theodora Fechnera, prvog eksperimentalnog psihologa iz 19. stoljeća. Cilj je bio ustvrditi hoće li ljudsko oko među sličnim likovima kao najljepši izdvojiti onaj u kojemu je implementiran omjer zlatnog reza. Iako je u ponovljenom eksperimentu bilo nekih nedostataka, rezultati su istovjetni Fechnerovim: najveći broj ispitanika "prepoznao" je ljepotu zlatnog reza. Pored opisa i rezultata eksperimenta u radu se diskutira kako učinkovito koristiti računalo kod pripreme ovakve vrste projekta, rada na projektu te predstavljanju rezultata.

I. UVOD

Istraživanje u središte stavlja problem kojega želimo riješiti ili naprosto bolje upoznati. Prirodna radoznalost učenika (djeteta) snažan je motivacijski faktor pa je različite oblike istraživanja dobro poticati već u najranijoj dobi. Koristi su neizmjerne, iako uobičajenim testovima nemjerljive; razvija se kreativnost, vještina komunikacije, odgovornost, kritičnost prema svom i tuđem radu, samosvijest. Učenici se s istraživačkim elementima mogu upoznati u okviru projektne nastave.

Projektna nastava takav je oblik nastave u kojem su učenici u grupama ili samostalno usmjereni na rješavanje odabrane problemske situacije ili obradu neke otvorene teme pri čemu put do cilja nije odmah dokučiv[1]. Zbog nedostatka vremena u tradicionalnom sustavu obrazovanja kakav je naš, ovakav oblik nastave u svoj svojoj punini teško je provediv. Budući da je tema otvorena, uloga nastavnika i učenika različita od one na koju su navikli, trajanje projekta teško je predvidjeti. Tijekom istraživanja može se pojaviti potreba za (čak i znatnijom) korekcijom plana izvedbe, pa i prekidom projekta ukoliko se pokaže da tema (ipak) nije u domeni iskustva učenika, da ne vladaju strategijama koje dovode do rješenja ili da naprosto nema dovoljno interesa za obradu teme. Stoga je osobito važan odabir teme koja će učenicima, a i nastavniku, biti dovoljno intrigantna da ustraju u istraživanju.

[1] Idealno bi bilo da to vrijedi za sve sudionike projekta, dakle i za nastavnika. Nastavnik koji "sve zna" nije autentični sudionik istraživanja. Rezultat do kojeg dođe, učenik neće dalje propitivati u sebi; zadovoljit će ga pozitivan odgovor nastavnika. U najboljem slučaju takav učenik može osjetiti zadovoljstvo dobro obavljenog posla, ali pravo zadovoljstvo pronalazača bit će mu uskraćeno.

Stara dilema o kojoj su rasprave vodili filozofi i matematičari, *je li matematika nešto stvarno ili nešto otkriveno*, odnosno *jesu li istine već tu samo ih još ne znamo*, može se uvijek nanovo aktualizirati kroz pojam zlatnog reza. Savršenu proporciju koju matematički opisuje broj Φ, nalazimo u ravnoteži, harmoniji i simetriji prirode. U toj se, kako je još nazivaju, božanskoj proporciji, dijelovi cjeline jedan prema drugom odnose isto onako kako se cjelina odnosi prema većem dijelu. Manifestacije zlatnoga reza vežu se i uz pojam *ljepote*. Ljepota nije nešto apsolutno i nepromjenjivo; različiti koncepti ljepote sukobljavali su se u različitim razdobljima, a ponekad i unutar iste kulture [1]. Zlatni rez i pojam lijepog teme su osobito prikladne za istraživanje u srednjoj školi. Razloga ima više: tom su uzrastu ovakve teme privlačne, većina je spremna i sposobna njima se intelektualno i emotivno posvetiti, a budući da je riječ o širokoj temi mogućnosti istraživanja su velike.

Polazište za naše istraživanje bio je rad Gustava Theodora Fechnera (1801.-1887.), kojega smatraju prvim eksperimentalnim psihologom [2]. On je mjerenjem ustanovio da brojni pravokutni predmeti koji ga okružuju (knjige, kutije, zgrade…) imaju omjer duljina stranica blizak broju Φ. Pretpostavio je da je to stoga što se većini ljudi taj omjer sviđa. Poslije je to dokazao eksperimentom: ispitanicima je pokazao niz pravokutnika različitih omjera, a oni su kao najljepši prepoznali baš taj omjer.

U načelu istim eksperimentom ispitali smo 79 ispitanika i potvrdili Fechnerove rezultate. Pored testa s pravokutnicima, istim ispitanicima pokazani su crteži ljudskih lica, tri ženska i tri muška lica. Par detalja na jednom ženskom i jednom muškom licu bilo je nacrtano u omjeru bliskom broju Φ. Rezultat je bio istovjetan prvom – lica s "ugrađenim" zlatnim rezom prepoznata su kao najljepša.

Projektna nastava je u načelu interdisciplinarna. Tako se i ovaj projekt mogao ostvariti u okviru matematike, likovne umjetnosti ili psihologije. No njemu je zasigurno mjesto i na satovima informatike budući da se ovakvim projektima učenici usavršavaju u korištenju informacijske tehnologije te uče kako tu tehnologiju upotrijebiti da bi priredili i prezentirali stručni rad. Ovim radom namjera je pokazati da i teme koje nisu vezane za područje informatike, ali su dovoljno interesantne, mogu biti korisne za razvijanje informatičke pismenosti učenika. Pogotovo je to slučaj s onom dubljom razinom pismenosti kod koje učenik ne samo da zna kako nešto u danoj aplikaciji napraviti, već i koju od mogućih opcija za određenu situaciju upotrijebiti. Naposljetku, tu je i

prikupljanje informacija preko Interneta. I mada se čini da je to područje kojim učenici jako dobro vladaju, praksa pokazuje da kod prikupljanja informacija za izradu stručnoga rada postoje problemi. Međutim, dobro definiranim zadacima nastavnik može pomoći učeniku da proradi informacije i s razumijevanjem ih prilagodi jer kao što upozorava K. P. Liessmann [4], u protivnom "većina informacija naprosto ostaje izvanjska". On pretpostavlja da je to stoga što lakoća pristupa (informacijama) sabotira tvorbu znanja. Tako dolazi do toga da studenti, a i drugi "sve više brkaju mehaničko kopiranje seminarske radnje s interneta sa samostalnim pisanjem takve radnje " [4].

II. OPĆI PRIKAZ TEME

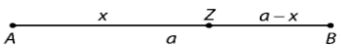
A. Eksperiment Gustava Fechnera

Uobičajena je sljedeća definicija zlatnog reza[10]: Točka Z dijeli dužinu $\overline{AB}$ u omjeru zlatnog reza ukoliko je omjer duljina cijele dužine $\overline{AB}$ i njezina većeg dijela $\overline{AZ}$ jednak omjeru duljina većeg dijela $\overline{AZ}$ i manjeg dijela $\overline{ZB}$ (Sl.1). Označimo li sa *a* duljinu dužine $\overline{AB}$, a sa *x* duljinu njezina većeg dijela $\overline{AZ}$, treba biti $a : x = x : (a-x)$. Kad ovo matematički izračunamo, dobijemo:

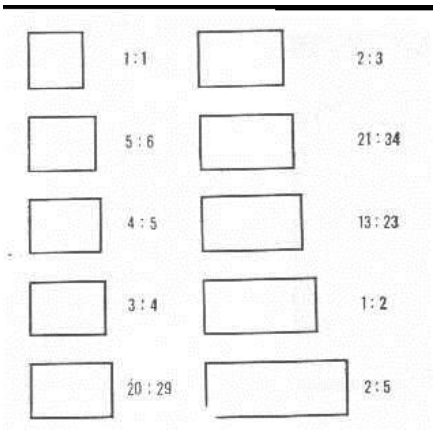
$$x = \frac{\sqrt{5}-1}{2}\, a \approx 0.6183a$$

Broj $\frac{\sqrt{5}-1}{2}$ označava se s Φ i zove se zlatni omjer.

1860. godine Gustav Theodor Fechner istražuje osjećaj ljepote i harmonije kod odraslih ljudi. Pripremio je niz od deset bijelih pravokutnika, među kojima je jedan imao omjer duljina stranica blizak omjeru zlatnog reza, 5:8, odnosno 0.625. Omjeri duljina stranica ostalih pravokutnika varirali su između 1:1 i 2:5. Sudionici eksperimenta (228 muškaraca i 119 žena) bili su zamoljeni da pogledaju pravokutnike i odaberu onaj koji im se najviše sviđa. Prema grafičkim rezultatima preuzetim iz [2] vidimo da se većini ispitanika (35%) "zlatni" pravokutnik najviše sviđa. Visok postotak ispitanika (blizu 40%) odabrao je pravokutnike omjera stranica bliskih "zlatnom" (2:3 i 13:23), dok se za svaki od preostalih pravokutnika odlučilo manje od 10% isitanika. Približno iste rezultate u ponovljenom su pokusu dobili i engleski znanstvenici 1958. god. [8]. Iako se jednostavno može konstruirati pravokutnik kojemu su stranice u omjeru broja Φ, u praksi se, pa tako i u Fechnerovom eksperimentu, pribjegava aproksimaciji (Sl.2.). Tako je primjerice "zlatni" pravokutnik negdje predstavljen omjerom 5:8, a negdje omjerom 21:34 (usp. [2], [6] i [8]). Iako se prema [2] i [8] dade naslutiti da su u Fechnerovom eksperimentu pravokutnici bili različitih površina, čini se da nije bilo tako. Prema [10] svi su pravokutnici imali



Slika 1. Zlatni rez dužine



Slika 2. Fechnerovi pravokutnici, omjer duljina stranica "zlatnog" pravokutnika je 21:34 (preuzeto iz [8])

jednake površine. Prema istom je izvoru nepoznato u kojem su položaju pravokutnici predočeni ispitanicima; koja je od stranica (veća ili manja) bila smještena vodoravno. Prema svim navedenim izvorima pravokutnici su bili "poslagani u red" u kojem se "zlatni" pravokutnik nalazio na sedmom mjestu.

Slične su eksperimente ponavljali i drugi psiholozi te dobivali katkad proturječne rezultate [7]. Michael Godkewitsch sa Sveučilišta u Torontu to objašnjava činjenicom da u zapadnom svijetu nije moguće verbalno iskazati što znači estetska sklonost prema određenom omjeru stranica pravokutnika, a da taj iskaz bude pouzdan [7]. Pitanje je prema kojem će kriteriju jedan omjer biti ljepši, privlačniji od drugoga? Kroz povijest estetika proporcije poprimala je različite i sve složenije oblike. Stari su Grci uz pojam lijepog vezali osobine prikladnosti i mjere, dok primjerice mediji potrošačkog društva 20. i 21. stoljeća kao uzor ljepote nude šminku i potpisanu odjeću. Stoga, premda dopušta da "onkraj različitih poimanja Ljepote postoje jedinstvena pravila za sve narode i stoljeća", Umberto Eco u knjizi Povijest ljepote [1] ne trudi se tražiti ih, već upravo suprotno, ističe razliku.

Pojam lijepog bila je i ostala velika i otvorena tema. I mada ne možemo točno odrediti što ljepota jest, istraživanja doživljaja lijepog, kao što je primjerice Fechnerov eksperiment, zasigurno imaju smisla.

B. Priprema projekta

Upoznavanje s temom dobro je početi popisom literature koju bi trebalo proučiti prije početka istraživanja. Budući da je riječ o srednjoj školi to mogu biti svega jedno ili dva djela (članak, odlomak neke knjige). Poželjno je da to bude neko zaista vrijedno djelo kao što su primjerice *Tajni kod* P.Hemenwaya [2] ili *Povijest ljepote*, knjiga koju je priredio Umberto Eco [1]. Iako je riječ o satovima informatike, ne bismo se smjeli ograničiti samo na multimedijalno čitanje; kvaliteta djela trebala bi biti jedini kriterij izbora.

Kad je na ovaj način dobio osnovne spoznaje o temi koja se obrađuje, učenika treba uputiti da prema ključnim

riječima istražuje Internet i pokuša doći do novih informacija i tako još više približiti danu temu.[2]

Za čitanje treba predvidjeti dovoljno vremena budući da se na taj način osim informacija prikupljaju ideje i zapažanja. Kad se dobije širi uvid, rasplet više nije predvidiv - i sam cilj istraživanja može se izmijeniti.

Neophodno je da učenici osnovni tekst shvate u cjelini i pojedinostima. Ukoliko ima problema dobro je pribjeći zajedničkom čitanju. Čitati treba polagano, pažljivo, strpljivo, pogotovo ako je riječ o težim dijelovima. Na trenutke treba prekidati čitanje kako bi se ono što se čita dovelo u vezu s iskustvom i ranije stečenim znanjem. Prema tekstu treba imati kritički odnos; treba vidjeti je li neka činjenica/ideja vezana za temu koja se obrađuje te je li neka ideja prihvatljiva ili nije. Dobar je trenutak da se učeniku objasni razlika između činjenice i ideje.[3]

Usporedo s čitanjem teksta treba pisati zabilješke. Prikladna je forma natuknica, a kako bi se izbjegao "copy-paste" efekt, poželjno je izabrati klasični medij – olovku i papir. Naposljetku se dobivene informacije zajednički diskutiraju. Treba biti siguran da se razumije njihovo značenje, a potom ih smisleno povezati i staviti u suodnose.

III. OPIS EKSPERIMENTA

A. Ponovljeni eksperiment Gustava Fechnera

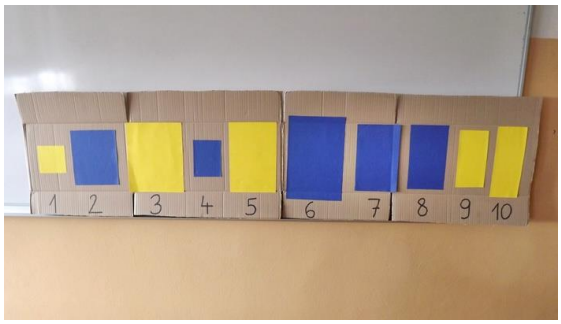
Pravokutnici zadanih omjera prikažu se na nekom mediju i predoče ispitanicima. Omjeri duljina stranica i redoslijed pravokutnika treba biti isti kao u Fechnerovom eksperimentu (prema [2]). Ispitanici će biti zamoljeni da bez razmišljanja odgovore koji im se pravokutnik najviše sviđa. Predviđeno je i dodatno pitanje – izdvojiti najružniji pravokutnik, dakle onaj koji bi prema estetskim kriterijima smjestili na dno ljestvice.

Hipoteza je da će se ispitanicima više sviđati pravokutnik koji ima odnos stranica blizak broju Φ. Hipotezu ćemo smatrati potvrđenom ukoliko većina ispitanika "zlatni" pravokutnik izabere kao najljepši. Očekuje se da će najmanji broj ispitanika ovaj pravokutnik prema estetskim kriterijima staviti na dno ljestvice.

B. Provedba i rezultati eksperimenta

Sugestija je bila da se pravokutnici prikažu na klasičnom mediju.[4] Učenici su ih nacrtali na debljim papirima u boji (plava i žuta boja), izrezali i zalijepili na veći karton (Sl.3).

Sudionicima eksperimenta rečeno je da se ispituju karakteristike broja Φ. Budući da su, s izuzetkom jedne starije osobe, sudionici bili učenici prvog razreda



Slika 3. Ponovljeni Fechnerov eksperiment; sedmi po redu je "zlatni" pravokutnik

TABLICA I. REZULTATI PONOVLJENOG FECHNEROVOG EKSPERIMENTA

Broj pravokutnika	Omjer stranica	Odabir najljepšeg u postocima
1	1:1	11.84
2	5:6	10.53
3	4:5	7.89
4	3:4	9.21
5	20:29	11.84
6	2:3	5.26
7	**5:8**	**14.47**
8	13:23	10.53
9	1:2	9.21
10	2:5	9.28

gimnazije te djeca starije vrtićke dobi [5] možemo pretpostaviti da gledajući pravokutnike nisu procjenjivali duljine stranica i njihove omjere. Zamoljeni su da bez puno razmišljanja izdvoje pravokutnik koji im se najviše sviđa. Poslije su na isti način odredili pravokutnik koji im se najmanje sviđa.

Budući da je većina ispitanika (Tab.1) kao najljepši izabrala pravokutnik kojemu su duljine stranica u omjeru zlatnog reza, hipotezu smatramo potvrđenom.

Usporedba s rezultatima Fechnerova pokusa prikazana je grafički na Sl. 4. Pravokutnik kojemu je omjer duljina stranica najbliži broju Φ, a to je pravokutnik na sedmom mjestu, u oba se eksperimenta najviše svidio ispitanicima. Međutim u našem se eksperimentu daleko manji broj opredijelio za taj pravokutnik; svega 14.47% ispitanika za razliku od 35% ispitanika koji su u Fechnerovom eksperimentu taj pravokutnik doživjeli kao najljepši.

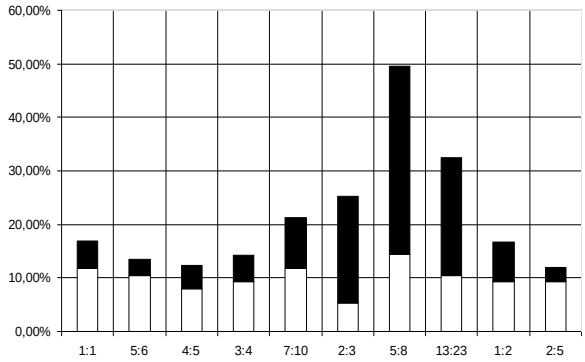
Za šesti i osmi pravokutnik koji su omjera bliskih "zlatnom" sedmom po redu pravokutniku, u našem se eksperimentu, u usporedbi s Fechnerovim, odlučio relativno mali broj ispitanika. Disproporciju nalazimo u

[2] Svakako treba izbjeći da učenik osnovne informacije o temi pokuša saznati samostalno preko Interneta. Praksa pokazuje da učenik posao obično odradi formalno, a nerijetko se "zabavi" i drugim sadržajima.

[3] Ideja je činjenica koja je na neki način interpretirana. Stoga je i njena istinitost upitna. Uočavanje ovih razlika na konkretnim primjerima usmjerava učenika na kritičko čitanje teksta.

[4] Učenici bi se trebali "igrati" različitim medijima, kritički odmjeravati i argumentirati prednosti i mane pojedinog medija.

[5] U istraživanju je sudjelovalo 11 osoba 1995. godište, 33 osobe 1997. godište, 1 osoba 1998. godište, 17 osoba 2006. godište, 13 osoba 2007. godište i jedna starija osoba čije godište nije poznato.



Slika 4. Usporedba rezultata našeg (svijetli dio) i Fechnerovog eksperimenta (tamni dio); sedmi po redu je "zlatni" pravokutnik

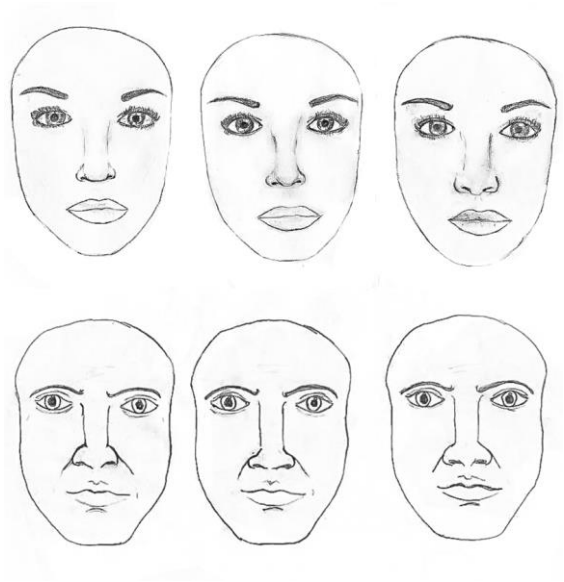
još nekim slučajevima. Primjerice kod pravokutnika koji su prvi i deseti po redu (omjeri 1:1 i 2:5), situacija je obrnuta; za njih se u našem eksperimentu odlučilo osjetno manje ispitanika nego u Fechnerovom.

Ovi rezultati ne začuđuju ukoliko uočimo boje pravokutnika: šesti, sedmi i osmi pravokutnik su plave boje, a prvi i deseti žute. Ispitanici su očito dali prednost žutoj (toploj) boji u odnosu na plavu (hladnu) boju. Isto se dogodilo i s ostalim pravokutnicima u nizu s izuzetkom drugog i četvrtog.

Osim plave boje "zlatni" je pravokutnik površinom bio među manjima i malo "savijen" jer se nalazio na pregibu kartonske podloge. Za pretpostaviti je da bi rezultat bio još uvjerljiviji da je eksperiment bio bolje pripremljen: da su pravokutnici bili iste boje (najbolje bijele – neutralne) te da se njihove površine nisu previše razlikovale.

C. Proširenje Fechnerova eksperimenta

Pored testa s pravokutnicima, istim ispitanicima pokazani su crteži ljudskih lica, tri ženska i tri muška lica. Lica su nacrtali učenici trudeći se da u jedno žensko i jedno muško lice "ugrade" omjer zlatnog reza. Riječ je o par detalja koji se nalaze na sredini lica. Žensko lice s elementima zlatne proporcije je drugo po redu, a muško treće. (Sl.5). Rezultat je bio istovjetan prvom – lica s "ugrađenim" zlatnim rezom prepoznata su kao najljepša. Žensko lice s takvim detaljima kao najljepše je doživjelo 57.89% ispitanika, a muško s istim karakteristikama 44.74% ispitanika.



Slika 5. Drugo po redu žensko lice i treće po redu muško lice ima par detalja u omjeru bliskom omjeru zlatnog reza

Rezultati su zanimljivi, ali da bi bili ozbiljno razmatrani, crteže je trebalo bolje pripremiti. Svakako bi dalja istraživanja valjalo nastaviti u tom pravcu.

D. Kako smo koristili računalo?

Računalo je univerzalno pomagalo koje koristimo u svakodnevnom životu nerijetko takvim automatizmom da to ni ne zamjećujemo. Na taj smo način i mi koristili računalo, primjerice za dobivanje informacija pretraživanjem web stranica, komunikaciju elektroničkom poštom ili onda kad smo programom za obradu teksta (MS Word) zapisivali planirane aktivnosti, pripremali anketna pitanja i slično. Dijelom smo na isti način koristili i program za rad s proračunskim tablicama (MS Excel) te program za izradu prezentacija (MS PowerPoint) budući da su korisnička sučelja osnovnih programa MS Office, učenici upoznali tijekom ranijeg školovanja. No u srednjoj nas školi ne bi trebala zadovoljiti činjenica da učenik zna tablično u Excelu prikazati podatke, obraditi ih i predočiti grafom ili kreirati slajdove PowerPoint prezentacije, već i da taj rad bude u skladu sa zahtjevima koje bi stručni rad trebao zadovoljiti. To je razlog da je dio rada na projektu ostvaren kao radionica u kojoj smo između ostaloga pokazali i diskutirali:

- koja je vrsta grafa prikladna za ilustraciju dobivenih rezultata;
- zašto je potrebno dokumentirati rad te kako to učiniti;
- kako činjenice i ideje predstaviti tako da oni kojima se prezentira razumiju što im se želi reći, da ih zainteresiramo za ono što govorimo i da sve to od njih ne iziskuje posebni napor;
- kako tehnički osmisliti prezentaciju, koji su fontovi/boje dobri, a koji nisu, koji podaci trebaju pisati na početnom slajdu…
- kako se pripremiti za predstavljanje rada, kako suzbiti tremu i slično.

U praksi često, u okviru različitih predmeta, učenici dio gradiva obrade sami, a potom izrade odgovarajuću prezentaciju. Dobar dio nastavnika smatra da je na taj način nastavu učinio modernijom, a argumenti su i da se učenici s kompjuterima dobro snalaze i da to vole raditi. Praksa pokazuje da učenici vole takve zadatke jer su s njima obično brzo gotovi, a zahvaljujući Internetu radovi se mogu obogatiti efektnim ilustracijama. Opasnost je da se bez dodatnog znanja, promišljanja i iskustva pri izradi takvih radova, dobivaju uradci upitne kvalitete koji su često ocijenjeni visokom ocjenom. Na taj se način njeguju loše navike korištenja računala.

Ljudi čitaju mnogo više nego što pišu. Brojne studije pokazuju ako se poveća kvantiteta pisanja to ne znači da će se nužno popraviti kvaliteta pisanja. Hipoteza da stil pisanja proizlazi iz čitanja, ne iz pisanja je u skladu s onim što je poznato kao jezična akvizicija: Bogatstvo jezika dolazi iz ulaza, ne izlaza, iz razumijevanja, a ne produkcije [3]. Slično, kvaliteta prezentacije neće se popraviti izradom većeg broja prezentacija, ukoliko izostane dubinsko razumijevanje onoga što se predočuje, promišljanje o svakom elementu koji će popratiti prezentaciju, bilo da će se naći na slajdu ili će biti izgovoren. Pretpostavka su i osnovna znanja o tome što stručni rad jest i kako bi trebala izgledati njegova prezentacija. Supervizija, kritički osvrt je neophodan.

Istraživački su projekti, poput projekta opisanoga u radu, motivirajući za učenike, stoga su dobra prilika da se učenike poduči što je stručni rad, kako bi ga trebalo napisati[6], a kako prezentirati.

IV. ZAKLJUČAK

U radu je predstavljen projekt ostvaren na satovima izborne nastave u trećem razredu prirodoslovno - matematičke gimnazije u kojem su učenici ponovili eksperiment psihologa Gustava Theodora Fechnera, a taj je ustanoviti hoće li ispitanici među pravokutnicima čije stranice imaju duljine različitih omjera, kao najljepši izdvojiti onaj kojemu su duljine stranica u omjeru zlatnog reza. Unatoč nedostacima koji su u ponovljenom eksperimentu bili na štetu "zlatnog" pravokutnika, upravo je taj pravokutnik, kao i u Fechnerovom eksperimentu, prepoznat kao najljepši.

Daljim istraživanjima valjalo bi ustanoviti da li, i u kojoj mjeri, dob i spol ispitanika utječu na rezultate eksperimenta. Na sličan bi način valjalo istražiti u kojoj se mjeri lijepima doživljavaju elementi zlatnog reza na složenijim objektima, primjerice građevinama, biljkama, ljudskom licu i slično.

Paralelno s prikazom eksperimenta u radu se diskutira na koji način voditi učenike kroz ovakav oblik istraživanja, posebno na koji način učinkovito koristiti računalo u fazi prikupljanja informacija, obrade podataka i predstavljanja rezultata.

Pomoću novih tehnologija možemo brzo doći do rezultata. I tu leži opasnost. U brzini mnoge stvari promaknu, pa se ponekad zadovoljimo izvanjskim, površnim znanjem koje se uglavnom temelji na činjenicama. A to je ono što svakako treba izbjeći. Želimo da učenik razumije činjenice te da iz tih činjenica dobije ideje. Svijet računala, poglavito Internet, globalna je i sveprisutna pojava, pa je tako i prisutnost novih tehnologija u obrazovanju neupitna. Zato je pogotovo važan način kako ćemo ovu snažnu tehnologiju koristiti da se učenik kao osoba razvije na najbolji mogući način.

[6] Učenike bi trebalo upoznati sa strukturom pisanog stručnoga rada, uputiti ih na razlike pisanog i govornog jezika, ukazati na važnost pravopisne i gramatičke korektnosti i slično.

LITERATURA

[1] U. Eco, "Povijest ljepote", Hena com, Zagreb, 2004.

[2] P. Hemenway, "Tajni kod: zlatni rez – tajanstvena formula koja vlada umjetnošću, prirodom i znanošću", V.B.Z., Zagreb, 2009.

[3] S. Krashen, "The power of reading: insights from research", Libraries Unlimited, Inc, Englewood, Colorado,1993.

[4] K. P. Liessmann, "Teorija neobrazovanosti. Zablude društva znanja", Jesenski i Turk, Zagreb, 2008.

[5] M. Benšić, "Projekt u nastavi matematike srednjih škola", Matematički odjel Sveučilišta u Osijeku, http://www.mathos.hr/~mirta/tekst_S.pdf (Preuzeto: 12.09.2013)

[6] D. Kisets, "Informational validity of Fechtner's experiments foutcomes", National Physical Laboratory, Israel, www.optimization-online.org/DB_HTML/2011/11/3235.html (Preuzeto: 15.03.2013.)

[7] M. Livio, "The golden ratio and aestetics. Is beauty in the eye of the beholder?", Plus internet magazine, http://www.plus.maths.org/content/golden-ratio-and-aesthetics (Preuzeto: 04.01.2014.)

[8] Virtual Museum of Harmony and Golden Section. Return to the Golden Section, autori projekta A. Stakhov, A. Sluchenkova, http://www.goldenmuseum.com/0502Return_engl.html (Preuzeto:0 9.02.2013.)

[9] Matematika plus, http://element.hr/plus/misc (Preuzeto: 20.01.2014)

[10] The Myth of the Golden. Phy in Phychology, http://goldenratiomyth.weebly.com/phi-in-psychology.html (Preuzeto: 04.01.2014.)

Primjena interaktivnog videa u obrazovnim nastavnim sredstvima

Danijel Vrbanc, dipl. ing.
Željeznička tehnička škola – Moravice, Vrbovsko, Hrvatska
danijel.vrbanc@skole.hr

Sažetak - U radu je prikazan način interpretacije električnih shema iz dvaju predmeta: Elektronički sklopovi i Učinska (energetska) elektronika. Prikaz je namijenjen učenicima i studentima koji su na nastavi, tj. predavanjima bili upoznati s radom određenih elektroničkih sklopova, a za bolje razumijevanje i zornost rada istih - trebao bi im poslužiti interaktivni video. Dok se pisani i slikovni sadržaj u udžbenicima ili prezentacijama uglavnom koristi na uobičajen način, interaktivni video u danim primjerima ne rabi pisanu riječ već samo slike i pokretne slike tijeka struje i iscrtavanje valnih oblika (uglavnom napona i struja). Time se paralelnim i slijednim promatranjem objašnjava princip rada određenog sklopa, a izostanak pisane riječi omogućuje jednostavnu internacionalizaciju. Princip rada ovog prikaza prikazan je na dva primjera. Za prvi primjer je odabran rad mosnog punovalnog ispravljača (Graetzov spoj), na kojem su pokazani mehanizmi rada videa. Drugi primjer je iz učinske elektronike, (opis rada ciklopretvarača), kojim se jedan složeniji elektronički sklop interpretira na način kojim ga je jednostavnije pojmiti. Dakle, naglasak rada je stavljen na pokazivanju nevidljivog vidljivim. Zna se da je električna struja usmjereno gibanje nosioca naboja, pa se ta definicija, zajedno s valnim oblicima, koristi u prikazivanju električne struje u danim primjerima.

I. UVOD

U današnje vrijeme se mogućnost ostvarivanja učenja na daljinu već praktički udomaćila. Način dislociranog učenja u svrhu usvajanja novih sadržaja ili utvrđivanja gradiva postaje sve više jedan od modela obrazovanja. Učenje na daljinu nije vezano samo uz pojavu računala. Već se nekoliko desetljeća koristi sustav učenja putem radija i televizije kroz obrazovne programe. Oni su se redovito svaki dan emitirali a i danas se emitiraju ali u smanjenom obimu. Pojavom i širenjem računalne tehnike te izgradnjom širokopojasnog interneta, informacija je postala dostupnija pa se tako i ta mogućnost iskoristila za proces obrazovanja.

Sadržaji koji su se emitirali ili se emitiraju putem radija i televizije su uglavnom slijedni i vezani uz jedan određeni termin emitiranja. Postoji mogućnost snimanja programa putem video snimača i naknadnog prikazivanja, ali takvi sadržaji opet imaju mogućnost samo slijednog (sekvencijalnog) gledanja. Danas su također medijski obrazovni sadržaji locirani na računalu ili mreži te se mogu koristiti na isti način. No računala imaju i drugu mogućnost koja se prije nije mogla ostvariti, a to je sustav usvajanja znanja na računalu kroz programske pakete koji služe za razne proračune, simulaciju laboratorijskih vježbi i slično. Međutim uporaba takvih računalnih paketa ima svoje nedostatke. Uglavnom njihova uporaba je vezana uz potrebu licenciranja, a osim toga potrebna je određena vještina kako bi se mogli koristiti u pravom obimu za koje su namijenjeni. Dakle to iziskuje znatna financijska sredstva kao i vještine za njihovu uporabu. U elektrotehnici je primjer takvih programa „Elektronic Workbench“, „SIMPLORER“ i slično.

II. MULTIMEDIJA U NASTAVI ELEKTROTEHNIKE

Dakle, za usvajanje osnova iz određenog predmeta potrebno je na početnoj razini objasniti suštinu. Da bismo u elektrotehnici i posebno elektronici, matematički gledano, podskupu elektrotehnike, mogli lakše proučiti pojave koje se događaju, postavljaju se razni modeli prikaza određenih fizičkih veličina poput struje, napona magnetskih i električnih polja (silnica). Predočavanjem kroz različite simbole i matematičkim prikazom valnih oblika, pojednostavnjuje se prikaz prirode raznih pojava u poluvodičima i strujnim krugovima, tj. povećava se zornost električkih pojava.

U elektronici postoji čitav niz aktivnih elemenata različitih osobina čije djelovanje u strujnim krugovima nije jednostavno objasniti, a time nije niti jednostavno usvojiti kod slušatelja. Ako pri tome uzmemo da postoje predavači koji nemaju podjednake kvalitete predavanja, kao i to da uvijek postoji mogućnost izostanka učenika i studenata sa predavanja, tada još više dolazi do izražaja problematika vezana uz prijenos i utvrđivanje znanja.

Udžbenici su, pored predavanja, glavna poluga u usvajanju znanja, ali je njihova nedostatnost (naročito zadnjih godina pa i desetljeća u Hrvatskoj), jedan od glavnih nedostataka. U doba kada nije postojala računalna i mrežna infrastruktura postojalo je kontinuirano obrazovno nakladništvo. Nažalost u ovo računalno „moderno“ doba, broj naslova koji se pojavljuje na hrvatskom obrazovnom području se smanjio.

Na sveprisutnom internetu se može pronaći niz obrazovnih sadržaja. Nažalost, pregledavajući razne postavljene datoteke na mreži (PowerPoint, PDF...), često se nailazi na radove koji sadrže tzv. „copy-paste“ električnih shema, valnih oblika kao i teksta iz knjiga koje su objavljene kao udžbenici, te koji su nerijetko, kršeći autorska prava, završili u raznim seminarskim radovima ili predavanjima objavljenih na internetu. Uporaba Web 2.0 alata je prisutna, ali u jednoj mjeri koju bi trebalo povećati. U nastavi se najviše koristi PowerPoint prezentacijski alat, dok se ostali alati koriste jedanput do dva puta tjedno i to najviše u svrhu razmjene audio ili video materijala [2]. Dakle razvoj računalne

tehnike i infrastrukture ne prati i kvalitetan obrazovni sadržaj koji je vrlo često kompilacija prethodno objavljenih, a često i kopiranih dijelova već objavljenih radova.

U ovom radu će se pokazati pristup kojim se želi napraviti jedan iskorak u procesu poučavanja a koristi danas sveprisutne računalne i mrežne mogućnosti. Na dva primjera, jedan iz elektroničkih sklopova a drugi iz učinske (energetske) elektronike pokazat će se primjena interaktivnog videa.

Pretpostavimo u obzir čimbenike u poučavanju i to redoslijedom kako je napisano:

- predavanje,
- udžbenik,
- rješavanje zadataka,
- računalni simulacijski program.

Predavanje kao takvo je vezano uz određeni prostor (ako nije predavanje na daljinu) i točno određeno vrijeme, dok se udžbenik, zadaci i simulacijski programi uglavnom mogu rabiti kao takvi u bilo koje vrijeme i uglavnom na bilo kojem mjestu.

U tom slijedu interaktivni video spada u nastavno sredstvo koje se može koristiti u predavanju ali i kao samostalni medijski sadržaj koji u dobroj mjeri može nadomjestiti predavanje. To je u biti osnova, polazište, kojim se objašnjava princip rada pojedinog elektroničkog sklopa. On nije računalni program s mogućnostima koje takav program daje, već daje osnovnu predodžbu funkcionalnosti određenog sklopa i sadrži početne uvjete spoznaje koji su uobičajeni u radu toga sklopa kako bi se što jednostavnije predočila njegova funkcionalnost.

Cilj obrazovanja u elektrotehnici je poimanje interakcije među fizikalnim veličinama i tvarima koji su golim okom nevidljivi. Te pojave ne možemo direktno promatrati jer nemamo receptore za takve pojave, ali možemo promatrati ili mjeriti učinke koji su iz toga proizašli. Zato se ukazuje potreba za predočivanjem električnih veličina (napon, struja, silnice magnetskog ili električnog polja i sl.) tj. da se te električne pojave - ljudskom oku prikažu prihvatljivim nadomjesnim simbolima. To se već usvojilo na razne više manje unificirane načine pri pisanju udžbenika i objavi električnih shema.

Za razliku od klasičnog udžbenika (koji može biti u elektronskom ili materijalnom obliku) interaktivni video koristi pokretnu sliku za objašnjenje funkcionalnosti određenog elektroničkog sklopa. U tu svrhu koriste se slijedeći elementi prikaza:

- Tijek struje
- Valni oblici strujnog kruga (većinom napona i struja)
- PiP (picture in picture) – slika u slici, tj. video u videu
- Prikaz pojedinih ili svih prikaza na određenoj električnoj shemi
    - supstitucija postojeće sheme sa nadomjesnom shemom
    - uključivanje samo nekih ili određenih prikaza valnih oblika
- Klasične funkcije kod prikaza videa (kreni, stani, skok na određenu poziciju, postavi video na početak, postavi video na kraj zapisa)

A. Tijek struje

Kao što je poznato električna struja je usmjereno gibanje nosioca naboja kroz medij. U ovom interaktivnom videu električna struja je prikazana kroz elementarnu česticu naboja u obliku stošca (slika 1).

Gledajući taj element s njegovog vrha, vidi ga se kao kružnica s točkom u sredini – dakle naboj se kreće prema promatraču. U suprotnom pogledu, vidi se baza stošca, koja na sebi ima križić pa je i u tom slučaju postignut sklad s definicijama smjera struje u vodičima.

Struja se sastoji od niza usmjerenih elementarnih naboja koji prolaze kroz medij. Prema tome se struja unutar vodiča u shemama elektroničkih sklopova prikazuje kao neprekidni niz elementarnih stožaca koji idu jedan iza drugog u određenom smjeru (slika 2).

Orijentacija stošca i njegovo kretanje određuju smjer struje.

Brzina kretanja struje u vodičima je veličina koja se može podudarati sa brzinom kretanja u interaktivnom videu pa je tim jednostavnije pojmiti kretanje struje kroz vodiče.

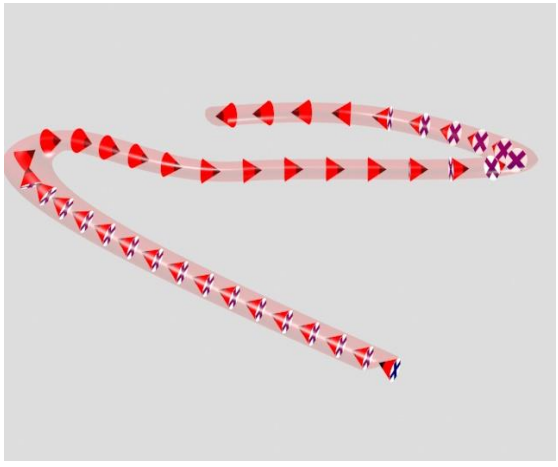
Struja u vodičima može imati različiti karakter. Može biti stalnog iznosa, pa kažemo da je istosmjerna, a može biti i promjenjivog karaktera. Najpoznatija struja promjenjive veličine je izmjenična struja sinusnog valnog oblika. Bilo da se radi o istosmjernoj struji stalnog iznosa ili izmjeničnoj struji promjenjivog iznosa, element struje će biti prikazan kroz njegovu veličinu. Dakle veća struja - veći elementarni naboj tj. veći stožac, manja struja manji simbol elementarnog naboja tj. stošca.

B. Valni oblici strujnog kruga

Prikaz tijeka električne struje kroz vodiče i elemente



Slika 1: Stožac kao prikaz elementarnog naboja u prostoru



Slika 2: Struja električnih naboja kroz vodič (primjer protjecanja struje)

elektroničkih sklopova je najvažniji i najbitniji element za razumijevanje rada sklopa. Napon kao razliku potencijala je na sličan način teže prikazati pa se za njegovo prikazivanje koriste simboli i valni oblici pridruženi tim simbolima.

Valni oblici, većinom napona i struja, prikazani su kroz grafičke prikaze funkcija koji su već prema željenom odabiru vidljivi ili sakriveni na videu.

C. *Slika u slici (PiP –Picture in picture)*

Kako bi interaktivni video bio to u svom punom značenju, koristi se prikaz slike u slici kako bi se što jednostavnije objasnila funkcionalnost pojedinog sklopa. U početku pregledavanja videa koristi se osnovna shema na kojoj se promatra način rada, a onda se dodaju dodatni sadržaji kao što su na primjer, nadomjesna shema, ili ostali video sadržaji koji pojašnjavaju rad sklopa.

D. *Prikaz pojedinih ili svih prikaza na određenoj električnoj shemi*

Pored glavne sheme spoja postoji slobodan prostor koji se rabi za prikaz valnih oblika. Zbog toga je važno da se koristi format videa 16:9 kako bi se što efikasnije i svrsishodnije iskoristila zadana površina. Pojedini prikazi tijekom prikazivanja videa mogu se po želji uključivati i isključivati kako bi gledaocu shema bila što prihvatljivija s njegovog osobnog stanovišta ili faze usvajanja principa djelovanja.

E. *Klasične funkcije kod prikaza videa*

Pri uporabi interaktivnog videa koriste se klasične funkcijske tipke i klizač („slider“) za navigaciju kroz video: kreni, stani, skoči na poziciju, skoči na početak, skoči na kraj. Svi video zapisi koji se nalaze unutar interaktivnog videa sinkronizirani su sa tim tipkama za navigaciju (slika 3).

III. PRIMJERI INTERAKTIVNOG VIDEA

U ovom radu obrađena su dva primjera interaktivnog videa. Prvi primjer je izvedba mosnog ispravljača (Graetzov spoj) koji je dobro poznat svim elektrotehničarima i kojim se pokazuju principi rada interaktivnog videa. Drugi primjer je iz učinske (energetske) elektronike gdje se objašnjava princip rada ciklopretvarača.

A. *Mosni punovalni ispravljač (Graetzov spoj)*

Kao prvi primjer na kojem je najjednostavnije pokazati funkcionalnost interaktivnog videa je jednostavan sklop za ispravljanje izmjenične struje u istosmjernu. (slika 4).

Na videu se zamjećuju putujući elementi koji predstavljaju električnu struju. Ovisno o amplitudi, elementi struje su veći ili manji. Također je definiran i smjer struje koji se za izmjeničnu struju također mijenja.

Prva pojava koja se zamjećuje je ta, što struja mijenjajući orijentaciju kretanja, odabire i druge vodove za svoj prolazak.



Slika 3: Navigacijske tipke u interaktivnom videu (skroz nazad, stani, reprodukcija, skroz naprijed – ispod tipki je navigacijski klizač

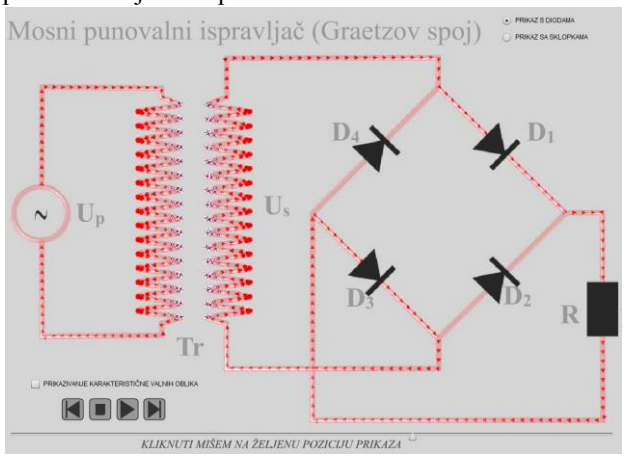
Zbog toga video ima opciju promjene aktivne elektroničke komponente. Može se pokazati nadomjesna shema gdje su diode zamijenjene s neupravljivim sklopkama (slika 5). Sklopke se otvaraju i zatvaraju ovisno o polaritetu napona koji se nalazi na njima.

Dakle, na nadomjesnoj shemi je prikazana sklopka umjesto diode i u ovom slučaju se jasno vidi da prolaskom jedne cijele periode električne struje iz transformatora (pozitivne i negativne poluperiode) – kroz otpor, prolaze samo pozitivne poluperiode. Ovime je postignut efekt punovalnog ispravljanja.

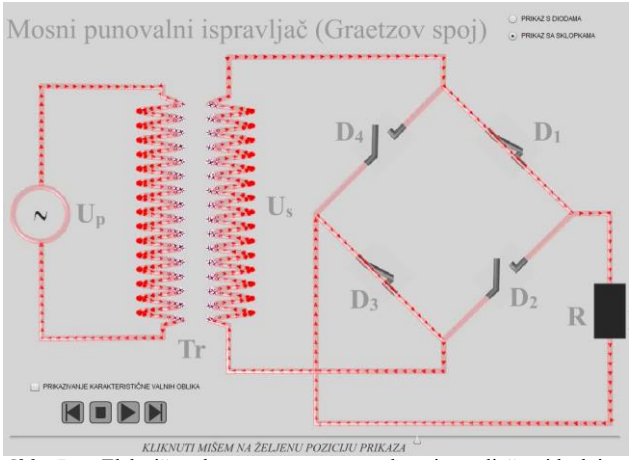
Kako plastična slika same sheme i tijeka elektrona kroz vodiče, pasivne i aktivne elemente nije dovoljna, potreban je grafički prikaz pojedinih električnih veličina. Interaktivni video omogućava prikaz pojedinih karakterističnih veličina struja i napona. To se radi postavljanjem kvačice u kvadratić „PRIKAZIVANJE KARAKTERISTIČNIH VALNIH OBLIKA“. (slika 5)

Odabirom na shemi može se sada kombinirati prikaz valnih oblika po vlastitom nahođenju. Grafički prikazi funkcija nisu stacionarni već su sinkronizirani s događanjima u električnoj shemi (slika 6).

Svaki od pojedinih valnih oblika se može isključiti i ponovno uključiti u prikazu.



Slika 4: Električna shema mosnog punovalnog ispravljača s poluvodičkim diodama (Graetzov spoj)



Slika 5: Električna shema mosnog punovalnog ispravljača s idealnim sklopkama koje zamjenjuju diode

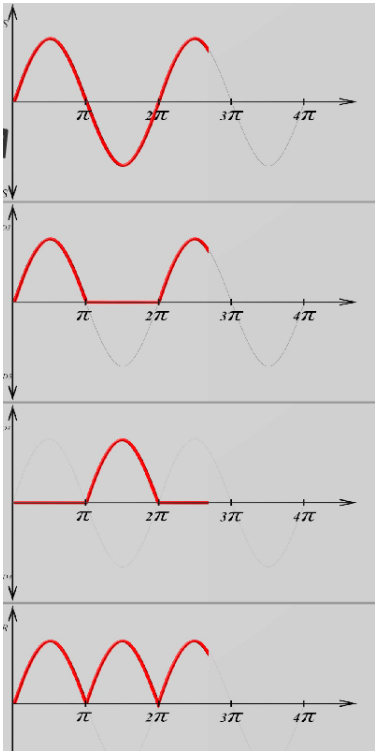
Kompletna slika rada punovalnog ispravljača je pokazana na potpuno popunjenom prikazu interaktivnog videa za taj sklop (slika 7).

Iz samog prikaza je vidljiva popunjenost zaslona sa mnogo podataka. Zbog toga je važno imati opciju za uključivanjem pojedinih prikaza po vlastitom nahođenju kako bi slika rada sklopa bila što razumljivija.

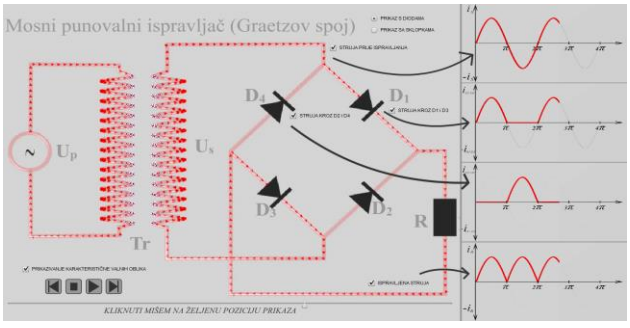
B. *Ciklopretvarač*

Mnogo kompliciraniji za objašnjenje principa rada je pretvornik frekvencije pod nazivom ciklopretvarač. U biti to je sklop učinske elektronike koji se koristi za transformaciju karaktera električne mreže na velikim snagama. U ovom primjeru se mijenja frekvencija.

Princip rada mu je na prvi pogled isti kao i kod Graetzovog spoja. Međutim ovdje se umjesto dioda u mostu uporabljuju upravljive sklopke (tiristori, GTO-i, IGBT). Obzirom da je potrebno objasniti princip rada ciklopretvarača, aktivni element učinske elektronike će se zamijeniti upravljivom sklopkom. Mehanizam upravljanja sklopkom nije predmet ovog razmatranja pa se neće niti spominjati.



Slika 6: Detalj prikaza valnih oblka na karakterističnim točkama Graetzovog spoja (odozgo prema dolje: izmjenična struja, poluvalno ispravljanje jedne grane, poluvalno ispravljanje druge grane, punovalni iznos struje na otporu R)



Slika 7: Kompletan izgled zaslona interaktivnog videa s primjerom punovalnog ispravljača (Graetzov spoj)

Shema spoja prema [1] je prikazana na slici 8.

Osnovno pitanje koje se postavlja je: kako od ulazne frekvencije ω_1 dobiti izlaznu frekvenciju ω_2? Za odgovor na ovo pitanje nam je potreban prikaz ulaznog napona i napona koji prolazi kroz otpor R.

Ciklopretvarači mogu konvertirati frekvenciju samo na niže ($\omega_2 < \omega_1$). Odgovarajućim uklapanjem i isklapanjem sklopki dobit će se jedan interesantan valni oblik (slika 9).

Na slici je vidljivo da se idući po osnovnoj sinusoidi ulazne frekvencije, izmjeničnim uklapanjem i isklapanjem sklopki dobiju dijelovi sinusoide koji na odsječku od 0-10π imaju uglavnom pozitivan predznak, a na odsječku od 10π-20π uglavnom negativan predznak.

Ako se hoće vidjeti srednja vrijednost napona za te odsječke onda treba u strujni krug staviti induktivitet koji zajedno sa otporom čini vremensku konstantu koja utječe na faktor glađenja („peglanja“) krivulje tj. smanjenje amplituda viših harmonika koji se ovakvim načinom pretvorbe frekvencije javljaju u izlaznom signalu.

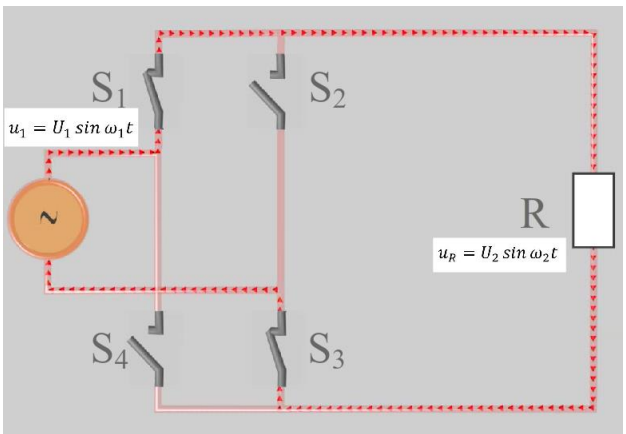
U tu svrhu se s radio gumbom promijeni shema prikaza (slika 10).

Isto tako se i upotpuni valni oblik s izravnatom („ispeglanom“) krivuljom (slika 11). Iz slike na interaktivnom videu je vidljivo da je ovime dobivena promjena frekvencije koja je na izlazu 10 puta manja.

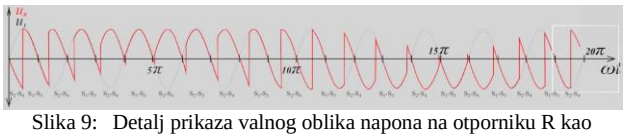
Naravno da je to idealizirana slika, te da prava slika rada ciklopretvarača treba biti verificirana kroz simulacijski program ili mjerenjem.

Kako je na zaslonu slika stvaranja valnog oblika relativno mala i nepregledna, to je kreiran jedan detalj slike koji se prati tijekom projiciranja videa. Dio sadržaja koji se posebno može pratiti je na slici 11 označen s bijelim kvadratom, a sam izgled dodatnog prozora je prikazan na slici 12.

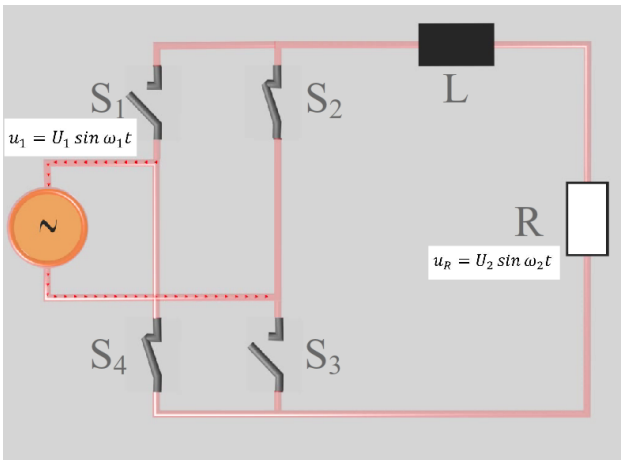
Kompletan prikaz interaktivnog videa s ciklopretvaračem i svim otvorenim prozorima te mogućnostima, prikazan je na slici 13.



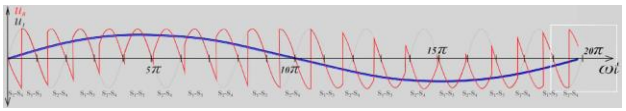
Slika 8: Osnovni prikaz sheme spoja ciklopretvarača uključujući idealne sklopke (S_1-S_3 i S_2-S_4 koje ukapčaju i iskapčaju u parovima) [1]



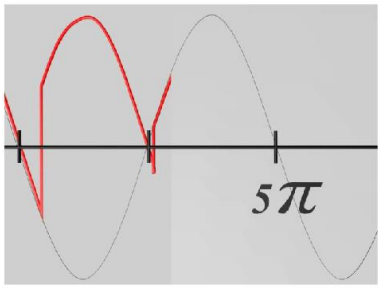
Slika 9: Detalj prikaza valnog oblika napona na otporniku R kao posljedica određenog ritma ukapčanja i iskapčanja parova sklopki S_1-S_3 i S_2-S_4



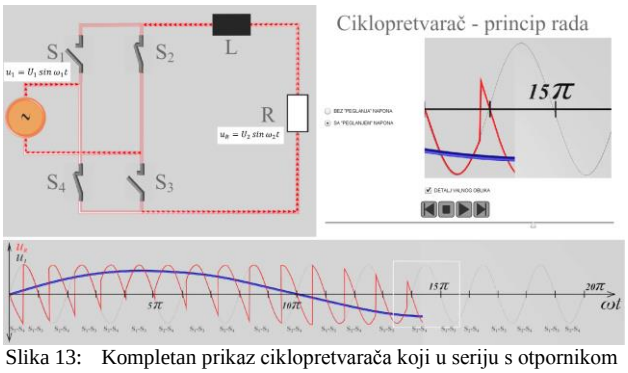
Slika 10: Detalj prikaza ciklopretvarača koji u seriju s otpornikom ima induktivitet L koji služi za izravnavanje krivulje



Slika 11: Detalj prikaza valnog oblika napona na otporu R (plava krivulja) kao posljedica „izravnavanja“ krivulje induktivitetom L



Slika 12: Detalj prikaza krivulje koja se iscrtava tijekom rada



Slika 13: Kompletan prikaz ciklopretvarača koji u seriju s otpornikom ima induktivitet L koji služi za izravnavanje krivulje

Na primjeru ciklopretvarača se ne spominju kvantitativni podaci o veličinama koje se koriste u prikazu. U ovom slučaju je npr. bitna vremenska konstanta L/R sklopa, koja mora biti u određenom odnosu s vremenom trajanja periode ulaznog i izlaznog signala, kako bi mogla efikasno (do određene mjere) izgladiti napon niže frekvencije u svrhu dobivanja izlaznog signala što sličnijeg sinusnom obliku.

Tako je predviđeno i u ostalim primjerima interaktivnog videa, čime se ispunjava osnovna zadaća ovog nastavnog sredstva koji je upućen na objašnjenje rada sklopa. Za egzaktne odnose fizikalnih veličina se mogu koristiti simulacijski računalni programi.

ZAKLJUČAK

U dosta primjera uporabe moderne računalne i mrežne tehnologije, ne vidi se toliki napredak iste za obrazovne svrhe [2]. Još uvijek se računalo koristi za pregledavanje i razmjenu medijskih sadržaja koji se koriste na sličan način kao da čitamo udžbenik ili gledamo televiziju. Dakle sadržaj je isti, način prezentiranja sadržaja isti, a jedina je razlika u pomagalu koji koristimo. Dok je to prije bila isključivo knjiga kao fizička jedinica – danas je to pored knjige: računalo, tablet, čitač knjiga, mobitel. Ili rečeno nastavnim rječnikom – nastavno sredstvo je ostalo isto, samo se promijenilo nastavno pomagalo (tu se izuzima uporaba simulacijskih programskih paketa koji imaju nemjerljivu ulogu danas u odnosu na nekada – ali imaju nedostatak, kao što je važno ponoviti iz uvoda, što su većinom vezani uz licencu, a trebaju i ne malu vještinu za njihovu uporabu).

U tu svrhu uporaba interaktivnog videa ima svoju opravdanost. Snaga računala koja se danas rabe, premašuju potrebe ove vrste prikaza, te njihovo uvođenje u nastavni proces ima smisla. Nameće se potreba za eventualna poboljšanja, nadogradnje i ostale izmjene kojima bi se zornost takvog načina prikaza poboljšala. Potrebu za unificiranjem tj. nekom vrstom standardizacije ovakvog načina prikaza, trebalo bi isto tako razmotriti, te time napraviti iskorak u smislu poučavanja.

I na kraju ostaje pitanje: kako kvalitetno opisati interaktivni video u pisanom obliku? Ovaj rad je samo pokušaj dočaravanja njegovih odlika koje treba doživjeti i ocijeniti na zaslonu prezentacijskog uređaja.

LITERATURA

[1] John G. Kassakian Martin F. Schlecht George C. Verghese, “Osnove energetske elektronike”, I dio, Graphis, Zagreb, 2000. stranice 5-30.

[2] M. Gligora Marković, M. Rauker Koch , Miro Frančić, „Uporaba Web 2.0 alata u poučavanju“, Mipro, 2012.

Primjena e-portfolija u nastavi

Stjepan Šalković
Srednja škola Krapina
Šetalište hrvatskog narodnog preporoda 6, 49000 Krapina
Republika Hrvatska
stjepan.salkovic@skole.hr
+385 49 382 100

Sažetak – U radu je opisana primjena e-portfolija u nastavi računalstva učenika strukovnih usmjerenja. Realizacija je temeljena na sustavu otvorenog koda Mahara. Aktivnost je bila usmjerena na domaće zadaće i seminare, no može imati puno širu primjenu. Pokazalo se da je motivacija za rješavanje problema bila veća na ovaj način, nego na klasične načine koji su dosad korišteni. Učenici su pozvani da portfolio razvijaju kroz cjelokupno školovanje, te da ga popunjavaju kako bi sačuvali svoje uratke. Završetkom školovanja, uz pretpostavku čuvanja radova iz svih predmeta struke, mapa struke može biti vidljiva budućim poslodavcima ili u daljem školovanju. Na kraju prve faze provedena je anketa, kako bi se uočile prednosti i nedostaci. Temeljem rezultata ubuduće će biti uvedene određene promjene.

I. UVOD

Ishodi učenja i kompetencijski sustav postavljeni su u prvi plan u Nacionalnom okvirnom kurikulumu, kao temeljnom dokumentu koji određuje bitne sastavnice odgojno-obrazovnog sustava. U suvremenom društvu nove kompetencije pojedinca usmjerene su na razvoj inovativnosti, stvaralaštva, rješavanja problema, razvoj kritičkog mišljenja, poduzetnosti, informatičke pismenosti, socijalnih i drugih kompetencija. To je potrebno realizirati kroz strukovne, predmetne ili školske kurikulume.

Neki od pokazatelja razine usvojenosti znanja i vještina su ocjene, rezultati državne mature, rezultati natjecanja, završni radovi i sl. No, konkretni uradci koji su nastajali kroz nastavne aktivnosti uglavnom ne ostaju sačuvani. Učenik koji će tražiti posao ili ići na fakultet na kraju školovanja tako ne može prikazati svoj cjelokupni slijed napredovanja u struci.

Uporaba sustava za udaljeno učenje donekle rješava taj problem. Uradci ostaju sačuvani, te je napredak učenika jasnije vidljiv. Ipak, korak dalje je sustav koji može omogućiti izradu elektroničkog portfolija, životopisa, te određeno druženje i funkcionalnost društvene mreže.

S ciljem unapređenja nastavnih aktivnosti, poboljšanja motivacije, te organizacije spremišta uradaka proveo sam aktivnosti rješavanja programskih zadataka, te njihovo spremanje u e-portfolio. Učenici tijekom ili nakon odrađene cjeline mogu pokazati koje su praktične vještine usvojili. Kako se ne bi previše opteretili, zadatak koji su imali bio je podijeliti dnevnik i dodati datoteke kao zapis u dnevnik.

Provedena je anketa kojom se nastojalo sagledati stavove učenika u vezi s elektroničkim portfoliom i domaćim zadaćama. Kroz anketu je ispitano što je učenicima predstavljalo najviše problema, čime su zadovoljni, a što bi trebalo poboljšati.

II. DOMAĆE ZADAĆE I PROJEKTI

Internet je ušao u sve pore današnjeg društva. Djeca ga koriste već od malih nogu. Nakon toga vezan je uz nastavni proces kroz osnovnu i srednju školu. Već u srednjoj školi, a pogotovo na fakultetu upotrebljava se i u struci. Neke su struke direktno, a većina indirektno vezane uz njegovu uporabu. Može se doći do zaključka da je uz današnje generacije Internet neraskidivo vezan.

Elektronička komunikacija koja kao medij koristi Internet ubrzava i olakšava komunikaciju, ali joj daje i nove dimenzije. Materijali za učenje uobičajeno su na web stranicama, a nastaju i e-udžbenici. Elektronička pošta se u nastavnom procesu sve više predmeta podrazumijeva kao uobičajeni medij komunikacije. Istovremeno se dio ili cijela nastava određenih predmeta provodi online uporabom sustava za udaljeno učenje.

Kroz svoje školovanje učenici rješavaju domaće zadaće ili rade na školskim projektima čiji rezultat je digitalni dokument, ili ostaje neki drugi digitalni trag. Radovi se šalju elektroničkom poštom, ostavljaju unutar određenog CMS-a ili dostavljaju nastavniku na mediju i preko mreže. Nažalost, kroz određeno vrijeme ti radovi se izgube, izbrišu ili pak ostaju arhivirani.

Učenici strukovnih usmjerenja kojima je temelj struke tehnika u više predmeta koriste računalo kao osnovni alat u nastavi. U računalstvu izrađuju dokumente u uredskim aplikacijama, programiraju u raznim programskim jezicima, izrađuju web stranice i multimediju. Unutar struke izrađuju sheme, predloške, crteže, simulacije, programiraju i dokumentiraju određene radove. Dio učenika u izvannastavnim i izvanškolskim aktivnostima, te samostalno usvaja dodatna znanja i stječe vještine.

Poželjno bi bilo na određeni način sačuvati uratke tijekom cjelokupnog školovanja, kako bi prilikom zapošljavanja poslodavac mogao sagledati kojim kompetencijama vlada njegov budući djelatnik. Ujedno bi visokoškolska ustanova na kojoj učenik nastavlja školovanje mogla procijeniti početno stanje kompetencija, te kurikulum i grupe urediti tako da je napredak studenta optimalan. Kako bih započeo aktivnosti sa navedenim ciljevima, odlučio sam uputiti učenike u rad s elektroničkim portfoliom.

III. ELEKTRONIČKI PORTFOLIO

Portfolio se u poučavanju i bilježenju učeničkih postignuća koristi već niz godina u više područja. Vizualni i literarni uradci ili pak primjena u arhitekturi tradicionalni su načini uporabe. Radovi su se spremali u mape i čuvali, no ograničenje im stvara prostorna udaljenost, a teško ih je i prilagoditi za određenu uporabu.

Digitalno doba donijelo je medije za pohranu i prijenos slika, zvuka, videa i dokumenata raznih vrsta. Razvojem

sustava za upravljanje sadržajima (eng. Content management system - CMS) poučavanje je dobilo nove mogućnosti. Materijali za učenje, zadaci i rješenja, te sama komunikacija izašli su iz tradicionalnih učionica.

U novom tisućljeću počeo je i veći zamah elektroničkih portfolija. Počele su se razvijati tehnologije, funkcionalnosti i osobine. Danas se e-portfolio smatra jednim od načina sakupljanja dokaza učeničkog napretka i razvoja obrazovnih postignuća unutar određenog predmeta, struke ili institucije[I].

Kako bi e-portfolio polučio uspjeh nužno je obratiti pažnju na određivanje ishoda učenja, razumijevanje učenikovih potreba, identificiranje dionika, određivanje aktivnosti u poučavanju, evaluaciju portfolija, uključivanje više oblika uradaka, te vrednovanje učinka e-portfolija[II].

Elektronički portfolio je vrijedan alat za proces poučavanja, ali i vrednovanja i ocjenjivanja. Može poslužiti kao administrativni alat za pomoć u vođenju procesa učenja, za prezentaciju riješenih zadataka, te bilježenje ciljeva, rada i postignuća. Uporaba tehnologije omogućuje uporabu bilo kad, bilo gdje, pojedinačno ili u grupi i sa povratnom informacijom.

Učenički portfolio kao skup radova moguće je upotrijebiti kao jedan od načina vrednovanja i ocjenjivanja postignuća u nastavnim aktivnostima. Moguće je pratiti napredak kroz školsku godinu tako da se određena aktivnost odradi na početku, sredinom i na kraju razdoblja. Na taj način nastavnici, roditelji i učenici dobivaju pokazatelj razvoja učenika.

Ukoliko pak učenik tijekom nastavnog procesa odabere svoje najbolje uratke, može ih također uključiti u e-portfolio. Takav skup radova može se ocijeniti, ali i sačuvati kao dokaz stečenih kompetencija u školovanju.

Različita iskustva uporabe e-portfolija u Republici Hrvatskoj, nastavne aktivnosti i pedagoške osnove uporabe prikazane su u priručniku nastalom iz projekta „Uporaba sustava e-portfolija u sveučilišnoj nastavi“[III].

IV. Uporaba Mahare za izradu učeničkog portfolija

Mahara je e-portfolio sustav temeljen na sustavu otvorenog koda. Omogućuje izradu plana učenja, obaveza i ciljeva, izradu i nadopunjavanje životopisa, izradu stranica, komunikaciju i prikupljanje postignuća. Omogućuje i pisanje blogova, te razvoj komunikacije kroz određenu vrstu društvene mreže.

Korisnik Mahare može urediti svoje osobno okruženje za učenje, te za praćenje vlastitog razvoja. Dijelove sustava koje želi može podijeliti s drugim korisnicima, nastavnicima ili poslodavcima.

Mahara počiva na tri stupa. Sakupljanjem i sastavljanjem izrađuje se osobni profil, dodaju se datoteke, piše životopis ili blog. Drugi stup podrazumijeva organizaciju i prezentaciju portfolija kroz oblikovanje stranica različitim elementima i medijima. Treći stup je komunikacija, dijeljenje, prezentiranje, komunikacija i diskusije.

V. Praktični zadatak

Učenici strukovnih usmjerenja završetkom školovanja stječu kvalifikaciju određene razine, obujma, profila i kvalitete. Ona se dokazuje javnom ispravom, no poželjno je da učenik ima i materijalne dokaze kompetencija. Oni nastaju u nastavnom procesu, no nakon ocjenjivanja, njihova važnost kao da se zaboravlja. Upravo time sam pokušao učenicima pokazati da im porfolio nije nametnut, nego da će od njega imati koristi.

Zadatak učenika je bio izraditi program u jeziku C na određenu temu. Teme uglavnom pokrivaju primjenu u struci, a učenici ih mogu birati sami. Ovisno o kompleksnosti algoritama, uloženom trudu i pokazanom napretku vrednuje se i rad.

Kako bi portfolio započet u srednjoj školi učenici mogli koristiti i u visokoškolskom obrazovanju, upotrijebljena je instalacija Mahare Centra za učenje na adresi https://moodle.srce.hr/eportfolio/. Jedna od prednosti ove instalacije je uporaba AAI@EduHr autentikacijske i autorizacijske infrastrukture. Time je administrativni dio za mene kao nastavnika bio sveden na minimum. Prvotna ideja bila je instalacija i podešavanje Mahare za potrebe škole, no takvo rješenje imalo bi više nedostataka nego prednosti.

Za potrebe komunikacije i dijeljenja sadržaja izrađene su grupe za Srednju školu Krapina i za pojedine razrede prema godini upisa. Nakon prijave učenici su se samostalno prijavljivali u navedene grupe. Mogli su popuniti stranicu „O meni“, no to im nije nametnuto kao obveza. Životopis je također dio portfolia za koji nije traženo da se popuni.

U anketi je uočeno da shvaćaju važnost životopisa, no kako trenutno nisu očekivali da će biti pregledavan, uglavnom nije bio ni popunjen.

Učenici su sadržaj dijelili pomoću stranice. Na linku „Moj portfolio“, unutar područja „Stranice“ izradili su novu stranicu. Ovisno o željama mogli su urediti izgled. Važno za dijeljenje je bilo povlačenje bloka za prikaz cijelog dnevnika, te njegovo dijeljenje na linku „Podijeli stranicu“. Ovlasti su mogli dodijeliti samo profesoru, razredu ili školi.

Uratke su učenici dodavali kao zapise u dnevnik. Zapisu su prilagali datoteke. Trenutni prostor kojim raspolažu je 20 MB. Savjetovano im je da prema potrebi koriste poveznice na svoj prostor u određenom oblaku (Google Drive, SkyDrive, Dropbox i sl.). Svaki novi uradak bit će novi zapis u dnevniku, za čije dodavanje učenik treba vrlo malo vremena.



Slika 1: dijeljenje sadržaja



Slika 2: povratna informacija

Iako su svi učenici dobili jasne upute, korak po korak, bilo je problema s postavljanjem i dijeljenjem sadržaja. Dosadašnje uratke uglavnom su ostavljali na mrežnom disku, za što je trebao minimum uloženog truda, pa je dio učenika pružao otpor početnim podešavanjima e-portfolia.

Nakon nekog vremena, tek nekoliko učenika svoje uratke nije postavilo na e-portfolio. Sukladno dogovoru, upisane su povratne informacije koje su uz obrazloženje ocjene saopćene i na nastavnom satu.

VI. REZULTATI ANKETE

Za istraživanje je korištena anketa provedena u Srednjoj školi Krapina. Anketirano je 77 učenika i učenica drugog i trećeg razreda strukovnih usmjerenja. Određena pitanja postavljena su tako da je korišteno obrnuto bodovanje kako bi ispitanici razmislili o svakom odgovoru.

Prva skupina pitanje odnosila se je na ocjenu vlastitog iskustva s e-portfoliom. Učenici su se uglavnom složili da e-portfoliom mogu pokazati usvojenost znanja, vještina i kompetencija, da omogućuje vidljivost uradaka nastavniku, te čuvanje uradaka za duže razdoblje na jednom mjestu. Navedeno mišljenje prevladalo je kod 70 do 80 % učenika. Kompleksnost početnih podešavanja nešto je veća u postotku od kompleksnosti uporabe. Navedeno je smatralo čak do 50% učenika. Povratnu vezu i mogućnost komunikacije neki nisu ni uočili, iako upotrebljavaju društvene mreže svakodnevno.

Prikupljanje radova velika većina (76%) je smatrala kao najvažniju mogućnost u e-portfoliju. Slijedi plan rada, obveza i ciljeva, izrada životopisa, dijeljenje iskustva, razmišljanja i informacija, te osobne stranice sa oko 50%. Najmanje smatraju da je važno pisanje dnevnika, odnosno bloga. Kako većina ovih aktivnosti nije ni tražena, teško je bilo bez vlastitog iskustva očekivati da će se moći procijeniti važnost aktivnosti.

Treća skupina pitanja odnosila se je na domaću zadaću. Uratke koji su postavljeni na Maharu, učenici su u određenom vremenskom razdoblju radili kod kuće. Osim tih uradaka, ranije su dobivali i druge domaće zadaće. Većini učenika (60 do 70%) zadaća pomaže u učenju, a istovremeno je zadaća primjerena sposobnosti, pomaže u razumijevanju gradiva obrađenog na satu i pomaže u stjecanju radnih navika. Nažalost, tek 16% učenika uživa raditi domaću zadaću. Oko 60% učenika je previše opterećeno domaćim zadaćama i piše ih zato da ne bi dobili kaznu ukoliko ih ne izrade. Nešto više od pola učenika misli da domaće zadaće ne treba zadavati. Ovime se vidi da su učenici svjesni važnosti domaćih zadaća, ali nedostaje motivacija. Na tome treba svakako poraditi u budućnosti.

Testiranjem razlika s obzirom na usmjerenje utvrđeno je da su tehničari za mehatroniku pokazivali veće slaganje s tvrdnjama, odnosno da su pozitivnije procjenjivali navedene aspekte.

Testiranjem razlika s obzirom na razred utvrđeno je da učenici trećih razreda vide kao veću prednost mogućnost dokazivanja usvojenosti svojih znanja i vještina (M=2,94, sd=0,74) od učenika drugih razreda (M=2,51, sd=1,06). Prikupljanju radova, projekata i sličnih dokumenata također daju učenici trećih razreda veću važnost (M=3,06, sd=0,78) od učenika drugih razreda (M=2,60, sd=0,98). Slična je još situacija s pisanjem dnevnika-bloga gdje treći razredi daju važnost (M=2,44, sd=0,86), a drugi razredi (M=1,93, sd=1,01). Može se doći do zaključka kako su stariji učenici već sakupili određenu količinu radova kojom bi se mogli pohvaliti, ali i da su sve svjesniji slijedeće stepenice u svojem životu u kojoj im mogu zatrebati određeni dokazi o kompetencijama.

U dijelu u kojem su učenici mogli samostalno dodavati komentare i prijedloge uočeno je nesnalaženje unutar sučelja i konstatacije o kompleksnosti upotrebe Mahare.

Nakon nekoliko godina korištenja vrijedilo bi postaviti pitanje "planirate li nastaviti koristiti e-portofolio", kako bi se mogla raditi regresijska analiza - da li se na osnovi tvrdnji može predvidjeti buduće korištenje e-portofolia

VII. ZAKLJUČAK

Elektronički portfolio može postati dio procesa cjeloživotnog učenja, svatko ga može prilagoditi svojim potrebama, učenik mu je u središtu, a web tehnologija omogućuje dostupnost bilo gdje i bilo kad.

Nove tehnologije sve se više ugrađuju u nastavni proces. Ponekad ih uključujemo možda i više iz razloga popularnosti kod mladih, nego prepoznatljive edukativne vrijednosti. Jedan od faktora koji je bitan kod planiranja e-porfolija su dosadašnje navike, odnosno kultura. Uvijek postoje otpori uvođenju novih tehnologija, čak i ako je vidljivo da donose mnogostruko više koristi od uloženog truda.

Generalni zaključak ankete može se svesti na činjenice da učenici prihvaćaju e-portfolio za čuvanje svojih uradaka, no na žalost ne uživaju raditi domaću zadaću. Moguće je ipak da će vrijeme pokazati da e-portfolio postaje nezaobilazan pratitelj učenika, studenta i njihovih edukatora.

LITERATURA:

[I] Leskes, A., i Wright, B. D.. *The Art and Science of Assessing General Education Outcomes*, Washington, DC: Association of American Colleges and Universities, 2005.

[II] Chen, H. L., Penny Light, T., *Electronic Portfolios and Student Success: Effectiveness, Efficiency, and Learning*, Washington, DC: Association of American Colleges and Universities, 2010.

[III] Kučina Softić S., Perišić Pintek T., Bekić Z., ur., *„E portfolio kao nastavna aktivnost“ – priručnik za nastavnike*, Zagreb, sveučilišni računski centar, 2013.

Bonton digitalnog doba

Marina Mirković, dipl. inženjer elektrotehnike
Tehnička škola, Požega, Republika Hrvatska
marina.mirkovic@po.t-com.hr

Sažetak

Informatičko/informacijska pismenost je nužnost današnjeg digitalnog doba. Ponašanje na internetu, u virtualnoj zajednici mora biti u skladu s kulturno-civilizacijskim i poslovnim normama. Usvajanje pravila bontona u raznim situacijama u kojima koristimo moderne tehnologije pomoći će nam da steknemo dobre manire.

Važno je definirati skup pravila poželjnog ponašanja u internet zajednici. Pridržavanje pravila u slanju i primanju elektroničke pošte nije samo stvar bontona već je neophodno i zbog očuvanja sigurnosti, stvaranja ugleda i održanja produktivnosti poslovanja.

Brojna su pravila kojih se moramo pridržavati pri primjeni mobilnih telefona, elektroničkoj komunikaciji putem elektroničke pošte, u virtualnim zajednicama kao što su News grupe i forumi, kao i društvenim mrežama.

U Tehničkoj školi Požega nizom godina učenici sudjeluju u projektu «Sigurnost djece na internetu» te su provedena istraživanja i prikupljanje podataka o iskustvima učenika prilikom korištenja interneta, kao i primjene pravila e-bontona. Analiza rezultata ankete pokazuje da u školama učenici moraju upoznati e-bonton i primjenjivati ga u svojoj digitalnoj komunikaciji. Pravila lijepog ponašanja u virtualnom svijetu objavio je Hrvatski Telekom izdavanjem e-bontona koji je podjednako obuhvatio mišljenje sudionika komunikacije na društvenim mrežama i mišljenje eksperata s iskustvom promatranja najraznovrsnijih oblika komunikacije kao međuljudske interakcije.

Osim djece i mladih, poslovni ljudi, zaposlenici u tvrtkama također moraju posebnu pažnju posvetiti digitalnom bontonu. Edukacijom zaposlenika u tvrtkama moguće je zaštititi sebe i tvrtku od nezgodnih pitanja odgovornosti. Važno je osigurati e-bonton politiku koja se odnosi na sve zaposlenike.

I. UVOD

Moderno doba i nove tehnologije donijele su nove izazove a time i promjene normi lijepog ponašanja. Osnovne karakteristike i dalje su: ljubaznost, poštovanje, obzirnost i tolerancija. Svojim ponašanjem utječemo na okolinu. Usvajanje pravila bontona u raznim situacijama u kojima koristimo moderne tehnologije pomoći će nam da steknemo dobre manire a ujedno učiniti naš privatni i poslovni život ugodnijim i uspješnijim. Lijepim ponašanjem naša zajednica postaje bolje mjesto življenja. Koliko se pridržavamo pravila bontona i shvaćamo li njegovu važnost kao i praktične koristi koje nam pruža? Kako se mladi, djeca, naši učenici odnose prema njemu, poznaju li i pridržavaju se pravila?

Ulaskom u Europsku unija, Hrvatska mora poraditi i na školskom sustavu. Prema anketi portala EduCentar, u kojoj je sudjelovalo 400 ispitanika, hrvatski građani bi među novim nastavnim predmetima u školski program najprije uveli bonton. Najviše ispitanika (39%) podržava uvođenje bontona u školstvo, jer problem s bontonom imaju odrasli, a time, nažalost i njihova djeca [1].

II. PONAŠANJE KOD PRIMJENE DIGITALNE TEHNOLOGIJE

Mnogi, a posebno mladi ljudi su postali sve ovisniji o mobitelima, neprestano su uz njih. Mobilni telefoni služe za kratku razmjenu informacija ili hitne pozive. Dugi i glasni razgovori uslijed dosade, uz prisutnost drugih osoba, posebno na javnim mjestima, nisu primjereni. Nepristojno je u prisutnosti drugih osoba neprestano provjeravati pozive, poruke ili surfati po internetu. Mnogi su u društvu, na zabavi ili jednostavnom druženju opsjednuti mobitelom. Nepristojno je dugotrajno pisanje poruka, na što mladi ne obraćaju pažnju. Ponekad ne isključuju mobitele tamo gdje su oni nepoželjni (predavanja, u kinu, kazalištu, crkvi, muzeju ili knjižnici). Mobitele često koriste za fotografiranje pri čemu je važno voditi računa o tome koga su slučajno «uhvatili» na fotografiji i u kojoj situaciji. Ne smijemo narušavati tuđu privatnost, posebno ako namjeravamo fotografije objaviti na društvenim mrežama [2].

Elektronička pošta putem interneta nije sasvim sigurna. Nužno je pažljivo postupati sa slanjem svojih privatnih informacija i voditi brigu o zaštiti sadržaja. Nikako nije poželjno prosljeđivati ili započinjati «lance sreće» jer oni nisu dopušteni na internetu. Zbog toga pristup mreži, forumu ili servisu može biti uskraćen [3].

Ukoliko šaljete velike količine podataka, moguće je «zakrčiti» elektronički poštanski sandučić ili povećati troškove korisnicima koji e-poštu otvaraju na privatnom pametnom telefonu. Iz istog razloga nepoželjno je odnosno zabranjeno nezatraženo e-mail oglašavanje (spam).

Pismenost je važan dio pristojnosti pri slanju elektroničkih poruka. Treba izbjegavati pisanje svih riječi velikim slovima. Korištenje slova u raznim bojama otežava čitanje i djeluje neozbiljno. Ne smijemo pretjerivati s emotikonima u porukama [4].

Forum je virtualna zajednica korisnika sa sličnim interesima. Pri tome se korisnici moraju pridržavati pravila online bontona. Poruka koju pišete mora biti jasna i jednoznačna. Poštujte temu o kojoj se raspravlja i ne pokušavajte ju mijenjati. Za privatne razgovore koristite privatne poruke ili elektroničku poštu. Važno je nastojati ne stvarati nesporazume u kojima se može uočiti arogancija ali i agresivnost.

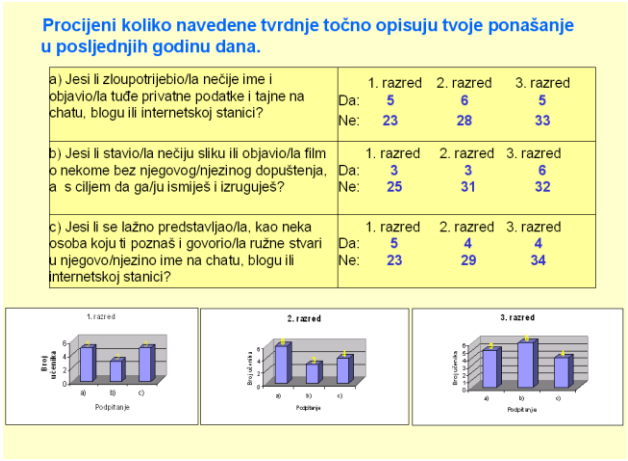
Putem društvenih mreža dijelimo iskustva i mudrosti na koje smo naišli ili naučili. Otkrivamo što se ljudima sviđa, kako se ponašaju, što dijele i komentiraju. Treba nastupiti prijateljski, oprezno prema osobama iz neke druge kulture. Ne treba otkrivati previše osobnih detalja. To nije mjesto gdje ćete iskazati svoje frustracije i bijes, javno se osvećivati ili rugati. Pazite što stavljate na društvene mreže. Sve ostaje zapisano, a netko to može iskoristiti protiv vas.

Facebook je nažalost postao rasadnik nekulturne mase koja na ovoj društvenoj mreži čini stvari koje se u stvarnom životu mogu usporediti sa psovanjem, vikom, tučnjavom i sličnim. Mnoštvo aplikacija korisnicima pomaže da budu što dosadniji i naporniji. Blokiranjem aplikacija riješiti ćete problem. Pretjerivanje u psovanju i vrijeđanju najodbojnija je stvar na društvenim mrežama. Česti su dosadni statusi, pojavljuju se najnaporniji i najneugodniji korisnici koji vas neprestano pozivaju na različite događaje i aktivnosti. Uz njih su naporni i hvalisavci.

Privatno i poslovno na društvenim mrežama mora biti odvojeno. Mnogi ne znaju ili se ne pridržavaju bontona na društvenim mrežama. Ne shvaćaju važnost primjene pravila uz više od milijun korisnika društvenih mreža u Hrvatskoj. Ponašanje u digitalnom svijetu odraz je i našeg ponašanja u pravom životu. Pokažimo da se znamo ponašati. Nepristojne korisnike na vašim društvenim mrežama bez ustručavanja obrišite s popisa prijatelja.

III. ANALIZA UČENIČKOG PONAŠANJA U DIGITALNOM SVIJETU

Učenici moraju uočiti povezanost između primjene pravila bontona u digitalnom svijetu i sprječavanja mogućnosti nasilja i doživljavanja neugodnih situacija u radu na internetu. U Tehničkoj školi Požega nizom godina učenici sudjeluju u projektu «Sigurnost djece na internetu» te je provedeno istraživanje i prikupljanje podataka o iskustvima učenika prilikom korištenja interneta. U anonimnom anketiranju o sigurnosti i zaštiti djece na internetu u veljači 2010. godine, putem upitnika na nastavi računalstva sudjelovali su učenici razrednih odjela: 1.f, 1.h, 2.a, 2.b, 2.e, 2.f, 2.h, 3.f, 3.h - sveukupno 100 učenika. U analizi odgovora, među ostalim, otkriveno je da dio učenika (13%) pokazuje znakove "internet nasilnika" (slika 1). Djeca i mladi moraju postati svjesni kakve posljedice može imati njihovo ponašanje na internetu. Potrebno je istaknuti kako su distribucija i prosljeđivanje nasilne, uvredljive, ponižavajuće poruke jednako loši i mogu imati jednako teške posljedice kao i ponašanje osobe koja je začetnik poruke.



Slika 1. Analiza odgovora jednog pitanja ankete učenika o sigurnosti i zaštiti djece na internetu

U svibnju 2013. godine 109 učenika Tehničke škole u Požegi (iz razrednih odjela: 1.h, 1.f, 2.a, 2.b, 2.f, 2.h. 3.b. 3.f. 3.h) sudjelovalo je u anonimnoj anketi o pisanju SMS i poruka elektroničke pošte [5]. Analiza njihovih odgovora, prema pitanjima u anketi prikazana je u tablici 1.

Analizirajući rezultate anketiranja, iznenađuje broj učenika koji šalju poruke s uvredljivim sadržajem, kao i uporaba elektroničke pošte za raspravu o povjerljivim informacijama. Dio učenika prosljeđuje lančana pisma i odgovara na spam. Velika brzina pri pisanju/tipkanju navodi ih da pišu dugačke rečenice u porukama, a neprestana prisutnost i povezanost s digitalnom tehnikom omogućuje odgovaranje po primitku poruke. Mnogi učenici ne razmišljaju o smislenom naslovu poruke elektroničke pošte.

Većina učenika pročita poruku prije nego ju pošalje, dok se dio učenika ne pridržava tog pravila. Pri tome šalju poruke s pravopisnim, gramatičkim i interpunkcijskim pogreškama. Ne smijemo biti brzopleti, važno je provjeriti tekst poruke jer ćemo tada poslati učinkovitiju poruku i izbjeći nesporazume i neprikladne komentare. Budući da je čitanje s ekrana teže nego čitanje s papira, struktura i raspored vrlo su važni za e-mail poruke. Kod uporabe oblikovanja u porukama, pošiljatelj ga možda neće vidjeti ili će vidjeti drugačije fontove od željenih.

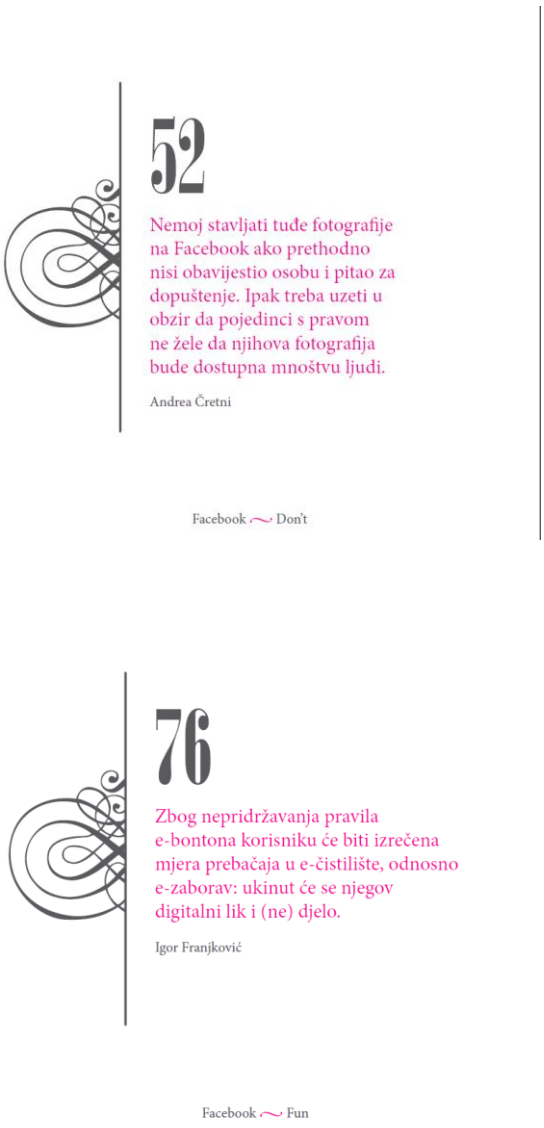
Rezultati istraživanja ukazuju na činjenicu da je potrebna daljnja edukacija učenika o važnosti primjene e-bontona, kao i opasnostima i načinima samozaštite djece prilikom korištenja računala i interneta.

Tablica 1. Rezultati ankete učenika Tehničke škole u Požegi o pisanju SMS i e-mail poruka

Anketno pitanje	Odgovor DA Broj učenika	DA postotak	Odgovor NE Broj učenika	NE postotak
Pišeš kratko i jasno	91	83 %	18	17 %
Paziš na gramatiku	47	43 %	62	57 %
Odmah odgovaraš na poruke	72	66 %	37	34 %
Koristiš nepotrebne privitke	21	19 %	88	81 %
Pišeš velikim slovima	17	16 %	92	84 %
Kod odgovora, uvijek uključiš prethodne poruke	56	51 %	53	49 %
Pročitaš poruku prije nego je pošalješ	91	83 %	18	17 %
Pažljivo koristiš kratice i emocionalne ikone	64	59 %	45	41 %
Pažljivo formatiraš poruke	47	43 %	62	57 %
Prosljeđuješ lančana pisma	14	13 %	95	87 %
Uključuješ tuđu poruku ili privitak bez dozvole	16	15 %	93	85 %
Koristiš e-poštu za raspravu o povjerljivim informacijama	26	24 %	83	76 %
Koristiš suvisli predmet – naslov (Subject)	29	27 %	80	73 %
Koristiš riječi HITNO i VAŽNO	40	37 %	69	63 %
Pišeš dugačke rečenice	57	52 %	52	48 %
Šalješ poruke s uvredljivim sadržajem	31	28 %	78	72 %
Prosljeđuješ hoax (lažna upozorenja)	11	10 %	98	90 %
Odgovaraš na spam	19	17 %	90	83 %

IV. POUČAVANJE O E BONTONU

Pravila lijepog ponašanja u virtualnom svijetu objavio je Hrvatski Telekom izdavanjem e-bontona koji je podjednako obuhvatio mišljenje sudionika komunikacije na društvenim mrežama i mišljenje eksperata s iskustvom promatranja najraznovrsnijih oblika komunikacije kao međuljudske interakcije [6]. Autori ovog priručnika su tako fejsbukovci, tviteraši, blogeri i forumaši, svakodnevni korisnici društvenih mreža koji su svojim prilozima pomogli osmisliti prvi hrvatski priručnik za lijepo ponašanje na društvenim mrežama kao što su Facebook i Twitter. Priručnik sadrži 130 odabranih pravila a dostupan je u digitalnom i tiskanom izdanju (slika 2).

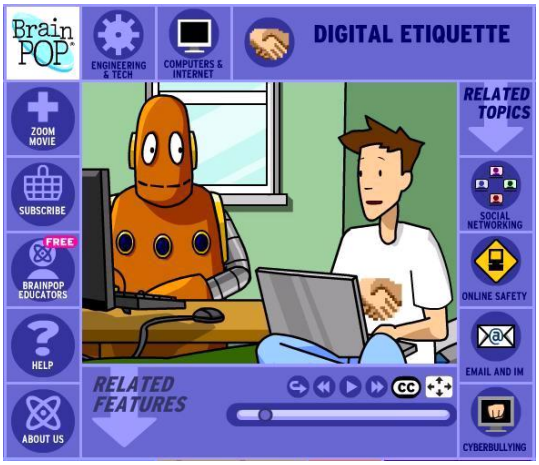


Slika 2. Pravila iz Prvog hrvatskog e-bontona čiji je izdavač Hrvatski Telekom

U primjeni pravila bontona kod učenika, pored nastavnika veliku ulogu imaju i roditelji. Premda se osjećaju zbunjeni kad je riječ o novim tehnologijama jer smatraju da djeca o tome znaju više nego oni, roditelji moraju biti svjesni da o životu znaju više od svoje djece i da moraju imati nadzor nad njihovim ponašanjem na internetu. Djetetovo pravo na privatnost ne smije spriječiti roditelje pri postavljanju granica, kao i naglašavanju primjene pravila online bontona.

Internet bonton učenici mogu proučiti i na CARNet-ovom portalu kroz pravila komunikacije pojedinaca i komunikacije u grupi [7].

Na portalu BrainPOP (obrazovne web stranice) učenici mogu na zanimljiv način, multimedijskim prikazom upoznati online pravila (slika 3), kao i online komunikaciju i zaštitu na internetu [8].



Slika 3. Edukativni sadržaji o digitalnom bontonu na portalu BrainPOP

Učenicima osnovne škole zanimljiv će biti alat Digitalna putovnica kojeg je pokrenuo Common Sense Media. To je interaktivna mrežna platforma za učenje, na engleskom jeziku, koja pomaže učiteljima u poučavanju i pripremi njihovih učenika kako bi postali dobri digitalni građani koji će se spremno uključiti u korištenje online i mobilnih tehnologija. Kroz pet interaktivnih modula učenici se upoznaju s osnovama sigurnosti i zaštite, osnovama lijepog ponašanja u digitalnom svijetu i razvijaju istraživačke vještine [9].

Dan sigurnijeg interneta, koji se obilježava svake godine u veljači, pruža mogućnost da se u školama pažnja posveti bontonu primjene digitalne tehnologije (slika 4). Kroz razne teme o sigurnosti i zaštiti djece na internetu, učenici upoznaju i kritički razmišljaju o svojim online pravima i odgovornostima. Analiziraju utjecaj novih medija na pojedinca i društvo, raspravljaju o prednostima i manama široke uporabe medija novih tehnologija u svakodnevnom životu. Pri tome upoznavaju računalnu etiku, dogovaraju pravila ponašanja na internetu te definiraju posljedice kršenja tih pravila.



Slika 4. Aktivnosti povodom obilježavanja Dana sigurnijeg interneta u Tehničkoj školi, Požega

V. ZAKLJUČAK

Novi komunikacijski kanali, moderne tehnologije, nedostatak vremena i težnja ka pojednostavljivanju i ubrzavanju protoka sve većeg broja informacija nužno utječu na prirodu komunikacije i pravila lijepog ponašanja. Svatko od nas ima pravo odabrati kako će se postaviti u primjeni digitalne tehnologije, ali okvir pravilnog online ponašanja ipak mora postojati. Proučavanjem e-bontona mnogi korisnici će možda napokon shvatiti svoje pogreške te ih uspješno početi ispravljati. Važno je ispoštovati e-bonton i unaprijediti online zajednicu, počevši od sitnih, ali bitnih pravila komuniciranja.

Osim djece i mladih, poslovni ljudi, zaposlenici u tvrtkama također moraju posebnu pažnju posvetiti digitalnom bontonu. Tri najvažnija razloga zbog kojih je važno paziti na e-poštu i digitalnu komunikaciju su profesionalizam (korištenje odgovarajućeg jezika utječe na sliku tvrtke), učinkovitost (jasno napisane poruke ubrzavaju poslovanje) i pravna zaštita (osviještenost o opasnostima e-pošte zaštitit će tvrtku od skupih sudskih parnica). Osobnom edukacijom i educiranjem zaposlenika možemo zaštititi sebe i svoju tvrtku od nezgodnih pitanja odgovornosti. Važno je osigurati e-bonton politiku koja se odnosi na sve zaposlenike. Sastavni dio etičkog kodeksa škole i tvrtki mora biti i provedba pravila online bontona. On se može definirati i u okviru školske Politike prihvatljivog korištenja interneta i računala u školi radi sigurnosti školskih računala, opreme i korisnika koji se moraju ponašati odgovorno i u skladu s etičkim načelima i u stvarnom i u virtualnom svijetu.

LITERATURA

[1] EduCentar, dostupno na http://www.educentar.net/

[2] Bonton 2 – Dobri maniri i pravila lijepog ponašanja (pristojnost i nove tehnologije), prilog Večernjeg lista, Zagreb, ožujak 2013. godine

[3] CSO: http://www.csoonline.com/read/060106/email_etiquette.html?source=csotechwatch

[4] Emailreplies.com: http://www.emailreplies.com/

[5] Bonton u elektroničkoj pošti (E-Mail Etiquette) i pristojnost na e-mailu u devet (i više) točaka, dostupno na http://www.sapmag.com.hr/show_article.php?id=361

[6] e-bonton, Pravila lijepog ponašanja na društvenim mrežama, Hrvatski Telekom d.d., 2011.

[7] Pravila ponašanja na Internetu – Internet bonton, dostupno na http://www.carnet.hr/pravila_ponasanja_na_internetu_internet_bonton

[8] Digital etiquette, dostupno na http://www.brainpop.com/technology/computersandinternet/digitaletiquette/

[9] Digital Passport, dostupno na https://www.digitalpassport.org/educator-registration

IKT, ljudski resursi te informacijska i računalna sigurnost u hrvatskom osnovnom školstvu

Mario Zovkić*, Tedo Vrbanec**
*Osnovna škola Barilović, Barilović & Osnovna škola "Katarine Zrinski" Krnjak, Krnjak, Hrvatska
**Učiteljski fakultet, Sveučilište u Zagrebu, Zagreb, Hrvatska
*mario.zovkic1@skole.hr, **tedo.vrbanec@gmail.com

Abstract - Nastavnici informatike u hrvatskim osnovnim školama često su "volonteri" u ulozi sistemskih administratora. Ta je uloga tehnički i tehnološki zahtjevna, a vremenski izuzetno zahtjevna. U školama se takav angažman uglavnom ne primjećuje. Sistemsku administraciju dodatno otežavaju sljedeće činjenice: (a) računala su često različitih hardverskih konfiguracija što onemogućava neke metode održavanja, (b) računala su u prosjeku stara osam godina pa imaju slabije hardverske mogućnosti (slabiji procesor te nedovoljna/minimalna količina radne memorije), zbog čega su spora. Računalne učionice u prosjeku sadrže 19 računala, što predstavlja problem ako su razredna odjeljenja većeg broja učenika, a razredna odjeljenja nije moguće dijeliti na manje grupe. U manjim školama nastavnici informatike često su pored računalne učionice zaduženi za održavanje i svih drugih računala u školi. Učenici se uobičajeno prijavljuju na računala u svojstvu administratora, sve se više raznih uređaja poput pametnih telefona, tableta i fableta spaja na školsku mrežu, a informatičari i ostali nastavnici zbog kroničnog nedostatka novaca za stručno usavršavanje nisu dovoljno educirani s novim prijetnjama i metodama zaštite u pogledu računalne i informacijske sigurnosti. Stoga edukacija o IKT sigurnosti ovisi o dobroj volji, slobodnom vremenu i entuzijazmu pojedinog nastavnika (informatike). Rad predstavlja pregled sigurnosti internet komunikacija te fizičke sigurnosti IKT u školskim računalnim učionicama.

Ključne riječi: e-sigurnost, IKT, računalna učionica.

I. UVOD

Računala u hrvatskim osnovnim školama krajem svake školske godine najčešće su puna virusa, pomiješanih/izbrisanih prečaca, promijenjenih ikona, izbrisanih programa ili programa koji javljaju čudne greške i/ili se jako dugo pokreću, nakrcana su raznim nepotrebnim programima (engl. bloatware), zbog svega toga su usporena i općenito se nalaze u jednom konfuznom stanju – reklo bi se da su "zrela za reinstalaciju" OS-a i svih potrebnih programa. Ukoliko se o sigurnosti ne vodi računa te ukoliko se u nj ne uloži potrebno vrijeme i trud, za godinu dana simptomi su isti.

Nastavnici informatike u hrvatskim osnovnim školama najčešće su i administratori školskih IT resursa. Prema CARNetu, [1] administrator je "u ime članice odgovoran za ispravnost i stabilnost funkcioniranja resursa koje je CARNet stavio na raspolaganje članici, kao i za opće tehničko i sigurnosno stanje računalno-komunikacijskog sustava članice spojene u CARNet mrežu." Realnost je da se s vremena na vrijeme, unatoč svim poduzetim mjerama zaštite, poneka računala zaraze malicioznim kodom koji se potom počne širiti - lokalnom mrežom, prijenosnim medijima (tipično USB memorijama) ili elektroničkom poštom. Bez obzira no to ima li škola pristup Internetu putem CARNeta ili putem nekog drugog ISP-a, isti može zbog pritužbi Abuse službi, biti prisiljen ukinuti pristup Internetu cijeloj ustanovi dok se problem ne riješi. Tko za taj propust i kome odgovara? Trenutno odgovara nastavnik informatike i to najprije ravnatelju, a onda potencijalno i CARNetu[1] koji može [2] "privremeno ili trajno uskratiti pravo korištenja mreže i njezinih servisa", odnosno "u slučaju težih prekršaja, povrede zakona, CARNet će prekršitelja prijaviti nadležnim organima".

U svakodnevnom radu, nastavnicima informatike se (prirodno) nameće zahtjev da njihova sredstva rada pomoću kojih održavaju nastavu i ona kojima se učenici služe tijekom nastave, budu upotrebljiva, funkcionalna, bez malicioznog koda i grešaka u radu. Kad god postoji problem s računalima u informatičkoj učionici, nepisano je pravilo da ga nastavnik informatike sam riješi. U protivnom je ostao bez dijela sredstava za rad (računala) zbog čega trpi nastava. To je to više izraženo što je veća disproporcija broja učenika i računala. Ne riješi li nastavnik informatike problem, (zbog nedostatka vremena, neznanja, dobre volje ili jednostavno zato jer to nije ni dužan raditi), mora čekati dozvolu ravnatelja odnosno raspoloživost financijskih sredstava škole kako bi se računalo poslalo u servis, te čekati na (uvijek skuplji negoli je potrebno) popravak. Što je škola po broju učenika manja, to je problem veći, jer u Republici Hrvatskoj škole primaju financijska sredstva za materijalne troškove prema broju učenika [3]. Nadalje, kolektiv očekuje, a ponekad čak i zahtijeva od nastavnika informatike, da rješava računalne probleme cijele ustanove. Ponekad ravnatelj takvo zaduženje i formalno dodijeli nastavniku informatike. Naizgled je logično da nastavnik informatike održava računala u školi, ali u suštini je nepravedno ukoliko ga se proporcionalno ne rastereti nastave. Ilustrirajmo to primjerom: želi li nastavnik informatike naučiti ili usavršiti strani jezik, po istoj bi ga logici tome trebao besplatno naučiti nastavnik stranog jezika; a za usavršavanje matematike, besplatno bi ga trebao podučiti nastavnik matematike, itd. Dakle, da li bi računalo drugih nastavnika trebao besplatno servisirati nastavnik informatike? Svi ti popravci i održavanje računala (uključujući i instalaciju sigurnosnih zakrpi, ažuriranja programa i slično), rade se izvan nastave, a svi ostali poslovi koje nastavnik ima u zaduženju odgađaju se za kasnije. Ipak, u praksi nastavnici informatike obično

[1] Gotovo sve škole u Hrvatskoj imaju besplatan pristup internetu preko CARNeta, a troškove pokriva MZOŠ.

rado educiraju svoje kolege, pomažu im savjetima i rješavaju njihove IT probleme, ali to je stvar prijateljstva ili kolegijalnosti te bi trebalo biti dobrovoljno.

U sklopu obrazovnog sustava, malo se toga može postići. Nastavnik informatike je u prošlosti dobivao uz satnicu i funkciju "voditelj informatičke učionice", pod uvjetom da je ravnatelj osnovne škole to zatražio od MZOS-a te dobio dozvolu. Time mu se oslobodilo dva do četiri sata nastave na ime održavanja računala. Posljednjim važećim kolektivnim ugovorom za zaposlenike u osnovnoškolskim ustanovama (koji je 27. prosinca 2013. odlukom Vlade Republike Hrvatske otkazan, a trenutno teče otkazni rok od tri mjeseca [4]) naziv "voditelj informatičke učionice" je izbačen. Škola koja do sada nije imala voditelja informatičke učionice teško da će je dobiti jer je funkcija vezana za određenu školu koja je tu dozvolu po starom kolektivnom ugovoru ishodila. Male škole (prema osobnom iskustvu autora) u pravilu ne dobiju odobrenje da se nastavniku informatike do popunjenja satnice odobri nekoliko sati tjedno za održavanje računala. U srednjim školama RH zakonski je uređeno zapošljavanje osoba kojima je jedini zadatak brinuti o ispravnosti, ažurnosti i sigurnosti računala i računalnih mreža [5]. U susjednoj Sloveniji "Pravilnikom o normativih in standardih za izvajanje programa osnovne šole" uvedeno je 2007. godine radno mjesto [6] "računalnikar - organizator informacijskih dejavnosti" kojem se priznaje određen dio sati u redovnu satnicu s obzirom na broj odjeljenja u školi, pa ako primjerice škola ima do 8 razrednih odjeljenja onda računalnikar dobije 0,20 udjela radnog mjesta, a ako škola ima 46 odjeljenja onda je udio radnog mjesta 1, odnosno puno radno vrijeme prema članku 11 [6].

U Njemačkoj se u srednjim gimnazijskim školama (5-9 razred) predaje informatika, a za održavanje se zapošljava druga osoba ili se informatičaru to ubraja u satnicu. U SAD-u, obrazovne institucije u cijeloj obrazovnoj vertikali (osnovne škole, srednje, fakulteti) zapošljavaju tzv. "Computer Technician" koji su zapravo dio radnog mjesta pod imenom "Chief Operating Officer"[2]. Svaki okrug (district) samostalno plaća osobu, a institucija samostalno određuje da li je ta pozicija striktno vezana za IT ili npr. mrežu koja ne uključuje posao poučavanja. "Building technicians" kako ih zovu u SAD-u prosječno provedu oko 1.5 dana svaki tjedan u manjim školama, dok u većim školama svaki tjedan provedu oko 4-5 dana. [7]

Primjeri na koje bi se u Hrvatskoj mogli ugledati jesu Makedonija i Indonezija (Sumatra) gdje postoje tzv. "Student Support Technician Clubs" (SSTCs) u kojima se učenici/studenti organiziraju u grupu/sekciju koja je zadužena za održavanje računala. [8] Skupine se sastoje od najuspješnijih učenika/studenata koji i poučavaju mlađe generacije o brizi za računalima.

U nastavku donosimo pregled osnovnih tema računalne sigurnosti u informatičkoj učionici: fizičke sigurnosti i sigurnosti komunikacije uz koje ćemo predložiti neka poželjna pravila korištenja računala.

[2] Drugi nazivi koji se još susreću su: School Technology Integration Coordinator, Technology Coordinator, Director of Educational Technology.

II. INSTITUCINALNI OKVIR ZA JAČANJE E-SIGURNOSTI U RH

U Hrvatskoj se pitanjima informacijske, komunikacijske, računalne, podatkovne i ostalih oblika sigurnosti te privatnošću i povjerljivošću uvelike bavi CARNet, posebno sa svojim projektom "Informacijsko-komunikacijska tehnologija u obrazovanju – ICT Edu" koji se provodi kroz 3 modula, po jedan za svaku školsku godinu. On ima za cilj [9] "upoznati sve učitelje, nastavnike, stručne suradnike i ravnatelje osnovnih i srednjih škola u Republici Hrvatskoj s brojnim sadržajima i uslugama koje im pruža MZOS i CARNet, te ih o nekim sadržajima detaljnije educirati". Sa stajališta ICT sigurnosti, projekt sadrži teme iz "Sigurnosti na internetu" gdje su dani savjeti za roditelje i učenike o štetnim programima te problemima komunikacije na Internetu.

MUP, Ministarstvo uprave, Hrvatska akademska i istraživačka mreža - CARNet, Tehničko veleučilište u Zagrebu, Udruga „Suradnici u učenju“, Poliklinika za zaštitu djece Grada Zagreba, Agencija za zaštitu osobnih podataka i drugi pridružujući partneri Ministarstva znanosti, obrazovanja i športa, Ministarstvo socijalne politike i mladih, Agencija za odgoj i obrazovanje, Učenički dom Hrvatskoga Radiše i Hrvatska agencija za poštu i elektroničke komunikacije su 05. lipnja 2012. godine potpisali Sporazum o strateškom partnerstvu na uspostavi Centra za sigurniji Internet [10] kojem je cilj zajedničko promicanje računalne i informacijske sigurnosti u RH, te se "zalagati za institucionalizirano uvođenje računalne i informacijske sigurnosti u obrazovni sustav, npr. kao dio nastavnog programa informatike za osnovne i srednje škole." Taj centar omogućava prijave eventualnih povreda on-line sigurnosti djece (svih nezakonitih, štetnih i opasnih sadržaja počinjenih na štetu djece i mladih) te je predstavljen 28.rujna 2013. u Zagrebu. Nastao je kao dio IPA twinning projekta pod punim nazivom „Jačanje kapaciteta u području suzbijanja seksualnog iskorištavanja i seksualnog zlostavljanja djece te pružanje pomoći policije ranjivim žrtvama kriminaliteta.“ Takvi centri za sigurniji Internet (engl. Safer Internet Center) su trenutno smješteni u svih 28 zemalja EU te u Islandu, Norveškoj i Rusiji [11] te su dio Insafe-INHOPE joint mreže.

U RH je CARNet 19. srpnja 1996. osnovao CCERT (CARNet Computer Emergency Response Team) - nacionalno središte za sigurnost računalnih mreža. Dobivanjem suglasnosti Ureda vijeća za nacionalnu sigurnost, u CARNetu je 13. svibnja 2009. počeo s radom Odjel za Nacionalni CERT osnovan u skladu sa Zakonom o informacijskoj sigurnosti RH i prema tom zakonu jedna od njegovih zadaća je [12] "obrada incidenata na Internetu, tj. očuvanje informacijske sigurnosti u RH", a misija "prevencija i zaštita od računalnih ugroza sigurnosti javnih informacijskih sustava u Republici Hrvatskoj." Nacionalni CERT preuzeo je ulogu i nasljednik je CARNet CERT-a.

Europska mreža centara za sigurniji Internet - Insafe i Međunarodna zajednica online linija za pomoć – INHOPE, obilježavaju Safer Internet Day - SID. SID [11]

organizira "Insafe" u veljači svake godine s ciljem promicanja sigurnije i odgovornije uporabe online tehnologije i mobilnih telefona, posebice među djecom i mladima diljem svijeta. Inicijativa je pokrenuta u okviru Programa za sigurniji Internet Europske komisije, a jednom godišnje poziva na uključivanje u aktivnosti za stvaranje sigurnijeg i boljeg Interneta za djecu i mlade. SID je prvi puta u svijetu obilježen 2004. g., u RH 2009-te, a svake godine se provode razne akivnosti na određenu temu.

Najnoviji projekt pod nazivom "Sigurnost djece na Internetu" ima za cilj ući u Hrvatski kvalifikacijski okvir, tj. pokušaj je da se MZOS-u dokaže da teme e-sigurnosti uđu u kurikulum predmeta (koje će ministarstvo tek izraditi za svaki pojedini predmet). Nositelj gore navedenog projekta je OŠ Veliki Bukovec, a partneri osnovne škole: OŠ Popovača, OŠ Mato Lovrak iz Nove Gradiške, OŠ Gripe iz Splita i OŠ Mladost iz Osijeka. Projekt će trajati od 19. kolovoza 2013. do 19. prosinca 2014. uz sudjelovanje učenika, učitelja, roditelja te lokalne i šire zajednice. [13]. U sklopu tog projekta učitelji iz nabrojanih pet osnovnih škola iz Hrvatske izradit će i zatim primijeniti školski kurikulum o sigurnosti djece na internetu u sklopu Hrvatskog kvalifikacijskog okvira kojim će direktno biti uključeno 200 učitelja, 220 učenika te 220 roditelja. Bit će izrađeni i priručnici, radni materijali, multimedijalni sadržaji i interaktivne igre za pomoć u edukaciji djece, roditelja i učitelja te prevenciji elektroničkog nasilja. Školski kurikulum bit će pripremljen za sve razrede osnovne škole, odnosno organiziran u četiri grupe prema uzrastu učenika. Svi pripremljeni sadržaji bit će besplatni i dostupni za korištenje svim školama u Hrvatskoj. Projekt financira IPA program Europske unije s 2,5 milijuna kuna. [13]

III. FIZIČKA SIGURNOST ICT-A U ŠKOLI

"Osnova fizičke sigurnosti je zaštita fizičkog dijela informatičke infrastrukture, zgrade u kojoj je ona smještena, kao i medija za pohranu podataka te komunikacijske opreme. Mjere fizičke sigurnosti obuhvaćaju sve obrambene mjere poduzete u svrhu zaštite računalne infrastrukture od prirodnih nepogoda, problema u okolini, nezgoda i namjernih oštećenja". [14] "Značaj fizičke zaštite sustava je distribucijom opreme i podataka te uključivanjem u Internet postao još važnijom, ali i znatno složenijom." [15]

Nastavnik informatike mora utvrditi tko ima izravan fizički pristup računalima, da li je to neophodno te kako zaštititi računala od oštećenja ili krađe. Oprema poput monitora ili kućišta može se pričvrstiti na radni stol s raznim kablovima ukoliko postoje utori na samom sklopovlju, običnim i jeftinim jednokratnim plastičnim vezicama (engl. "tie straps") ili se pričvršćuju vijcima za radni stol (LCD i LED monitori). Manje je predmete, poput miša i tipkovnice, teže osigurati od krađe, no u tom slučaju poželjno je na takve stvari staviti inventarni broj ili neko drugo obilježje (npr. logo/natpis ustanove). Na taj se način obeshrabruju lopovi, a ujedno je situacija uvijek spremna za popis inventure.

Za neophodnu razinu fizičke sigurnosti, preporučljivo je zaključati komunikacijske ormariće i prostorije u kojima se nalazi mrežna oprema i poslužitelji. Ukoliko u informatičkoj učionici nema video nadzora, kada se ona koristi uvijek bi trebao biti nazočan netko od zaposlenika škole, a kada se računala ne koriste prostoriju bi trebalo zaključavati.

A. Videonadzor

Video nadzor učionice, uz jasne obavijesti o njegovu postojanju, vrlo je poželjna zaštitna mjera koja obeshrabruje potencijalne lopove i vandale. Računalo skromnih performansi i web kamera [16] dovoljni su za izradu jednostavnog video nadzora "iako je uporaba video-nadzora u Republici Hrvatskoj u samom začetku i nedovoljno normativno uređena" obzirom [17] "na ograničenu uporabu tehničkih snimki kao dokaza u postupcima (kazneni i prekršajni)". Zakonom o kaznenom postupku [18] donesenim 2008. godine započela je primjena načela pretežitog interesa kod kojeg postoji mogućnost korištenja tehničkih snimki te drugih dokaza pribavljenih na nezakoniti način ako interes pravde i pravičnosti preteže nad interesom zaštite dobara. "Što se tiče valjanosti tehničkih snimki u prekršajnom postupku, Prekršajni zakon propisuje da su bilješke i tehničke snimke i formalno dokazi u prekršajnom postupku ukoliko ih je ovlaštena službena osoba tijela državne uprave sačinila u okvirima svojih nadležnosti za provođenje nadzora provedenog neposrednim opažanjem. U tim situacijama takve snimke mogu poslužiti kao dokaz u prekršajnom postupku, ali i tada sud može provjeravati njihovu vjerodostojnost te slobodno ocjenjivati njihovu dokaznu vrijednost." [19]

Što se tiče videonadzora u školama, postavljanje videokamera predstavlja važnu odluku za položaj radnika, jer se radi o mjeri u svezi sa sigurnošću na radu, o kojoj se poslodavac mora savjetovati s radničkim vijećem što je propisano [20] u odredbi čl. 149. st. 1. Zakona o radu. Prema Agenciji za zaštitu osobnih podataka, po pitanju videonazdora u školama [21], a na temelju odredbi Zakona o zaštiti osobnih podataka [22] članaka 2. stavak 1. točka 1. razvidno je da video snimka dobivena video nadzornom kamerom predstavlja osobni podatak fizičke osobe – ispitanika. Člankom 9. [22] propisana je obveza informiranja ispitanika prije prikupljanja bilo kojih osobnih podataka. Dakle škole bi trebale "adekvatnim internim aktom regulirati uporabu istog (svrha, opseg tj. zahvat podataka koji se prikupljaju, način i vrijeme čuvanja te uporaba snimljenih podataka kao i zaštita prava ispitanika) a s činjenicom kojeg postavljanja bi na odgovarajući način pravovremeno trebalo upoznati učenike (roditelje) i zaposlenike." [21] Nadalje, "potrebno je ukazati da predmetni video nadzor (postavljen u svrhu sigurnosti) treba pokrivati samo zajedničke prostore škole (hodnike, ulazni prostor i sl.) kao i moguće dvorište škole (bez zahvaćanja u okolni javni prostor), dok nije prihvatljivo da video nadzorom budu pokrivene učionice, svlačionice i toaletni prostori. Također, kod uvođenja (postavljanja) video nadzora potrebno je uočljivo i nedvosmisleno označiti (slikom i tekstom) da je odnosni prostor pod video nadzorom, koja informacija mora biti pružena svim trećim osobama koje dolaze u školu, za

ostvarivanje spoznaje da ih se snima i na taj način prikupljaju njihovi osobni podaci." [22]

Škola može angažirati pravnu i fizičku osobu registriranu za obavljanje poslova tehničke zaštite, na način i pod uvjetima utvrđenim [23] Zakonom o privatnoj zaštiti koja će izraditi projektnu dokumentaciju prema Pravilniku o uvjetima i načinu provedbe tehničke zaštite [24] koji određuje tko ga smije postaviti i određuje uvjete, poput elaborata, održavanja. Na ovaj zadnji način je snimanje legalno.

B. Sigurnosne mjere za korisnike

Prema M. Bači [14] "najveća ranjivost svakog računalnog sustava i najveća prijetnja računalnom sustavu su ljudi koji s njim imaju vezu, kroz svakodnevni rad i kroz povremeno održavanje." Prema K. Rhodesu [25] zaposlenici u organizaciji imaju ključnu ulogu u prevenciji i detektiranju sigurnosnih incidenta. Dok su sistem administratori odgovorni za sigurnu konfiguraciju vatrozida i poslužitlja, korisnici su odgovorni za vježbanje sigurnosnih protumjera poput odabira i zaštite odgovarajućih lozinki. Stoga, za efektivnu sigurnost, korisnici moraju napraviti svjesne odluke da bi se pridržavali organizacijske sigurnosne politike i da bi prihvatili računalno sigurnosno ponašanje.

Nastavnik informatike često puta je jedina osoba koja u školi ima (dostatno?) znanje o računalima i računalnoj sigurnosti. Učitelji uglavnom koriste USB memorijske štapiće često zaražene virusima koji se šire na računalo u zbornici ili preko mreže, često jedni drugima "posuđuju" osobna korisnička imena i lozinke za prijavu na službene servise (npr. E-matica, CMS školskog web sjedišta), koje koriste na javnom računalu za koje se ne može uvijek pretpostaviti da je čisto od virusa i drugih malicioznih programa.

U novije vrijeme su učestali slučajevi da neki učenici pomoću aplikacije za udaljeni pristup računalu (npr. TeamViewer), u odsutnosti učitelja instaliraju aplikaciju, a kasnije se pomoću pametnog telefona spajaju na računalo te ga kontroliraju. Takva je nepodopština kažnjiva prema člancima 226. i 227. Kaznenog zakona [26] na što je 4.10.2013. reagiralo i ravnateljstvo policije RH poslavši dopis MZOS-u, a potonji proslijedio dopis svim osnovnim i srednjim školama te učeničkim domovima [27].

CARNet je temeljem dokumenta CDA0035 [2] točke 1. stavka 1. propisao: "Obveza je svake ustanove članice propisati sigurnosnu politiku na svojoj ustanovi...". Ustanova članica CARNeta ima obavezu izraditi sigurnosnu politiku te s njome upoznati svoje djelatnike i korisnike o dužnosti poznavanja i primjenjivanja iste. Do danas tek je nekolicina osnovnih škola izradila sigurnosnu politiku iako je trebala biti donesena još u rujnu 2006-te, odnosno šest mjeseci od dana spajanja u CARNet mrežu [2]. Provedba sigurnosnih mjera među korisnicima koji teško prihvaćaju promjene i ne shvaćaju opasnosti korištenja ICT-a teško je izvediva, no ovdje je riječ o visoko obrazovanim korisnicima i možemo reći da je žalosno što je tako.

IV. SIGURNOST INTERNET KOMUNIKACIJA U ŠKOLI

Internet je medij putem kojega korisnici žele komunicirati s osobama kojima vjeruju kao i onima kojima ne vjeruju, a isto vrijedi i za uređaje posredstvom kojih bi se ta komunikacija trebala odvijati. Na žalost ne postoji pouzdani način razlučivanja jednih od drugih. Kriptografija u koju se polagalo puno nade, nije uspjela iznaći pouzdane metode zaštite računala i korisnika od mreže u kojoj je četvrtina računala pod utjecajem bot mreža, a polovina e-maila neželjena pošta [28].

U školama uobičajen problem predstavlja nedozvoljeno posjećivanje društvenih mreža, stranica sumnjivog (ilegalnog) sadržaja, pa čak i stranica za odrasle. Doduše ako su računala spojena na Internet putem CARNetove mreže, ona imaju djelomičnu "zaštitu" od neželjenih i neprimjerenih sadržaja - "content filtering" baziran na CA eTrust Secure Content Manager content filter sustavu. O učinkovitosti ovog filtera bi se dalo raspravljati, jer ponekad blokira i sasvim legalne stranice i poveznice za preuzimanje legalnih programa, no zaštita se neprestano razvija i poboljšava. Još jedno rješenje za osnovno filtriranje sadržaja na internetu je projekt OpenDNS koji blokira sadržaje na DNS razini, a po potrebi se razina filtriranja može detaljnije konfigurirati. Konfiguriranje se vrši unutar postavki korisničkog računa, a osim već unaprijed definiranih kategorija po kojima se vrši filtriranje, moguće je dodavati svoje unose u listu dozvoljenih i zabranjenih domena.

Unutar škole bi bilo poželjno imati dvije (pod)mreže: jednu za učenike, a drugu za djelatnike. Na taj način rasterećujemo mrežni promet, a ako dođe do nekog sigurnosnog incidenta ili malicioznog programa koji se počne širiti jednom mrežom, druga će podmreža ostati netaknuta ukoliko mreže međusobno štitimo i filtriramo promet između njih. S 2011/2012. nastavnom godinom CARNet je krenuo u praktičnu realizaciju projekta e-dnevnik te preporuča prilikom planiranja bežičnih mreža dati učenicima mogućnost spajanja na nju. Projekt e-dnevnik je web aplikacija za vođenje razredne knjige u elektroničkom obliku, koji je iste godine uveden kao pilot projekt u tri srednje škole: XV. Gimnaziji u Zagrebu, Gimnaziji Požega i u Medicinskoj školi Ante Kuzmanića u Zadru. Dakle službena evidencija se vodi u aplikaciji, a nastavnici u razrede ulaze s tablet računalima putem kojih pristupaju aplikaciji. "E-dnevnik je izradio CARNet te ona ima sve funkcionalnosti postojeće razredne knjige, uz dodatne funkcionalnosti koje omogućava uporaba informacijsko-komunikacijskih tehnologija (ICT)." [29] Od školske godine 2013./2014. se za njegovo korištenje prijavilo preko 250 škola uključujući područne škole.

U Europskoj uniji se nastoji izglasati novi zakon pod imenom "Pravo na zaborav informacija". [30] Ukoliko to osoba zatraži, neugodni, netočni ili jednostavno osobni podaci morati će biti izbrisani s Interneta ili bilo kakvih baza podataka kompanija. Europska Komisija smatra da se pojedincu mora omogućiti da zatraži brisanje podataka o njoj/njemu, ukoliko ne postoje pravne prepreke za to. Također, pojedinac ima pravo da se njegovi podaci ne skladište u podatkovnim centrima te ne koriste za obradu, ukoliko to ne želi [30]. Nedavno je i bivši izvršni direktor Googlea Eric Schmidt izjavio: "Ovaj manjak gumba za brisanje na Internetu je u biti važno pitanje." [31] "Pravo

na zaborav" još nije doneseno kao dio EU prava niti je prošlo ili stavljeno u praksu. Ovaj zakon (ako bude donesen) će u praksi biti teško provesti. "Jednom kad se osobna informacija objavi, nemoguće ju je tehničkim mjerama spriječiti, ili čak promatrati/slijediti, stvaranje neautoriziranih kopija te informacije. U otvorenom sistemu kao što je Internet, pravo na zaborav se ne može ojačati samo tehničkim mjerama. Ojačanje mora polaziti na kombinaciji tehničkim i međunarodnih pravnih osiguranja." [32]

V. ZAKLJUČAK

Nastavnici informatike u RH su de facto "sistem administratori u malom" koji svoj posao rade uglavnom volonterski, ali ponegdje/ponekad i prisilno (poštivanjem odluka ravnatelja), snose pravnu i moralnu odgovornost kako za ispravnost mreže tako i za računala. Njihova motivacija je u njihovom pozivu i osobnoj satisfakciji u onome što rade. Gotovo uvijek sami rješavaju različite probleme vezane za informatiku iz učionice ili zbornice. Educiraju svoje učenike i kolege i ukupno gledajući su u neravnopravnom položaju u odnosu na svoje kolege iz drugih predmeta. Na žalost tako će i ostati neko vrijeme u RH barem što se tiče situacije u osnovnoj školi. Jedan od mogućih smjerova u budućnosti (pod pretpostavkom još većih problema s financiranjem škola/opreme) jest da se problem održavanja prebaci na učenike, tako da učenici donose sami svoja prijenosna računala, tablete, fablete i pametne telefone pa ih i održavaju. (U Americi škole su već počele primjenjivati takav način rada poznat pod skraćenicom (BYOD - bring your own device) gdje učenici donose sami svoje uređaje no kako ne bi bilo još više problema (npr. zaražena računala) trebalo bi prije toga svakako educirati i učenike o sigurnosti računala. Jedan od preduvjeta takvom načinu rada jest pokrivenost Internetom u cijeloj školi, pa posebnu pažnju treba posvetiti konfiguraciji mreže. Slijedeće riješenje koje autori predlažu jest da se na razini općina ili županija odaberu (kvalifikacijskim ispitom ili formalnom potvrdom kvalificiranosti) osobe koje bi bile svojevrsni helpdesk za računala u školama. Nastavnici bi mogli preko on line obrazaca ispuniti kratki opis problema, a hepldesk bi uz pomoć telefona i/ili videokonferencije i spajanja na nastavnikovo računalo rješavao problem (ako kvar nije fizičke prirode ili ne zahtijeva intervenciju "in vivo"). Takvih osoba vjerojatno ne bi ni trebalo biti puno. U budućnosti bi se mogla provesti reorganizacija posla da osobama kojima nedostaju nastavni sati u školama dobiju dopunu do pune satnice na način da satnicu dopune radom u helpdesku. Moguća druga rješenja su zaključavanje postavki računala, virtualizacija ili brisanje promjena na računalu (npr. VHD ili sandbox tehnologijom).

REFERENCES

[1] Carnet, "Administratori škola," 02-Jan-2013. [Online]. Available: http://www.carnet.hr/ustanove_clanice/administratori_skola. [Accessed: 03-Dec-2012].

[2] CARNet, "Odluka o prihvatljivom korištenju CARNet mreže," 15-Jun-2012. [Online]. Available: ftp://ftp.carnet.hr/CARNet/docs/rules/CDA0035.pdf. [Accessed: 14-Mar-2013].

[3] Narodne novine, "Odluka o kriterijima i mjerilima za utvrđivanje bilančnih prava za financiranje minimalnog financijskog standarda javnih potreba osnovnog školstva u 2012. godini," NN 52/12. [Online]. Available: http://narodne-novine.nn.hr/clanci/sluzbeni/2012_05_52_1276.html. [Accessed: 23-Dec-2013].

[4] Vlada RH, "Odluka o otkazivanju Kolektivnog ugovora za zaposlenike u osnovnoškolskim ustanovama." [Online]. Available: https://www.vlada.hr/hr/content/download/282525/4165491/file/134.%20-%2012.c.pdf. [Accessed: 30-Dec-2013].

[5] Narodne novine, "Pravilnik o stručnoj spremi i pedagoško-psihološkom obrazovanju nastavnika u srednjem školstvu," 1996. [Online]. Available: http://narodne-novine.nn.hr/clanci/sluzbeni/263720.html. [Accessed: 23-Oct-2013].

[6] Ministarstvo za šolstvo in šport, "Pravilnik o normativih in standardih za izvajanje programa osnovne šole," 09-Jan-2007. [Online]. Available: http://www.mizks.gov.si/fileadmin/mizks.gov.si/pageuploads/zakonodaja/pdf/OS/Pravilnik_normativi_OS_7_5_07.pdf. [Accessed: 20-Mar-2013].

[7] Doug Johnson (Director of Libraries and Technology for the Mankato (MN) Public Schools), "School administrators", poruka elektroničke pošte, 18-Sep-2013.

[8] Eric Rusten, Josh Woodard, "Enabling, Accelerating & Sustaining Quality Education," 2011. [Online]. Available: http://siteresources.worldbank.org/EDUCATION/Resources/278200-1099079877269/547664-1099079956815/547670-1128086743752/BBL_4_4_2011_AED.pdf. [Accessed: 12-Mar-2013].

[9] CARNet, "Ict edu," 20-May-2013. [Online]. Available: http://www.carnet.hr/ictedu/o_projektu. [Accessed: 28-Mar-2013].

[10] "Centar za sigurniji Internet," 2013. [Online]. Available: http://www.sigurnijiinternet.hr/. [Accessed: 28-Jun-2013].

[11] Insafe, "Safer Internet Day," 22-Oct-2013. [Online]. Available: http://www.saferinternet.org/safer-internet-day. [Accessed: 30-Jun-2013].

[12] "CERT," 2013. [Online]. Available: http://www.cert.hr/. [Accessed: 28-Jun-2013].

[13] Darko Rakić, "Projekt 'Sigurnost djece na internetu' za edukaciju i prevenciju elektroničkog nasilja," Sigurnih pet za sigurniji net, 25-Oct-2013. [Online]. Available: http://www.petzanet.hr/Blog/ArticleID/7. [Accessed: 12-Jan-2014].

[14] Miroslav Bača, Uvod u računalnu sigurnost. Narodne novine, 2004.

[15] Ksenija Klasić, "Zaštita informacijskih sustava u poslovnoj praksi," Sigurnost, vol. 49., no. 1, 2007.

[16] Neil Mohr, "How to set up a webcam security system," PC Plus, 17-Nov-2012. [Online]. Available: http://www.techradar.com/news/computing/how-to-set-up-a-webcam-security-system-1112863. [Accessed: 10-Mar-2013].

[17] Franjo Podhraški, Stjepan Tršinski, Krešo Kancir, "Prijedlog unapređenja rada policije uvođenjem videonadzora javnih prostora," Policija i sigurnost, vol. 17., no. 3–4., 2009.

[18] Narodne novine, "Zakon o kaznenom postupku," 01-Jan-2013. [Online]. Available: http://www.zakon.hr/z/174/Zakon-o-kaznenom-postupku. [Accessed: 28-Jun-2013].

[19] Vlada RH, "Prijedlog zakona o prekršajima protiv javnog reda i mira," listopad-2012. [Online]. Available: www.sabor.hr/fgs.axd?id=21935. [Accessed: 21-May-2013].

[20] Narodne novine, "Zakon o radu," 2013. [Online]. Available: http://www.zakon.hr/z/307/. [Accessed: 21-May-2013].

[21] Agencija za zaštitu osobnih podataka, "Videonadzor u školi," 2013. [Online]. Available: http://www.azop.hr/news.aspx?newsID=229&pageID=25. [Accessed: 21-May-2013].

[22] Narodne novine, "Zakon o zaštiti osobnih podataka," NN, 2012. [Online]. Available: http://www.zakon.hr/z/220/Zakon-o-za%C5%A1titi-osobnih-podataka.

[23] Narodne novine, "Zakon o privatnoj zaštiti," 2010. [Online]. Available: http://www.zakon.hr/z/291/Zakon-o-privatnoj-za%C5%A1titi. [Accessed: 28-Jun-2013].

[24] Ministarstvo unutarnjih poslova (MUP), "Pravilnik o uvjetima i načinu provedbe tehničke zaštite," 05-Sep-2013. [Online]. Available: http://www.mup.hr/UserDocsImages/SAVJETOVANJE/2013/zozni/PRAVILNIK_O_TEH_ZASTITI_.pdf. [Accessed: 30-Dec-2013].

[25] K. Rhodes, "Operations security awareness: the mind has no firewall," Computer Security Journal 18 (3), 2001.

[26] Narodne novine, "Kazneni zakon," 01-Jan-2013. [Online]. Available: http://www.zakon.hr/z/98/. [Accessed: 09-Oct-2013].

[27] Ravnateljstvo policije, "Neovlašteni pristup računalnim sustavima," 30-Sep-2013. [Online]. Available: http://os-barilovic.skole.hr/upload/os-barilovic/newsattach/793/neovlasteni_pristup_racunalnim_sustavima.pdf. [Accessed: 14-Dec-2013].

[28] Diffie W., "Information security: 50 years behind 50 years ahead," Communications of the ACM, vol. 51, no. 1, pp. 55–57, 2008.

[29] CARNet, "E-dnevnik," 19-Mar-2013. [Online]. Available: http://www.carnet.hr/e-dnevnik. [Accessed: 30-Jun-2013].

[30] Viviane Reding, Vice President of the European Commision, "The EU Data Protection Reform 2012: Making Europe the Standard Setter for Modern Data Protection Rules in the Digital Age 5," 22-Jan-2012. [Online]. Available: http://europa.eu/rapid/pressReleasesAction.do?reference=SPEECH/12/26&format=PDF. [Accessed: 28-Mar-2013].

[31] Leonard N. Stern School of Business, "Eric Schmidt intervju," 2013. [Online]. Available: http://www.stern.nyu.edu/experience-stern/news-events/google-schmidt-cohen. [Accessed: 28-Jun-2013].

[32] Druschel P., Backes M., Tirtea R., "The right to be forgoten-between expecations and practice," European Network and Information Security Agency (ENISA), 18-Oct-2011. [Online]. Available: http://www.enisa.europa.eu/activities/identity-and-trust/library/deliverables/the-right-to-be-forgotten/at_download/fullReport. [Accessed: 12-Sep-2013].

Učinkovitost elektroničkog učenja u nastavi

Martina Sertić
Industrijsko-obrtnička škola Slavonski Brod, Republika Hrvatska
Studentica PDS Informacijskih i komunikacijskih znanosti Filozofskog fakulteta u Zagrebu, Republika Hrvatska
sertic_79@yahoo.com

Sažetak - Učinkovitost elektroničkog učenja u obrazovanju možemo prikazati korištenjem modernih tehnologija današnjice. Upotreba ovih modernih sredstava u nastavi vodi nas do nekoliko pitanja na koja trebamo odgovoriti. Neka od pitanja su: povećava li upotreba moderne tehnologije učenikovo znanje i vještine, i ako da, na kojoj razini i koliko? Može se pretpostaviti da će u budućnosti razne računalne aplikacije i računalni sadržaji stvoriti novu generaciju učenika. Za te nove generacije tehnologija će postati neophodan sadržaj i sredstvo za učenje, zabavu i komunikaciju.

I. UVOD

Ovim radom pokušat ću prikazati i predstaviti primjenu Bloomove taksonomije u nastavi. Bloom je podijelio odgojno-obrazovne aktivnosti na tri područja: kognitivno (znanje), psihomotorno (vještine) i afektivno (stavovi). U radu ću se ponajviše orijentirati na kognitivno područje i primjenom manjeg istraživanja korištenjem eksperimenata pokušati dokazati kako korištenje digitalnih nastavnih sredstava (računala, projektora, interneta) povećava razinu znanja kod učenika. Eksperiment ću provesti pod nastavom filozofije kroz kontrolnu i eksperimentalnu skupinu. Uzorak će biti učenici dva četvrta razreda, oba spola. Navest ću i ljestvice mjerenja znanja, standardiziranu prema hrvatskom sustavu ocjenjivanja na pet ocjena.

II. PROCJENA ZNANJA

Procjena znanja zapravo je sustavan proces u kojem nastavnik prikuplja podatke, analizira ih i tumači kako bi odredio u kojoj su mjeri učenici savladali obrazovne ciljeve.

Procjenjivanje znanja uključuje nekoliko postupaka:

- opažanje i s njim povezane tehnike (npr. usmeno ispitivanje) ,
- mjerenje, odnosi se na proces određivanja brojčane mjere nečijeg postignuća ili osobine. (Marina je riješila 15 od 20 zadataka),
- vrednovanje – davanje vrijednosti procijenjenom znanju, tj.ocjenjivanje.

Izmjereno se znanje izražava ocjenom koja ima određenu poruku. [6]

U našem sustavu to su ocjene:

- 1 (nedovoljan),
- 2 (dovoljan),
- 3 (dobar),
- 4 (vrlo dobar),
- 5 (odličan).

U dolje prikazanim tablicama može se vidjeti grafički prikaz odnosa znanja i ocjena kod nastavnika. Tablica I. prikazuje realnu ljestvicu znanja kod sustava ocjenjivanja od jedan do pet. Tablica II. prikazuje ljestvicu znanja, tj. ocjenjivanje po tom istom sustavu kod strožeg nastavnika, dok tablica III. prikazuje ljestvicu znanja, tj. ocjenjivanje kod nastavnika blažih kriterija. [6]

Tablica I. Prikaz realne ljestvice znanja

ZNANJE				
1	2	3	4	5

Tablica II. Prikaz ljestvice znanja kod strožeg nastavnika

ZNANJE				
1	2	3	4	5

Tablica III. Prikaz ljestvice znanja kod blažeg nastavnika

ZNANJE				
1	2	3	4	5

Svako ocjenjivanje, svaki ispit, test ili usmena provjera zapravo spadaju u mjerenje znanja. Tim postupcima nastavnik pokušava izmjeriti znanje učenika.

Svako mjerenje ima tri uvjeta:

- predmet mjerenja (olina, veličina, objekt),
- instrument mjerenja,
- tehnika mjerenja.

A. *Tehnike ispitivanja znanja*

Dvije su glavne tehnike ispitivanja znanja:

- usmeno ispitivanje znanja,
- pismeno ispitivanje znanja.

U daljnjem tekstu navedene su prednosti i nedostatci ovih tehnika ispitivanja znanja:

Usmeno ispitivanje znanja:

Prednosti usmenog ispitivanja znanja su:

- nastavnik može reagirati na odgovore učenika i tako primjerenije ispitati stvarno znanje učenika,
- nastavnik može bolje ispitati razumijevanje učenika,
- tijekom ispitivanja znanja učitelj može uočiti i reagirati na odgovore koji su približno točni ili one koji su potpuno apsurdni,
- učenici češće uče s razumijevanjem ukoliko očekuju usmeno ispitivanje. [6]

Nedostatci usmenog ispitivanja znanja su:

- vremenski je neekonomično,
- subjektivno, nastavnik je mjerni instrument različito reagira na različite učenike,
- učenici dobivaju i različita pitanja pa je teško usporediti znanje te ovise o sreći hoće li dobiti pitanje na koje znaju odgovor,
- obuhvaća manje gradiva. [6]

Pismeno ispitivanje znanja:

Prednosti pismenog ispitivanja znanja su:

- manji utjecaj subjektivnosti,
- svim učenicima su postavljena ista pitanja,
- isti je režim rada isti uvjeti za sve,
- ocjenjivanje je objektivnije jer je moguće lako izračunati postotak riješenih zadataka,
- može obuhvatiti više gradiva,
- vremenski je ekonomično jer se mogu u jednom satu ispitati svi učenici. [6]

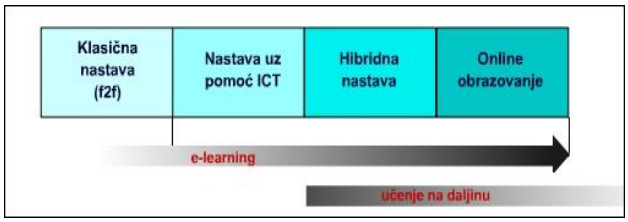
Nedostatci pismenog ispitivanja znanja su:

- često se traži samo reprodukcija materijala, proces pamćenja i sjećanje činjenica,
- ne uspijeva zahvatiti razumijevanje ili primjenu znanja,
- najčešći oblik jest pismeno postavljanje pitanja i ispravljanje zadaća,
- međutim, taj način da se redom pregledaju, procjene i zatim ocjene zadaće najmanje je valjan.
- ako je pak očekivao manje nego što je dobio u prvim zadaćama, nastavnik će dići kriterij pa će prve zadaće biti u boljem položaju,
- bilo bi dobro malo pregledati zadaće prije ocjenjivanja da se dobije raspon i da se odredi zajednički kriterij. [6]

III. VRSTE UČENJE

Osnovna klasifikacija učenja, tj. nastave, može se podijeliti na:

- klasično učenje (*f2f* ili *face-to-face*),
- učenje uz pomoć ICT-a (*ICT supported teaching and learning*) tj. elektroničko učenje,
- hibridnu ili mješovitu nastavu (*hybrid/blended learning*),
- online nastavu (*fully online*) . [1]



Slika 1. Prikaz vrsta učenja tj. nastave

A. *Klasično učenje*

Klasično učenje predstavlja model učenja u kojem nastavnik predaje, a učenik sluša i zapisuje predavanje nastavnika. To je takozvano klasično učenje licem u lice tj. *f2f* ili *face-to-face.*

B. *Elektroničko učenje*

Elektroničko je učenje interaktivan ili dvosmjeran proces između nastavnika i učenika uz pomoć elektroničkih medija pri čemu je naglasak na procesu učenja dok su mediji samo pomoćno sredstvo koje upotpunjuje taj proces.

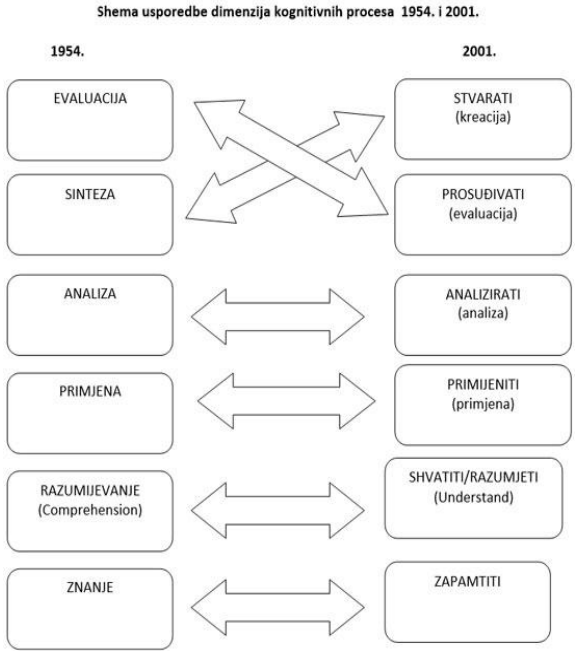
Elektroničko učenje je bilo koji oblik učenja, poučavanja ili obrazovanja koji je potpomognut uporabom računalnih tehnologija, a posebno računalnih mreža temeljenih na internet tehnologijama. [2]

IV. MJERENJE ZNANJA PREMA BLOOMOVOJ TAKSONOMIJI

Jedan od najkorištenijih teorijskih okvira za planiranje, pripremu i vrednovanje osnovnoškolskog, srednjoškolskog i visokoškolskog obrazovanja te tečajeva za odrasle na području obrazovanja i industrije Bloomova je taksonomija znanja (Kirkpatrikov model vrednovanja učenja (*Kirkpatrick learning evauation model*), 1959. podjednako je raširen). Nastala je tijekom 50. tih godina 20. stoljeća na temelju analiza intelektualnih ponašanja uz pomoć kojih učenici stječu akademska znanja. [5]

A. *Bloomova taksonomija*

Krajnji je cilj školskog učenja stjecanje trajnih i upotrebljivih znanja i umijeća. Znanje i umijeće su produkt mišljenja. Mišljenje se odvija u mozgu osobe, a u njemu su pohranjeni i znanje i umijeće, te nisu direktno opažljivi i mjerljivi. O njima zaključujemo na temelju onoga što učenik pokaže prema van: na temelju opažljivog ponašanja (Slika 2). [5]



Slika 2. Shema usporedbe dimenzija kognitivnih procesa prema Bloomu 1954. i 2001. [3]

Bloom i suradnici na učenje gledaju kao na umijeće ponašanja. Cilj njihovog rada bilo je sistematiziranje kategorija ponašanja koja se koriste tijekom učenja kako bi učiteljima pomogli pri planiranju i procjeni školskog učenja. Ciljevi učenja i ponašanja koja učenik koristi tijekom učenja podijeljeni su u tri kategorije koje su međusobno povezane i preklapaju se. [5]

Kategorije su područja učenja:

1. kognitivno (intelektualna sposobnost, tj. znanje ili mišljenje).
2. afektivno (osjećaji, tj. odnos ili stav),
3. psihomotorno (fizičke vještine, tj. ono što osoba može činiti) (Grafikon I.).

Bloomova taksonomija godinama je znanstveno provjeravana, korištena i dorađivana. 1990. godine revidirali su je Anderson i Krathwohl. U početku je bila orijentirana samo na kognitivno područje. Kasnije je definirano i afektivno i psihomotorno područje. Sva tri područja znanja jednako su vrijedna. Unutar svakog područja obrazovni su ciljevi razvrstani u kategorije koje predstavljaju razine znanja i hijerarhijski su poredane na temelju težine ili složenosti, od najjednostavnijih do najsloženijih. [5]

Grafikon I. Tri osnovna područja ili domene odgojno-obrazovnih aktivnosti prema Bloomu



B. *Razine znanja i ciljevi učenja prema Bloomovoj taksonomiji*

Bloomova taksonomija ciljeva u afektivnom (stavovi) odgojnom-obrazovnom području.

Podijeljena je na 5 kategorija:

1) percepcija,
2) reagiranje,
3) usvajanje vrijednosnih stavova,
4) organiziranje vrijednosnih stavova,
5) primjena vrijednosnih stavova na praktičnoj razini. [3]

Bloomova taksonomija ciljeva u psihomotornom (vještine) odgojno-obrazovnom području.

Ciljevi su razvrstani u 4 kategorije :

1) gruba koordinacija,
2) fina koordinacija,
3) neverbalne komunikacije,
4) govorno ponašanje. [3]

Razine znanja i ciljevi učenja na kognitivnom području su:

Tablica IV. Ciljevi učenja prema Bloomu [5]

Razine	CILJEVI (ISHODI) UČENJA Značenje razine
I.	**DOSJETITI SE (Znanje)** Dosjetiti se, prepoznati ili reproducirati informaciju, ideju i princip u približno onakvom obliku u kojem su naučeni.
II.	**SHVATITI (Razumijevanje)** Uočiti i povezati glavne ideje. Prevesti, razumjeti, objasniti ili interpretirati naučeni sadržaj. Izvesti logičan zaključak iz dostupnih informacija. Zaključiti o uzroku i predvidjeti posljedice.
III.	**PRIMIJENITI (Primjena)** Rješavati probleme primjenom naučenog u kontekstu učenja ili u novoj situaciji na nov način. Odabrati i primijeniti podatke i principe za rješavanje problema ili zadatka u drugom području.
IV.	**ANALIZIRATI (Analiza)** Raščlanjivati informacije kako bi se utvrdili dijelovi cjeline, njihovi međusobni odnosi, organizacijski principi, uzroci i posljedice, izveli dokazi i zaključci i podržale generalizacije. Prepoznati skriveno značenje. Razlikovati činjenice i zaključke.
V.	**PROSUĐIVATI (Evaluacija)** Usporediti i pronaći sličnosti i razlike među idejama. Procijeniti valjanost ideja i/ili kvalitete uratka na temelju poznatih kriterija. Dokazati vrijednost. Izabrati mogućnost i argumentirano obrazložiti.
VI.	**STVARATI (Sinteza)** Kreativno koristiti postojeće znanje za stvaranje nove cjeline. Stvarati nove ideje i rješenja. Izvoditi generalizacije na temelju dobivenih podataka. Povezati znanje iz različitih područja.

V. PRIMJENA ELEKTRONIČKOG UČENJA U NASTAVI

Znanje mjerimo kada je maksimalno smanjen utjecaj subjektivnih čimbenika o kojima smo govorili pri procjeni znanja. Objektivno ispitivanje znanja provodi se pomoću nizova zadataka objektivnog tipa (ZOT) i pomoću testova znanja. Zadaci objektivnog ispitivanja znanja primarno služe za praćenje napredovanja učenika, a testovi znanja na formalnim ispitima nakon određenog obrazovnog perioda (npr. na kraju školske godine). Po svojoj se formi ZOT i testovi znanja ne razlikuju jer sadrže iste vrste zadataka objektivnog tipa. [7]

Za ispitivanje znanja u ZOT i testovima znanja koriste se različiti zadaci objektivnog tipa. Za sve je karakteristično da odgovori mogu biti jednoznačno interpretirani i objektivno vrednovani kao točni ili netočni.

Vrste zadatka objektivnog tipa:

a) zadaci dosjećanja i dopunjavanja,
b) alternativni zadaci, tj. alternativne tvrdnje,
c) zadaci višestrukog izbora,
d) zadaci s dva tipa kriterija izbora,
e) zadaci ispravljanja,
f) zadaci sređivanja,
g) zadaci povezivanja,
h) esejski zadaci,
i) zadaci rješavanja problema. [7]

A. *Provjera znanja iz novovjekovne filozofije (Gordano Bruno, Rene Descartes i Gottfried Leibnitz)*

ZADACI OBJEKTIVNOG TIPA

Škola:_____
Razred:_____
Spol: _____ M/Ž

1. Giordano Bruno vjeruje da je planet Zemlja?

a) u središtu sunčevog sustava
b) zvijezda
c) planet koji se kreće po kružnoj stazi oko Sunca
d) planet koji se kreće po eliptičnoj stazi oko Sunca

2. Rene Descartes poznat je po svojoj rečenici:

a) mislim dakle postojim
b) mislim dakle jesam
c) homo homini lupus
d) _____

3. Giordano Bruno poznat je po svojem izrazu za čovjeka:

a) mislim dakle postojim
b) zoon politikon
c) homo homini lupus
d) _____

4. Gottfried Leibnitz napisao je:

a) Monadologiju
b) Mondologiju

5. Navedi dva filozofa predstavnika racionalističke filozofije 17. stoljeća:

_____
_____

6. Zaokruži djela koja je napisao Descartes, a podvuci djela koja je napisao Leibnitz:

Rasprava o metodi

Rasprava o metafizici

O principu pojedinačnog

Načela filozofije

O meteorima

7. Gottfried Leibnitz osmislio je modernu formu:

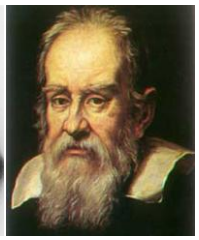
a) heksadekadskog računa
b) binarnog računa
c) dekadskog računa
d) oktalnog računa

8. Po kojem filozofu je kartezijev sustav dobio ime

a) Galileu
b) Descartesu
c) Leibnizu
d) Kepleru
e) _____

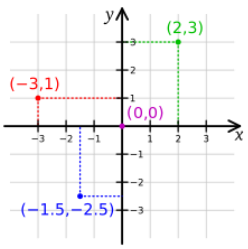
9. Kako još nazivamo kartezijev sustav _____ _____ i gdje ga koristimo.

10. Ispod slika napiši imena filozofa, znanstvenika.


_____ _____ _____

11. Ispod slike napiši što slika predstavlja.

_____ _____

12. Pitanje za eksperimentalnu skupinu: Skiciraj Brunov planetarni sustav i Keplerov planetarni sustav korištenjem računala.

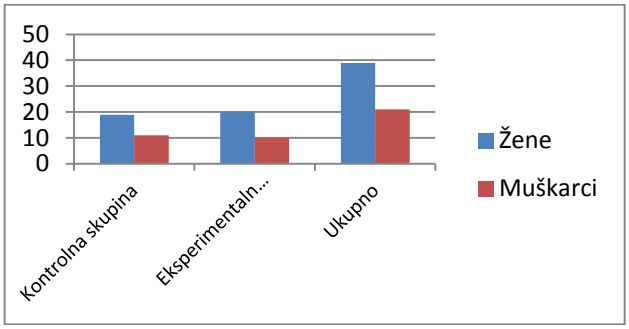
Pitanje za kontrolnu skupinu: Skiciraj Brunov planetarni sustav i Keplerov planetarni sustav.

B. Rezultati istraživanja

Istraživanje je provedeno na uzorku od 60 učenika, od toga je bilo 39 učenica ženskog spola i 21 učenik muškog spola. Učenici su pohađali nastavu filozofije u četvrtom razredu po nastavnom programu za gimnazije. Učenici su u prosjeku imali 18 godina (Tablica V.).

Istraživanje je provedeno kroz dvije skupine – kontrolnu i eksperimentalnu skupinu. U kontrolnoj skupini sudjelovalo je 30 učenika jednog razreda, a u eksperimentalnoj skupini sudjelovalo je 30 učenika drugog razreda. Kontrolna skupina slušala je predavanja od tri nastavna sata iz novovjekovne filozofije korištenjem klasičnog tipa predavanja. Eksperimentalna skupina slušala je ista ta predavanja iz novovjekovne filozofije korištenjem informatičke opreme (računala, projektora, interneta, digitalnih sadržaja, multimedijalnih prezentacija, animacija i video isječaka (Grafikon II.).

Grafikon II. Prikaz broja ispitanika po skupinama i spolu



Svaka od skupina (kontrolna i eksperimentalna) rješavala je po dva testa (zadaci objektivnog tipa) na početku predavanja i na kraju predavanja (nakon odrađene tri nastavne jedinice). Na ovaj sam način pokušala izmjeriti znanje učenika prije i poslije održanih predavanja kod obje skupine.

Rezultati istraživanja su očekivani i zadovoljili su obje hipoteze.
Prva hipoteza je glasila: elektronička nastava povećat će razinu znanja kod učenika.
Druga hipoteza je glasila: postignuta razina znanja u eksperimentalnoj skupini bit će veća nego u kontrolnoj skupini.

Tablica V. Prikaz broja ispitanika po skupinama i spolu

	Žene	**Muškarci**	**Ukupno**
Kontrolna skupina	**19**	**11**	**30**
Eksperimentalna skupina	**20**	**10**	**30**
Ukupno	**39**	**21**	**60**

Iz prikazanoga je vidljivo da su učenici četvrtih razreda, u skladu s očekivanjima, bili puno više zainteresirani za sadržaje u elektroničkom tj. digitalnom obliku, nego za sadržaje vezane uz klasičnu nastavu (udžbenik, bilježnica).

VI. ZAKLJUČAK

Današnje mlade generacije sve su sklonije korištenju interneta i računala. Cijeli njihov svijet svodi se na ta dva medija. Zato je postalo sve očitije da nove mlade generacije učenika koriste i žele koristiti računalo i njegove mogućnosti ne samo za zabavu i druženje, već i za učenje.

Elektroničko učenje potiče nove generacije na kreativnost i radoznalost koja često sa sobom nosi i neke opasnosti novog globalnog svijeta. Elektroničko je učenje postalo sastavni dio današnjih škola, fakulteta i ostalih institucija za obrazovanje ljudi.

Digitalne nastavne jedinice mnogo su razvijenije te se sastoje od individualnih lekcija, nastavnih cjelina ili tečajeva u interaktivnom obliku. Korištenjem digitalnih nastavnih jedinica jednostavnije je i lakše kreirati mnoge različite digitalne obrazovne sadržaje; primjerice, lekcije iz fizike, filozofije, geografije, matematike te mnogih drugih predmeta u nastavi.

Sve više se čini da je pokrenuta informacijsko-komunikacijska revolucija nezaustavljiva i da će uskoro većina našeg obrazovanja i razmijene znanja postati dio virtualnog svijeta, a to je ono što nove generacije učenika pokreće ka kreativnom stvaranju i pamćenju potrebnih sadržaja te kreiranju novih ideja.

LITERATURA

[1] Elliott, M. i McGreal, R. Tehnologije online učenja (e-learning), Edupoint 37 (V). 2005. Dostupno na: http://edupoint.carnet.hr/casopis/37/clanci/4.html (20.1.2014.).

[2] E-learning (Electronic Learning) - Elektroničko učenje. Dostupno na: http://www.srce.unizg.hr/proizvodi-i-usluge/obrazovanje/centar-za-e-ucenje/e-ucenje/ (21.2.2014).

[3] L. W. Anderson, D. R. Krathwohl i suradnici, A Taxonomy for Learning, Teaching, and Assessing; A Revision of Bloom's Taxonomy of Educational Objectives, 2001.

[4] Maravić, J. (2003), Cjeloživotno učenje. CARNet časopis. br.17, god. III. Zagreb: Edupoint. Dostupno na: http://www.carnet.hr/casopis/17/clanci/5 (3.2.2014.).

[5] Nimac, Elvira. Primjena Bloomove taksonomije u nastavi. Dostupno na: http://www.azoo.hr/index.php?option=com_content&view=article&id=2526:razine-uenikih-postignua-bloomova-taksonomija-znanja (1.2.2014.)

[6] Pavlović, Ante. Praćenje, vrjednovanje i ocjenjivanje učenika u procesima učenja i poučavanja; načela, metode, instrumenti. 2009.

[7] Tečić, A. Ocjenjivanje napretka i vrednovanje postignuća učenika u školama. Šibenik: Exp. Edit 2006.

Analiza uspješnosti u usvajanju algoritamskog načina razmišljanja kod učenika u prvom razredu novog strukovnog kurikuluma za zanimanje Tehničar za računalstvo

K. Blažeka, Tehnička škola Čakovec, Čakovec, Hrvatska
kristinka.blazeka@ck.t-com.hr

Sažetak: U školskoj godini 2013./2014. u dijelu srednjih škola krenula je eksperimentalna provedba nekoliko novih strukovnih kurikuluma. Učenici upisani u prvi razred za zanimanje Tehničar za računalstvo također su krenuli po novom bitno osuvremenjenom programu. Već u prvoj godini školovanja uveden je novi nastavni predmet Algoritmi i programiranje sa naglaskom na usvajanje znanja i razvoj vještina u području analize problema i dizajna njegovog rješenja prilikom izrade računalnih programa. Predmet se osim u prvom razredu treba realizirati i u drugom. Prema ishodima učenja u početnim je mjesecima stavljen naglasak na usvajanje važnosti algoritama u procesu rješavanja problema i razvoj algoritamskog razmišljanja s ciljem odabira i osmišljavanja odgovarajućeg algoritma na što se nastavlja učenje i savladavanje sintakse odabranog programskog jezika. Kako već površne analize pokazuju da su logičko zaključivanje, kreativno stvaralaštvo i algoritamsko razmišljanje, iako nužne općenito za rješavanje problema, na žalost vještine prečesto zanemarene u osnovnoškolskoj nastavi, potrebno je uložiti prilično napora kako bi učenici dobili upotrebljive temelje za uspješnije stvaranje računalnih programa. S ciljem preispitivanja postignuća učenika prije učenja sintakse programskog jezika istraženo je koliko su učenici bili uspješni u usvajanju osnova algoritamskog razmišljanja. Dobiveni rezultati trebali bi pomoći u uočavanju aspekata u kojima učenici trebaju dodatnu pomoć, za eventualne modifikacije kurikuluma, a mogu poslužiti i kao inspiracija nastavnicima koji obavljaju složen zadatak razvoja vještina algoritamskog razmišljanja i programiranja kod učenika.

I. UVOD

Unatoč dobronamjernim pokušajima integriranja u nastavni kurikulum, učenici nakon završene osnovne škole većinom pokazuju nedovoljnu razvijenost algoritamskog načina razmišljanja i strategija rješavanja problema, neophodnih vještina potrebnih kod učenja programiranja računala. Prečesto i sami nastavnici izborne nastave informatike zaključuju (bez opravdanih razloga što pokazuju brojna istraživanja) kako je učenje algoritamskog razmišljanja radi konstruiranja algoritama i razvoja vještina programiranja nešto nepristupačno i neshvatljivo gotovo svim mlađim učenicima.

Primjenom danas dostupnih, vrlo primjenjivih alata, olakšano je razvijanje i usavršavanje logičkog i matematičkog razmišljanja kod učenika koji, upotrebljavajući ove često vrlo izazovne alate, pokazuju ne samo da mogu usvajati algoritamsko razmišljanje i učiti programiranje već da u tome uživaju. Pri tom nije zanemariva ni činjenica potpore roditelja koji i sami potiču interes učenike svjesni svrhovitosti i značajne upotrebljivosti razvoja ovih vještina.

U mnogim je državama tako trend da se moguća zabavna i zanimljiva narav programiranja mlađim učenicima približava kroz vizualne okoline kao što su Scratch i Kodu (npr. kurikulum u Engleskoj i Njemačkoj)[1]. Na žalost, još uvijek nije rijetkost nedostatak kvalitetno pripremljenih nastavnika koji mogu uspješno poučavati osnove računalnih znanosti, s obzirom da često nemaju nužno dovoljno poznavanje programskih koncepata koji počivaju u pozadini ovih okolina. Tek na takav način oni mogu olakšati razvoj i usvajanje osnovnih principa programiranja umjesto da upotrebu ovih moćnih i izazovnih alata svode na puku upotrebu aplikacije.

Osim toga, nisu rijetki ni slučajevi da se nastava početnog učenja programiranja (obično u smanjenom opsegu) koncentrira na sintaksu programskog jezika stavljajući u drugi plan aktivnosti usmjerene stjecanju iskustva u analizi rješavanja problema i razvoju algoritamskog razmišljanja. Poželjan način učenja dizajniranja i pisanja programa trebao bi u dovoljnoj mjeri naglašavati veliku važnost učenja sustavnog načina pristupa traženju rješenja i prevođenja tog rješenja u funkcionalni program. Ta sustavnost podrazumijeva rigorozno slijediti korake koji uključuju analizu problema, kreiranje plana rješavanja, konstruiranje algoritma (npr. dijagramom tijeka), implementacija algoritma (u izabranom programskom jeziku) te konačno provjeravanje i ispravljanje algoritma. Kao raširena metoda, dijagrami tijeka pomažu početnicima vizualizirati i razumjeti proces, od njegove formulacije preko traženja do oblikovanja rješenja, te olakšavaju uvježbavanje logike osnovnih upravljačkih struktura kao i uočavanje mogućih problematičnih točaka u razvoju programa.

Bez obzira na pristup i izabranu metodu, preporučljivo je u početku davati odgovarajuće, lakše primjere. Tako se, primjerice, za početni, uvodni, jednostavan proces ili slijed koraka može tražiti uočavanje osnovih koraka kod čestih životnih situacija kao što su pripremanje čaja ili sendviča, odlazak u školu, zbrajanje brojeva kalkulatorom i slično. Za učenje uvjetnih struktura koje uključuju potrebu donošenja odluka ovi se problemi dodatno mogu zakomplicirati određenim (realnim) situacijama (npr.

zaslađivanje čaja šećerom ili medom, odabir prijevoznog sredstva kod odlaska u školu ili potreba da se svrati po pecivo i dr.). Petlje, odnosno ponavljajuće strukture također se bez problema mogu uvoditi uspoređivanjem sa nekim svakodnevnim situacijama (npr. izračunavanje prosječne ocjene svih predmeta jednog, odnosno više učenika) [2].

Nastavak teksta podijeljen je na cjelinu u kojoj se govori o pristupu novog strukovnog kurikuluma za stjecanje kvalifikacije Tehničar za računalstvo koji u prvoj godini učenja (a nastavlja se u drugoj) uvodi predmet naziva Algoritmi i programiranje. Slijedi analiza predznanja učenika i usvojenosti predviđenih ishoda nakon završetka prvog polugodišta za prvu generaciju upisanih učenika u promatranoj školi. U zaključku su dana završna razmatranja, komentari i prijedlozi mogućih praćenja, odnosno mjera za poboljšanje stanja, kao i moguća buduća istraživanja.

II. NOVI STRUKOVNI KURIKULUMI I NASTAVNI PREDMET: ALGORITMI I PROGRAMIRANJE

Vezano uz predmet razmatranja ovog rada, izvođenje predmeta Algoritmi i programiranje, u uvodnom standarda kvalifikacije Tehničar za računalstvo navodi se kako će učenici planirano usvojiti slijedeće kompetencije i ishode učenja: upotrijebiti i primjenjivati temeljene programske i podatkovne strukture i algoritme (nizove, polja, petlje, grananja) u izradbi programskih rješenja manjeg stupnja složenosti korištenjem metoda i alata sekvencijalnog programiranja, što treba biti nužan i dovoljan temelj kako bi mogli i projektirati i izraditi složenije računalne programe sa grafičkim sučeljem za prezentaciju rezultata i stanja, uz korištenje algoritama pretraživanja te komunikaciju s bazom podataka, korištenjem objektno orijentiranih razvojnih alata i programa što bi trebali usvajati kroz nastavu drugih stručnih predmeta odgovarajućeg izabranog modula. Cilj i svrha nastave predmeta Algoritmi i programiranje je razvoj algoritamskog načina razmišljanja u rješavanju problema i ovladavanje osnovama korištenja programskog jezika kao alata za rješavanje problema. Sukladno tome definirani su slijedeći ishodi učenja [6]:

1. objasniti važnost algoritama u procesu rješavanja problema
2. razmotriti tipične strukture podataka i pripadnih algoritama te prepoznati važna svojstva algoritama
3. odabrati i osmisliti algoritme za rješavanje jednostavnijih problema
4. koristiti naredbe za upis i ispis podataka
5. koristiti naredbe za grananje i ponavljanje dijelova programa
6. koristiti algoritme za pretraživanje i sortiranje
7. kreirati i koristiti funkcije u programu
8. koristiti pokazivače u radu s poljima i funkcijama
9. kreirati i koristiti strukture podataka
10. kreirati i koristiti datoteke

III. RAZVOJ ALGORITAMSKOG RAZMIŠLJANJA

Algoritam možemo pojednostavljeno definirati kao postupak za rješavanje problema koji se sastoji od točno određenih uputa. Uspješno nalaženje odgovarajućeg algoritma usko je povezano sa razvijenim algoritamskim razmišljanjem koje predstavlja jednu od najvažnijih kompetencija u području računarstva. Na nju se izravno nadovezuje pisanje računalnih programa kao vježba logičkog razmišljanja u samom temelju računarstva s obzirom da su računala neupotrebljiva bez skupa naredbi koji im objašnjava što trebaju raditi, bez obzira na vrstu aplikacije koja se na računalu izvodi.

Zbog znatne složenosti poučavanja vještina algoritamskog razmišljanja koje inkorporira kritičko razmišljanje, a u konačnici i razvoj programske logike, tehnologije i sintakse, nastavnici ove discipline moraju biti izvrsni u vještinama koje poučavaju i pri tom uzimati u obzir specifične karakteristike i potrebe mladih i početnika. Njihov interes može se pojačavati naglašavanjem činjenica kako su algoritmi snažan alat primjenjiv u nizu zanimljivih aktivnosti prema postizanju značajnog napretka kroz nove ideje i to ne samo u području računarstva i programiranja računala.

Algoritamsko razmišljanje objedinjava niz sposobnosti vezanih uz stvaranje i razumijevanje algoritama. Tu je neizostavno potrebno uključiti sposobnost preciznog definiranja i analiziranja danog problema, sposobnost uočavanja i nalaženja osnovnih problemu odgovarajućih postupaka, sposobnost stvaranja korektnog algoritma za dani problem upotrebom osnovnih postupaka, sposobnost razmišljanja o svim mogućim posebnim i uobičajenim slučajevima problema, sposobnost evaluacije i poboljšavanja korektnosti i učinkovitosti algoritma.[3] Mnogi humani kognitivni čimbenici utječu na sposobnost algoritamskog razmišljanja i tu se prije svega može istaknuti apstraktno i logičko razmišljanje, disciplinirano sustavno i formalno razmišljanje, razmišljanje u strukturama, kreativnost i kompetencija rješavanja problema.

Uz to, algoritamsko razmišljanje u određenoj je mjeri neprirodno. Dok u svakodnevnom životu probleme rješavaju ljudi, algoritamsko razmišljanje učenik ipak mora posebno uvježbati jer je povezano s nalaženjem rješenja koje će biti ukodirano u računalo, koje mora biti „razumljivo“, izvedivo na stroju kodirano njegovim nizom instrukcija iz skupa naredbi koje može izvoditi.

Očito je da takva složenost nužno čini algoritamsko razmišljanje teškim za usvajanje i objašnjava potrebu za izbjegavanjem neuspjeha primjenom dobrih didaktičkih pristupa posebno kod početnika. Razloga za neuspjeh kod usvajanja vještina programiranja može biti više, a uključuju: nedostatak zrelosti kod učenika, nedostatak vještina kritičkog razmišljanja, nedostatak praktičnog rada u laboratoriju, nedostatna matematička predznanja ili sposobnost logičkog razmišljanja,neodgovarajući zadaci za uvježbavanje potrebnih vještina, prebrzo napredovanje kroz gradivo, a javlja se i strah od programiranja kao slijed straha od matematike.

IV. KAKO POUČAVATI ALGORITME I ALGORITAMSKO RAZMIŠLJANJE?

Više je načina na koje se može stimulirati razmišljanja o algoritmima. Odgovarajući pristup može biti razvijanje neformalnih (ali učinkovitih i korektnih) algoritama za rješavanje bliskih i razumljivih problema i zagonetki, nealgoritamske zagonetke koje potiču rigorozno logičko zaključivanje i analize slučajeva, jednostavni zadaci u kojima učenici prate i slijede dobro definirani niz uputa, zadaci u kojima učenici ispituju i istražuju snagu i slabosti danog algoritma kroz proučavanje njegovog ponašanja (traženjem približnog broja koraka ili scenarija najlošijeg slučaja) potičući kod učenika tako razmišljanja o računalnoj složenosti. Također, učenicima se mogu zadavati zadaci u kojima trebaju naći grešku u nizu naredbi ili moraju upotpuniti iscrpan skup ulaznih scenarija kako bi ispitali korektnost danog algoritma, čime se uvježbavaju u usvajanju važnih karakteristika algoritma prezentiranog vlastitim programskim kodom: ispravnosti, učinkovitosti i robustnosti. Zadaci koji se zadaju učenicima mogu podržavati jedan, ali i kombinaciju spomenutih pristupa pazeći da rješavajući ih učenici razvijaju nužan sustavan pristup razvoju algoritma, prateće tehnike manipulacije podacima, brzinu i djelotvornost.

Kako bismo svu složenost učenja algoritamskog razmišljanja približili učenicima njihovi zadaci i akcije koje uključuju trebaju učenicima biti bliski iz svakodnevnog života, a algoritmi opisani prirodnim opisnim jezikom i izvedeni na sustavu koji će izvoditi algoritam, a na kojem učenici istovremeno mogu i eksperimentirati kako bi preko kvalitetne povratne veze ostvarili neposredna iskustva učenja. U usvajanju algoritamskog razmišljanja pomaže i rješavanje što većeg broja problema, no posebno je za početnike važno da su ti problemi izabrani vrlo pažljivo i da ih je moguće opisati prihvatljivim načinom, npr. pseudo kodom ili vizualizacijskim alatom (npr. dijagramom tijeka). Problemi ne bi smjeli biti prejednostavni, no formulacija problema treba biti lako razumljiva. Upotreba složenih problema također je smislena jer što su problemi složeniji, više je prostora za razvijanje kreativnosti i neuobičajenih rješenja. No bez obzira na složenost, nastavnici ne bi trebali prezentirati gotova rješenja već poticati učenike u njihovom procesu učenja motivirajući ih u napretku prema nalaženju rješenja.

Pri razumijevanju principa algoritama korisne su animacije koje izvode računalni programi, no isto tako poželjni su pristupi sa zadacima i scenarijima koji isključuju upotrebu računala [4] ili [5].

V. ANALIZA REZULTATA ISPITIVANJA UČENIKA

Uzimajući u obzir konstruktivističku teoriju učenja koja učenje objašnjava kao proces u kojem se smisao novih informacija stvara kroz spektar postojećih znanja, najbolje je kad su tvorci i izvoditelji kurikuluma svjesni prethodnih znanja i iskustva učenika pa im kroz nadogradnju izazivaju postojeće ideje. Poželjno je da se nastavnik u tom procesu postavlja u ulogu vođe i dizajnera obrazovnih iskustava, a ne tek pukog predavača. [7]. Dodatno, kurikulum bi trebalo sagledavati i kroz konstrukcionističku teoriju učenja i poučavanja građenu na konstruktivističkim idejama pri čemu se postavlja kako ljudi najbolje uče razvojem ideja i stvaranjem [11].

Konstruktivistička perspektiva učenja traži od nastavnika da prije nego pristupi nastavnom procesu poučavanja jasno identificira ideje koje učenici razvijaju o algoritamskom razmišljanju, s posebnim naglaskom na dva aspekta: algoritme i programiranje. Te ideje nastavniku omogućavaju da nastavu razvija u vidu nadogradnje pri čemu učenici napredovanjem u učenju usvajaju određen skup ideja i vještina omeđeno donjom točkom sidrenja (koja opisuje ideje koje učenici imaju prije pohađanja nastave) i gornjom točku sidrenja (koja opisuje očekivane ciljeve ili ishode koje bi učenici trebali ostvariti završetkom nastave). Između tih točaka obično se nalazi više međurazina.

S ciljem dobivanja uvida u prethodno znanje i izloženost osnovnim idejama i konceptima algoritamskog načina razmišljanja, odnosno postavljanja donje točke sidrenja, provedeno je ispitivanje na uzorku od 25 učenika prvog razreda u zanimanju Tehničar za računalstvo.

Anketa je pokazala kako su svi upisani učenici učili programski jezik Basic u osnovnoj školi, dakle trebali su biti formalno izloženi usvajanju algoritamskog načina razmišljanja što je izrijekom napisano u opisu ciljeva izbornog predmeta Informatike: „... učenici moraju naučiti djelotvorno upotrebljavati računala i biti sposobni ugraditi osnovne zamisli algoritamskog načina razmišljanja u rješavanje svakodnevnih problema.“[8].

Programski jezik Basic bio je gotovo jedini odabir nastavnika, a učenici su ga učili najviše u izbornoj nastavi Informatike koju je pohađalo 88% upisanih učenika. U nastavi Tehničke kulture učilo ga je 20% učenika. S obzirom na trajanje učenja, u izbornoj nastavi učenje se svelo na manje od jednog polugodišta kod 45% učenika, dok ih je 31% taj programski jezik učilo više od jedne godine .

Iako su učili programiranje tek je 7 odnosno 28% učenika odgovorilo da su prije srednje škole čuli za algoritam kao pojam, no nitko od učenika nije znao imenovati, niti navesti neki algoritam za koji su čuli ili ga koristili.

Nadalje, učenicima je na početku nastavne godine dan zadatak prepoznavanja problema i oblikovanja uputa za njegovo rješavanje. Svi su učenici pronašli neki svoj problem iz svakodnevnog života, no 4 učenika (16%) nisu postupak njegovog rješavanja uspjeli opisati ni na najjednostavniji način. 8 (32%) učenika je u svom problemu posredno postavilo strukturu grananja, a tek 2 učenika (8%) u opis rješavanja unijelo je i ponavljajuću programsku strukturu.

Oblikovanje i izvođenje naredbi korak po korak jedna je od temeljnih vještina u području algoritamskog razmišljanja. Ova se aktivnost može kod početnika uvježbati tako da jedan učenik prema zadanoj slici daje upute drugom kako nešto nacrtati [5]. Svrha te aktivnosti je pokazati učenicima i značaj osiguravanja preciznih uputa drugima i ne malu izazovnu prirodu takvog zadatka kad se to svede isključivo na pisane instrukcije (upućene stroju).

Krajem prvog polugodišta učenici su najviše kroz vizualizaciju algoritama dijagramima tijeka ili kroz pseudokod već trebali bolje usvojiti osnovne pojmove vezano uz algoritamsko razmišljanje i predstavljanje rješenja za bilo koju instancu zadanog problema. Oblikovanje postupka za rješavanje jednostavnog problema izrazito su dobro riješila tek 2 učenika (8%), 24% učenika pri tom je pokazalo manje nepreciznosti, dok ih je najveći broj , njih 15 ili 60% prilično površno i neprecizno. 2 učenika taj su zadatak nisu na zadovoljavajući način riješila.

Kod ispitivanja poznavanja pseudokoda, svi su ispitani učenici riješili zadatak slijednog pridruživanja vrijednosti varijablama, dok ih je 68-88% uspjevalo riješiti zadatak sa strukturom grananja. Najmanji postotak riješenosti imao je zadatak sa ponavljajućom strukturom. Taj je zadatak točno riješilo tek 40% učenika.

Pokazuje se kako je dobar pristup razvoju ideja o algoritamskom načinu razmišljanja kroz aktivnosti bez računala i one na računalu. Naime, kad ne koriste računala učenici razvijaju koncepte u kontekstu svakodnevnih životnih situacija. I kurikulum ovog predmeta aktivnosti sa računalom predviđa tek naknadno, u drugom polugodištu. U radovima se može naći niz različitih vrlo uspješnih pristupa poučavanju programiranja [npr. 9].

Kako bi se kurikulum potpunije evaluirao ove rezultate bilo bi korisno usporediti sa rezultatima koje će učenici ostvariti na kraju nastavne godine.

VI. ZAKLJUČAK

Skup vještina i kompetencija poželjan za zapošljavanja konstantno evoluira. Poslodavci danas uz vještine komuniciranja, etičke, interpersonalne i personalne navode i znanja programiranja [Aasheim, 2012, Gruba 2004] relevantna prilikom rješavanja suvremenih problema. No, iako bi učenici, bar oni uključeni u izbornu nastavu Informatike, trebali već tijekom osnovnog školovanja (kad im se može pokazati kako je računalo razmišljanje moguće integrirati i u druge predmete) usvojiti osnovne zamisli algoritamskog načina razmišljanja pokazuje se kako taj cilj uglavnom nije ostvaren.

Analizom dobivenih podataka uočava se kako je izvođenjem nastave predmeta Algoritmi i programiranje ostvaren napredak u usvajanju algoritamskog načina razmišljanja kod učenika. Ovim istraživanjem dobivene informacije mogu se iskoristiti kako bi se usmjerilo ili korigiralo izvođenje inicijalnog eksperimentalnog kurikuluma. Bilo bi dobro istraživanjem prikupiti podatke i iz drugih sredina i usporediti ih međusobno.

"Programming is exciting, stimulating, fun and develops new ways of thinking". [10]

LITERATURA

[1] Dagiene V., Jevsikova T., Schulte C., Sentence S. Thota N.**,: A comparison of current trends within Computer Science teaching in school in Germany and the UK**, Informatics in Schools: Local Proceedings of the 6th International Conference ISSEP 2013, Selected PapersOldenburg, Germany, 2013

[2] 1Qureshi B., Ghalem O.: **Writing Computing Routines using Flowcharts**, College of Computer and Information Science, Prince Sultan University, Saudi Arabia, 2013

[3] Futschek, G.: **Algorithmic thinking: the key for understanding computer science.** In:Proceedings of the 2nd International Conference on Informatics in Secondary Schools: Evolution and Perspectives (ISSEP2006), 159–168., 2006.

[4] Bischof E., Mittermeir, R. **Informatik erleben**, Institut fur InformatikSysteme, Alpen-Adria University Klagenfurt, http://informatik-erleben.uni-klu.ac.at/, 2008.

[5] Fellows M., Bell T., Witten I. **Computer Science Unplugged**, http://csunplugged.org/ ., 2005.

[6] **Standard kvalifikacije tehničar za računalstvo**, Ministarstvo znanosti, obrazovanja i sporta, 2011.

[7] Dwyerz H., Hilly C., Carpenterz S., Harlowz D., Frankliny D.: **Identifying Elementary Students' Pre-Instructional Ability to Develop Algorithms and Step-By-Step Instructions,** Department of Computer Science Gevirtz Graduate School of Education, UC Santa Barbara, SIGCSE 2014.

[8] **Nastavni plan i program za osnovnu školu**, Ministarstvo znanosti, obrazovanja i športa, 2006.

[9] Babb J., Longenecker, Jr., H. E., Baugh J., Feinstein D. **Confronting the Issues of Programming In Information Systems Curricula: The Goal is Success,** Proceedings of the Information Systems Educators Conference, San Antonio, Texas, USA, 2013

[10] http://www.code.org (posjećeno veljača 2014.)

[11] S. Papert. **Constructionism: research reports and essays,** 1985-1990. Ablex Publishing Corporation, 1991..

iPad u nastavi matematike i fizike

Mirela Babić*
* OŠ Vežica, Rijeka, Hrvatska
mirelababic@me.com

Sažetak - U suvremenom svijetu informacijsko-komunikacijske tehnologije imaju velik utjecaj na svakodnevni život , pa ti alati poprimaju sve veću i važniju ulogu i u obrazovanju. Otvaraju se brojne mogućnosti i perspektive unutar obrazovnog procesa kao što su vizualizacija, simulacija, čitanje digitalnih knjiga, slušanje zvuka, provjera vlastitog izgovora na stranom jeziku. Za razliku od tradicionalne nastave u kojoj je učitelj sveznajući prenositelj informacija, on se pretvara u vodiča i trenera u svijetu informacija. Učenici pak postaju aktivni učesnici obrazovnog procesa, sve je veća interakcija u odnosu učitelj-učenik i učenik- učenik. Iskorak prema "školi budućnosti" učinjen je u OŠ Vežica u kojoj se već drugu godinu provodi nastava pomoću tableta 1:1. Početna istraživanja pokazuju da smo na dobrom putu.

I. UVOD

Prije nekoliko godina nove tehnologije ušle su kroz vrata naših učionica u vidu prezentacijskih slajdova i pretraživanja interneta. U našoj smo školi otišli korak dalje kada su prije dvije godine za učenike jednog odjeljenja 8. razreda pribavljeni iPad uređaji. Započelo se s edukacijom učitelja koja se kontinuirano i dalje provodi, oformljen je tim za tehničku podršku , škola je tehnički opremljena i bili smo spremni za nove izazove.

II. NASTAVA FIZIKE

Osobno sam učiteljica matematike i fizike, a rad u „iPad“ odjeljenju započela sam predavajući fiziku. U mojoj nastavi iPad nije isključivo nastavno sredstvo, već dodatna vrijednost i dobrodošla pomoć u nastavi. Pronalazila sam alate, odgojne-obrazovne aplikacije za koje sam smatrala da će mojim učenicima i meni olakšati rad i učiniti nastavu zanimljivijom i dinamičnijom. Neke od tih aplikacija su: Xperica, SerialBulbs, ParallelBulbs, RayOptics, PhysicalSci. Web stranice fizika-os.webs.com, stevespanglerscience.com omogućile su uvid u simulacije i demonstracije eksperimenata uvijek i ovdje (u školi ili kod kuće) . Dakle, potraga za informacijama odvija se u realnom vremenu.

Prilog1.

Priprema za izvođenje nastave

Nastavna cjelina: Valovi
Nastavna jedinica: Zvuk. Ton i šum.
Tip sata: obrada
Cilj nastavne jedinice: upoznati svojstva zvuka

Ishodi učenja:

- Objašnjavati nastajanje i rasprostiranje zvuka
- Nabrajati izvore zvuka
- Definirati područje čujnosti
- Razvijati sposobnost promatranja i zaključivanja
- Stjecati naviku istraživanja promatranjem

Oblici rada: frontalni, individualni

Nastavne metode: rasprava, demonstracija, praktični rad, rad na ipad uređaju

Nastavna sredstva i pomagala: ploča, kreda, ipad, aplikacije PhysicalSci Science By CPO

Ključni pojmovi: zvuk, ton, šum

Tijek nastavnog sata

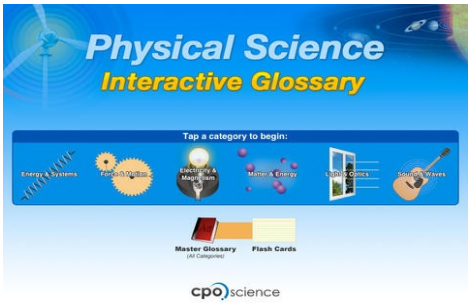
Uvodni dio sata:

- na početku sata učiteljica reproducira zvuk klavira koristeći aplikaciju PhysicalSci By CPO Science

https://itunes.apple.com/hr/app/physicalsci/



Koristi se dio:Sound and Waves



Središnji dio sata:

Učiteljica otvara problem pitanjima: Kako čujemo zvuk? Što je uzrok zvuka na klaviru? Kako ljudi proizvode zvuk dok pričaju?

Kroz raspravu dolazi se do zaključka da zvuk nastaje titranjem tijela u elastičnom sredstvu, a tijela koja proizvode zvuk su izvori zvuka. Zvuk je longitudinalni val frekvencije od 20-20000 Hz. Učenici zapisuju bilješke.

Učiteljica daje upute za otvaranje aplikacije PhysicalSci By CPO Science, uz pomoć koje učenici slušaju zvukove iz različitih izvora(padanje pera, ljudski šapat, razgovor dvoje ljudi, bušilica, rock koncert). Zvukove koje su učenici čuli ,pokušaju razvrstati na tonove i šumove. Definirati ton kao pravilno titranje , a šum kao nepravilno titranje tijela u zraku, te buku kao skup tonova i šumova.

Učenici se podijele u dvije skupine , jedna na mrežnim stranicama http://m.youtube.com/home pronalazi tonove, a druga šumove. Nakon 2 minute nekoliko učenika sa svog uređaja reproducira pred cijelim razredom pronađeni zvuk.

III. NASTAVA MATEMATIKE

Ove školske godine u našoj školi nastava s tabletima 1:1 odvija se u 5 odjeljenja, tri 5. razreda i dva 6. razreda. Predajem matematiku u dva 5.razreda. Okosnicu nastave čini društvena mreža za obrazovanje Edmodo. Pokazalo se da je Edmodo omogućio povezivanje s učenicima na njima vrlo prihvatljiv način. Društvena mreža je zatvorena jer se učenici priključuju preko koda koji im dodjeljuje učitelj. Kreirala sam grupe i podgrupe pojedinih odjeljenja, kako za redovnu tako i za dodatnu nastavu. Nastava putem Edmoda ubrzo je pokazala mnoge vrijednosti i prednosti: direktna razmjena informacija, materijala i obavijesti, kreiranje kvizova, praćenje rada pojedinog učenika, ažuriranje biblioteke, trenutna povratna informacija učenicima o njihovoj uspješnosti, povezivanje s drugim učiteljima diljem svijeta. Osim Edmoda u nastavi koristimo brojne matematičke aplikacije poput: Math Zombies, Geometry, Math Drill, Dragon Box, BasicFraction, SketchExplorer. Kroz neke od njih učenici kroz igru mogu usavršavati vještinu računanja, vizualno doživljavati razlomke, uvježbavati tehniku rješavanja jednadžbi i sl. Najveći dio aplikacija dozvoljava rad na više nivoa težine čime se proces učenja prilagođava individualnim potrebama učenika.

Prilog2.

Priprema za izvođenje nastave

Nastavna cjelina: razlomci

Nastavna jedinica: osnovno o razlomcima

Tip sata: uvježbavanje

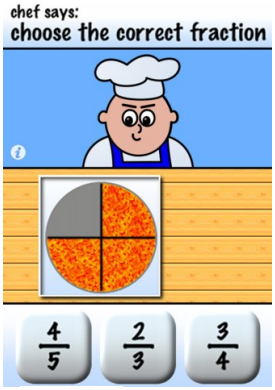
Cilj nastavne jedinice: usvojiti pojam razlomka

Ishodi učenja:

- Definirati pojam razlomka, brojnika , nazivnika i razlomačke crte
- Povezivati razlomak s računskom operacijom dijeljenja
- Prepoznati razlomak na konkretnom primjeru
- Prepoznati uporabu razlomaka u svakodnevici
- Razvijati pozitivan odnos prema matematici, odgovornost za svoj uspjeh i napredak te vijest o svojim matematičkim postignućima



Nastavna sredstva i pomagala: ploča, kreda, ipad, aplikacije Pizza Fractions 1By Brian West,Fraction lite by rana nir, zMath fraction by Zenitech, društvena mreža Edmodo. Tijek nastavnog sata



Uvodni dio sata:

Na početku sata ponove se osnovni pojmovi vezane uz pojam razlomka i povezuju se s uporabom u svakodnevici (npr. kruh, čokolada, pizza)

Učiteljica daje upute za otvaranje aplikacije Pizza Fractions 1By Brian West



(https://itunes.apple.com/hr/app/pizza-fractions)



Učenici dijelove pizze povezuju s odgovarajućim razlomkom i na zanimljiv način uvježbavaju pojam razlomka.



Središnji dio sata:

Za uvježbavanje u središnjem dijelu sata učenici prema

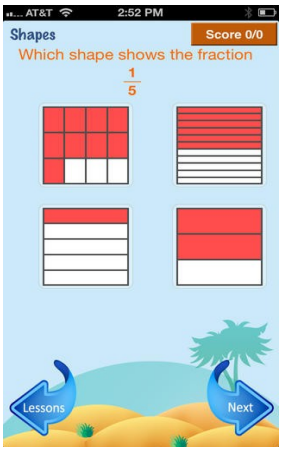
uputama učiteljice koriste aplikacije: Fraction lite by rana nir

(https://itunes.apple.com/hr/app/fraction-lite/

IzMath fraction by Zenitech)

(https://itunes.apple.com/hr/app/zmath-grade-3-fractions-lite/)

U ovim aplikacijama učenici kroz različite oblike (krug, kvadrat, mnogokut…) uvježbavaju pojam razlomka na više razina složenosti. Svaki učenik samostalno bira razinu na kojoj će započeti rad. Ukoliko se neka razina pokaže preteškom za učenika, vraća se na nižu.



Završni dio sata:

Razina usvojenosti sadržaja provjerava se kroz petminutnu pismenu vježbu koju učenici rješavaju preko aplikacije Edmodo: https://itunes.apple.com/hr/app/edmodo/id

Aplikacija pruža mogućnost brze povratne informacije koju učenici dobivaju u roku od nekoliko sekunda.

IV. ZAKLJUČAK

Suvremeno obrazovanje postalo je nezamislivo bez primjene informacijsko-komunikacijske tehnologije. Generacije naših današnjih učenika postavljaju pred nas učitelje izazove s kojima se moramo i želimo uspješno nositi na zadovoljstvo svih sudionika odgojno obrazovnog procesa. Pred nama učiteljima je ogromna odgovornost, ali i osjećaj da nismo bespomoćni , da postoje načini i mogućnosti za promjene koje od nas iziskuje galopirajući razvoj tehnologije. Lao Ce, kineski filozof iz 6.st.pr.Kr. rekao je:"Učenje je kao veslanje uzvodno, čim prestane, odmah se kreće unatrag." Budućnost će od naših učenika tražiti da ubrzano "veslaju uzvodno", a mi ih za to moramo pripremiti.

LITERATURA

[1] https://itunes.apple.com/hr/app/physicalsci/

[2] https://itunes.apple.com/hr/app/pizzafractions

[3] https://itunes.apple.com/hr/app/zmath-grade-3-fractions-lite/

[4] https://itunes.apple.com/hr/app/edmodo/id

#codeEU

Ivana Ružić
I. osnovna škola Čakovec,
Ulica kralja Tomislava 43, Čakovec, Hrvatska
ivana.ruzic@skole.hr

#codeEU je inicijativa pokrenuta od Europske komisije sa ciljem da se što više ljudi diljem Europe potakne da uče programirati. Programiranje je kreativno područje koje svima nama nudi mnogobrojne mogućnosti da ostvarimo svoje ideje i zamisli. Inicijativa je usmjerena na djecu i mlade, nezaposlene osobe ali i na sve nas nudeći svima nama mogućnost da postanemo aktivni sudionici u kreiranju budućnosti čovječanstva ili osiguramo vlastitu egzistenciju. Ovaj rad opisuje inicijativu koja ruši državne granice, jezične barijere ili bilo koji oblik diskrimacije, ukazujući svima na potrebe zajednice u bližoj budućnosti, nudeći mogućnost izbora, priliku za usvajanje novih znanja i vještina iz područja ICT te ukazuje na važnost cjeloživotnog učenja.

I. UVOD

#codeEU je inicijativa koju je pokrenula Neeile Kroes uz podršku DG Connect u Europskoj komisiji sa ciljem da se ljude potakne da uče programirati, da se svima nama pruži prilika da ostvarimo vlastite ideje i promijenimo svijet. Tehnologija oblikuje naše živote ali manjima odlučuje što ćemo i kako koristiti. Mi možemo učiniti puno više od korištenja, dijeljenja i „lajkanja“. Mi možemo ostvariti naše „lude“ ideje i izgraditi usluge i stvari koje će donijeti radost drugima. Nikada nije bilo lakše napraviti vlastitu aplikaciju, izgraditi vlastitog robota ili izmisliti leteće automobile, zašto ne pokušati? To nije lagan put ali to je putovanje puno kreativnih izazova, zajednica koja pruža podršku i mnogo zabave. Jeste li spremni prihvatiti izazov i postati stvaratelj?



Slika 1. Logo inicijative #codeEU

II. ZAŠTO PROGRAMIRANJE?

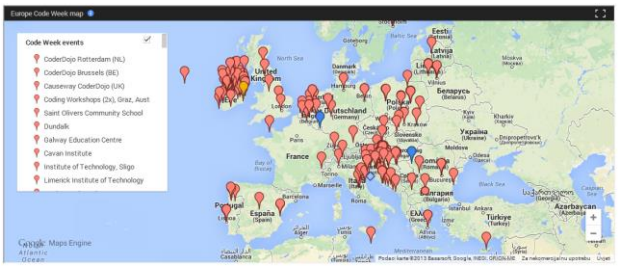
Programiranje je svuda oko nas. Svaka interakcija između čovjeka i računala je podržana programskim kodom. Programiranje je puno više od mnoštva različitih alata, to je način gledanja na svijet: razdijeliti zadatke u sve manje podskupine. To je zabavno i kreativno područje ispunjeno velikim ljudima i strastvenom zajednicom. Treba li spomenuti da će 2015. u EU nedostajati 900.000 ICT stručnjaka?

Programiranje zaslužuje veću pažnju javnosti i stoga je organiziran prvi Europski tjedan programiranja od 25. do 30. studenoga 2013. sa željom da se organizira niz prigodnih događaja diljem Europe - sve sa ciljem upoznavanja građana s programiranjem kao i učeći ih ovu vještinu.

Cilj inicijative je poboljšati programiranje na europskoj razini, motivirati ljude da počnu učiti programirati te povezati pojedince, skupine, ustanove, organizacije i tvrtke koje su spremni pomoći ljudima doživjeti radost i relevantnost programiranja. Između ostalog cilj inicijative je i pokretanje velike lavine koja će se vidjeti izvan naših uobičajenih krugova, lavine toliko velike i snažne da ne može biti ignorirana od strane medija, bez obzira na najnovije celebrity tračeve. Znamo koliko je zabavno graditi stvari, a mi smo spremni dijeliti našu strast. Programiranje je zabavno. Kreativno. Njime se bavi mnogo poznatih ljudi i odlikuje se snažnom zajednicom. Jeste li se već okušali u usvajanju znanja i vještina programiranja, jeste li pomogli ljudima oko sebe da upoznaju ljepotu programskog koda? Uvijek je zabavno graditi, stvarati.

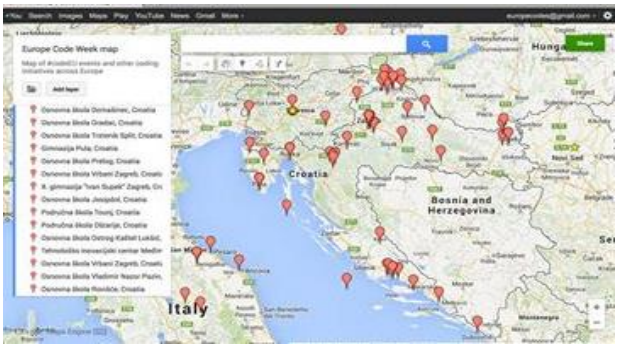
III. AKTIVNOSTI U EUROPSKOM TJEDNU PROGRAMIRANJA

Između 25. i 30. studenog 2013. godine diljem Europe organizirano je preko 300 različitih prigodnih događanja u kojima se okupilo nekoliko desetaka tisuća sudionika. Time je dokazano da je Europski tjedan programiranja i službeno postigao veliki uspjeh, a programiranje je na svojevrstan način ujedinilo Europu. Najveći interes zabilježen je u Irskoj, a Hrvatska kao najmlađa članica Europske unije po brojnosti događanja i sudionika zauzela je drugo mjesto. Karta svih prijavljenih prigodnih događanja dostupna je na: http://codeweek.eu/map/



Slika 2. Karta inicijative #codeEU

U Hrvatskoj su se u ovoj inicijativi okupile ukupno 64 osnovne i srednje škole, fakulteti i udruge sa gotovo 4000 sudionika.



Slika 3. Karta inicijative #codeEU u Hrvatskoj

Održane su različite radionice programiranja za djecu i mlade svih uzrasta od osnovnoškolskog do akademskog obrazovanja. Sudionici su imali prilike upoznati programske jezike C++, Phyton, Logo, Small Basic, HTML, CSS, JavaScript, Game Maker, Kodu, Scratch,… ispisane su tisuće linija koda, radošću stvaranja ozarila su se mnoga lica, ostvarene su mnoge suradnje i vjerujemo da se zakotrljao veliki kotač koji će mnogim sudionicima inicijative ili onima koji će te to postati do slijedeće godine i novog Europskog tjedna programiranja donijeti mnogo novih znanja i vještina.



Slika 4. Radionica programiranje Lego robota u I. OŠ Čakovec

Sadržaji radionica i aktivnosti bili su zaista raznoliki i prilagođeni uzrastu i predznanju polaznika radionica, osmišljeni na način da se polaznike dodatno motivira i potakne na kontinuirano učenje i usvajanje novih vještina, samostalno istraživanje, uključivanje u obrazovnu online zajednice istih i/ili sličnih interesa.

Poseban interes pokazan je za programiranje robota što uključuje sastavljanje robota od kockica, motora i senzora, na računalu izradu programa za upravljanje robotom, prijenos programa usb sučeljem s računala na robot te pokretanje robota tj. gledanje njegovog izvršavanja postavljenih naredbi.

S posebnom pažnjom pripremale su se aktivnosti radionica programiranja za najmlađe polaznike, učenike prvih i drugih razreda osnovnih škola. Tynker je nova računalna platforma dizajnirana u Silicijskoj dolini namijenjena poučavanju djece vještinama programiranja na zabavan i maštovit način. To je jednostavna platforma smještena u oblaku koja ne traži nikakve posebne zahtjeve, a temelji se na otvorenim web standardima poput JavaScripta, HTML5 i CSS3. Tynker omogućuje djeci da uče koncept računalne znanosti kroz otkrića, kreativne aktivnosti i eksperimentiranja. Najmlađi sudionici Europskog tjedna programiranja usvojili su logički slijed događaja i modeliranje stvarne situacije, razvili su svoje algoritamsko razmišljanje i sposobnosti dizajna.



Slika 5. Radionica programiranja u Gospodarskoj školi Buje

Vjerujemo da računalno razmišljanje i računalno programiranje treba biti dio jezgre nastavnog plana i programa u obrazovanju sa ciljem da svako dijete izgradi čvrste temelje STEM-a (znanosti, tehnologije, inženjerstva i matematike) kako bi ih pripremili za karijere 21. stoljeća .

Interes za uključivanje u ovu inicijativu iskazale su mnoge fizičke i pravne osobe, stručnjaci iz područja ICT koji su volonterski održali niz radionica, predavanja i sličnih aktivnosti ili su donacijama potpomogli održavanje aktivnosti. Uključivanjem stručnjaka iz privrede u inicijativu pridonijelo je većem interesu i motiviranosti učenika za sudjelovanje u aktivnostima, a jasno je pokazan interes i potreba privrede za ICT stručnjacima.

Inicijativa je u Hrvatskoj pridobila veliki interes javnosti, regionalnih i državnih medija, šire obrazovne zajednice i lokalnih čelnika što je doprinjelo još većem značaju i popularizaciji inicijative i samog programiranja. Aktivnosti vezane uz #codeEU inicijalizirale su osnivanje prvog Coderdojo kluba u Hrvatskoj sa sjedištem u I. OŠ Čakovec, organiziranje ljetnih radionica programiranja za osnovnoškolce u odmaralištu grada Čakovca u Dramlju. Razvoj svijesti o važnosti programiranja te poticanje na kontinuirano usvajanje novih znanja i vještina iz područja

programiranja na europskoj razini nastavlja se pripremama i organiziranjem drugog Europskog tjedna programiranja koji će se obilježiti od 11. do 17. listopada 2014.

IV. ZAKLJUČAK

Međunarodne inicijative i projekti mogu obogatiti cjelokupan nastavni proces, povećati motiviranost učenika za usvajanje novih znanja i vještina. Ukoliko se inicijativa predstavi i prepozna kao korisna i u svakodnevnom životu primijenjiva, interes i motiviranost za sudjelovanje u projektnim aktivnostima je veći. ICT je nezaobilazna u svakoj sferi života, a upoznavanje i savladavanje novih programskih alata ovisno o interesima sudionika osigurava zadovoljstvo učenika i veću angažiranost.

LITERATURA

[1] Bates, A. W.: Upravljanje tehnološkim promjenama: Strategije za voditelje visokih učilišta.. Zagreb, Hrvatska: CARNet/Benja, 2004.

[2] Jenkins, N.: A Project Management Primer or a guide on how to make projects work, 2005.

[3] http://codeweek.eu/code/

[4] http://www.tynker.com/

[5] http://www.youtube.com/watch?v=MfxWn4YHXt4

Prijedlog entitetsko-relacijskog modela podataka registara Hrvatskog kvalifikacijskog okvira

M. Balković *, N. Bačun Lebarović ** i H. Balen***
* *Visoko učilište Algebra – visoka škola za primijenjeno računarstvo*
Ilica 242, Zagreb, Hrvatska
e-mail: mislav.balkovic@racunarstvo.hr
** *Visoko učilište Algebra – visoka škola za primijenjeno računarstvo*
Ilica 242, Zagreb, Hrvatska
e-mail: nives.bacunlebarovic@algebra.hr
*** *Visoko učilište Algebra – visoka škola za primijenjeno računarstvo*
Ilica 242, Zagreb, Hrvatska
e-mail: hrvoje.balen@racunarstvo.hr

Sažetak - Republika Hrvatska je jedna od 146 država u svijetu koje su u nekoj od faza uvođenja nacionalnog kvalifikacijskog okvira. S tim ciljem je u 2013 donesen Zakon o Hrvatskom kvalifikacijskom okviru, te su u izradi dva pravilnika koji će regulirati područja priznavanja informalno i neformalno stečenih znanja i vještina te ustroj registara Hrvatskog kvalifikacijskog okvira (HKO). Ovim člankom se sukladno Zakonu o HKO i objavljenom prijedlogu pravilnika o registrima HKO nastoji predložiti logički entitetsko-relacijski model podataka registara HKO koji će uključiti opis atributa koji opisuju skupine ishoda učenja, standarde zanimanja, standarde kvalifikacija te obrazovne programe kao i programe za vrednovanje neformalnog i informalnog učenja. Predloženi logički relacijski model podataka uzet će u obzir odnose između pojedinih entiteta, ali isto tako i predviđene procedure izmjena i dopuna podataka te praćenje verzija, kako ih predlažu zakonska i podzakonska rješenja. Članak stoga daje moguću podlogu za razvoj budućeg informacijskog sustav koji će predstavljati potporu punoj implementaciji Hrvatskog kvalifikacijskog okvira, sukladno aktualnom Zakonskom rješenju.

I. UVOD

Poticaj razvoju nacionalnih kvalifikacijskih okvira u tranzicijskim državama i državama u razvoju često je uvjetovan potrebom povećanja kvalitete ljudskog kapitala i željom za povećanjem konkurentnosti. S druge strane, potaknuto aktualnom gospodarskom krizom i recesijom koja je pogodila brojna svjetska gospodarstva, uvođenje nacionalnih kvalifikacijskih okvira je postalo jedan od važnih alata koji doprinose političkom cilju povećanja zaposlenosti i zapošljivosti. Kao rezultat navedenih tendencija nacionalni kvalifikacijski okvir je do sada uvelo ili uvodi 146 država [1], među kojima je i Republika Hrvatska. Iako je ideja o uvođenju Hrvatskog kvalifikacijskog okvira (HKO) bila prisutna i ranije, službeno je proces u Hrvatskoj započeo u ožujku 2006 osnivanjem Povjerenstva za uvođenje HKO od strane Ministarstva znanosti, obrazovanja i sporta. U srpnju 2007 Vlada Republike Hrvatske prihvatila je polazne osnove HKO [2] da bi iste godine bilo osnovano i povjerenstvo Vlade RH za razvoj HKO. Konačno je u svibnju 2010 imenovano i novo povjerenstvo Vlade RH za uvođenje HKO koje je dovršilo rad na Zakonu o HKO [3] koji je Hrvatski Sabor donio u veljači 2013. Predmetnim Zakonom i pripadajućim Pravilnikom o registru HKO [4], koji je u trenutku pripreme ovog članka dostupan u javnoj raspravi, definirani su osnovni elementi HKO i njihovi međusobni odnosi. Dodatno, člankom 11 Zakona [3] je propisana izgradnja informacijskog sustava za vođenje registara HKO, dok Pravilnik [4] propisuje elemente temeljnih dokumenata koji će biti pohranjivani u registre i time definira svojstva budućeg modela podataka. Pored navedenog pravilnika, Zakon [3] predviđa i donošenje pravilnika o priznavanju i vrednovanju neformalnog i informalnog učenja koji će postaviti dodatne zahtjeve na model podataka i na pripadajući informacijski sustav.

Imajući u vidu navedenu zakonsku regulativu te podzakonske akte koji su još u nastajanju, kao i europsku ESCO taksonomiju [5] i pripadajući informacijski sustav koji je za nju razvijen, u nastavku će biti predložen mogući logički podatkovni model registara HKO koji bi omogućio buduću integraciju informacijskog sustava HKO sa ESCO i EURES [6] platformama.

II. ELEMENTI SUSTAVA KVALIFIKACIJA

Sukladno članku 2 Zakona [3], HKO se uspostavlja kao instrument uređenja sustava kvalifikacija u Republici Hrvatskoj koji osigurava jasnoću, pristupanje stjecanju, utemeljeno stjecanje, prohodnost i kvalitetu kvalifikacija kao i povezivanje razina kvalifikacija u Republici Hrvatskoj s razinama kvalifikacija u Europskom kvalifikacijskom okviru (EQF) [7] te posredno s razinama kvalifikacija kvalifikacijskih okvira u drugim zemljama. Kako bi ostvario navedenu funkciju uređenja sustava kvalifikacija, HKO definira osnovne elemente sustava, pri čemu se koriste i kratice kako slijedi:

- **Standard zanimanja** (SZ): popis svih poslova koje pojedinac obavlja u određenom zanimanju i popis kompetencija potrebnih za njihovo uspješno obavljanje.
- Metodološka podloga za izradu standarda zanimanja: predstavlja skup dokumenata i postupaka njihove izrade koji osiguravaju dokaze strateške, sektorske i analitičke utemeljenosti za izradu standarda zanimanja. Dio metodologije predstavlja i upitnik za istraživanje potrebnih kompetencija po zanimanjima i radnim mjestima [8].

- Kompetencije: znanja i vještine te pripadajuća samostalnost i odgovornost. Grupirane su u skupove.
- **Standard kvalifikacije** (SK): sadržaj i struktura određene kvalifikacije koja uključuje sve podatke koji su potrebni za određivanje razine, obujma i profila kvalifikacije te podatke koji su potrebni za osiguravanje i unapređenje kvalitete standarda kvalifikacije. Sastoji se od skupova ishoda učenja.
- **Skup ishoda učenja** (SIU): najmanji cjelovit skup povezanih ishoda učenja iste razine, obujma i profila.
- Ishodi učenja : kompetencije koje je osoba stekla učenjem i dokazala nakon postupka učenja.
- Sektor: skupina kvalifikacija jednog obrazovnog područja i zanimanja koja koriste ishode učenja tih kvalifikacija na radnim mjestima.
- Program za stjecanje kvalifikacija: postupci provedbe obrazovanja (kurikulum, studijski program) koji dovode do stjecanja ishoda učenja koji su navedeni u standardu predmetne kvalifikacije.
- Program za stjecanje i vrednovanje skupova ishoda učenja: postupci provedbe obrazovanja (kurikulum, studijski program) i postupci provedbe provjere znanja koji dovode do dodjele potvrde o posjedovanju skupa ishoda učenja.
- Program za vrednovanje skupova ishoda učenja: postupci provedbe provjere znanja koji dovode do dodjele potvrde o posjedovanju skupa ishoda učenja.

Navedeni elementi sustava kvalifikacija definirani su sa svojim svojstvima (atributima) i predstavljaju osnovu za izradu logičkog podatkovnog modela HKO. U naravi se radi o dokumentima ili skupinama dokumenata koje osiguravaju jasnoću, pristupanje stjecanju, utemeljeno stjecanje, prohodnost i kvalitetu kvalifikacija. Pored navedenih dokumenata, HKO uvodi i upravljačku strukturu koja se sastoji od niza organizacijskih jedinica i pravnih osoba sa propisanim odgovornostima i ovlastima. Upravljačka struktura HKO se realizira kroz:

- Ministarstvo znanosti, obrazovanja i sporta te Ministarstvo rada i mirovinskog sustava i Ministarstvo nadležno za regionalni razvoj. Navedena ministarstva obavljaju poslove koordinacije primjene i razvoja HKO-a.
- Nacionalno vijeće za razvoj ljudskih potencijala koje je središnje strateško tijelo Republike Hrvatske za razvoj HKO-a.
- Sektorska vijeća koja su dionički sastavljena savjetodavna i stručna tijela koja skrbe o razvoju ljudskih potencijala u skladu s potrebama tržišta rada unutar pojedinih sektora. Ustrojiti će se ukupno 13 sektorskih vijeća zaduženih za strukovna i 12 vijeća zaduženih za opća područja te jedno interdisciplinarno vijeće, ukupno 26.
- Stručne agencije koje djeluju u području obrazovanja bit će zadužene za provođenje vanjskog osiguravanja kvalitete HKO-a, uz nadležnost koja ovisi o obrazovnoj razini. Stručne agencije zadržavaju i dosadašnje zakonske ovlasti vezane uz inicijalnu akreditaciju programa.

Pored navedenih institucija i organizacija koje su dio organizacijsko – upravljačke strukture, u kvalifikacijskom sustavu će djelovati i druge pravne osobe koje omogućuju primjenu HKO. Važnije su:

- Akreditirane obrazovne ustanove koje svoj program za stjecanje kvalifikacije žele kroz inicijalnu akreditaciju i kasniji trajni nadzor kvalitete (npr. kroz reakreditaciju) uvrstiti u HKO.
- Akreditirane pravne osobe koje svoj program za stjecanje ili vrednovanje skupa ishoda učenja žele kroz inicijalnu akreditaciju i kasniji trajni nadzor kvalitete (npr. kroz reakreditaciju) uvrstiti u HKO.
- Pravne osobe koje će sukladno akreditiranim programima provoditi vrednovanje neformalno i informalno stečenih ishoda učenja, a koje će također zadovoljiti uvjete inicijalne akreditacije i trajnog nadzora kvalitete i time biti uvrštene u sustav HKO.

Kvalifikacijski okvir se izgrađuje kao instrument koji će građanima omogućiti jasnoću i kvalitetu stjecanja kvalifikacija, što će posredno doprinijeti između ostalog i poslodavcima te društvu u cjelini. U tom smislu ključni dionici i korisnici kvalifikacijskog okvira, koji nisu do sada navedeni i nisu formalno pozicionirani kroz zakonsku regulativu su i građani zainteresirani za stjecanje kvalifikacija ili certifikata o posjedovanju skupova ishoda učenja, ali isto tako i poslodavci, karijerni savjetnici, agencije za zapošljavanje, roditelji i društvo u cjelini.

Hrvatski kvalifikacijski okvir nije obvezni instrument uređenja sustava kvalifikacija što znači da će uz programe koji su uneseni u kvalifikacijski okvir i koji odgovaraju SK moći postojati i drugi programi koji ne zadovoljavaju te uvjete. Tu se ne radi samo o neformalnim programima nego i o dijelu formalnih koje danas postoje temeljem aktualnih propisa.

III. LOGIČKA STRUKTURA HKO

HKO svoju opisanu funkciju ostvaruje kroz uvođenje razmjerno složenog mehanizma odobravanja i osiguravanja kvalitete standarda zanimanja, skupova ishoda učenja i standarda kvalifikacije, što predstavlja podlogu za inicijalnu akreditaciju programa (nekog od tri navedene vrste). Ključna promjena paradigme koju kvalifikacijski sustav uvodi jest javno vođenje registara standarda zanimanja i kvalifikacija koji predstavljaju jasne i nedvosmislene opise ishoda učenja odnosno kompetencija koje su nužne za rad u određenom zanimanju ili su uključene u pojedinu kvalifikaciju, te inzistiranje na potpunoj implementaciji tih ishoda učenja u programe. Analizirajući postojeći način izrade i inicijalne akreditacije programa na nekoj razini obrazovanja, obično nalazimo nekoliko faza koje uglavnom nisu regulirane, iako često koriste propisane obrasce s jasno navedenim atributima:

- Istraživanje potreba društva kao podloga za definiranje sadržaja programa izvodi se obično kroz ankete poslodavaca, fokus grupe, praćenje razvoja područja i tehnološkog napretka, smjernice za sastavljanje obrazovnih programa koje mogu davati strukovna udruženja i drugo.

- Potrebe identificirane u prethodnoj fazi obično služe kao podloga za definiranje obrazovnih ciljeva, ishoda učenja i sadržaja obrazovnog programa (obično na razini nastavnog plana).
- Konačno se u fazi izvedbe programa u nekoj konkretnoj ustanovi i od strane konkretnog nastavnika planira se izvedbeni plan koji u sebi obično ima i raspored odvijanja nastave, potrebnu literaturu, materijalno - tehničke uvjete, popis tema koje su uključene u svaku nastavnu cjelinu itd.

Model inicijalne akreditacije programa unutar HKO značajno je složeniji jer osim propisanih formi pojedinih dokumenata uvodi i osiguravanje kvalitete u svim ključnim točkama provedbe postupka te u velikoj mjeri uključuje neovisna vanjska tijela (ponekad dionički sastavljena) ovlaštena procjenjivati kvalitetu predložene dokumentacije.

U logičkom smislu postupak započinje u trenutku kada zainteresirana pravna ili fizička osoba odnosno tijelo državne uprave koji za to imaju opravdani interes žele izvršiti upis standarda zanimanja, standarda kvalifikacije ili skupa ishoda učenja u registar. Ukoliko radi usporedbe sa već opisanim mehanizmom akreditacije programa pretpostavimo da se radi o zahtjevu obrazovne ustanove koja ima krajnji cilj akreditirati program za stjecanje kvalifikacije (obrazovni ili studijski program) postupak bi bio slijedeći:

A. Podloge za izradu standarda zanimanja

Budući da sukladno članku 12 Pravilnika [4] predlagatelj mora dokazati utemeljenost novog standarda zanimanja odnosno skupova kompetencija obrazloženjem strateške, sektorske i analitičke utemeljenosti zahtjeva, on ima mogućnost koristiti već postojeće stručne podloge za izradu standarda zanimanja u skladu s metodologijom koju je razvilo Ministarstvo nadležno za rad ili te podloge može samostalno razvijati. Stručne podloge obuhvaćaju Profile sektora i analizu potrebe i regulacije tržišta rada za ona zanimanja za koja se predlaže standard, a ako ih predlagatelj samostalno razvija one se baziraju na upotrebi već postojećeg on-line upitnika za poslodavce o potrebnim kompetencijama po zanimanjima te analizom i grupiranjem njihovih rezultata u formi tablice kompetencija. U dijelu formalne provjere predanog zahtjeva standarda zanimanja, Ministarstvo nadležno za rad provodi i vrednovanje kvalitete stručnih podloga.

B. Izrada standarda zanimanja

Standard zanimanja je nužni dio dokumentacije u postupku inicijalne akreditacije programa isključivo ukoliko se program, i prije njega standard kvalifikacije, akreditira za potrebe tržišta rada. Standard zanimanja, između ostalog, sadrži i skupove kompetencija potrebnih za rad na odgovarajućim radnim mjestima a koje su razvijene temeljem stručnih podloga. Provjera predloženog SZ izvodi se u dva stupnja. Prvi stupanj čini formalna provjera predane dokumentacije koju provodi Ministarstvo nadležno za rad. Temeljem odabira odgovarajućeg sektora, zahtjev se upućuje sektorskom vijeću/vijećima koje u području svoje nadležnosti provodi stručno vrednovanje uključujući i ocjenu opravdanosti pokretanja postupka za upis u Registar. Sektorsko vijeće je neovisni element osiguravanja kvalitete koji strukturom svojih članova osigurava dionički utjecaj ali i stručni razvoj sektora [9]. Pozitivno mišljenje sektorskog vijeća uvjet je upisa SZ u podregistar SZ, koji se u formalnom smislu provodi odlukom Ministra. Republika Hrvatska do uvođenja HKO nije koristila SZ, osim dokumenta sličnog sadržaja koji su od 2009 prisutni u srednjoškolskom strukovnom obrazovanju.

C. Izrada standarda kvalifikacije i skupova ishoda učenja

Standard kvalifikacije izrađuje se i predlaže sa tri moguće svrhe: za tržište rada, za nastavak obrazovanja i za ostale potreba pojedinaca i društva. Samo je u prvom slučaju SZ obvezni element za razvoj SK, dok u druga dva slučaja opravdanost predloženog SK treba dokazati na drugi način. SK se u osnovi sastoji od popisa SIU koji se u slučaju kvalifikacije za tržište rada izrađuju temeljem postojećeg SZ i tamo navedenih skupova kompetencija. Kako bi osigurao jasnoću, transparentnost i upravljivost u sustavu kvalifikacija HKO propisuje obvezu korištenja već postojećih SIU za izgradnju novih kvalifikacija kada je to moguće. Konkretno, ukoliko se izrađuje SK temeljem SZ koji sadrži više skupova kompetencija, kod izrade SK je prvo nužno za svaki od skupova kompetencija provjeriti postoji li odgovarajući SIU u javnom registru SIU. Ukoliko postoji nužno je koristiti taj postojeći kako bi se nove kvalifikacije izgrađivale od jednakih gravidnih elemenata kao i one već postojeće, osiguravajući na taj način modularnost i prenosivost ishoda učenja za građane. Za one skupove kompetencija za koje ne postoji odgovarajući SIU predlaže se dodavanje novog SIU u podregistar SIU. Svaki SIU raspoređuje se u odgovarajući od 26 sektora i u odgovarajuću od 8 razina i 4 podrazine HKO-a [3], a njegov upis u podregistar SIU izvodi se bilo pojedinačnim zahtjevom ili skupno u trenutku unosa novog SK u podregistar SK. Kao i kod SZ, pozitivno stručno vrednovanje sektorskog vijeća podloga je Ministru za upis SIU ili SK u odgovarajuće podregistre HKO.

Pojedini SIU u sebi sadrži, između ostalog, zapise 5-10 ishoda učenja te primjere njihovog vrednovanja (provjere znanja) i on može imati najmanje jedan kreditni bod (npr. ECTS). Cilj uvođenja kvalifikacijskog okvira je između ostalog svaki SIU koji izgrađuje bilo koji SK koristiti i za druge SK ako je primjenjiv. Dodatno, ako se neki SIU radi za kvalifikaciju koja je vezana za neki SZ, nužno je postojanje logičke poveznice između SIU i odgovarajućeg skupa kompetencija u SZ članak 29 Pravilnika [4]. Treba naglasiti da ne postoji obveza da svi ishodi učenja u pojedinom SIU odgovaraju u potpunosti kompetencijama u povezanom skupu kompetencija, nego SIU može biti i širi od skupa kompetencija. Usklađenost SIU i skupa kompetencija određuje sektorsko vijeće u postupku stručnog vrednovanja SK ili SIU.

D. Inicijalna akreditacija programa sukladno metodologiji HKO

Inicijalna akreditacija programa se do uvođenja HKO izvodila temeljem odgovarajućih normativnih akata specifičnih prije svega za razinu obrazovanja i vrstu programa koji se akreditira. U tom postupku su uglavnom

korišteni ulazni (engl. Input) parametri odnosno izvodila se provjera zadovoljenja formalnih uvjeta u smislu potrebnog prostora, opreme, nastavnog kadra i slično. Kod pojedinih programa postojalo je i stručno vrednovanje sadržaja programa, ali ono nije bilo vezano uz konkretan i javno objavljen standard koji odgovara točno tom programu i koji program mora zadovoljiti u smislu usklađenosti ishoda učenja. Kako bi osigurao kvalitetu, transparentnost i jasnoću kvalifikacija, HKO proširuje postupak inicijalne akreditacije programa uvođenjem dodatnih elemenata koji se odnose prije svega na usklađenost predloženog programa odnosno ishoda učenja koje izgrađuje sa izabranim SK ili SIU. Konkretno, sukladno članku 34 Pravilnika [4] propisuje se da je preduvjet za upis programa u Registar postojanje odgovarajućeg SK, odnosno odgovarajućih SIU upisanih u Registar i pozitivno mišljenje nadležne agencije o provedenom vrednovanju usklađenosti prijedloga programa s odgovarajućim SK, odnosno s odgovarajućim SIU. Programi za stjecanje kvalifikacija se upisuju temeljem odluke Ministra u podregistar SK, dok se programi za stjecanje i vrednovanje skupova ishoda učenja i programi za vrednovanje skupova ishoda učenja upisuju u podregistar SIU. U postupku vrednovanja nadležna stručna agencija u području obrazovanja više nema obvezu samo skrbiti o generalnim „input“ parametrima propisanim postojećom regulativom, nego joj je dana dodatna obveza provjeriti usklađenost ishoda učenja ugrađenih u program sa pripadnim SK te provjeriti zadovoljenje specifičnih zahtjeva za stjecanje i vrednovanje navedenih u SK i SIU. Predlagač koji program želi akreditirati prema HKO slobodan je, kao i prije HKO, samostalno oblikovati program u smislu njegovog izvedbenog plana, rasporeda izvođenja, literature i slično, ali je dužan u njega ugraditi i sve ishode učenja uključene u pripadni SK, te dodatno zadovoljiti uvjete za stjecanje i vrednovanje propisane unutar SK i svakog njegovog SIU. Postupak inicijalne akreditacije programa koji se predlažu radi nastavka obrazovanja ili ostalih potreba pojedinaca i društva biti će nešto kraći i jednostavniji jer on za zahtjeva izradu stručnih podloga niti standarda zanimanja. Ipak, kod primjerice kvalifikacija koje se predlažu za nastavak obrazovanja bit će nužno identificirati odgovarajuće takve kvalifikacije u HKO.

Logička organizacija HKO i veze pojedinih elemenata opisane su unutar ovog poglavlja, dok je fizička implementacija registra HKO predviđena kroz izgradnju ukupno tri podregistra kako slijedi:

- Podregistar SIU koji sadrži: SIU, programe za stjecanje i vrednovanje SIU, programe za vrednovanje SIU, popis ovlaštenih pravnih i fizičkih osoba za dodjelu potvrde o stečenim SIU i drugo;
- Podregistar SZ koji sadrži: SZ, skupove potrebnih kompetencija za odgovarajuća zanimanja i drugo;
- Podregistar SK koji sadrži: SK, programe za stjecanje kvalifikacija, popis ovlaštenih pravnih osoba za dodjelu isprave o stečenoj kvalifikaciji i drugo.

IV. ER MODEL PODATAKA HKO

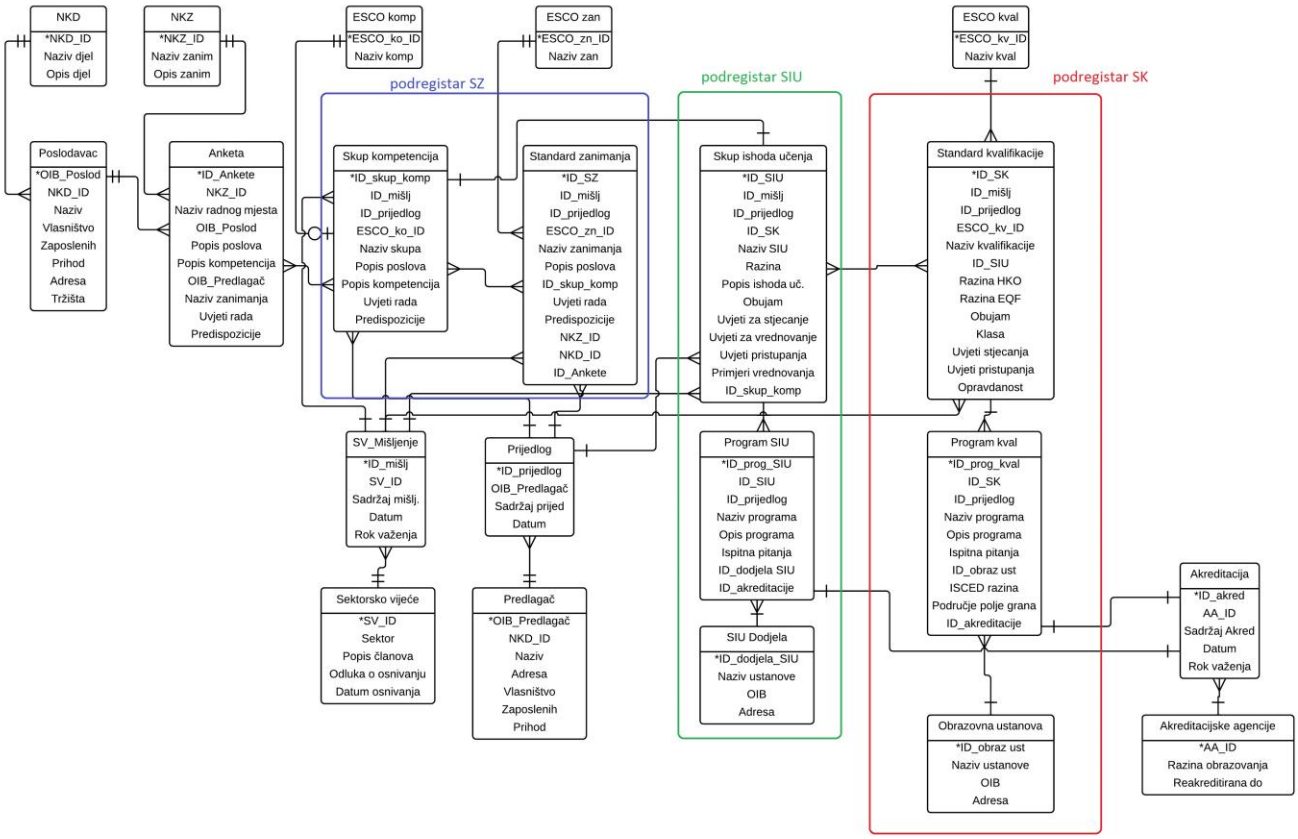
A. *Metodologija izrade*

Za prikaz podatkovnog modela biti će korišten entitetsko – relacijski dijagram (ERD) temeljen na Crow's Foot notaciji [10] koji počiva na analizi potreba definiranoj zakonskim rješenjima i u osnovi opisanoj u prethodnim poglavljima. Potporu procesu modeliranja entiteta i njihovih veza dali su propisani obrasci za predlaganje SZ, SIU i SK jer vrlo jasno navode set svojstava odnosno atributa koji će biti ugrađeni u dijagram. Budući da je dio zakonske regulative još uvijek u nastajanju, razvoj će biti ograničen samo na logički model podataka dok bi se implementacija fizičkog modela trebala realizirati u trenutku izrade specifikacije informacijskog sustava HKO. Također, iako su Zakonom [3] definirana tri podregistra unutar Registra HKO, predloženi model podataka biti će izgrađen na temelju šireg shvaćanja HKO koji u sebi uključuje i moguće buduće vanjske reference sustava (npr. ESCO ili registar ljudskih potencijala), te vodeći računa o osiguravanju kvalitete i institucijama koje su za njega zadužene. To konkretno znači da on neće biti ograničen samo na tri navedena podregistra nego da će nastojati prepoznati i povezati sve entitete ključne za uspostavu HKO.

B. *Popis entiteta*

Entiteti prikazani ER modelom dani su tablicom 1.

Oznaka	***Opis***
Standard zanimanja	Standard zanimanja
Skup kompetencija	Skup kompetencija
SIU	Skup ishoda učenja
Standard kvalifikac	Standard kvalifikacija
NKZ	Nacionalna klasifikacija zanimanja
Program za kval	Program za stjecanje kvalifikacije
Program za SIU	Program za vrednovanje ili stjecanje i vred. SIU
Prijedlog	Prijedlog za odobrenje Skup ko, SIU, SK ili SZ
SIU dodjela	Akreditirana osoba za dodjelu potvrde o SIU
Obrazovna ust.	Akreditirana osoba za stjecanje i dodjelu SK
ESCO	ESCO komp i ESCO kval taksonomija naziva
Sektorsko vijeće	Sektorsko vijeće
AA	Akreditacijska agencija
Akreditacija	Odluka o akreditaciji
Poslodavac	Poslodavac
Anketa	Podloga za izradu SZ (anketni rezultati)
NKD	Nacionala klasifikacija djelatnosti
SV Mišljenje	Odluka o vrednovanju skup komp. SZ, SK, SIU
Predlagač	Predlagač novog skup komp., SZ, SK ili SIU



Slika 1. Prijedlog logičkog ER modela podataka HKO

C. *Opis predloženog ER logičkog modela*

Sukladno slici 1, predlaže se logički model podataka koji osim zakonom propisanih podregisara koristi i brojne vanjske referentne izvore podataka i javne registre na razini Republike Hrvatske i EU. Prikaz modela je radi ograničenosti dostupnog prostora i preglednosti donekle apstrahiran, tako da nisu prikazane pojedine relacije kojima se uključuju atributi iz javnih registara, ali je njihova poveznica jasno prepoznatljiva kroz navođenje identičnih i intuitivnih naziva atributa. Osnovni domaći vanjski registri koji se koriste su Nacionalna klasifikacija djelatnosti (NKD) i Nacionalna klasifikacija zanimanja (NKZ). Predviđeno je korištenje NKD i NKZ šifre sa punim dostupnim brojem znamenki. Nadalje, predviđeno je i korištenje sva tri ključna elementa ESCO taksonomije radi koherentnosti naziva skupina kompetencija, zanimanja i kvalifikacija sa EU taksonomijom koja omogućuje i lakše buduće povezivanje na EURES portal tržišta rada. Ova poveznica prema EURES portalu nije predložena unutar modela sa Slike 1 budući da ona predstavlja definiranje dodatnog registra ljudskih potencijala čije uvođenje jest predviđeno strateškim dokumentima, ali ne postoji za sada provedbena regulativa niti su poznati temeljni atributi.

Izrada novog prijedloga moguća je od strane predlagača koji sukladno Zakonu [3] može biti pravna ili fizička osoba. Ukoliko se radi o prijedlogu skupa kompetencija ili SZ, predlagač je obvezan pripremiti odgovarajuće metodološke podloge koje su opisane u modelu podataka i prikazane unutar entiteta Anketa. Ove podloge predlagač će obraditi koristeći primjerice pristup tablice kompetencija [8] i na temelju njih predložiti skupove kompetencija ili standard zanimanja. Nakon provođenja osnovne formalne provjere koja nije prikazana modelom podataka, prijedlog predlagača u obliku neodobrenog skupa kompetencija, SZ, SIU ili SK upućuje se na odgovarajuće sektorsko vijeće koje daje svoje mišljenje. Sukladno prikazanom modelu podataka, poveznicom mišljenja i sektorskog vijeća osiguran je dostup ostalim podacima koji su nužni dio SIU, SZ i SK a koja opisuju sektorsko vijeće kroz dokument o njegovom osnivanju, mandat i članove. Skupovi kompetencija su gradivni elementi SZ, dok su s druge strane oni i direktno povezani sa odgovarajućim SIU vezom 1:1. Naime, kvalifikacijski okvir i njegov informacijski sustav mora dati jasnu informaciju o tome koji su sve skupovi kompetencija koji se primjerice traže na tržištu rada pokriveni sa SIU, što znači da su pokriveni nekim vidom stjecanja i/ili vrednovanja, a koje obrazovni sustav trenutno ne pokriva. Objekti unutar entiteta SIU predstavljaju također i temeljne gradivne elemente SK i to poveznicom više na više budući da svaki SIU može sudjelovati u više SK i da svaki SK može imati više SIU.

Končano, pojedini SK i pojedini SIU mogu imati više odgovarajućih programa za stjecanje i/ili vrednovanje što je prikazano vezom jedan na više. Ovi programi provode se od strane akreditiranih ponuđača, što je evidentirano kroz zapis o adekvatnoj akreditaciji i njenom roku važenja. Budući da je predviđen parametar vremena trajanja pojedinog objekta kao što su SK, SZ, SIU, Program i slično, očekuje se od informacijskog sustava trajno pohranjivanje svih objekata uz postojanje informacije o njihovoj aktivnosti. Tako je moguće da će SIU istog naziva koji se koristi u više SK u nekom trenutku biti promijenjen. Zapis starog objekta ostati će trajno pohranjen u bazi podataka kako bi se u bilo kojem trenutku mogao rekonstruirati sadržaj bilo kojeg SK, ali će biti označen kao neaktivan uz oznaku perioda njegove aktivnosti.

Povezanost pojedinih entiteta trebala bi u konačnici omogućiti primjerice dobivanje brzog i automatskog odgovora na pitanje usklađenosti pojedine kvalifikacije koju ima pojedinac sa zahtjevima nekog radnog mjesta. Kako bi sustav moga dati odgovor na takvo pitanje krenuo bi od usporedbe skupova kompetencija koje se nalaze u SZ koji je vezan uz pojedino radno mjesto, te bi preko njih došao do popisa SIU koji su sa njima jednoznačno povezani. Poznavanjem šifri svih SIU koji odgovaraju skupinama kompetencija, razmjerno je lako pronaći SK koji u sebi sadrži njihov najveći broj i pripadajuće programe za stjecanje SK. Odgovor koji bi zainteresirani građanin mogao dobiti na tako postavljeni upit bio bi popis skupova kompetencija za koje ne postoje povezani SIU (dakle kompetencija koje nije moguće trenutno akreditirati jer ne postoje za njih programi), te popis SIU koji nedostaju osobi koja ima određenu kvalifikaciju kao i popis programa i ponuditelja kod kojih je takve SIU moguće steći ili certificirati.

V. ZAKLJUČAK

Zakonom o HKO [3] te pravilnikom o registru HKO [4] Republika Hrvatska je napravila značajan iskorak u implementaciji nacionalnog kvalifikacijskog okvira te si je postavila zadatak ne samo izgraditi zakonsku regulativu HKO nego i prateći informacijski sustav. Budući da se radi o složenom i sveobuhvatnom instrumentu koji objedinjuje ne samo sustav obrazovanja nego i tržište rada i zapošljavanje, predmetni informacijski sustav planira se razvijati u suradnji Ministarstva nadležnog za obrazovanje i Ministarstva nadležnog za rad, koristeći pri tome postojeće javne registre poput NKZ i NKD te imajući u vidu obvezu budućeg povezivanja HKO sa Europskim kvalifikacijskim okvirom ali i ESCO inicijativom i EURES portalom. Već samo navođenje pojedinih atributa unutar podzakonskih akata, koji predstavljaju vanjske reference unutar nekih entiteta dalo je jasan poticaj da se preloženi ER logički model podataka ne veže samo uz tri propisana podregistra, nego da se obuhvati nešto širi kontekst. Na ovaj način cilj je izbjeći izgradnju novog informacijskog sustava bez njegovog povezivanja na postojeću infrastrukturu, budući da je i HKO zamišljen kao neobvezujući evolutivni razvojni instrument koji ne narušava postojeće stanje nego ga nastoji dodatno unaprijediti. Ovakav pristup razvoju HKO, pa time i njegovog informacijskog sustava, nastao je na temelju dioničkog rada na razvoju sustava ali i niza stranih negativnih iskustava koja su pratila nagle i nepripremljene reforme ili kopiranje tuđih rješenja koja nisu uklopiva u pojedini nacionalni kontekst [11].

Tako su definirani i entiteti poput sektorskih vijeća ili akreditacijskih agencija koje predstavljaju upravljačke elemente sustava kvalifikacija, dok su svoje mjesto pronašli i nužni dokumenti poput akreditacijskih preporuka ili mišljenja sektorskih vijeća ili obveznih metodoloških podloga.

Predloženo rješenje nastalo je na temelju promišljanja trenutno postojećih zakonskih rješenja, uzimajući u obzir kroz njih propisane entitete i atribute te njihove relacije, ali bez implementacije rješenja u stvarno okruženje. Zbog toga očekujemo da model predložen ovim člankom može biti korisna vodilja kod pripreme fizičkog modela podataka, ali je vrlo izvjesno da će u implementaciji stvarnog rješenja biti potrebno uvesti dodatne entitete i niz internih atributa koji će osigurati potpuno uklapanje svih nužnih internih i eksternih elemenata. Konačno, imajući u vidu kako se radi o nacionalnoj podatkovnoj infrastrukturi koju će koristiti brojni građani i institucije i koja će osigurati jasnoću i prepoznavanje naših kvalifikacija i u inozemstvu, kod dizajniranja baze podataka biti će potrebno posebni naglasak staviti na nužnu buduću interoperabilnost, racionalizaciju i optimizaciju upita i strukture baze podataka.

VI. LITERATURA

[1] European Training Foundation, Cedefop, UNESCO, UNESCO Institute for Lifelong Learning, “Global National Qualifications Frameworks Inventory”, Kuala Lumpur, ožujak 2013

[2] Vlada RH, “Hrvatski kvalifikacijski okvir – uvod u kvalifikacije”, Zagreb, srpanj 2007

[3] Hrvatski Sabor, “Zakon o Hrvatskom kvalifikacijskom okviru”, Narodne Novine br. 22/13, Zagreb, veljača 2013

[4] Ministarstvo znanosti, obrazovanja i sporta, “Pravilnik o registru Hrvatskog kvalifikacijskog okvira – Nacrt”, Zagreb, siječanj 2014 *http://public.mzos.hr/Default.aspx?art=12988&sec=1933*

[5] The European Commission, Directorate General Employment, Social Affairs and Inclusion and Directorate General Education and Culture, ESCO (2013) SEC 038 FINAL ESCO (2013) SEC 038 FINAL “Introduction to ESCO”, Brisel, travanj 2013 *https://ec.europa.eu/esco/about-esco*

[6] European Comission, “Three tools to facilitate online job matching throughout Europe (ESCO, EURES, Mathch & Map)”, Publications Office of the European Union, Luxembourg, 2010

[7] European Parliament - Council, "Recommendation of the European Parliament and of the Council on the establishment of the European Qualifications Framework for lifelong learning", Brisel, travanj 2008

[8] M. Balkovic, T. Kristof, “Methodology of qualification standard development in ICT based on competences matrix model”, Annals of DAAAM for 2011 & Proceedings of the 22nd International DAAAM Symposium, Volume 22, No. 1, ISSN 1726-9679, pp. 0009-0010, Beč, studeni 2011

[9] Peters, M., Meijer, K., Nuland, E., Viertelhauzen, T. & Ad, R. “Sector Councils on Employment and Skills at EU level – A study into their feasibility and potential impact”, EcorysNederlands BV in cooperation with KBA, Rotterdam, 2010

[10] Teorey, T, “Database modeling & design”, Academic Press, 3rd edition, USA, 1999, pp. 45-56

Chakroun, B. “National Qualifications Frameworks: from policy borrowing to policy learning” European Journal of Education, vol. 45, No. 2, 2010 pp. 199 – 217.

Istraživanje o odlukama o odabiru karijere maturanata i brucoša na platformi Facebooka

Tomislav Krištof,
Visoko učilište Algebra - Visoka škola za primijenjeno računarstvo
Zagreb, Hrvatska
tomislav.kristof@algebra.hr

Hrvoje Balen,
Visoko učilište Algebra - Visoka škola za primijenjeno računarstvo
Zagreb, Hrvatska
hrvoje.balen@algebra.hr

Vanja Šebek
Otvoreno učilište Algebra
Zagreb, Hrvatska
sebek.vanja@gmail.com

Autori ovog rada proveli su istraživanje na populaciji maturanata i brucoša u Hrvatskoj na temu odabira obrazovnog puta i karijere. Istraživanje je obuhvatilo 4968 anketiranih koji su se izjašnjavali o odabiru svoje karijere, željama u daljnjem obrazovanju i planovima za budućnost. Autori istraživanja poslužili su se raspoloživim (besplatnim) informacijskim tehnologijama: Facebook i *open source* platformama – PHP i MySQL. Kombinacijom tri navedena alata napravili su aplikaciju za istraživanje te na efikasan i efektivan način dosegli ciljanu publiku.

Istraživanje je bilo segmentirano u dvije skupine: Maturanti i brucoši 2013., a sastojalo se od 25 pitanja po skupini. Kako ispitanici pripadaju generacijski sukcesivnim i analognim populacijama, pored statičkih podataka, voditelji projekta stekli su uvid i u dinamiku istih varijabli sličnih populacija.

Cilj istraživanja bio je dobiti uvid u motivaciju za odabir studija i proces donošenja odluke o odabiru studija kod maturanata te zadovoljstvo upisanim studijem kod brucoša. Rezultati istraživanja namijenjeni su tome da povećaju efikasnost komunikacije između fakulteta i maturanata zainteresiranih za upis studija.

I. UVOD

Polazeći od ciljeva istraživanja – uvid u motivaciju za odabir studija i proces donošenja odluke o odabiru studija kod maturanata te uvid u zadovoljstvo upisanim studijem kod brucoša – organizatori su se našli pred nekoliko ključnih izazova od kojih su neki u proturječju, kao primjerice doseg što većeg broja ispitanika i ograničen budžet za provedbu istraživanja.

Izazovi su kako slijedi:

1. Definiranje segmenata ispitanika
2. Na temelju definiranih segmenata odabir medijskih kanala za dosezanje tih segmenata.
3. Definiranje parametara ankete i definiranje pitanja na temelju tih parametara
4. Motiviranje ispitanika da popune anketu iskreno i u potpunosti.
5. Pronaći tehnologijsko rješenje (platforma, programski kod, baza podataka, korisničko sučelje)
6. Pronaći mehanizam provjere valjanosti ispunjene pojedinačne ankete
7. Projekt razviti u što kraćem roku i uz minimalni raspoloživi budžet
8. Podatke dobivene kao rezultat istraživanja organizirati u korisne informacije te iste prezentirati javnosti na efikasan način

Kako bi riješili prikazane izazove organizatori istraživanja podijelili su ih na organizacijske, marketinške i tehnologijske te definirali poslove prema toj podjeli. Na taj način su mogli pokrenuti paralelne procese kako bi se skratilo vrijeme izvedbe projekta. Najveći segment prikazanih izazova potpali su pod marketinški segment koji je uključivao dva odvojena aspekta. Prvi; dizajn i arhitektura cijelog istraživanja i drugi; odabir medija i partnera u projektu te odnosi s javnošću. Sljedeći ne manje važan segment izazova obuhvaćen je tehnologijskim procesima koji su uključivali razvoj programskog koda, baze podataka te povezivanje navedenog sa sustavom Facebook aplikacije, kako će kasnije u tekstu biti preciznije razjašnjeno. Posljednji proces, ali ne manje važan, odnosi se na organizaciju i povezivanje različitih tipova izvođača u projektu

(marketing, mediji, programiranje, dizajn) te brigu oko budžetiranja i same efikasnosti projekta.

I. ISTRAŽIVAČKI PROCESI U PROJEKTU

Iz definiranih izazova u uvodu proizlazi da su poznavanje marketinga i stručna primjena marketinških znanja i vještina ključni za ovaj projekt. Dva su bitna aspekta:

1. Poznavanje procesa segmentacije i istraživanja tržišta kao jednog od osnovnih funkcija marketinga[1].
2. Poznavanje medijske komunikacije i poglavito poznavanje digitalnog marketinga kao interdisciplinarnog područja koje obuhvaća i tehnologijske aspekte projekta.

A. Proces segmentacije i arhitektura istraživanja

Prema definiciji istraživanja, metodologiji i tehnikama ovo istraživanje je deskriptivno s ciljem opisivanja osobina određene populacije (maturanti i brucoši), njihovih stavova, uvjerenje i ponašanja. Instrument istraživanja je bila anketa. Istraživanje također nosi i kauzalne elemente jer dovodi u vezu odabir studija, poziciju odabira i zadovoljstvo odabranim. Tehnike korištene u anketi su bile zaokruživanje i nadopunjavanje.

Polazeći od ciljeva istraživanja prvi korak je bio precizno odrediti segment ispitanika u populaciji. Kako se radi o tematici odlučivanja o upisu studija te zadovoljstvu upisanim studijem, samo definiranje segmenata nije predstavljao problem. U vrijeme provođenja istraživanja (lipanj, 2013.) segmenti su definirani po demografskom principu: maturanti i brucoši za školsku i akademsku godinu 2012./2013., odnosno preciznije za maturante svi koji su namjeravali upisati studij akademske godine 2013./2014., dakle i osobe koji su maturirali i ranijih godina.

Od demografskih varijabli prikupljene su sljedeće: godina rođenja, spol, završena srednja škola, uz obavezno navođenje prvih pet odabranih fakulteta. Tražena je i jedna geografska varijabla a to je grad završene srednje škole.

Složeniji proces koji je uslijedio bilo je precizno definiranje pitanja ključnih za ispunjenje osnovnog cilja istraživanja. Kako se ovdje radi o psihografskim varijablama (motivi, namjere i odluke) te o bihevioralnim varijablama (upis) potrebno je bilo napraviti kratko pred-istraživanje na malom uzorku studenata u vidu intervjua. To pred-istraživanje poslužilo je kao orijentacijski okvir za definiranje pitanja. Kao uzorak „poslužili“ su studenti 1. godine studija primijenjenog računarstva na Visokom učilištu Algebra – Visokoj školi za primijenjeno računarstvo.

Nadalje, kako je administracijski proces kvalifikacije i upisa na studij strogo definiran, ta je činjenica poslužila kao baza za ona pitanja koja se odnose na administrativne aspekte upisa poput segmentacije izbora studija i sustava bodovanja[2].

Dio ovog procesa koji je zahtijevao složeniji pristup bio je aspekt procesa odlučivanja i definiranje parametara koji utječu na odluku. Pored intervjua u pred-istraživanju organizatori su uzeli u obzir i mnoga istraživanja koja su rađena na adolescentima, a tematika im je bila upravo donošenje odluka[3].

U spomenutim istraživanjima ustanovljene su karakteristike adolescenata (društvene, kognitivne, emotivne, bihevioralne) koje određuju načine na koje adolescenti donose odluke, a razlikuju se, u manjem ili većem iznosu u odnosu na donošenje odluka kod „odraslih“ ljudi.

Adolescencija je razdoblje psihološkog, kognitivnog, društvenog i emotivnog rasta i promjena. Zbog intenzivnijih mijena adolescenti će biti skloniji rezoniranju koje je u suprotnosti sa činjenicama, skloniji rezoniranju temeljenom na lažnim premisama, više će se oslanjati na procjenu nego na pronalaženje i analiziranje podataka iz okoline.

Kako adolescenti u prosjeku provode podjednako vremena sa svojim vršnjacima kao i sa svojim roditeljima za pretpostaviti je da će utjecaji jednih i drugih na odluke pojedinca biti podjednako jak, a ti utjecaji mogu međusobno biti i proturječni. Jači utjecaj vršnjaka primijećen je kod mlađih adolescenata dok se kod starijih adolescenata umanjuje. Pokazalo se također da je kod društvenih utjecaja na odluku važnije koga adolescent identificira i odabire kao „utjecatelja“ (eng. *influencer*), tek onda taj određeni „utjecatelj“ ima više upliva na odluku pojedinca. S emotivnog aspekta prepoznato je da se tijekom samog procesa odluke važno emotivno stanje. Ako je ono pozitivno adolescent je manje sklon da sagleda negativne posljedice loše odluke. Utjecaj vršnjaka je i ovdje prisutan jer se u doba adolescencije ispitanici izjavljuju da je najbolje osjećaju kad su s vršnjacima – o odnosu na vrijeme kad su sami – tada se osjećaju dosadno.

Adolescenti su populacija koja dosta konzumira medije, prema istraživanjima[4], provode 4.5 sati gledajući televiziju, 1.5 sati na računalu, a 7.5 sati konzumiraju medije za zabavu. Posljedica toga je da se ne može zanemariti utjecaj medija na donošenje odluke.

Imajući sve navedeno u vidu organizatori su sastavili anketu kako bi dobili što bolji uvid tko i što utječe na odluke o upisu studija, u stvari na odabir karijere.

S obzirom da se radi istraživanje na dvije različite populacije – maturanti i brucoši, oformljene su dvije slične, ali ne i iste ankete. Cilj ovakvog načina istraživanja je da se pokuša prepoznati odnos između procesa odluke i zadovoljstva istom te da se prepoznaju trendovi na analognim populacijama.

Poseban naglasak anketnih pitanja stavljen je na već spomenute psihografske čimbenike koji se tiču stava o određenim mogućnostima studiranja (privatne naspram državnim studijima), motivima upisa i vanjskim utjecajima na odluku.

B. *Odabir medija i partnera*

Prilikom odlučivanja o načinu provođenja ankete organizatori su si postavili dva općenita, ali važna cilja – efikasnost i efektivnost. Efikasnost se trebala manifestirati kroz maksimalni doseg ankete (po broju ispitanika i kvalitetnoj segmentaciji). Efektivnost se trebala manifestirati kroz maksimalni broj kvalitetno ispunjenih anketa. Kako samo Visoko učilište Algebra nema adekvatne medijske alate, za pronalaženje načina postizanja željenog dosega organiziran je *brainstorming* s postavljenim zahtjevima efikasnosti i efektivnosti.

Već u ovoj fazi donesena je odluka da će platforma za provođenje ankete biti Facebook, temeljila se na javno dostupnim istraživanjima[5] i spomenutom pred-istraživanju koje je u vidu intervjua provedeno prije. Ona se temeljila na sljedećim prednostima:

1. Facebook koristi iznimno velik postotak adolescenata.
2. Adolescenti su dobro „umreženi“ i u prosjeku imaju 300 prijatelja na Facebooku.
3. Facebook nudi vrlo fleksibilne mehanizme za povezivanje prilagođenih aplikacija sa svojom bazom.
4. Facebook nudi mogućnost preuzimanja podataka od korisnika (uz njihovu dozvolu) koji također pomažu u stvaranju psihografske slike populacije nad kojom se istraživanje vrši.
5. Medijska moć Facebooka leži i u činjenici da se akcije njegovih korisnika dijele i mogu ih pratiti ostali umreženi članovi – prijatelji.

Nadalje, organizatori su se u ovoj fazi suočili s izazovom – kako dosegnuti ciljanu populaciju, kako bi ih uopće usmjerili na anketu smještenu na Facebooku. Odabrana je strategija da se pronađe odgovarajući medij koji bi za ovu svrhu imao najpreciznije ciljanje. Dodatni zahtjev je bio da s obzirom na vrlo ograničen budžet taj medij ne bude preskup i u obzir su zbog svoje interaktivnosti i povoljne cijene dolazili samo *online* mediji.

Analizom različitih medija, od vodećih hrvatskih portala do specijaliziranih blogova odabran je jedan portal čija je tematika vrlo bliska problematici projekta – srednja.hr. Prednosti ovog portala su visoka posjećenost srednjoškolaca motiviranih za upis studija kao i vezana Facebook stranica idealna za inicijalni medijski doseg. U vrijeme provođenja istraživanja gotovo 20000 *fanova*. Dogovorom između Visokog učilišta Algebra i portala srednja.hr definiran je partnerski odnos u projektu i osiguran je aspekt efikasnosti provođenja istraživanja.

Sljedeći korak bio je osigurati što veću efektivnost, tj. motivirati ispitanike da u što većem broju ispune anketu savjesno i do kraja. Dio ovog zahtjeva zadovoljile su sljedeće karakteristike projekata: Kako adolescenti prema istraživanju[6] koriste društvene mreže na dnevnoj bazi, njih 3/4, velika je vjerojatnost da će zamijetiti ponuđenu anketu odnosno preciznije percipirat će njeno početno otvaranje kao minimalan napor. Anketa je vizualno „ugrađena“ u samu Facebook stanicu stoga im i okruženje anketiranja neće biti strano. Dodatna karakteristika projekta koja je povećavala vjerojatnost da će anketa biti efektno ispunjena se skrivala u tome da je tema ispitanicima bila bliska (maturant i brucoši), i da su bili motivirani sudjelovati u tematici upisa na studij.

Ipak sama familijarnost sa korisničkim sučeljem te bliskost s tematikom ne bi bili dovoljni. Organizatori su procijenili da je bila potrebna dodatna motivacija kako bi ispitanici ispunili anketu do kraja. S odabranim partnerima i sponzorima dogovorena je suradnja na način da su svim ispitanicima nudili mogućnost da osvoje nagrade. Partneri su za uzvrat dobili medijsku prisutnost u projektu.

Zadatak medijskog partnera, srednja.hr bio je da u relativno kratkom roku (14 dana) mobilizira što veći broj ispitanika. Glavni kanali promoviranja ankete bili su sami portal te pripadajuća Facebook stranica.

II. TEHNOLOGIJSKI PROCESI U PROJEKTU

Nakon postavljanja zahtjeva, ciljeva i strategija projekta te nakon izrade arhitekture istraživanja pristupilo se odabiru tehnologija na kojima će se istraživanje vršiti. Zahtjevi koje su trebala zadovoljiti tehnologijska rješenja kretali su primarno od korisnika, a onda se razmišljalo o resursima potrebnim za izradu:

1. Kvalitetno korisničko sučelje, minimalan napor za ispitanika i jednostavnost korištenja.
2. Kvalitetno programsko rješenje, baza podataka i server.
3. Jednostavnost povezivanja s platformom Facebooka i Graph APIjem.

Kod izrade korisničkog sučelja dizajner se vodio smjernicama dobrog UXa (*User eXperience*) i odlučio se na minimalistički pristup kako bi jasno istaknuo pitanja, polja za upis i ostala polja obrasca. Kako je anketa opširna bilo je potrebno veliki obrazac za unos i odabir odgovora podijeliti na više manjih dijelova, njih pet. Na taj način se smanjila vjerojatnost da će korisnik odmah odustati od ispunjavanja. Dodatno, u svakom koraku naznačeno je točno koliko je koraka odrađeno i koliko je još do kraja kako bi korisnik u svakom trenutku znao gdje se nalazi.

Nadalje, specifičnost obrasca je bila njegova kauzalna izmjenjivost upisnog polja. Primjerice, ispitanik je odabirom grada završene srednje škole pokrenuo pozadinski proces koji je iz baze podataka ispitaniku ponudio srednje škole koje se nalaze samo u tom gradu. Također isti princip je primijenjen kod odabira fakulteta, jer se za svaki odabrani fakultet kauzalno nudio studij na

tom fakultetu. Ovaj pristup je zakomplicirao izvedbu programskog koda, ali je za rezultat imao smanjenje napora za ispitanika što je navedeno kao primaran cilj korisničkog sučelja. Ovo je tipični primjer prioriteta u procesu donošenja odluka o dizajnu korisničkog sučelja. Oko odabira programske platforme nije bilo dileme, jer je zbog raspoloživih resursa odabrana kombinacija PHP programskog jezika i MySQL baze podataka. I jedno i drugo rješenje razvijeno je na *open source* tehnologiji.

Kako je PHP popularan jezik, SDK (*software development kit*) elementi postoji i za Facebook[7] tako da nije bilo dodatnih zahtjeva za programiranjem procesa komunikacije između programskog koda istraživanja i samog Facebooka. Korišteni Facebook SDK obuhvaća gotove unaprijed pripremljene funkcije koje služe za spajanje na Facebook Graph API[8]. Ovim rješenjima programer koristi gotove Facebook funkcije ugrađene u vlastiti kod kako bi slao parametre s ciljem dodavanja, izmjene, brisanja i čitanja podataka iz Facebook baze podataka.

Čitanje i općenito pristup podacima iz Facebook baze je strogo kontroliran proces i Facebook „pazi“ da se ti podaci ne mogu čitati bez odobrenja vlasnika pojedinog profila. Iz tog razloga nezaobilazan korak u pristupanju anketi je bilo odobrenje korisnika(ispitanika) kako bi organizator ankete mogao raspolagati podacima unesenim u Facebook profil. Facebook često mijenja ili bolje rečeno dodaje nove značajke svojoj Graph aplikaciji stoga je preporučljivo pratiti novosti na Facebook službenim stranicama za *developere*[9].

Podaci koji se ovim putem automatizmom preuzimaju od korisnika su sljedeći: ime, prezime, spol, e-mail adresa, datum rođenja, završena srednja škola, stranice koje je kontaktirani korisnik *lajkao*. Prednosti ovog rješenja su velike. Za samog ispitanika to znači da nema potrebe za ispunjavanjem navedenih polja, a za organizatora ankete uvid u profil korisnika koji nadilazi klasične demografske karakteristike. Prvenstveno se mogu analizirati afiniteti i društvene karakteristike vlasnika profila.

III. ORGANIZACIJSKI PROCESI U PROJEKTU I BUDŽETIRANJE

Produkcija i organizacija samog projekta navedeni su kao poseban proces jer obuhvaćaju angažiranje vanjskih suradnika (mediji, partneri, dobavljači tehnologija) kroz prizmu ograničenih resursa (budžet, vrijeme). Mediji i partneri mogli su biti kontaktirani tek u fazi projekta kad su bili detaljno razrađeni plan nastupa (komunikacija prema ciljnoj skupini), arhitektura, metodologija i tehnike istraživanja. U ovom procesu dakle, pored stručnih znanja i organizacijskih sposobnosti bile su potrebne i persuazivne pregovaračke vještine za pridobivanje medija i partnera u projekt u uvjetima niskog raspoloživog budžeta. U konačnici je zahvaljujući partnerskom odnosu i ravnopravnom investiranju vremena i resursa svih uključenih strana dogovoren kompenzacijski model gdje se moneta razmjene manifestirala kroz uložene radne sate, medijsku vrijednost *brandova* (srednja.hr, sponzori) i stručnost u provedbi i organizaciji projekta (Visoko učilište Algebra – Visoka škola za primijenjeno računarstvo).

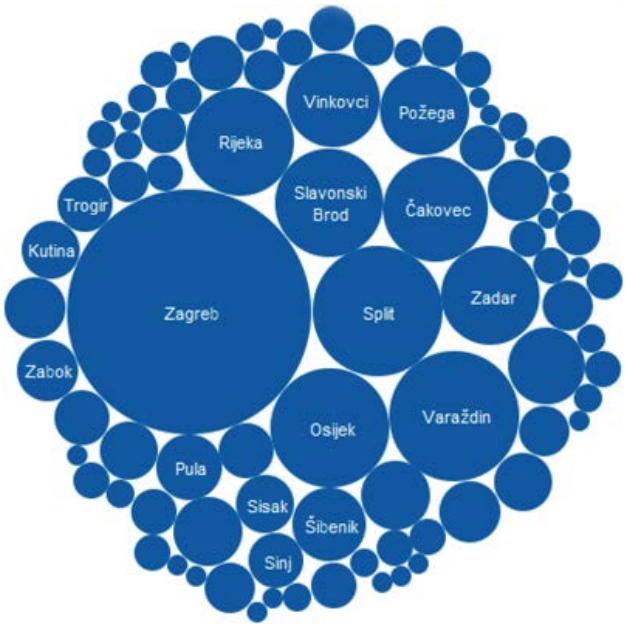
IV. REZULTATI ISTRAŽIVANJA

Po završetku anketiranja nastupio je proces obrade podataka i pronalaženje vrijednih informacija iz baze. Prije sistematičnog pristupa podacima baza se morala iščistiti od neispravno unesenih podataka, dvostrukih ili višestrukih unosa, do lažnih unosa podataka. Kao posebni parametar za provjeru valjanosti ankete uzet je *timer* postavljen s ciljem da se izmjeri vrijeme provedeno uz anketu. Odabirom svih ispitanika po parametru provedenog vremena izbačeni su svi koji su anketu ispunili za manje od 3 minute, koliko je bilo otprilike potrebno da se najbržim klikanjem bez čitanja pitanja završi anketa.

A. *Demografski i geografski podaci*

Starost ispitanika određena je godinama rođenja prema distribuciji: 63% rođenih 1994., 19% rođenih 1993., 18% rođenih 1995. godine. 58% ispitanika bilo je ženskog, a 42% ispitanika muškog spola. Distribucija ispitanika po gradovima prikazana je grafikonom proporcija na slici 1.

Ispitanici su u prve izbore odabirali studije u Zagrebu, a tek je jedan osječki fakultet (filozofski) dospio u prvih deset najpoželjnijih (TABLICA I.).



Slika 1. Prikaz proporcija frekvencije gradova u kojima su sudionici pohađali srednju školi

B. Psihografski i biheviоralni podaci

Prvo važno pitanje koje se tiče procesa odlučivanja bilo je pitanje motiva upisa odabranog studija. Od više ponuđenih odgovora izdvojit će se koncentracije iz kojih se mogu izvući zaključci.

1. Kao najvažniji parametar odluke iskristalizirao se osobni afinitet prema tematici studija – 81% maturanata 2013. godine izjavilo kako je to svakako važan motiv
2. Društveni aspekt pokazao se utjecajan kod važnosti upoznavanja novih kolega sa sličnim pogledima na svijet i profesionalnu karijeru – 89,85% izjavilo je da im je to važno ili relativno važno
3. Upitani pak za motiv u vidu poznanika ili prijatelja koji već studiraju na odabranom fakultetu, tek je 10,66% izjavilo da im je to svakako važno, a čak 56,78% da im nije uopće važno
4. Racionalni aspekti motivacije poput visine plaće i perspektive zapošljavanja u struci navedeni su većinom kao relativno važni – u rasponu od 40%-50% za afirmativne odgovore.
5. Situacijski parametar motivacije u smislu ne postojanja prilike za zapošljavanje u trenutku polaganja mature ili odabira studija većinom je odbačen kao važan – 82,06% nije se složilo s ovakvom motivacijom

Pored motivacije organizatori istraživanja željeli su ustanoviti tko i koliko može utjecati na odluku o odabiru studija – karijere (društveni utjecaj). Na pitanje tko je utjecao na njihov odabir studija odgovori su otkrivali sljedeće:

1. Na odluku o upisu najviše su utjecali roditelji i prijatelji/poznanici koji rade u struci odabranog studija, a ponajmanje vršnjaci (poznanici 52%, roditelji 58%)
2. Kod analize odgovora na ovo pitanje ustanovljeno je da je kod onih studenata koji su nezadovoljni odabranim studijem veći utjecaj imali su roditelji nego kod onih studenata koji su zadovoljni odabirom.

Kod pitanja što je utjecalo na odabir studija, gdje su ponuđeni razni parametri od troškova studiranja, preko blizine stanovanja do certifikacijskih pogodnosti tijekom studija i utjecaja medija, odgovori su pokazali sljedeće:

1. Očekivano su najveći utjecaj imali troškovi studiranja, ali i blizina stanovanja (oko 2/3 ispitanika)
2. Posjedovanje diplome priznate i u inozemstvu, te međunarodni certifikati imali relativno manji utjecaj na odluku (gotovo 50% ispitanika izjavilo je da ova stavka nema utjecaj na odluku)

TABLICA I. NAJPOŽELJNIJI FAKULTETI 2013. GODINE

Redni broj	Fakultet	1. odabir
1	Filozofski fakultet (Zagreb)	6,16%
2	Ekonomski fakultet (Zagreb)	4,68%
3	Fakultet elektrotehnike i računarstva (Zagreb)	4,05%
4	Tehničko veleučilište u Zagreb	3,46%
5	Pravni fakultet (Zagreb)	3,33%
6	Učiteljski fakultet u Zagrebu	3,04%
7	PMF (Zagreb)	3,00%
8	Medicinski fakultet (Zagreb)	2,78%
9	Zdravstveno veleučilište u Zagrebu	2,57%
10	Filozofski fakultet (Osijek)	2,53%

3. Utjecaj medija naveden je samo kao djelomičan, više od 50% ispitanika navelo je da utjecaja medija nema, a tek oko 10% ispitanika je navelo da svakako utječe.

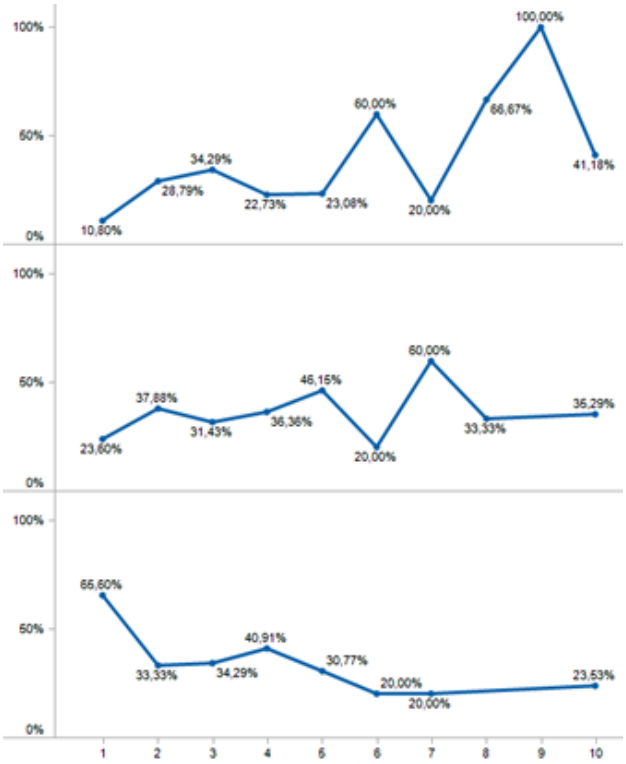
Na pitanje vezano na pripreme za maturu 52% ispitanika pohađalo je neku vrstu priprema, bilo da su pripreme bile organizirane u školi ili su na njih išli u specijalizirane institucije, 10% ispitanika pohađali su individualne instrukcije.

U istraživanju jedno je pitanje bilo naglašeno vezano uz STEM područje (studiranje prirodnih i tehničkih znanosti), direktno je bilo usmjereno na činjenicu da se relativno mali broj maturanata odlučuje na upis studija iz STEM područja. Dominantan odgovor je bio da su takvi studiji teški, a većina želi diplomu bez previše truda (50% ispitanika).

Kako je već ranije navedeno, istraživanje je obuhvatilo i populaciju brucoša, a ta populacija bila je pitana i koliko je zadovoljna odabirom studija, odnosno bi li ispitanici iz te populacije mijenjali svoj odabir. U korelaciju organizatori su stavili varijable ranga odabira studija (1-10) i sklonost promjeni (da, možda, ne) te su dobili očekivani rezultat – studenti koji su upisali studij prvih izbora bili su više zadovoljni svojom odlukom, odnosno bili su manje skloni promjeni odabira. (Slika 2.)

Kako su se organizatori u istraživanju koristili i podacima iz Facebookove Graph aplikacije, pored osnovnih podataka dobili su i dopuštenje samih korisnika da zavire u njihove profile i time steknu uvid u kulturne aspekte života starijih adolescenta u Hrvatskoj, stoga su ovdje navedeni trivijalni podaci kao demonstracija snažnog društvenog aspekta jedne internetske aplikacije – Facebooka.

U obrađenom uzorku prosječan broj *lajkanih* stranica je 553, a najmotiviraniji *lajker* imao ih je 4724. U lipnju 2013. promatrana populacija najviše je lajkala Hrvatsku .



Slika 2. Odnos ranga odabira studij i sklonosti promjeni studija

rukometnu reprezentaciju (12.46%), Slavena Bilića (12.34%) i nogometni klub Barcelonu (11.65%), za ilustraciju dakle, ovdje je odabrano područje sporta.

V. ZAKLJUČAK

Pregledom informacija dobivenih iz ovog istraživanja doneseni su sljedeći opći zaključci:

1. Gotovo polovica ispitanih studenata koji su prošle godine upisali fakultet, razmišlja o promjeni studija, dok bi svaki peti sigurno odabrao nešto drugo da može ponovno birati.
2. Maturanti prilikom odabira studija najviše cijene mišljenje poznanika koji rade u struci, a oni koji se vode savjetima roditelja u velikom broju budu nezadovoljni upisanim studijem.
3. Unatoč činjenici da su maturanti prilično samosvjesni prilikom odabira fakultete, ipak značajno raste utjecaj poznanika, obitelji, prijatelja i profesora – analiza pokazuje da je najkorisnije slijediti savjete pojedinaca koji već rade u struci kojom se želite baviti, a da savjeti roditelja mogu biti i prilično neobjektivni.
4. Tri najpoželjnija fakulteta u državi su zagrebački Filozofski fakultet, Ekonomski fakultet i Fakultet elektrotehnike i računarstva.

S obzirom na velik broj dosegnutih ispitanika, gotovo 5000, a uz investirana minimalna financijska sredstva organizatori zaključuju da su postignuti opći ciljevi efikasnosti i efektivnosti. Facebook kao platforma, te *open source* tehnologije PHP/MySQL pokazali su se kao snažan alat za dobivanje relevantnih informacija koje su stavljene na raspolaganje prvenstveno organizatoru, ali i zainteresiranim javnostima za donošenje odluka.

LITERATURA

[1] P. Kotler, K.L.Keller, Marketing Management, 14th ed., p.cm,, Prentice Hall, p 213.

[2] https://www.postani-student.hr, 15.6.2013.

[3] Baruch Fischhoff, Nancy A. Crowell, and Michele Kipke, Adolescent Decision Making,, Implications for Prevention Programs, Summary of a Workshop, National Research Council, Institute of Medicine, NATIONAL ACADEMY PRESS, Washington, D.C., 1999

[4] Henry J. Kaiser Foundation, "Generation M2: Media in the Lives of 8-18 Year Olds," January 2010

[5] http://www.pewinternet.org/Reports/2013/Teens-Social-Media-And-Privacy/Summary-of-Findings.aspx

[6] http://www.pewinternet.org/Reports/2013/Teens-Social-Media-And-Privacy/Main-Report/Part-1.aspx

[7] https://developers.facebook.com/docs/reference/php/

[8] http://en.wikipedia.org/wiki/Facebook_Platform#High-level_Platform_components

[9] https://developers.facebook.com/

Nastavni plan i program za predmet programiranje u prirodoslovno -matematičkim gimnazijama

Julijana Novaković
III.gimnazija, Matice hrvatske 11, Split, Republika Hrvatska
Telefon: 021-384 933 Fax: 021-322 754 E-mail: julijana.novakovic@skole.hr

Sažetak - Usprkos sve raširenijoj primjeni e-učenja u obrazovanju programera, postoje brojne barijere koje udaljavaju učenike od pravilnog pristupa programiranju. Problemi koji se pojavljuju pri podučavanju programiranja su ili fokusirani na razvojnu okolinu ili na sintaksu više programskih jezika kojih tijekom obrazovanja učenici usvajaju te se tako u potpunosti gubi osnovni smisao programiranja, a to je usvajanje osnovnih koncepata algoritamskog načina rješavanja problema. Iako je u nizu programskih jezika danas zaista teško izabrati pogodan za učenje programiranja smatram da treba napraviti redizaj postojećih nastavnih sadržaja. U ovom radu ću ponuditi prijedlog nastavnih sadržaja koji potiču razvijanje apstraktnog i algoritamskog načina razmišljanja.

Redizaj nastavnih sadržaja potaknut je, između ostaloga, željom da se unaprijedi kvaliteta učenja programiranja te time odgovori na zahtjeve natjecanja u kategoriji algoritmi. Članak donosi praktična iskustva autora, aktivnosti i rezultate implementacije u dodatnoj nastavi informatike. Rezultati su vidljivi u motivaciji učenika i povećanoj aktivnosti te potrebi da se na nivou škole sistematizirano i metodički oblikovano vodi dodatna nastava informatike. U tijeku je izrada pilot-projekta „Teorija grafova u natjecateljskim zadatcima" namijenjen učenicima kao dodatni sadržaj, a realizirat će se u suradnji s Prirodoslovno-matematičkim fakultetom Split. Na temelju tih iskustva planira se izrada nastavnog programa dodatne nastave programiranja za školsku godinu 2014./15.

I. UVOD

Nastava u III.gimnaziji obiluje brojnim i raznolikim inovacijama u procesu učenja i poučavanja, a posebno se ističe inovativnim primjenama uporabe informacijsko komunikacijskih tehnologija u odgojno-obrazovnom procesu. Izradi nastavnih sadržaja kojima se bavi ovaj rad predhodila je SWOT analiza (Sl. 1).

Izradi novih nastavnih sadržaja za drugi, treći i četvrti razred prethodio je projekt izveden tijekom školske godine 20012./13. u dodatnoj nastavi informatike koja se bavi pripremom učenika za natjecanje u kategoriji algoritmi. Trenutno je izrada idejnog rješenja nastavnih sadžaja koje ima za cilj planiranje, izvođenje te postavljanje novih i redizajniranje postojećih nastavnih programa.

Osnovni cilj je stvaranje poticajnog okruženja za razvoj i sistematičnu poduku programiranja kako bi se uspješno kod učenika razvijao algoritamski pristup rješavanja problema koji je do sada uvelike zapostavljen. Programiranje se nerijetko svodi na učenje sintakse određenog programskog jezika.

Projekt se također bavi uvođenjem u nastavni proces eučenja koje bitno doprinosi unapređenju kvalitete toga procesa i kvalitete ishoda obrazovanja. Analizom tehničkih i pedagoških aspekata e-okoline LMS Moodle sustava za administraciju nastave (engl. *Learning Management System*, LMS) utvrđeno je da sustav Moodle podržava jezgru nastavnih aktivnosti i omogućuje olakšano komuniciranje i intenzivniju suradnju između učenika i nastavnika te između učenike i učenika. Učenici stječu nove vještine kao što su timski rad, komunikacijske vještine, istraživački pristup rješavanju problema i sl., a važne za njihovo kasnije obrazovanje i rad.



Sl.1 SWOT analiza

II. E-OKOLINA–TEHNIČKI OKVIR IMPLEMENTACIJE PROJEKTA

Moodle platforma je izabrana kao tehničko rješenje. Na školskom serveru je instaliran sustava Moodle. Popularno nazvan eTrema (Sl. 2).



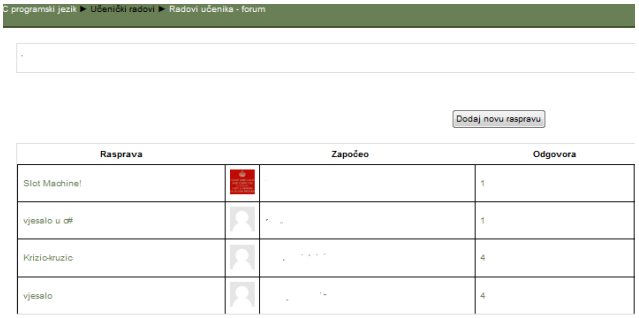
Sl.2 eTrema-portal za eučenje III.gimnazije Split

Moodle je programska aplikacija za izradu i održavanje *online* kolegija putem Interneta i može služiti za samostalno učenje ili kao nadopuna klasičnoj nastavi. Riječ Moodle je akronim od *Modular Object-Oriented Dynamic Learning Environment* što u prijevodu znači modularna objektivno-orijentirana dinamična okolina za učenje. To je projekt otvorenog koda (*open source*) što znači da je korisnicima omogućen uvid u izvorni kôd, uz mogućnost promjene aplikacije i prilagodbe vlastitim potrebama. Softver je besplatan i može se preuzeti sa http://www.moodle.org.

Moodle je web aplikacija, te je kao takva dostupna iz nekog od standardnih preglednika weba, poput Internet Explorera. U okvirima aplikacije integrirane su alati za kreiranje web stranice premeta i alati za kontrolu pristupa, koji omogućavaju jednostavan način postavljanja nastavne građe, vođenje online rasprava i razgovora, izradu kvizova i anketa, te prikupljanje i predaju zadataka.

Moodle platforma se koristi za :

- distribuciju materija za vježbu
- „spremanje“ vježbi kako bi učenici mogli vrijeme učenja jednim dijelom prilagoditi svom tempu
- timski rad koji je potaknuo natjecateljski duh i samostalno učenje (Sl. 3 Radovi učenika postavljeni na forum)



Sl. 3 Radovi učenika postavljeni na forum

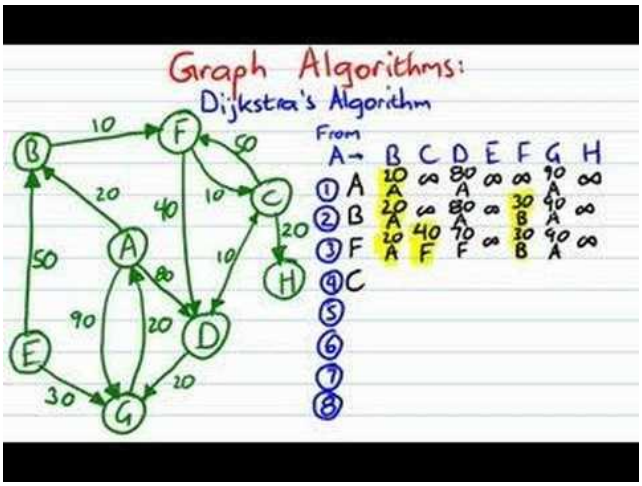
III. TEORIJSKI OKVIR

U procesu pripremanja zadataka s natjecanja počevši od školske razine zaključeno je da nastavno gradivo ne prati nastavne sadržaje redovite nastave programiranja. Tom zaključku je uslijedila detaljna analiza izvedbenog nastavnog programa. Naime, programiranje u prirodoslovno-matematičkim gimnazijama se uči od drugog razreda. U prvom razredu unutar nastavne cjeline osnove programiranja učenici se upoznaju s algoritamskim pristupom rješavanju problema potpuno neopterećeni učenjem sintakse programskog jezika.

Učenici su bili uključeni u veći dio aktivnosti. Te aktivnosti otvorile su čitav niz informacija koje su bitno doprinijele implementaciji i vođenju projekta. Unutar njih iščitavani su interesi učenika, njihove želje ne samo za nastavnim sadržajima koji prate gradivo na natjecanjima nego i novim načinima poučavanja i učenja.

Aktivnosti koje su provedene u okvirima teorijske implementacije projekta:

- Upoznavanje učenika sa novim tehnologijama e-učenja, konkretno upoznavanje s Moodle platformom
- Testiranje redovitog nastavnog programa. Tad je došlo do bitnih informacija, a to je da učenici oko 70% nastavih sadržaja ne poznaju i ne koriste redovitoj nastavi. Primjerice, strukture podataka kao što su liste, stabla, algoritme sortiranja.
- Skupljanje informacija o interesima učenika za učenje naprednog programiranja koje je u korelaciji s matematikom, kombinatorikom i teorijom grafova (Sl.4 Dijkstrin algoritam).



Sl.4 Dijkstrin algoritam

IV. PRIJEDLOG IZVEDBENOG PLANA I PROGRAMA NASTAVE INFORMATIKE (PROGRAMIRANJE)

Danas je u nizu programskih jezika teško odabrati onaj koji je pogodan za učenje programiranja. Takav bi jezik s jedne strane morao biti fleksibilan i jednostavan, kako bi prve programerske korake učinio što bezbolnijim. S druge strane on bi trebao biti dobra podloga za nastavak programiranja u bilo kojem drugom programskom jeziku. Naše je iskustvo da većina učenika ima poteškoća u usvajanju osnovnih koncepata algoritamskog načina rješavanja problema zbog savladavanja detalja pojedinih progra mskih jezika te posebice zbog uporabe različitih programskih jezika tijekom obrazovanja. Uobičajeno se na početku koriste jezici (LOGO, BASIC,Pascal, C), a kasnije kod podučavanja naprednijih koncepcija programiranja koriste se jezici za izradu velikih programskih projekata (C++, C#, Java). Navedeni jaz između „jednostavnih“ i „moćnih“ jezika zasigurno je smanjen pojavom Pythona. Radi se o jeziku koji s jedne strane može biti vrlo jednostavan i kao takav može poslužiti početnicima,a istovremeno je Python i moćan jezik koji podržava objektno-orijentirani princip programiranja [1].

Na prijedlog da se u III.gimnaziju Split uvede programiranje u Phytonu većina kolega aktiva informatike se nije pozitivno izjasnila. Još jedna od od prednosti koju

ima učenje Phytona je postojanje udžbenika koji na sistematičan i metodički oblikovan način vodi nastavnike kroz gradivo.

Ovaj rad se neće baviti prednostima programskih jezika tako da se izvedbeni plan može primijeniti na bilo koji objektno-orijetirani programski jezik.

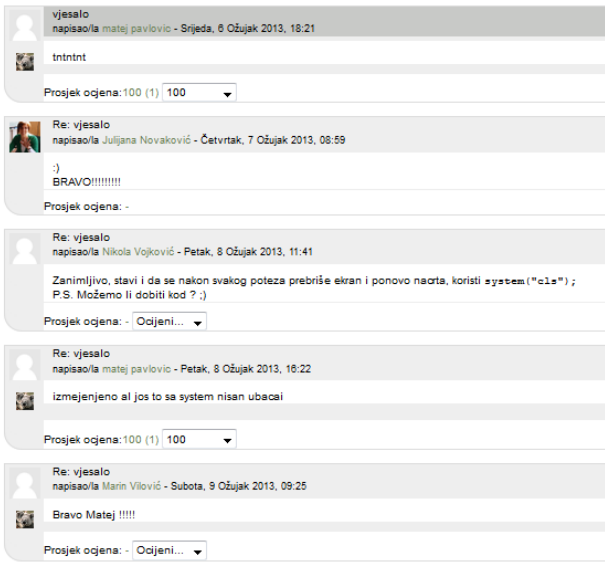
A. Programiranje pseudo jezik – algoritamski pristup

Pseudo jezik se sastoji od kratkih izraza na govornom jeziku koji opisuju i ukratko objašnjavaju pojedine radnje algoritma.
Pri zapisivanju algoritma pseudo jezikom treba nastojati zadatak razložiti na što manje radnje. Te bi radnje trebale biti tako male da se za svaku od njih može naći odgovarajuća naredba pseudo jezika. Kako u pseudo jeziku ne postoje ograničenja koja svaki programski jezik ima, pseudo jezik omogućuje mnogo pregledniji i jednostavniji prikaz načina na koji se zadatak može riješiti.

B. Nastavne metode i oblici rada

Nastava se odvija u informatičkoj učionici, dva puta tjedno i svaki od učenika radi sam. Računala imaju pristupi internetu.

Od verbalnih nastavnih metode koriste se sve metode počevši od objašnjavanja, razgovora i pripovijedanja, jer se novi pojmovi i nastavna građa na upoznaju i uvježbavaju na nastavi. Moodle platforma ne koristi se za samostalno učenje. Preko nje učenici ostvaruju komunikaciju koja nije vođenja od strane nastavnika (Sl.5 primjer komunikacije).



Sl.5 primjer komunikacije

Od neverbalnih nastavnih metoda koristi se metoda demonstracije.

C. Nastavni sadržaji

Dodatna nastava programiranje kao dodatni nastavni sadržaj podučava učenike:

- rješavanje problema
- pisanje algoritama
- razumijevanju i kritičkoj ocjeni prikupljenih informacija
- donošenju zaključaka na temelju prikupljenih informacija.

Zadaće ove nastave su upoznavanje s:

- standardnim numeričkim algoritmima (djeljivost, prosti brojeve, savršeni brojevi,...)
- s poljima (pretraživanje, sortiranje,...)
- s rekurzivnim potprogrami
- sa stringovima (palindromi, parsiranje,...)
- sa složene strukture podataka (vezana lista, red, stog,stablo,...)

U tablici (Tablica 1) je prikazan izvedbeni paln i program dodatne nastave naprednog programiranja za školsku godinu 2012./13. i 2013./14.

Naziv nastavne cjeline (kompleksa)	Sadržaji
1. STRUKTURE PODATAKA I ALGORITMI	Uvodni sat 1.1. Pojam algoritma i način zapisivanja 1.2. Pseudojezik i dijagram toka 1.3. Neki algoritmi za prirodne brojeve 1.4. Nizovi i sortiranje 1.5. Još neki matematički algoritmi
2. NAPREDNIJI C++-u	2.1. Pokazivači 2.2. Strukture 2.3. Elementarni algoritmi 2.4. Biblioteke 2.5. Funkcije 2.6. Klase i objekti 2.7. Standard Template Library
3. LISTE	3.1. Općenita lista 3.2. Stog 3.3. Red
4. STABLA	4.1. Uređeno stablo 4.2. Binarno stablo
5. GRAFOVI	5.1. Modeliranje pomoću grafova 5.2. Zapis grafa 5.3. Flood fill algoritam 5.4. Dijkstrin algoritam 5.5. BFS 5.6. Eulerovska staza
6. OBLIKOVANJE ALGORITAMA	6.1. Metode izbrajanja 6.2. Modeliranje rekurzivnim relacijama 6.3. Metoda „podijeli pa vladaj 6.2. Dinamičko programiranje 6.3. „Pohlepni“ pristup 6.4. Backtraking

Tablica 1.

D. Okvir prijedloga novih nastavnih sadržaja redovite nastave

U tablici 2 je izvedbeni plan i program za 2. i 3. razred prirodoslovno-matematičke ginazije

Tablica 2

Naziv nastavne cjeline (kompleksa)	Sadržaji
1. OSNOVE PROGRAMSKOG JEZIKA C++	1.1. Osnovna struktura programa 1.2. Komentari 1.3. Ključne riječi 1.4. Kompajliranje i pokretanje programa
2. KONSTANTE I VARIJABLE – OPERATORI I IZRAZI	2.1. Konstante i varijable 2.1.1. Imena varijabli 2.1.2. Tipovi podataka 2.1.3. Tip void 2.1.4. Veličine podataka 2.1.5. Konstante 2.1.6. Deklaracije 2.2. Operatori i izrazi 2.2.1. Aritmetički operatori 2.2.2. Unarni operatori 2.2.3. Relacijski i logički operatori 2.2.4. Operatori pridruživanja 2.2.5. Uvjetni operator: ? 2.2.6. Prioriteti i redoslijed operatora
3. ULAZ I IZLAZ PODATAKA	3.1. Funkcije getchar i putchar 3.2. Funkcije iz datoteke <iostrem> 3.3. Funkcije gets i puts 3.4. Funkcija cin 3.4.1. Učitavanje cijelih brojeva 3.4.2. Učitavanje realnih brojeva 3.5. Funkcija cout 3.5.1. Ispis cijelih brojeva 3.5.2. Ispis realnih brojeva 3.5.3. Ispis znakovnih nizova
4. KONTROLA TOKA PROGRAMA	4.1. Izrazi i naredbe 4.2. Naredbe uvjetnog grananja 4.2.1. If naredba 4.2.2. if-else naredba 4.2.3. Višestruka if-else naredba 4.3. switch naredba 4.4. while petlja 4.5. for petlja 4.6. do-while petlja 4.7. Naredbe break i continue
5.POLJA	5.1. Definicija i inicijalizacija polja 5.2. Polja znakova 5.3. Biblioteka string 5.3. Funkcije za rad sa stringovima 5.5. Sortiranje
6. FUNKCIJE	5.1. Definicija funkcije 5.2. Deklaracija funkcije 5.3. Prijenos argumenta 5.4. Inline funkcije 5.5. Rekurzivne funkcije 5.6. Funkcije s varijabilnim brojem argumenata 5.7. Polje kao argument funkcije (prenos po adresi)
7. BIBLIOTEKE	7.1. Naredba include 7.2. Naredba #define 7.3. Uvjetno uključivanje 7.4. Predefinirani makroi.
8. Standard TemplateLibrary	8.1. Vector 8.2. List 8.3. Map
9. OSNOVE OBJEKTNO USMJERENOG PROGRAMIRANJA	9.1. Klase i objekti
10. SLOŽENI ALGORITMI I TRAJANJE NJIHOVA IZVOĐENJA	
10. ALGORITMI SORTIRANJA	10.1. Buble sort 10.2. Mergesort 10.3. Quicksort

E. Vrednovanje

Kako je dodatna nastva napredovala, sadržaji i aktivnosti su se obogaćivale tako su učenici pokazivali sve više interesa za timskim radom koji bi se izvodio online.

Učenici su bili zadovoljni jer se na nastavi nastojalo rješiti što više natjecateljskih zadatka.

Pripreme za nastavu su bile opsežne i kompleksne. Trebalo je obuhvatiti gradivo od struktura podataka do teorije grafova. Nije uvijek sve išlo dobro. Najveći problem je nedostata sistematičnih i metodičko-oblikovanih nastavnih sadržaja.

Literatura

[1] P. Brođanac, L.Budin "Python – jezik za podučavanje algoritamskog pristupa rješavanju problema," MIPRO 2010

Koliko programiranja i kada nam treba u kurikulu za osnovne i srednje škole?

Marina Čičin-Šain*, Snježana Babić**
* Ekonomski fakultet Rijeka/Rijeka, Hrvatska
** Veleučilište u Rijeci, Rijeka, Hrvatska
*marinacs@efri.hr
** sbabic@veleri.hr

Sažetakt – Danas je učenicima razvoj znanja, vještina i sposobnosti iz područja programiranja omogućen na različim obrazovnim razinama u različitom obimu kroz formalni, neformalni i informalni oblik obrazovanja. Međutim obrazovna praksa pokazuje da je programiranje za većinu učenika poput matematike, uglavnom slabo prihvaćeno i nepopularno. Stoga je važno pronaći načine koji će olakšati process učenja i poučavanja programiranja, te potaknuti kod učenika želju za razvojem u području programiranja. S druge strane, važno je pronaći odgovarajuću razinu zastupljenosti programiranja u kurikulu za osnovne i srednje škole. (A moglo bi biti lakše.). Ovaj rad opisuje jednu mogućnost kako se to može postići.

I. UVOD

Činjenica je da današnji učenici koriste računala, da provode svakodnevno često i više sati za računalom, tabletom, pametnim telefonom i drugim uređajima uz pomoć kojih vrše različite aktivnosti, od privatnih, poslovnih do obrazovnih.

Tako danas kao jednu od glavnih komponenata obrazovnog sustava imamo generaciju koju je Prensky [1] još 2001. godine nazvao "*digitalni urođenici*". Autor navodi, da nova generacija učenika "… *ubrzano prima informacije, voli paraleno obrađivati i raditi nekoliko poslova istovremeno, da najbolje rade u interentskom okruženju, da su više skloniji igrama nego ozbiljnom radu, ...*". Današnja praksa, nažalost, potvrđuje u velikom broju slučajeva da su upravo njihove online aktivnosti češće gubitak vremena nego aktivnosti koje doprinose njihovom razvoju, edukaciji i napretku.

S druge strane, u okviru formalnog osnovnoškolskog i srednješkolskog obrazovanja učenicima je omogućeno stjecanje znanja, vještina i sposobnosti iz područja informatike/računalstva kroz informatičke predmete ili pak kroz primjenu računala u drugim nastavnim predmetima. Pod utjecajem različitih trendova, sadržaji kurikula za osnove i srednje škole iz područja informatike se kontinuirano mijenjaju.

Tako se danas javlja potreba za razvojem pismenosti više razine u području informatike vezane uz programiranje. Premda je programiranje u svijetu informatike vezano uz početak samog razvoja znanstvenog područja informatike, danas upravo postoji potreba za razvojem kompetencija učenika u tom smjeru.

Stoga se u nastavku ovoga rada upravo vraćamo djelomično u prošlost koja u današnjim okvirima ima ponovno veliki potencijal u procesu učenja i poučavanja mladih ljudi u području programiranja.

II. ZAŠTO PROGRAMIRATI?

Studija na razini europske zajednice [2] upozorava da možemo biti zadovoljni digitalnom pismenosti naših učenika, ali da u većini zemalja uopće ne može biti govora o zadovoljstvu što se tiče informatičke edukacije, pod kojom se podrazumijeva ozbiljnije poznavanje rada računala (algoritmi, programiranje, paralelizam i sl.) U istoj studiji naglašava se da se stanje u tom području značajno pogoršava nakon pionirskih pokušaja iz 1970. i 1980. godine. Oni koji su stariji sjetit će se da je zaista kod početka uvođenja računala u škole programiranje bilo znatno više zastupljeno nego danas!

I konačno, ista studija upozorava europljane, da je digitalna pismenost nedovoljna, ako ne želimo biti svedeni na čiste potrošače novih tehnologija, a ne njihove kreatore i autore, za što je naravno potrebno znanje na višoj razini.

Navedeni zaključci koji vrijede na razini Europe za Hrvatsku vrijede još i više. U našoj zemlji glavni su problemi nezaposlenost i odljev mozgova. Zanimanja koja su u bogatijim zemljama deficitarna su upravo zanimanja za koje je potrebna informatička edukacija. No dok za druga zanimanja "rad za druge zemlje" znači i selidbu u te zemlje, za informatička zanimanja to uopće ne mora biti tako. Informatičari mogu fizički ostati u Hrvatskoj, a raditi za naručitelja iz bilo kojeg dijela svijeta.

III. KADA ZAPOČETI?

Da bi se postigla dovoljna razina znanja ne može se sa ozbiljnijim temama započeti tek na fakultetu, nego je potrebno započeti ranije, dakle u osnovnoj i srednjoj školi.

Prva je autorica ovog teksta počela programirati nakon studija teoretske matematike na PMF-u kao programer pripravnik u Ini prije 45 godina. U ono vrijeme bila je rijetkost da netko uopće uči programirati ili koristiti računalo na bilo koji način, a učenje o računalima nije uopće bilo uključeno u nastavu u školama. Dakle, ako se nešto o računalima i učilo, to je bilo nakon formalnog školovana. To sigurno nije bilo dobro i već onda se znalo da je bolje početi ranije.

Pitanje je koliko ranije? Odgovor je čim ranije, onda kada se i druge stvari uče, primjerice računati, čitati, raspoznavati boje, pisati, strani jezici i slično. Djeca koja su kao malena naučila strani jezik, igrati tenis ili skijati bolje to rade od svojih vršnjaka koji su počeli tek kada narastu. Sa programiranjem je ista stvar, samo način na koji djeca uče mora biti prilagođen uzrastu djece. Zato za sasvim malu djecu ni jedan postojeći programski jezik nije dosta dobar. Za sasvim malu djecu oni su svi preteški. Djeca mogu učiti koristiti kompjutor, ali ne i programirati na nekom od postojećih jezika, čak ni ako se radi o Logu ili Basicu.

Pa ipak, učiti programirati može se učiti tako prirodno i lako, kako se pomoću slikovnica može učiti čitati, pomoću igre Čovječe ne ljuti se brojati ili pomoću crtanih filmova i dječjih stihova engleski jezik. Prva autorica ovih redaka je prije više od trideset godina razvila metodu pomoću koje su djeca mogla napisati program, provjeriti kako program radi, naučiti koliko je važan redoslijed naredbi, kako se ponašaju varijable u memoriji računala, a sve su to radila pomoću igre i igračke koju su sami izrađivali. Igračku smo nazvale "Mema".

IV. TKO TREBA PODUČAVATI I NA KOJOJ OPREMI?

Ako se radi o osnovnoj školi programiranje po metodi koja se predlaže može predavati učitelj ili učiteljica bez nekog posebnog znanja iz područja programiranja.

Možda je to i glavna prednost ove metode, da zapravo ne zahtjeva nikakvu dodatnu edukaciju učitelja ili učiteljica!

Druga prednost je da za ovu metodu nije potrebna nikakva dodatna oprema, osim za učitelja ili učiteljicu računala s mogućnosti spajanja na Internet kako bi mogli pristupiti nastavnim materijalima tijekom pripreme za nastavni sat.

V. POČETNICA PROGRAMIRANJA

Metoda koju preporučujemo opisana je još 1990. u knjizi "Kompjutersta početnica", koju je izdala Školska knjiga [3]. Knjiga se sastojala iz dva dijela: prvog, koji je bio namijenjen radu sa djecom u prvom ili drugom razredu osnovne škole i nije zahtijevao računalo, te drugog dijela koji se nadovezivao na prvi u kojem su bile opisane osnove u radu na programskom jeziku Basic.

Upravo prvi dio knjige je ono na što želimo upozoriti, jer je on danas aktualan isto kao i daleke 1990. Za razliku od drugog dijela koji se može smatrati zastarjelim.

Novost je da je koautorica ovog rada, ove godine taj prvi dio priredila za korištenje na računalu i da je dostupan na adresi:

www.magistra-nova.hr/dokuwiki_pocetnica/doku.php

Važno je naglasiti sljedeće:

- Metoda je namijenjena učitelju ili učiteljici, koji su voljni učenike osnovne škole pripremiti za programiranje.
- Učitelj ili učiteljica pri tom ne moraju znati mnogo o računalima, zapravo nije potrebno nikakvo predznanje.
- Zamišljeno je da učitelji zajedno sa svojim učenicima uče, korak po korak.

Na online stranicama [4] učitelji mogu pronaći nastavno gradivo kao i upute za svaki nastavni sat.

U okviru nastavne metode u prvoj fazi "*učenja programiranja bez računala*" koristi se igračka nazvana "Mema" koju djeca sama na nastavnom satu izrađuju od uobičajenih stvari, kao što su: bojice, ljepilo, papir i slično (vidjeti: sliku 1).



Slika 1. Igračka Mema koju djeca sama izrađuju

Početnica programiranja se sastoji iz sljedećih dijelova:

- Kada početi programirati?
- Opis metode
- Kako početi?
- Izrada "Meme"
- Ulaz
- Izlaz
- Obrada
- Igrokaz "Moj kompjutor"
- Napredniji rad s "Memom"
- Sljedbenik i prethodnik
- Ponavljanje gradiva
- Slikovnica "ANDOR"
- Priprema za LOGO
- Igre na dvorištu
- Umjesto zaključka: što smo sve naučili?
- Zadaci: 1, 2, 3, 4, 5,6, 7
- Pitanja za grupu B
- Pitanja za grupu C.

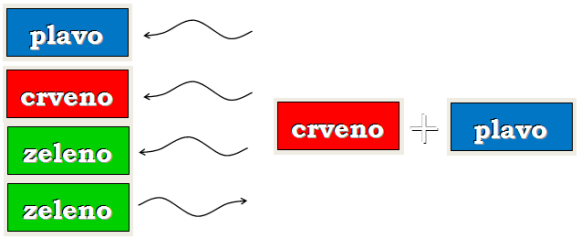
Tijekom primjene metode učenici neće biti dodatno opterećeni, učit će ono što ionako uče u prvom razredu: zbrajati i oduzimati do dvadeset, raspoznavati boje, crtati, razvijati motoriku.

Pri tome će sasvim usput učenici čuti za neke pojmove iz informatike, ne na razini definicija, nego upravo čemu te stvari služe. Kao primjer neki od pojmova su:

- korisnik

- program
- naredba
- varijabla
- redoslijed naredbi
- ulaz
- obrada
- izlaz
- podaci
- konstante
- kompatibilnost
- kapacitet polja
- adresa i još mnogi drugi

Način na koji će se usvajati navedeni pojmovi iz područja informatike, učitelj ili učiteljica doznaje radom iz nastavnog sata u sat. Tako na primjer, u nekom trenutku učenik ili učenica upoznaju program za zbrajanje kao što je prikazano na slici 2.



Slika 2. Program za zbrajanje

Nove tehnologije omogućuju komunikaciju postavljanjem pitanja kad nešto nije jasno ili upozorenjem ako smo u nečemu pogriješile. Učitelji i učiteljice pozivaju se da nas obavijestite kako napreduju u svom radu, da nam pošalju koju fotografiju svojih učenika i učenica dok rade, koji crtež koji su izradili, neku njihovu primjedbu koju smatraju zanimljivom i sl. Već sama činjenica da netko razmišlja o tome da radi po ovoj metodi dovoljna je da nam se javi s eventualnim primjedbama.

O korisnicima metode će ovisiti koliko će metoda ponovni uspjeh.

Pri radu sa učiteljima ili učiteljicama često se nailazi na pitanje: "*A što ako mi učenik ili učenica postave pitanje na koje ne znam odgovoriti*?"

Ne treba strahovati od takvih pitanja, ona su sasvim prirodna. Djeca puno puta postavljaju pitanja na koja odrasli nemaju odgovor. Po našem mišljenju takva pitanja i takvi učenici ili učenice su prava blagodat; zahvaljujući njima i sami ćemo uvijek biti prisiljeni učiti. Nastavnik se ne treba pretvarati da zna odgovor, ako ga slučajno ne zna. Bolje je pohvaliti pitanje i pokušati naći odgovor na njega do slijedećeg sata. Možda i u tome možemo pomoći.

VI. Što dalje

Nakon prvih koraka treba nastaviti. Paralelno sa nastavom matematike, može se u nižim razredima osnovne škole i za programiranje još neko vrijeme koristiti Mema. Sadržaji koje bi tu trebalo uvesti svakako su grananje i petlje, što u Početnici nije obuhvaćeno.

Naravno da u višim razredima treba početi programirati i na računalu. Kada su učenici u osnovnoj školi usvojili osnovna znanja, njima je znatno lakše koristiti neki pravi programski jezik [5].

Svjedoci smo da studenti primjerice na fakultetu, koji nikad ranije nisu programirali vrlo teško usvajaju ta znanja, odnosno usvajaju ih samo oni koji su za to talentirani. Tvrdimo je da i prosječan učenik može naučiti programirati, samo ako počne dosta rano s edukacijom.

VII. Zaključak

Korištenje računala na tipično korisnički način je lakši, ali i lošiji način. Odustajanje od programiranja pokazalo se kao loša praksa.

Pa ako je učenje programiranja važno, čak jako važno za jednu zemlju poput Hrvatske, onda je važno i kako ta znanja učiniti dostupna našim učenicima.

Učenicima će biti lakše ako uče postepeno i počnu čim ranije. Metoda i Internet stranica iz ovoga rada mogu pomoći učiteljima i učiteljicama da osnovna znanja prenesu svojim malim učenicima.

Literatura

[1] Prensky, Marc. "Digitalni urođenici, digitalni pridošlice." Edupoint, časopis o primjeni informacijske tehnologije u obrazovanju, www. carnet. hr/časopis/40/članci/3. www. carnet. hr/časopis/40/članci/3, 2005. [1.2.2014.]

[2] W.Gandler, A. Petit, G. Berry, B. Demo, J. Vahrenhold, Informatics education: Europe cannot afford to miss the boat, April 2013

[3] M. Čičin-Šain, Kompjuterska početnica, Školska knjiga, Zagreb 1990 .

[4] www.magistra-nova.hr/dokuwiki_pocetnica/doku.php

[5] Čičin-Šain, Marina, Utjecaj predznanja na usvajanje novih informatičkih sadržaja u osnovnoj školi, Sim-pozij "Kompjutor na sveučilištu", Cavtat 1987.

Webinar u funkciji promicanja prava na pristup informacijama

Tamara Horvat i Renata Pekorari
Digitalni informacijsko-dokumentacijski ured Vlade Republike Hrvatske
Zagreb, Hrvatska
thorvat@hidra.hr, pekorari@hidra.hr

Sažetak - Pravo na pristup (službenim) informacijama, jedno od osnovnih ljudskih prava, u Republici Hrvatskoj je Ustavno i zakonsko pravo. Rad će se baviti jednim od mogućih načina promoviranja tog prava. Digitalni informacijsko-dokumentacijski ured proaktivnom objavom službenih dokumenata tijela javne vlasti na svojim mrežnim stranicama direktno osigurava pravo na pristup informacijama svima zainteresiranima. Promicanju prava na pristup informacijama daje svoj doprinos i putem edukacijskih programa namijenjenih državnim i javnim službenicima i zainteresiranoj javnosti. Kako je nemoguće doći „uživo" do svakog potencijalno zainteresiranog službenika ili građanina, Ured razvija i izvodi i *online* edukacijske programe. Nakon prvog asinkronog *online* programa *Naučite kako, brzo i lako, do pravnih propisa!* ponuđenog svima u slobodnom pristupu na mrežnim stranicama, Ured se usredotočio na razvoj i izvedbu *webinara* koji su se u zadnjih nekoliko godina razvili u vrlo popularan edukacijski alat. U radu će se opisati projektni plan jednog od prvih *webinara* izrađenih u sklopu državne uprave kao i prvi pokazatelji njegova uspjeha. Također, rad će predstaviti i daljnje planove za korištenje *webinara* u svrhu razvijanja svijesti o pravu na pristup informacijama.

I. UVOD

A. Pravo na pristup informacijama

Pravo na pristup (službenim) informacijama, jedno od osnovnih ljudskih prava, u Republici Hrvatskoj je Ustavno i zakonsko pravo. U ožujku 2013. godine stupio je na snagu novi Zakon o pravu na pristup informacijama [1] (u daljnjem tekstu Zakon) koji, za razliku od prethodnog iz 2003. godine [2], snažno naglašava važnost proaktivne objave informacija u ostvarivanju prava na pristup. Smisao je informaciju, koja može biti zanimljiva i većem broju zainteresiranih, čim prije objaviti na mrežnim stranicama i tako je učiniti dostupnom. Zainteresirani korisnik moći će je pronaći i koristiti i bez postavljanja pojedinačnog zahtjeva za istom informacijom tijelu javne vlasti. Digitalni informacijsko-dokumentacijski ured Vlade Republike Hrvatske [3] (u daljnjem tekstu Ured) proaktivnom objavom službenih dokumenata tijela javne vlasti na svojim mrežnim stranicama direktno osigurava pravo na pristup informacijama svima zainteresiranima.

Promicanju prava na pristup informacijama Ured doprinosi i edukacijskim programima namijenjenima tijelima javne vlasti, dakle državnim i javnim službenicima koji trebaju provoditi Zakon, ali i zainteresiranoj javnosti u svrhu upoznavanja sa zakonskim pravima. Kako je nemoguće doći „uživo" do svakog potencijalno zainteresiranog službenika ili građanina, Ured u zadnje dvije godine, uz klasične, razvija i izvodi i *online* edukacijske programe. Nakon prvog asinkronog *online* programa *Naučite kako, brzo i lako, do pravnih propisa!* [4] ponuđenog svima u slobodnom pristupu na mrežnim stranicama, Ured se usredotočio na razvoj i izvedbu *webinara* koji su se u zadnjih nekoliko godina razvili u vrlo popularan edukacijski alat.

B. Webinar kao edukacijski alat

Webinar je izraz koji je preuzet iz engleskog jezika i skraćenica je od sintagme *web based seminar*. U hrvatskom jeziku ne postoji prijevod ovog izraza. Predstavlja alat kojim se seminari prenose putem mreže. Osim seminara to mogu biti razne duže ili kraće prezentacije, predavanja, radionice, konferencije i ostali događaji koji se prenose u realnom vremenu putem interneta.

Njihova osnovna prednost je jednostavnost korištenja, nema troškova i velikih zahtjeva za korisnike. Sustav za *webinare* uključuje video, audio i tekstnu komunikaciju između korisnika, razmjenu i zajednički rad s datotekama, anketiranje sudionika, preuzimanje nadzora nad udaljenim računalima i drugo. *Webinari* omogućuju davanje, primanje i komentiranje informacija, što omogućava interakciju između predavača i sudionika koji se nalaze na više različitih fizičkih lokacija [5].

Za potrebe prvog *webinara* Ureda korištena je usluga koju je uspostavio Centar za e-učenje pri Srcu [6]. Usluga korisniku omogućava otvaranje virtualne sobe u kojoj priprema i održava *webinar* te definira način pristupa, radnu površinu, sudionike i njihove uloge. U svakoj sobi standardno su unaprijed definirane tri radne površine: za razmjenu, za suradnju i za diskusiju, a korisnik usluge može otvoriti i izraditi nove radne površine i kombinirati dostupne alate u sustavu.

U sustavu postoji više uloga, a osoba kojoj se odobri otvaranje sobe za *webinar* ima ulogu domaćina. Sudionike *webinara* domaćin poziva u sobu pozivnicom preko e-pošte koja sadrži dan i sat/vrijeme održavanja te URL-adresu sobe, a sudionike se može i unaprijed upisati ili im odobriti ulazak prije samoga početka. Sustav također omogućava snimanje *webinara* radi njegova eventualnoga kasnijeg prikazivanja. Ovaj alat za e-učenje sigurno će biti zanimljiv svima koji drže nastavu na raznim lokacijama ili žele održati seminare ili

konzultacije s osobama koje se nalaze bilo gdje u svijetu. Osim što omogućavaju prisutnost i slikom i glasom, *webinari* ujedno štede vrijeme i visoke troškove putovanja

II. PRVI *WEBINAR* U DIGITALNOM INFORMACIJSKO-DOKUMENTACIJSKOM UREDU

A. Koraci u izradi webinara Predstavljanje tražilice pravnih propisa Republike Hrvatske - CADIAL

Ured već nekoliko godina provodi prezentacije uživo na kojima predstavlja tražilicu pravnih propisa Republike Hrvatske CADIAL. Sadržaj prezentacija uživo pretočen je i u prvi *webinar* Ureda, ali i jedan od prvih i na razini cijele državne uprave, pod nazivom *Predstavljanje tražilice pravnih propisa Republike Hrvatske – CADIAL.*

S obzirom na sve veće zahtjeve korisnika za održavanjem prezentacija izvan Zagreba, odnosno izvan mjesta sjedišta Ureda, kao i limitirana sredstva za službena putovanja, krenulo se u premošćivanje takve situacije. Osim korisnika koji su dislocirani, povećao se broj zainteresiranih za edukaciju onih kojima radne obveze ne omogućuju duže izbivanje s radnog mjesta. Sve to vodilo je prema osmišljavanju novih načina da Ured što kvalitetnije i što sveobuhvatnije nastavi promovirati svoje informacijske resurse i informacijske usluge i tako ispunjavati svoju misiju, odnosno ostvarivanje prava na pristup informacijama za sve zainteresirane.

Kada je donesena odluka da se osmisle i počnu izvoditi *online* programi, pristupilo se odabiru prikladnih alata imajući u vidu profil korisnika i željeni ishod učenja koji će omogućiti da zainteresirani korisnici, nakon što prođu edukaciju, mogu suvereno koristiti tražilicu i vrlo lako pronalaziti pravne propise koji im trebaju [7].

Odabran je *webinar* kao najprikladniji alat s obzirom da je profil potencijalnih korisnika informacijskih usluga Ureda raznovrstan, a fizički su dislocirani. Također, *webinar* kao alat nije tehnički zahtjevan, a brz je i jednostavan za instalaciju i početak korištenja. Sudionik *webinara* mora na svom računalu napraviti vođenu provjeru kojom će utvrditi jesu li mu računalo i pretraživač propisno spojeni s programom *Adobe Connect*. Nema ostalih posebnih tehničkih zahtjeva, sudioniku su od dodatne opreme potrebni samo zvučnici, a predavaču zvučnici i mikrofon, te po želji web kamera. Sve to upućuje da alat nije skup niti za sudionika niti za predavača. Osim toga, alat je intuitivan i jednostavan za korištenje te se svatko s malo iskustva u komunikaciji putem interneta može lako snaći. Jedna od velikih prednosti *webinara,* kao što je već naglašeno, jest i njegova interaktivnost. Naime, predavača može biti više te također ne moraju biti na jednome mjestu, a mogu međusobno i sa sudionicima komunicirati u realnom vremenu. Sudionici su također prisutni uživo i ponuđena im je mogućnost da se, osim što slušaju predavanje, mogu u bilo kojem trenutku uključiti u razgovor putem *chata* ili pak dati signal predavaču da uspori, ubrza predavanje, ohrabriti ga virtualnim pljeskom ili podizanjem ruku te postavljanjem pitanja.

Webinar pruža i opciju kojom se sudionicima omogućuje javljanje glasom i slikom, međutim, zbog vjerojatne inkompatibilnosti opreme i predvidivih smetnji, uobičajeno je da im se omogući javljanje samo *chatom*.

Ovaj alat se sve više koristi unazad nekoliko godina, pogotovo u znanstvenoj zajednici i u svim skupinama koje teško mogu pronaći adekvatno fizičko mjesto i vrijeme za razmjenu mišljenja i ideja.

Ideja o korištenju *webinara* kao edukacijskog alata u Uredu je potvrđena i pohađanjem nekoliko *webinara* koje je organizirala Inicijativa Partnerstvo za otvorenu vlast [8]. Naime, razne radne grupe ove Inicijative [9], koriste već nekoliko godina *webinare* za širenje informacija o svom radu, različitih analiza i rezultata istraživanja, preporuka te za komunikaciju sa sudionicima iz zemalja članica Inicijative diljem svijeta. Koriste isti *Adobe Connect* program s tim da redovito *webinare* snimaju kako bi ih svi zainteresirani koji nisu mogli sudjelovati uživo mogli pregledavati i preslušavati u vrijeme kada njima odgovara [10]. Njihovi su *webinari* poslužili kao dobar primjer i ukazali na sve prednosti jednog takvog *online* programa.

Nakon odabira alata stupilo se u kontakt s Centrom za e-učenje pri Srcu koji je na zahtjev, besplatno, otvorio virtualnu sobu namijenjenu *webinaru Predstavljanje tražilice pravnih propisa Republike Hrvatske - CADIAL*. Na predavačima je bilo osmisliti sadržaj odnosno prilagoditi već postojeći koji se provodi uživo te testirati sustav.

U rujnu 2013. godine održan je pilot *webinar* na koji su pozvani eksperti iz područja e-učenja, odnosno stručnjaci koji se bave sličnim sadržajima. Zamoljeni su da daju svoje mišljenje i o sadržaju i o kvaliteti tehničke izvedbe. Povratne su informacije bile odlične te se krenulo u redovite mjesečne izvedbe prvog *webinara* Ureda.

B. Realizacija webinara

Planom edukacijskih programa Ureda, predviđeno je održavanje *webinara* jednom mjesečno. Nekoliko dana prije zadanog termina održavanja *webinara*, na izabrane ciljane adrese šalju se obavijesti o održavanju, datumu i vremenu u kojem se održava te načinima prijave sudionika. Nakon što se sudionik prijavi, šalju mu se upute koje sadržavaju adresu *webinara*, način prijave u virtualnu sobu (sudionici se prijavljuju kao gosti, a lozinka nije potrebna), poveznicu za testiranje veze, napomenu da su od dodatne opreme potrebni samo zvučnici te naputak što učiniti u slučaju problema s priključivanjem. Naime, najčešći su problem sigurnosne postavke na serverima sudionika na kojima njihovi informatičari trebaju omogućiti izlazak na standardne portove Srca. Također, na mrežnim stranicama Ureda objavljene su stalno informacije o *webinaru* te *online* prijavni obrazac s naznačenim datumima održavanja *webinara* putem kojega se

mogu prijaviti i svi ostali zainteresirani. Po prijavi dobivaju istovjetne upute kao i oni prethodni. Na sam dan *webinara*, svim prijavljenim sudionicima šalje se kratki podsjetnik da *webinar* na koji su se prijavili uskoro počinje i ponovo se šalje adresa *webinara* kako ne bi morali pretraživati „stare" mailove.

Predviđeno trajanje *webinara* je 60 minuta, a pri otvaranju virtualne sobe postavljen je od strane Srca limit od 100 sudionika po jednoj izvedbi *webinara*. Predavači se prijavljuju u sustav pola sata prije zakazanog početka, te do početka programa komuniciraju sa sudionicima putem *chata*. Kada u zakazano vrijeme započne *Predstavljanje*, predavači se uključuju zvukom i objašnjavaju sudionicima pravila rada te započinju s edukacijskim programom. Edukacijski program se sastoji od *PowerPoint* prezentacije i sadržaja koji se pokazuju direktno na mrežnim stranicama Ureda. Sudionicima je ostavljena mogućnost da se cijelo vrijeme uključuju pitanjima i komentarima putem *chata.* S obzirom da su predviđena dva predavača za ovaj sadržaj, koji rade na odvojenim računalima, vrlo je lako posvetiti se predavanju sadržaja i komunikaciji sa sudionicima u isto vrijeme. Dok jedan predavač govori i može se u potpunosti koncentrirati samo na taj dio posla, drugi odgovara na pitanja i komentare sudionika. Prema potrebi, može se i drugi predavač glasovno uključiti kako bi potaknuo prvoga da na traženje sudionika nešto još bolje pojasni ili ponovi, a može i sam, kad je potrebno, nadopuniti izlaganje prvoga.

Pilot izvedba *webinara* održana je početkom rujna 2013. godine. Nakon dobivenih odličnih povratnih informacija, održano je do kraja siječnja 2014. godine još pet *webinara*. U Tablici I. prikazan je broj prijavljenih polaznika ukupno te broj prijavljenih polaznika izvan Zagreba. Veliki skok u broju prijavljenih u siječnju zasigurno ohrabruje. U prosincu je prvi put poslan poziv na prijavu na *webinar* i svim jedinicama lokalne i područne samouprave što je rezultiralo velikim povećanjem broja prijavljenih izvan Zagreba u siječnju.

Tablica I.

datum webinara	*broj prijavljenih*	*broj prijavljenih izvan Zagreba*
11.09. (pilot)	17	
20.09.	27	3
25.10.	29	8
29.11.	9	2
20.12.	10	5
28.01.	41	18

Broj prijava za *webinar* u veljači 2014. godine, iako još nije poslan službeni poziv za taj termin, ukazuje na trend povećanja broja sudionika u narednom periodu.

Pitanja i komentari postavljeni putem *chata* predstavljaju vrlo vrijedne povratne informacije. Ukazuju jasno što je potrebno od sadržaja obuhvatiti opsežnije, a što temeljitije izložiti.

III. PLANOVI ZA IZRADU NOVIH *WEBINARA*

S obzirom da Ured prema Zakonu direktno osigurava proaktivnu objavu službenih dokumenata kroz uspostavu i vođenje Središnjeg kataloga službenih dokumenata (u daljnjem tekstu Središnji katalog), postoje dodatni zahtjevi za edukacijskim programima, mimo onih koje Ured već provodi.

Naime, kako je Zakon relativno novijeg datuma, te obvezuje tijela javne vlasti dostavljati službene dokumente u Središnji katalog radi njihove formalne i sadržajne obrade te trajne pohrane i dostupnosti (čl. 10. stavak 2.), nameće se potreba za naputcima tijelima kako i što dostavljati u Središnji katalog. Kako bi Ured osigurao da Središnji katalog, kao središnje mjesto pristupa službenim dokumentima, bude učinkovit, potrebno je uspostaviti dobru komunikaciju s tijelima kako bi ona pravodobno dostavljala sve propisane dokumente.

Sva se tijela, obveznici dostave dokumenata u Središnji katalog, susreću s obvezom koja u Zakonu nije precizno razrađena, pa postavljaju pitanja Uredu o tome što i kako dostavljati. Kao pomoć pri odgovaranju na ta pitanja Ured je izradio dokumente u kojima se preciznije objašnjavaju obveze tijela. Ipak, s obzirom na različit profil tijela i osoba zaduženih za kontakt s Uredom i Središnjim katalogom, praksa je pokazala potrebu za dodatnim edukacijama koje će pomoći u osiguranju provedbe Zakona i time ostvarivanju prava na pristup informacijama. Problem predstavlja brojka obveznika dostave u Središnji katalog. Radi se o preko 5.000 tijela javne vlasti koja su dislocirana.

Kako Ured nije u mogućnosti u razumnom roku osigurati kvalitetnu edukaciju za sva tijela uživo, a što bi znatno utjecalo na dostavu dokumenata u Središnji katalog i na osiguranje kvalitetnog pristupa informacijama, odlučilo se za daljini razvoj *online* edukacije. Ponovno je, zahvaljujući svim prednostima i iskustvu dobre prakse, kao alat izabran *webinar*.

Planira se tijekom 2014. godine osmisliti *webinar* o načinu i sadržaju dostave službenih dokumenata u Središnji katalog, koji će biti namijenjen tijelima javne vlasti. Isti tim koji je izradio već predstavljeni *webinar* odredit će trajanje i sadržaj novog edukacijskog programa te se opredijeliti za platformu na kojoj će se uspostaviti virtualna soba.

IV. ZAKLJUČAK

Ured ima vrlo značajno mjesto u ostvarivanju prava na pristup informacijama. Poželjno je stoga pronalaziti neprestano nove načine kako bi se to pravo osvješćivalo i time se u sve većoj mjeri ostvarivala provedba Zakona. Kako je spektar usluga koje Ured nudi vrlo širok, treba koristiti što više kanala koji na bilo koji način mogu pomoći ostvarivanju postavljenih ciljeva Ureda.

Svaku ponuđenu uslugu potrebno je promovirati. Ukoliko se to ne učini, često će izostati željeni rezultat. S obzirom na doba u kojem računala kontinuirano povezuju

zainteresirane neovisno o mjestu na kojem se njihovi korisnici nalaze, puno je lakše pronaći prave kanale. Edukacijski alati koji se nude uvelike mogu olakšati način rada odnosno poslovanja.

Webinar se kao alat pokazao više nego primjerenim za trenutne potrebe rada Ureda, kao što je i asinkroni *online* program 2012. godine ispunio svoju svrhu. Iako korisnici kojima su edukacije namijenjene teže prihvaćaju nove načine usvajanja znanja i vještina, ipak se vide pomaci, koji su najviše vidljivi po sve većem interesu za izrađeni *webinar*, kao i odobravanju koje stiže nakon najava o održavanju *webinara* i o drugim srodnim sadržajima.

Webinar Predstavljanje tražilice pravnih propisa Republike Hrvatske – CADIAL jedan je od prvih webinara u okviru državne uprave. Trebao bi poslužiti kao primjer dobre prakse. *Webinar* kao alat jednostavan je za korištenje, tehnički i financijski nezahtjevan, omogućuje prijenos seminara, prezentacija, predavanja, radionica, konferencija i ostalih događaja koji se prenose u realnom vremenu putem interneta. Omogućuje davanje, primanje i komentiranje informacija, što omogućava interakciju između predavača i sudionika koji se mogu nalaziti na više različitih lokacija. Sve su navedene prednosti izuzetno značajne, posebice u današnje vrijeme nedostatnih financijskih sredstava i kroničnog nedostatka vremena.

Za prvi *webinar* Ureda postojao je predložak u vidu prezentacije uživo, dok će *webinar* o dostavi dokumenata u Središnji katalog zahtijevati više napora u osmišljavanju sadržaja. Ipak, upravo se njime planira doprijeti do svih onih koji na dnevnoj bazi moraju surađivati s Uredom kako bi svi zajedno, kreirajući Središnji katalog, radili na ostvarivanju prava na pristup informacijama.

LITERATURA

[1] Zakon o pravu na pristup informacijama, NN 025/2013

[2] Zakon o pravu na pristup informacijama, NN 172/2003, NN 37/2011, NN 77/2011

[3] www.hidra.hr, www.digured.hr, pristupljeno 04.02.2014.

[4] http://blog.hidra.hr/?lang=hr, pristupljeno 04.02.2014.

[5] M. Horvacki, Z. Martinović, D. Patafta i T. Perišić Pintek, Priručnik za rad sa sustavom za *webinare*: Adobe Connect 8.2. Zagreb: Srce, Centar za e-učenje, 2012.

[6] http://www.srce.unizg.hr/proizvodi-i-usluge/obrazovanje/centar-za-e-ucenje/, pristupljeno 04.02.2014.

[7] R. Pekorari i T. Horvat, „Otvoreni edukacijski *online* program u funkciji kvalitetnije dostupnosti službenih informacija“, MIPRO Proceedings, Hrvatska udruga za informacijsku i komunikacijsku tehnologiju, elektroniku i mikroelektroniku – MIPRO, 2013., pp. 960-963. http://docs.mipro-proceedings.com/ce/ce_057_1965.pdf, pristupljeno 04.02.2014.

[8] http://www.opengovpartnership.org/, pristupljeno 04.02.2014.

[9] http://www.opengovpartnership.org/get-involved/join-working-group, pristupljeno 04.02.2014.

[10] http://www.opengovpartnership.org/topics?f%5B0%5D=type%3Awebinar, pristupljeno 04.02.2014.

Izračun viših momenata distribucija pomoću VBA programskog jezika u Excelu

M. Gardijan*, B. Šego* i T. Škrinjarić*

* Ekonomski fakultet Sveučilišta u Zagrebu, Katedra za matematiku, Zagreb, Hrvatska
mail: mgardijan@efzg.hr, bsego@efzg.hr, tskrinjaric@efzg.hr

Sažetak - Standardni Markowitzev model optimizacije portfelja temelji se na dva parametra pojedinačnih dionica i samog portfelja – na očekivanoj vrijednosti te varijanci i kovarijanci kao mješovitom momentu. Izračun u Microsoft Excelu sastoji se od nekoliko koraka, pri čemu se koristi alat Rješavatelj. Međutim, ekonomska teorija i empirijska istraživanja pokazali su kako je u analizu potrebno uključiti i više momente: treći moment – koeficijent asimetrije, te četvrti moment – koeficijent zaobljenosti. Uvođenje viših momenata znatno otežava optimizaciju samog portfelja jer je potrebno računati sve mješovite momente trećeg i četvrtog reda. Ovaj rad nastoji, prvi puta u svijetu, prikazati kako se izračun samih momenata može znatno olakšati programiranjem u *Visual Basic for Application* (VBA) za Excel. Korištenjem VBA, dizajnirana je radna knjiga u kojoj se pokreću makronaredbe s funkcijama koje na temelju unesenih podataka o cijenama računaju prva četiri momenta distribucije prinosa te dobivene podatke prikazuju u radnim listovima na pregledan način. Funkcije zahtijevaju interakciju s korisnikom za što su implementirane korisničke forme i dijaloški prozori. Korišteni kod je jednostavan i koristi logičke petlje koje se lako mogu prilagoditi i za izračun viših momenata uvođenjem dodatnih varijabli.

I. UVOD

Tržišta kapitala, kao i druga financijska tržišta, dugi su niz desetljeća od velikog interesa investitorskoj i akademskoj zajednici. Upravljanje rizicima i optimalno upravljanje portfeljem jedna su o ključnih pitanja ekonomskih istraživanja, posebice u doba različitih turbulencija na financijskim tržištima, kao i u doba krize čije se posljedice još uvijek osjećaju. Markowitzev model (Markowitz 1952) kao temeljni model moderne teorije portfelja danas je dobro poznat u području financijskog investiranja. Temelji se na dva parametra distribucija prinosa dionica: prvi moment (očekivana vrijednost) te drugi moment (varijanca, te kovarijanca kao mješoviti drugi moment). Markowitz je, dakle, pretpostavljao da su jedino prva dva momenta distribucija prinosa važna za donošenje odluka o formiranju portfelja. Međutim, od samog nastanka ovog modela javljaju se kritike na račun istog. Pokazano je na teorijskoj razini, ali i u empiriji kako investitori uzimaju u obzir i treći te četvrti moment prinosa portfelja. Uvođenjem viših momenata u analizu, znatno se otežava sama optimizacija. Empirijski radovi koji se bave ovakvim modelima pojavljuju se tek 1991. godine (Lai 1991), pri čemu je u analizu najprije uključen treći moment, a četvrti je tek prije nekoliko godina uveden. Analiza inicijalnog Markowitzevog modela u Microsoft Excelu već je dobro poznata (pogledati, primjerice, Aljinović, Marasović i Šego 2008, 2011), dok s druge strane, izračun viših momenata distribucija prinosa, kao i sama optimizacija portfelja koji ih koriste nije. Ne postoji komercijalni *software* koji bi temeljem unesenih podataka o kretanju prinosa računao sve potrebne momente i mješovite momente te ih koristio dalje u optimizaciji. Ono što se smatra kompliciranim za izračun u pravilu predstavlja limitirajući faktor za daljnja istraživanja, što uzrokuje da su i pogledi na određena područja limitirani. Stoga je cilj ovoga rada prikazati kako pomoću Excela i programskog jezika VBA modelirati izračun viših momenata kao prvi korak u analizi optimizacije portfelja koji uzimaju u obzir spomenute momente, i proširiti mogućnosti za istraživanje na području optimizacije portfelja

II. MARKOWITZEV MODEL I KRITIKE

Markowitzev model u okvir ekonomike izbora smješta izbor portfelja temeljem očekivanog prinosa portfelja s jedne, i varijance koja mjeri rizik s druge strane. Pretpostavlja se da investitor raspolaže s podacima o cijenama za $i \in \{1,2,...,I\}$ dionica u posljednjih $t \in \{1,2,...,T\}$ razdoblja. Prinosi na dionicu i u vremenskom razdoblju t računaju se kao kontinuirani prinosi formulom

$$R_i(t) = \ln\left(\frac{P_i(t)}{P_i(t-1)}\right), \tag{1}$$

pri čemu $R_i(t)$ predstavlja prinos na dionicu i za razdoblje t, a ln prirodni logaritam. Očekivani prinos $E(R_i)$ dionice i, $i \in \{1,2,...,I\}$, računa se kao aritmetička sredina prinosa na tu dionicu $R_i(t)$, $t \in \{1,2,...,T\}$:

$$E(R_i) = \frac{1}{T}\sum_{t=1}^{T} R_i(t), \tag{2}$$

dok je varijanca prinosa σ_i^2 dionice i dana formulom

$$\sigma_i^2 = E\left[R_i(t) - E(R_i)\right]^2 = \frac{1}{T}\sum_{t=1}^{T}\left[R_i(t) - E(R_i)\right]^2. \tag{3}$$

Očekivana vrijednost prinosa pojedinačne dionice predstavlja nagradu koju investitori očekuju ulaganjem u dionicu. S druge strane, varijanca prinosa predstavlja rizik koji je pritom potrebno preuzeti. Kovarijanca kao mješoviti moment mjeri povezanost kretanja prinosa jedne dionice s kretanjima prinosa druge. Kovarijanca prinosa između dviju dionica i i j dana je izrazom

$$cov(R_i, R_j) = E\left[\left(R_i(t) - E(R_i)\right)\left(R_j(t) - E(R_j)\right)\right] =$$
$$= \frac{1}{T}\sum_{t=1}^{T}\left(R_i(t) - E(R_i)\right)\left(R_j(t) - E(R_j)\right). \tag{4}$$

Označimo li s w_i relativni udio dionice i u portfelju P, očekivani prinos portfelja $E(R_P)$ računa se kao ponderirana aritmetička sredina prinosa na pojedinačne dionice koje ga sačinjavaju, pri čemu su ponderi udjeli w_i:

$$E(R_P) = \sum_{i=1}^{I} w_i E(R_i). \tag{5}$$

Očekivani prinos portfelja konstruira se kao vagana aritmetička sredina pojedinačnih očekivanih prinosa dionica, dok se varijanca portfelja konstruira temeljem pojedinačnih varijanci prinosa dionica,

te kovarijanci kako bi se konstruirao portfelj kojeg će karakterizirati struktura koja će ostvarivati maksimalni mogući prinos za danu razinu rizika, odnosno minimalnu varijancu (rizik) za danu razinu prinosa.

Varijanca portfelja dana je izrazom

$$\sigma_P^2 = \sum_{i=1}^{I} w_i^2 \sigma_i^2 + 2 \sum_{\substack{i,j=1 \\ i \neq j}}^{I} w_i w_j cov\left(R_i, R_j\right) = \sum_{i=1}^{I} \sum_{j=1}^{I} w_i w_j \sigma_{ij}. \quad (6)$$

S obzirom da je w_i relativni udio dionice u promatranom portfelju, mora vrijediti

$$\sum_{i=1}^{I} w_i = 1, \quad (7)$$

što znači da zbroj relativnih udjela svih dionica koje sačinjavaju portfelj mora biti jednak jedinici (samo te dionice čine portfelj i niti jedne druge). Ako se pretpostavi da postoji mogućnost kratke prodaje, ne nameću se nikakve restrikcije na predznak udjela w_i. U suprotnom, isključi li se kratka prodaja, uvode se ograničenja:

$$w_i \geq 0, \forall i. \quad (8)$$

Temeljem uvedenih oznaka za očekivani prinos pojedine dionice, varijancu te same udjele dionica u portfelju, model pronalaska efikasne granice uz isključenje mogućnosti kratke prodaje simbolički pišemo ovako:

$$\max_{w_i} \quad E\left(R_P\right) = \sum_{i=1}^{I} w_i E\left(R_i\right)$$

uz ograničenja

$$\sigma_P^2 = \sum_{i=1}^{I} \sum_{j=1}^{I} w_i w_j \sigma_{ij} \leq c \quad \text{(I)}$$

$$\sum_{i=1}^{I} w_i = 1$$

$$w_i \geq 0, \quad i \in \{1,\ 2,\ ..,\ I\}.$$

Varijable odlučivanja su udjeli dionica w_i, dok je c proizvoljno zadana razina rizika. Uz pretpostavku postojanja mogućnosti kratke prodaje, posljednjih I ograničenja o nenegativnosti udjela svih dionica u portfelju nestaju. Izračun svih potrebnih elemenata te sama optimizacija ovog modela u Excelu prikazana je u Aljinović, Marasović i Šego (2011).

Iako su Markowitzevi doprinosi omogućili znatan napredak u području investiranja, od samog nastanka modela javljaju se kritike na pretpostavke na kojima se temelji. Jedna od najviše kritiziranih pretpostavki modela jest ona o normalnoj distribuciji prinosa dionica. Ako je zadovoljena pretpostavka o normalnoj distribuciji prinosa dionica u portfelju, dovoljno je samo prva dva momenta koristiti u optimizaciji. Međutim, financijske serije karakteriziraju nesimetrične leptokurtične distribucije. Negativna ili pozitivna asimetrija odnose se na broj nastupa ekstremnih prinosa, s jedne strane ispodprosječnih, a s druge nadprosječnih. Ako je prisutna pozitivna asimetrija distribucije prinosa, nadprosječne prinose karakterizira veći broj nastupa u odnosu na ispodprosječne. S druge strane, negativno asimetrična distribucija ukazuje kako postoji veći broj nastupa ispodprosječnih ekstremnih prinosa (gubitaka). Uvođenjem trećeg momenta u analizu, koeficijenta asimetrije i koasimetrije, uzimaju se u obzir takvi ekstremni događaji, pri čemu se nastoji maksimizirati pozitivna asimetrija portfelja. Leptokurtične distribucije karakteriziraju šiljatiji vrhovi u odnosu na normalnu distribuciju, stoga su repovi tih distribucija deblji, dok platikurtične distribucije karakteriziraju plosnatiji vrhovi i tanji repovi u odnosu na normalnu distribuciju. Kako se debljina repova distribucije veže uz vjerojatnosti nastupa ekstremnih događaja, važno je uključiti u analizu i četvrti moment distribucija – koeficijent zaobljenosti i kozaobljenosti. Pri tom se teži minimizaciji zaobljenosti portfelja, kako bi se smanjila vjerojatnost nastupa ekstremih događaja jer su, po pretpostavci, investitori neskloni riziku (Markowitz 1991, Fabozzi i Markowitz 2002).

Prikažimo izračun viših momenata samih prinosa dionica te portfelja. Raspolaže li investitor s prinosima i očekivanim prinosima pojedine dionice, koeficijent asimetrije jednak je[1]:

$$s_{iii} = E\left[\left(R_i - E\left(R_i\right)\right)^3\right]. \quad (9)$$

Promatranjem zajedničke distribucije prinosa R_i i R_j, koasimetričnost s_{iij} se definira kao[2]

$$s_{iij} = E\left[\left(R_i - E\left(R_i\right)\right)\left(R_i - E\left(R_i\right)\right)\left(R_j - E\left(R_j\right)\right)\right]. \quad (10)$$

Analogno izračunu koasimetričnosti s_{iij}, definira se i koasimetričnost s_{ijj}:

$$s_{ijj} = E\left[\left(R_i - E\left(R_i\right)\right)\left(R_j - E\left(R_j\right)\right)\left(R_j - E\left(R_j\right)\right)\right]. \quad (11)$$

Važno je uočiti simetričnost u izračunu koasimetričnosti. Naime, vrijedi $s_{iij}=s_{iji}=s_{jii}$, kao i što u slučaju za više od dvije dionice vrijedi:

$$s_{ijk} = E\left[\left(R_i - E\left(R_i\right)\right)\left(R_j - E\left(R_j\right)\right)\left(R_k - E\left(R_k\right)\right)\right], \quad (12)$$

pri čemu je $s_{ijk}=s_{ikj}=s_{jik}=s_{jki}=s_{kij}=s_{kji}$, što je posljedica zakona komutacije za množenje realnih brojeva. Na taj način, zbog simetričnosti je u slučaju I dionica potrebno izračunati $\binom{I+2}{3}$ različitih koeficijenata asimetrije i koasimetrije, koristeći formulu:

$$s_{ijk} = \frac{1}{T}\sum_{t=1}^{T}\left(R_i(t) - \bar{R}_i\right)\left(R_j(t) - \bar{R}_j\right)\left(R_k(t) - \bar{R}_k\right),$$
$$i \in \{1,2,...,I\},\ j \in \{1,2,...,I\},\ k \in \{1,2,...,I\}. \quad (13)$$

Koeficijent zaobljenosti za pojedinačnu dionicu se može definirati formulom

$$k_{iiii} = E\left[\left(R_i - E\left(R_i\right)\right)^4\right]. \quad (14)$$

Općeniti zapis kozaobljenosti svodi se na:

$$k_{ijkl} = E\left[\left(R_i - E\left(R_i\right)\right)\left(R_j - E\left(R_j\right)\right)\left(R_k - E\left(R_k\right)\right)\left(R_l - E\left(R_l\right)\right)\right]. \quad (15)$$

Kao i u slučaju izračuna koasimetrije, i prilikom izračuna kozaobljenosti prisutna je određena simetrija. U slučaju I dionica potrebno je izračunati $\binom{I+3}{4}$ različita koeficijenata zaobljenosti i kozaobljenosti, koristeći formulu:

[1] Sve formule i rezultati u ovome dijelu rada izvode se i detaljno objašnjavaju u Škrinjarić (2013b).
[2] R_i i R_j su zapravo $R_i(t)$ i $R_j(t)$.

$$k_{ijkl} = \frac{1}{T}\sum_{t=1}^{T}\left(R_i(t)-\bar{R}_i\right)\left(R_j(t)-\bar{R}_j\right)\left(R_k(t)-\bar{R}_k\right)\left(R_l(t)-\bar{R}_l\right),$$

$$i \in \{1, 2, ..., I\}, j \in \{1, 2, ..., I\}, k \in \{1, 2, ..., I\}, l \in \{1, 2, ..., I\}. \quad (16)$$

Uvedeni pojmovi sintetiziraju se u model optimizacije portfelja uzimajući u obzir više momente[3]:

$$\max_{w_i}\left[E(R_P) = \sum_{i=1}^{I} w_i E(R_i)\right]$$

$$\min_{w_i}\left[\sigma_P^2 = \sum_{i=1}^{I}\sum_{j=1}^{I} w_i w_j \sigma_{ij}\right]$$

$$\max_{w_i}\left[S_{Pr} = \frac{\sum_{i=1}^{I}\sum_{j=1}^{I}\sum_{k=1}^{I} w_i w_j w_k s_{ijk}}{\sigma_P^3}\right]$$

$$\min_{w_i}\left[K_{Pr} = \frac{\sum_{i=1}^{I}\sum_{j=1}^{I}\sum_{k=1}^{I}\sum_{l=1}^{I} w_i w_j w_k w_k k_{ijkl}}{\sigma_P^4}\right] \quad \text{(II)}$$

uz ograničenja

$$\sum_{i=1}^{I} w_i = 1$$

$$w_i \geq 0, \quad i \in \{1, 2, ..., I\}.$$

Dakle, radi se o pristupu optimizacije četiri funkcije cilja, pri čemu se maksimizira očekivan prinos portfelja te koeficijent asimetrije portfelja, dok se minimiziraju rizik (varijanca) te koeficijent zaobljenosti portfelja. Navedeni pristup sastoji se od višekriterijske optimizacije četiri funkcije cilja na skupu mogućih rješenja.

III. TEORIJSKA I EMPIRIJSKA OPRAVDANJA KORIŠTENJA VIŠIH MOMENATA

Athayde i Flores (2004) obrazlažu da investitori preferiraju veće vrijednosti neparnih momenata portfelja (tu ubrajamo očekivani prinos i asimetričnost), dok teže manjim vrijednostima parnih momenata portfelja (rizik te zaobljenost) kako bi na taj način smanjili gubitke i povećali vjerojatnosti ostvarivanja dobitaka. U posljednjih nekoliko godina interes za proučavanjem viših momenata u okviru optimizacije portfelja i vrednovanja imovine sve je veći, i to iz razloga koje spominju Jurzcenko i Mallet (2006): rastući interes i briga investitora za ekstremne rizike, uspjesi koje su ostvarivali investicijski fondovi čija distribucija imovine nije bila normalna, ali i teorijski, ekonometrijski i statistički napredak u izgradnji modela koji uzimaju u obzir više momente. Sam Markowitz (Fabozzi i Markowitz 2002) priznaje da promatranjem empirijskih distribucija prinosa dionica možemo uvidjeti je li opravdano ili ne koristiti samo prva dva momenta, ili pak i više ako distribucije prinosa nisu normalne.

Već se 1953. godine javlja interes za višim momentima vrijednosnica u radu Kendalla i Hilla (1953), koji su utvrdili prisutnost asimetrije kao i zaobljenosti u empirijskim distribucijama prinosa dionica. Uslijedili su i radovi Mandelbrota (1963a, 1963b) koji je zaključio kako su vremenske serije promjena cijena šiljatijih vrhova i debljih repova u odnosu na normalnu distribuciju. Cootner (1964) i Fama (1965) donose slične zaključke za distribucije vrijednosnica. 1967. godine Arditti (1967) je pokazao kako je pojedini investitor skloniji većoj asimetriji portfelja, što je u skladu s pojmom smanjujuće apsolutne averzije rizika. Alderfer i Bierman (1970) proveli su istraživanje u kojemu su ispitanicima nudili nekoliko investicijskih alternativa koje su bile slične što se tiče rizika i prinosa, ali su bile veoma različite što se tiče asimetrije distribucije prinosa. Željeli su potvrditi istraživačku hipotezu kako viši prinosi distribucija prinosa, s naglaskom na asimetriju, imaju značajan učinak na odabir vrijednosnica u portfelj. Rezultati istraživanja su potvrdili njihovu hipotezu, to jest investitori su preferirali pozitivnu asimetriju prinosa, čak i uz manji očekivani prinos!

U sedamdesetim godinama prošlog stoljeća dolazi do jačanja interesa za proučavanjem viših momenata u teoriji portfelja. Samuelson (1970) je prvi razmatrao očekivanu korisnost investitora temeljem tri momenta portfelja, a uskoro je uslijedio i doprinos Rubinsteina (1973) koji je proširio model procjenjivanja kapitalne imovine uzimanjem u obzir trećeg momenta distribucija[4]. U teorijskoj analizi se, dakle, najprije uvodi treći moment portfelja u analizu, zbog relativno jednostavnijeg izračuna u odnosu na četvrti moment. Tek 1988. godine Graddy i Homaifar (1988) proširuju analizu portfelja na četvrti moment. Zaobljenost je dugo zanemarivana u analizi, jer je s porastom dimenzije momenata portfelja koji se analiziraju teže dati njihovu interpretaciju (Jurczenko i dr. 2006). Uvođenje viših momenata u analizu znači kako funkcije korisnosti investitora neće više biti kvadratne (ne ovise samo o prva dva momenta), već u njih je potrebno inkorporirati preferencije prema asimetričnosti i zaobljenosti. Müller i Machina (1987) dokazali su teorem o funkciji korisnosti koja to uzima u obzir: „maksimizator korisnosti kojeg karakterizira neprekidna von Neumann-Morgensternova funkcija korisnosti $U(\cdot)$ rangirat će distribucije vjerojatnosti temeljem njihovih prvih m apsolutnih momenata ako i samo ako je $U(\cdot)$ polinom stupnja manjeg ili jednakog m."

Empirijski radovi koji rješavanju problem optimizacije portfelja uz više momente javljaju se od 90-ih godina prošloga stoljeća. Najčešći pristup jest provođenje kompromisnog programiranja, i tu ubrajamo radove Lai (1991), Chunhachinda i ostalih (1997), Prakasha i ostalih (2003), Sun i Yan (2003), Yang i Hung (2010) te od domaćih radova Škrinjarić (2013a). Svi navedeni autori, osim posljednje navedene, optimiziraju akademske primjere, dakle, koriste veoma mali broj dionica u analizi, što iziskuje relativno malo posla u Excelu. Međutim, povećavanjem opsega problema, kao što je to učinjeno u Škrinjarić (2013a), sam izračun se znatno otežava, s obzirom da ne postoji ustaljena procedura koja bi omogućila elegantno dobivanje rješenja. Stoga se iduće poglavlje bavi rješavanjem ovakvoga problema s kojim bi se investitori mogli suočiti u budućnosti.

IV. IZRAČUN MOMENATA U EXCELU

Računala omogućuju implementaciju i testiranje mnogih matematičkih modela u ekonomiji koji su desetljećima bili osuđeni na čekanje trenutka kada će primjena moćnijih računala omogućiti testiranje njihove valjanosti. No, većina korisnika računala oslanja se na korištenje već gotovih računalnih programa, što uzrokuje problem kada za istraživanje nekog problema nije dostupan već razvijeni softver. To je bio slučaj i u ovom problemu kod koje je dobra ideja trebala biti ostvarena korištenjem značajnog broja izračuna.

Prva četiri momenta za pojedinu dionicu nije problem računati, ali izračun komomenata zahtijeva repetitivan posao koji se značajno komplicira s višim momentima, što je vidljivo iz (13) i (15). Ako je I broj dionica, T broj mjeseci, a r red komometa kojeg treba izračunati, za svaki r-ti komoment prvo treba izračunati očekivani prinos za I dionica (izraz (2)), odstupanja ostvarenih od očekivanog prinosa $\left(R_i - E(R_i)\right), \forall i \in \{1,..,I\}$, umnoške svih mogućih kombinacija

[3] Izvod svih formula i objašnjenja u modelu dan je u Škrinjarić, T. (2013b).

[4] Uvođenje bezrizične imovine u portfelj analizirali su Jean (1971, 1972, 1973)[4] i Ingersoll (1975), dok su Arditti i Levy (1975) proširili analizu portfelja s tri momenta na višeperiodnu analizu. Kraus i Litzenberger (1976) nastavili su svoj rad na ideje Rubinsteina iz 1973. godine, pri čemu je njihovo istraživanje prvo empirijsko testiranje troparametarskog CAPM modela.

izračunatih *I* odstupanja od *r* elemenata (bez ponavaljanja) (12), te konačno prosjek izračunatih vrijednosti po mjesecima (13). Ako bi se takav posao radio ručno, izvršitelj mora razviti određeni sustav praćenja i evidentiranja kombinacija koje treba izračunati, za što je osim vremena potrebna i jako dobra koncetracija. Primjenjivost svakog modela za optimizaciju portfelja značajno ovisi brzini rješavanja problema. Zbog toga je bilo nužno napraviti program koji će u kratkom roku izračunati podatke koji su potrebni za rješavanje problema. Microsoft Excel je alat koji se često koristi za obradu i analizu financijskih podataka, pa se programiranje makronaredbi u Visual Basic for Application (VBA) za Excel nametnuo kao praktičan izbor.

Istraživanje je bilo usmjereno za traženje načina za brži izračun trećih i četvrtih komomenata distribucije potrebnih za rješavanje problema optimizacije modela (II), ali radna knjiga Excela, koja je dizajnirana za potrebe istraživanja ovog problema, sadržava i markonaredbe koje omogućavaju izračun svih momenata potrebnih za optimizaciju kako bi korisnik najjednostavnije došao do svih potrebnih podataka. Radni list je oblikovan tako da usmjerava korisnika na pravilno unošenje podataka o cijenama dionica kako bi ih funkcije mogle koristiti. Korisnik unosi ulazne podatke o izabranim dionicama u radni list „*cijene*“ počevši od ćelije B2. Svaka je dionica tako određena rednim brojem stupca u kojoj su njeni podaci uneseni. Na dnu radnog lista „*cijene*“ postavljeno je više *ActiveX* kontrola (slika 1) koje aktiviraju makronaredbe za izračun prinosa, matrice kovarijanci te viših momenata. Korisnik može birati za koliki broj svih dionica odnosno razdoblja, počevši od prve dionice i prvog razdoblja, želi izvršavati makronaredbe unoseći te podatke u dijaloški okvir. Na listu su postavljene i kontrole koje pokreću kopiranje izračunatih momenata u jedan radni list, kako bi se ti podaci lakše koristili u daljnoj optimizaciji.

Klikom miša na kontrolu s naslovom „*Izračunaj prinose*“ izvršava se makronaredba *lnReturn* koja računa prinos. Nakon što korisnik unese podatke o broju dionica (*nbSt*) i broju mjeseci (*nbMo*), procedura briše listove koji imaju naslov *lnPrinosi#* te stvara novi radni list *lnPrinosi#*:

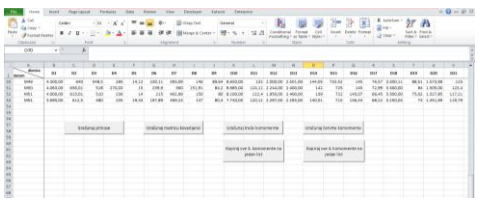
```
For Each ws In ActiveWorkbook.Sheets
   If Left(ws.Name, 10) = "LnPrinosi#" Then ' brisi postojece listove
      Application.DisplayAlerts = False
      ws.Delete
      Application.DisplayAlerts = False
   End If
Next

Set ws = Sheets.Add(After:=Sheets(Worksheets.Count)) 'stvori novi
list
ws.Name = "LnPrinosi#"
Worksheets("LnPrinosi#").Activate.
```

U određenom rasponu aktivnih listova (definiranom prema vrijednostima *finishRow=nbMo* i *lnfinishCol=nbSt*+1) svakoj se ćeliji (*I, J*) dodjeljuje svojstvo, tj. vrijednost formule za izračun prinosa (1) koja vuče vrijednosti s relativnih pozicija iz radnog lista „*cijene*“ koristeći dvije petlje FOR…NEXT :

```
   With Selection
      For J = 2 To lnfinishCol
         For I = 2 To finishRow
            Cells(I, J).Value = "=+LN(cijene!R[1]C/cijene!RC)"
         Next I
      Next J
   End With.
```

Nakon što se petlje FOR…NEXT izvrše, funkcija u zadnjem retku



Slika 1 – Prvi list radne knjige u Excelu

(*averRow*) radnog lista računa prosjek vrijednosti izračunatih u svakom stupcu, a koji se odnose na mjesečne prinose za svaku dionicu:

```
averRow = nbMo + 1
For J = 2 To lnfinishCol 'izracun prosjeka
   Cells(averRow, J).Select
   ActiveCell.FormulaR1C1 = "=+AVERAGE(R2C:R[-1]C)"
Next J.
```

Funkcije za brisanje i stvaranja novih radnih listova, te za računanje prosjeka su uz prilagodbu korištene i u ostalim makronaredbama. Za izračun matrice kovarijanci preuzeta je funkcija koja je prilagođena ovom radnom listu [4]. Funkcija se aktivira kao dio makronaredbe preko kontrole „*Izračunaj matricu kovarijanci*“ te stvara novi radni list u kojem se ispisuje matrica kovarijanci za željeni broj dionica i broj razdoblja.

Za izračun trećih komomenata napisana je makronaredba *momenti3* koja se aktivira klikom miša na kontrolu „*Izračunaj treće komomente*“. Makronaredba poziva nekoliko podprocedura i funkcija, između ostalog računa prinose za zadani broj dionica i razdoblja, te za svaku od *I* dionica kreira novi radni list u kojem se računaju koasimetričnosti. U svakom *i*-tom radnom listu s 3. komomentima, $i \in \{1,..,I\}$, računaju se koasimetričnosti s_{ijk} (12) za *i*-tu dionicu po mjesecima. U ovoj fazi se definiraju 3 varijable koje kontroliraju izmjenu članova umnoška iz formule (12): varijabla *I* koja poprima vrijednosti od 1 do *nbSt* te kontrolira redni broj dionice, odnosno prvi član umnoška od tri člana kao i otvaranje novog radnog lista, varijabla *J* koja poprima vrijednosti od *I* do *nbSt* te kontrolira izmjenu drugog člana umnoška i, konačno, varijabla *K* koja poprima vrijednosti od *L* do *nbSt* te kontrolira izmjenu zadnjeg člana umnoška. Svakoj aktivnoj ćeliji u prvom retku raspona (*nbSt*, *nbMo*) u aktivnom radnom listu se dodijeljuje formula s različitim vrijednostima varijabli *I, J* i *K* pomoću petlji FOR…NEXT te svojstva *formulaR1.C1* koje vuče vrijednosti s relativnih pozicija radnog lista *lnPrinosi3*. Tako se kontrolira da se u svakoj ćeliji računa različita koasimetričnost s obzirom na različite kombinacije između dionica *J, K, L*. Svaka sljedeća definirana varijabla (dakle, *J* i *K*) poprima najmanju vrijednost jednaku vrijednosti prethodne varijable što onemogućava neželjeno ponavljanje izračuna kombinacija, pa je broj izračuna u svakoj *i*+1 iteraciji manji nego u prethodnoj *i*-toj iteraciji. Kod je sljedeći:

```
For I = 1 To nbSt
  Range(Cells(2, 2), Cells(finishRow, finishCol)).Select
    With Selection ' racunaj momente
        For J = I + 1 To nbSt + 1
        If J > (nbSt + 1) Then
            J = 2
          Else: J = J
        End If
         For K = J To (nbSt + 1)
          ActiveCell.FormulaR1C1 = "=(LnPrinosi3!RC" & I + 1 &
"-LnPrinosi3!R" & averRow & "C" & I + 1 & ")*(LnPrinosi3!RC" &
J & "-LnPrinosi3!R" & averRow & "C" & J & ")*(LnPrinosi3!RC"
& K & "-LnPrinosi3!R" & averRow & "C" & K & ")"
           ActiveCell.Offset(0, 1).Range("A1").Select
            Next K
        Next J
If I = nbSt Then 'pozicija zadnjeg stupca
      lastCol = 2
    Else
      lastCol = Cells(2, 2).End(xlToRight).Column
 End If
      Range(Cells(2,2),Cells(2,lastCol)).AutoFill
Destination:=Range(Cells(2, 2), Cells(finishRow, lastCol)),
Type:=xlFillDefault.
```

U zadnjem retku radnog lista funkcija računa prosjek izračunat u svim razdobljima promatranja:

```
Range(Cells(averRow, 2), Cells(averRow, lastCol)).Value =
"=+AVERAGE(R2C:R[-1]C)".
```

Procedura se zatim obavlja za sljedeću vrijednost varijable *I*. Nakon što se prodecura izvrši za svih *I* dionica, u radnoj knjizi je stvoreno upravo *I* novih radnih listova u kojima su u zadnjem retku izračunate koasimetričnosti. Izračun koasimetričnosti pomoću svojstva *ActiveCell.FormulaR1C1* je slijedio iz dva razloga: prvo, korisnik u svakom trenutku može vidjeti na koji način i za koje dionice je izračunata koasimetričnost u određenoj ćeliji, a drugo, mogućnost korištenja funkcije *AutoFill* koja kopira formulu za izračun u svim preostalim recima odnosno stupcima, što je doprinijelo brzini izračunavanja potrebnih vrijednosti.

Prosjeci, koji su izračunati u zadnjem korištenom retku u svakom kreiranom radnom listu s 3. komomentima su zapravo jedini inputi koji su potrebni za daljnju analizu modela (II), pa je u radnom listu „*cijene*“ postavljena *ActiveX* kontrola s oznakom „*Kopiraj sve 3. komomente na jedan list*“ kojom se stvara novi radni list s ispisanim svim očekivanim komomentima (prosječne vrijednosti) na jedan list („*3komomenti-prosjek*“).
Za izračun četvrtih momenata se obavlja slična prodecura, samo što je u kodu dodana još jednu varijabla-brojač koja kontrolira izmjenu četvrtog člana umnoška iz (15) od četiri člana, odnosno dodatne petlje u kodu.

ZAKLJUČAK

Na opisani način konstruirana je radna knjiga s makronaredbama koje su pristupačne korisniku i na prosječnom osobnom računalu u relativno kratkom vremenu (u usporedbi s vremenom potrebnim za ručni izračun u Excelu) računaju potrebne ulazne podatke za rješavanje problema optimizacije portfelja dionica uzimajući u obzir i treće i četvrte komomente distribucije. Ilustracije radi, za manualni izračun trećih i četvrtih momenata na uzorku od 24 dionice kroz 52 mjeseca koristeći Excel i „razvlačenje formula“ potrebno je izgubiti značajan broj sati rada, odnosno dana rada (prema našem iskustvu 3 tjedna), a u tom periodu su korišteni podaci već zastarjeli i nove dionice postaju kandidati za ulazak u portfelj. Na prosječnom računalu s 64-bitnim operativnim sustavom, *i*3 procesorom s 2,13GHz i 3Gb RAM-a, makronaredba za izračun 3. momenata se izvršava za približno 33 sekunde, a makronaredba za izračun 4. momenata u prosjeku za nešto više od minutu po dionici. Također, funkcija koja kopira sve vrijednosti na jedan list izvršava se jako brzo, a kod kopiranja 4. momenata je ipak potrebno nešto manje od minute. Smatra se da će takva radna knjiga u budućim istraživanjima olakšati analizu financijskih podataka i drugim potencijalnim korisnicima, te da će druge korisnike potaknuti i uputiti na dizajniranje programa prema vlastitim potrebama. Kako i optimizacija s dobivenim podacima zahtijeva značajne napore u izračunu, sljedeći korak je dizajniranje makronaredbe u Excelu koja korištenjem ovih podataka uz pomoć Rješavača rješava problem optimizacije portfelja.

LITERATURA

1. Alderfer, C. P, Bierman, H. (1970) Choices with Risk: Beyond the Mean and Variance, Journal of Business, Vol. 43, Issue 3, pp. 341-353.
2. Aljinović, Z., Marasović, B., Šego, B. (2008) Financijsko modeliranje, Zgombić&Partneri, Zagreb.
3. Aljinović, Z., Marasović, B., Šego, B. (2011) Financijsko modeliranje, drugo dopunjeno i izmijenjeno izdanje, Ekonomski fakultet u Splitu, Split.
4. Arditti, F. (1967) Risk and the Required Return on Equity, Journal of Finance 22, pp. 19–36.
5. Arditti, F., Levy, H. (1975) Portfolio Efficiency Analysis in Three Moments: The Multiperiod Case, Journal of Finance 30, pp. 797–809.
6. Athayde, G., Flores, R. (2004) Finding a Maximum Skewness Portfolio: A General Solution to Three-Moments Portfolio Choice, Journal of Economic Dynamics&Control 28, pp. 1335-1352.
7. Chunhachinda, P., Dandapani, K., Hamid, S., Prakash A. J. (1997) Portfolio Selecion and Skewness: Evidence from International Stock Market. Journal of Banking and Finance 21, pp.143-167.
8. Cootner, P. (Ed.) (1964) The Random Character of Stock Market Prices, MIT Press.
9. Fabozzi, F. J., Markowitz, H. M. (editors) (2002) The Theory and Practice of Investment Management, John Wiley & Sons, Inc., Hoboken, New Jersey.
10. Fama, E. (1965) The Behaviour of Stock-Market Prices, Journal of Business 38, pp. 34–105.
11. Fama, E. F. (1965) The Behaviour of Stock Market Prices, Journal of Business 38, pp. 34-105.
12. Graddy, D., Homaifar, G. (1988) Equity Yields in Models Considering Higher Moments of the Return Distribution, Applied Economics 20, pp. 325–334.
13. Ingersoll, J. (1975) Multidimensional Security Pricing, Journal of Financial and Quantitative Analysis, 10, pp. 785-798.
14. Jean, W. (1971) The Extension of Portfolio Analysis to Three and More Parameters, Journal of Financial and Quantitative Analysis 6, pp. 505–515.
15. Jean, W. (1972) Distribution Moments and Equilibrium: Reply, Journal of Financial and Quantitative Analysis 7, pp. 1435–1437. Jean, W. (1973) More on Multidimensional Portfolio Analysis, Journal of Financial and Quantitative Analysis 8, pp. 475–490.
16. Jurczenko, E., Maillet, B., Merlin, P. (2006) Hedge Funds Portfolio Selection with Higher-order Moments: A Non-parametric Mean-Variance-Skewness-Kurtosis Efficient Frontier in Multi-moment Asset Allocation and Pricing Models. Ed. by E. Jurczenko, B. Maillet. Sussex, John Wiley.
17. Jurczenko, E., Mallet, B. (ed.) (2006) Multi-Moment Asset Allocation and Pricing Models, John Wiley & Sons, Ltd, England.
18. Kendall, M., Hill, B. (1953) The Analysis of Economic Time-series – Part I: Prices, Journal of the Royal Statistical Society Series A, 116 (1), pp.11–34.
19. Lai, T. (1991) Portfolio with Skewness: A Multiple-objective Approach, Review of Quantitative Finance and Accounting 1, pp. 293–305.
20. Mandelbrot, B. (1963a) The Variation of Certain Speculative Prices, Journal of Business 36, pp. 394–419.
21. Mandelbrot, B. (1963b) New Methods in Statistical Economics, Journal of Political Economy 71, pp. 421–440.
22. Markowitz, H. (1952) Portfolio Selection, The Journal of Finance, Vol. 7, No. 1, pp. 77-91.
23. Markowitz, H. (1991) Foundations of portfolio theory, Journal of Finance 46 (2), pp. 469–478.
24. Müller, S. M., Machina, M. J. (1987) Moment preferences and polynomial utility, Economics Letters 23, pp. 349-353.
25. Prakash, A.; Chang, C. H. Pactwa, E. (2003) Selecting a Portfolio with Skewness: recent Evidence from US, European, and Latin America Equity Markets. Journal of Banking and Finance 27, pp. 1375–1390.
26. Rubinstein, M. (1973) The Fundamental Theorem of Parameter-preference Security Valuation, Journal of Financial and Quantitative Analysis 8, pp. 61–69.
27. Rubinstein, M. (1973) The Fundamental Theorem of Parameter-preference Security Valuation, Journal of Financial and Quantitative Analysis 8, pp. 61–69.
28. Samuelson, P. (1970) The Fundamental Approximation of Theorem of Portfolio Analysis in Terms of Means, Variances and Higher Moments, Review of Economic Studies 37, pp. 537-542.
29. Sun, Q., Yan, Y. (2003) Skewness persistence with optimal portfolio selection, Journal of Banking & Finance 27, pp. 1111-1121.
30. Škrinjarić, T. (2013a) Portfolio Selection with Higher Moments and Application on Zagreb Stock Exchange, Zagreb International Review of Economics & Business, Vol. XVI, No 1., pp. 65-78.
31. Škrinjarić, T. (2013b) Modeli optimizacije portfelja uz više momente u burzovnom trgovanju, Specijalistički

polijediplomski rad, Ekonomski fakultet, Sveučilište u Zagrebu.
32. Yang, C. W., Hung, K. (2010) A Generalized Markowitz Portfolio Selection Model with Higher Moments, International Review of Accounting, Banking and Finance, Vol. 2, No. 1, pp. 1-7.

VAŽNOST UVOĐENJA NOVIH TEHNOLOGIJA U OBRAZOVNE INSTITUCIJE ZA RAZVOJ LOKALNOG GOSPODARSTVA

Mr.sc.Anica Vukašinović*
*Brodsko-posavska županija/Upravni odjel za obrazovanje šport i kulturu, 35 000 Slavonski Brod, Hrvatska
e-mail:anica.vukasinovic@gmail.com

Sažetak - U kontekstu velikih globalizacijskih izazova poput znanstvenog i tehnološkog napretka,demografskih promjena,transformacije nacionalnih gospodarstava, zahtjeva za brzim stjecanjem znanja i vještina potrebnih sve zahtjevnijem tržištu rada te potrebi za cjeloživotnim učenjem, sustav odgoja i obrazovanja u Hrvatskoj potrebno je mijenjati i usklađivati s gospodarstvima razvijenih zemalja utemeljenim na znanju. Na taj način moguće je pokrenuti kotač gospodarske aktivnosti u Republici Hrvatskoj prema naprijed s ciljem osiguranja zapošljivosti radno aktivnog stanovništva, osiguranju kvalitete života i kulture življenja njenih stanovnika. Trend pada BDP-a u 2013.godine ogleda se u daljnjem smanjenju vrijednosti proizvedenih roba i usluga te zaposlenosti. Uloga obrazovnog sustava tu je ključna, posebice u dijelu strukovnog obrazovanja koje u razvijenim europskim zemljama vrlo uspješno odgovara na izazove današnjice uvođenjem novih tehnologija, stvaranju novih kompetencija, razvoju ljudskih potencijala s ciljem gospodarskog rasta i razvoja a time i održive zapošljivosti. Obzirom na višegodišnja relativno mala izdvajanja iz državnog proračuna za obrazovanje u Hrvatskoj u odnosu na druge razvijene zemlje i tendenciji daljnjeg smanjenja izdvajanja, teško možemo govoriti o kvalitetnim i brzim promjenama u obrazovnom sustavu. Fiskalni kapaciteti jedinica lokalne uprave i samouprave nisu ujednačeni i u pravilu rijetko osiguravaju viši standard obrazovnim institucijama na svom području. Korištenjem sredstva fondova EU obrazovne institucije u Hrvatskoj otvaraju vrata izazovima razvijenog društva, koristeći nove tehnologije, unapređujući odgojno obrazovni rad, provodeći brojne aktivnosti kojima se proširuju kompetencije nastavnika strukovnih škola i stvaraju materijalni preduvjeti za najsuvremeniju strukovnu nastavu, a u skladu s aktualnim potrebama gospodarstva. Strukovne obrazovne ustanove imaju mogućnost obrazovati učenike koji će biti konkurentni na tržištu rada, koji će svojim znanjima i vještinama biti pokretači gospodarskih aktivnosti na lokalnoj razini.
Na primjeru ulaganja u strukovne škole na području Brodsko-posavske županije pokazati ćemo važnost razvoja novih tehnologija u obrazovnim institucijama i njihov utjecaj na razvoj gospodarstva Brodsko-posavske županije.

Ključne riječi: nove tehnologije, obrazovne institucije, razvoj ljudskih resursa, tržište rada, kompetencije nastavnika

I. UVOD

U uvjetima globalne gospodarske krize, povećanja nezaposlenosti posebice mladih ljudi posjedovanje znanja, vještina i inovativnosti presudni su za razvoj pojedinca i sredine u kojoj djeluje. Razina stečenih znanja i vještina u formalnom obrazovanju u mnogim razvijenim zemljama, pa tako u Hrvatskoj pokazala se nedovoljnom za uključivanje mladih ljudi u svijet rada. Planiranje i razvoj kvalifikacija potrebnih tržištu rada već dulje vrijeme nije primarna aktivnost obrazovne politike u Republici Hrvatskoj, što je rezultiralo velikim brojem mladih nezaposlenih osoba. Gospodarski rast i razvoj kako lokalne tako i nacionalne razine odraz je neusklađenosti obrazovnog sustava i potreba gospodarstva.

Područje strukovnog obrazovanja najizravnije i najuže je povezano sa tržištem rada. U njegovoj je prirodi omogućiti stjecanje stručnih kompetencija koje su relevantne, suvremene i što je najbitnije, koje podupiru gospodarstvo i razvijaju ga na suvremen i konkurentan način.

I. OBRAZOVANJE I GOSPODARSKI RAST

Planiranje razvoja kvalifikacija ključno je pitanje svakog obrazovnog sustava u Europi i svijetu.
Istraživanja tržišta rada širom svijeta ukazuju na ključnu ulogu strukovnog obrazovanja u društvenom i gospodarskom razvoju. Preduvjet tomu je osmišljavanje relevantnih nastavnih planova i programa koji se temelje na identifikaciji i analiziranju potreba na tržištu rada. U prvoj fazi prema [1] identifikacija potreba na tržištu trebala bi se promatrati sa tri aspekta:

1. Poslovnih odnosa uključujući ciljeve, resurse i međunarodnu orijentaciju,
2. Izbor ljudskih resursa, tehnološke i menadžerske čimbenike
3. Trendove na tržištu rada, ponude i potražnje, buduće potrebe i ekonomske politike.

U drugoj fazi prilikom izrade kurikuluma treba uzeti u obzir:

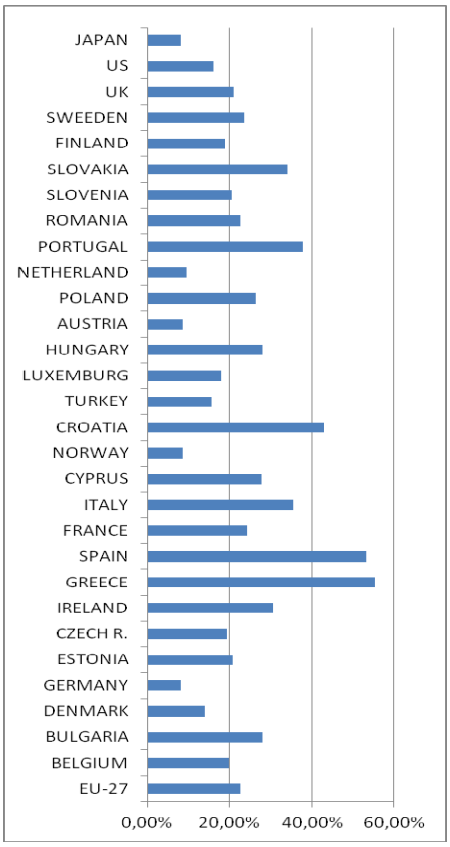
1. Učenike i njihove potrebe, posebno njihovo prethodno obrazovanje, sposobnosti i mogućnosti zapošljavanja
2. Stav učenika prema obrazovanju općenito,
3. Utvrđivanje svrhe škole

Utvrđivanje stvarne svrhe svake strukovne škole jedan je od važnih zadataka u budućem planiranju unapređenja strukovnog obrazovanja, koliko one zadovoljavaju sadašnje i buduće potrebe tržišta rada i s njima povezuju razvoj vještina i kompetencija kroz nastavni plan i program. Obrazovanje, povećanjem produktivnosti rada također doprinosi gospodarskom rastu. Produktivnost se prema [4] povećava s rezultatima obrazovanja i povećanjem zarada radne snage sa višom razinom obrazovanja. Relevantna literatura na temu pokazuje da se gospodarski rast ubrzao u zemljama gdje je razina obrazovanja viša. Doprinos obrazovanja gospodarskom rastu u nekim je slučajevima više od fizičkog kapitala kada se uzme u obzir stručnost radne snage i potrebe gospodarstva. U zemljama gdje je stupanj obrazovanja i gospodarski rast viši stopa izdvajanja iz nacionalnog dohotka za obrazovanje je viša, veća je kvaliteta rada, kao i mogućnosti dugoročnog zapošljavanja.

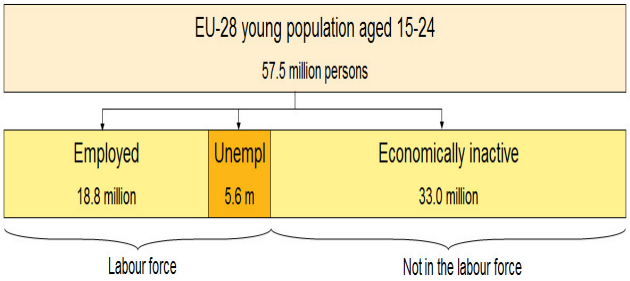
II. UČINCI EKONOMSKE KRIZE NA TRŽIŠTE RADA

Ekonomska kriza na tržištu rada EU odražava se kroz stopu nezaposlenosti. U prosincu 2013.godine 5.544.000 mladih osoba (do 25 godina) u EU-28 bilo je nezaposleno, od čega 3.500.000 u eurozoni (Slika 2). U odnosu na prosinac 2012.godine (Slika 1) nezaposlenost mladih smanjila se za 184 000 u EU-28, a za 77 000 u eurozoni. U prosincu 2013, stopa nezaposlenosti mladih iznosila je 23,2% u EU-28 i 23,8% u eurozoni, u usporedbi s 23,6% i 23,9%, koja je zabilježena u prosincu 2012. U prosincu 2013 najniže stope zabilježene su u Njemačkoj (7,4%) i Austriji (8,9%), a najviše u Grčkoj (59,2% u listopadu 2013), Španjolskoj (54,3%), i Hrvatskoj (49,2% u četvrtom kvartalu 2013 godine). Ukoliko promatramo tržište rada u SAD-u u prosincu 2013, stopa nezaposlenosti bila je 6,7 %, što je pad od 7,0% u odnosu na prethodni mjesec. U prosincu 2012. godine stopa nezaposlenosti u SAD-u iznosila je 7,9%.

Analize tržišta rada eurozone ukazuju na važnost posjedovanja potrebnih obrazovnih kvalifikacija radi lakšeg pronalaska zaposlenja. Ta karakteristika zabilježena je u svim zemljama članicama EU u 2012. godini. Stopa nezaposlenosti osoba sa nižim srednjim obrazovnim kvalifikacijama u 2012.godini kod osoba između 25 i 64 godine života, iznosila je 16,8% dok je kod onih koji su završili tercijalno obrazovanje niža i iznosila je 5,5%. Globalizacija i tehnološki napredak imaju sve veći utjecaj na svakodnevni život, a potražnja za različitim poslovima i radnim vještinama razvija se brzim tempom. Poduzeća u nastojanju da poboljšaju

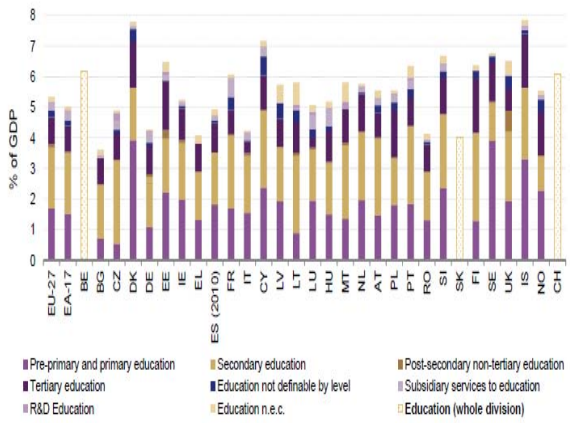


Slika 1. Stopa nezaposlenosti mladih do 25 godina u 2012godini. Izvor: Eurostat

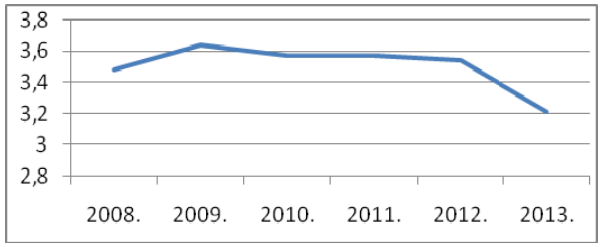


Slika 2.Mladi od 15-24 godine, EU-28, 2013. Izvor: Eurostat

svoju produktivnost, konkurentnost i inovativnost, trebaju sve fleksibilniju radnu snagu. Strategija Europa 2020 koju je donijela Europska komisija navodi viziju europske socijalne tržišne ekonomije za 21. stoljeće. U njoj su sadržane aktivnosti koje idu ka rješavanju uzroka globalne ekonomske krize, velike stope nezaposlenosti mladih ljudi, stvaranju potrebnih znanja i vještina, socijalnim politikama, uključujući i prijedloge usmjerene na institucije za obrazovanje i osposobljavanje, odnosno mjere za stvaranje (radnog) okruženja pogodnog za više stope aktivnosti i veću produktivnost rada. Tu su i inicijative usmjerene na povećanje stope mladih na tržištu rada.



Slika 3. Izdaci opće države za obrazovanje,% od BDP-a,2011
Izvor: Eurostat



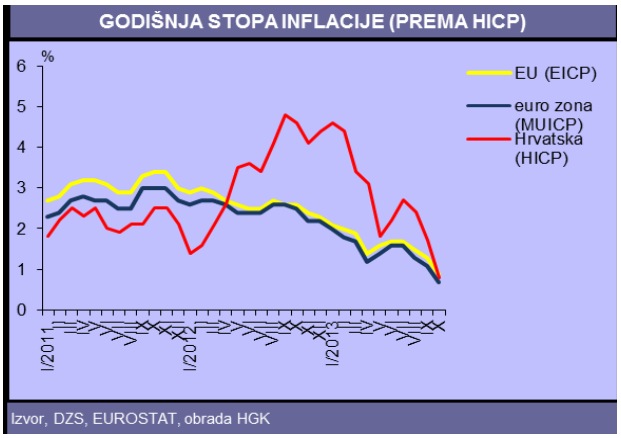
Graph 1. Izdvajanja iz proračuna RH za obrazovanje
Izvor: Ministarstvo financija, 2013

A. Neujednačenost proračunskih izdvajanja za obrazovanje

Iznos proračunskih izdvajanja za obrazovanje utječe na kvalitetu i učinkovitost obrazovnog sustava. Proračunski udio u BDP-u za obrazovanje EU-27 u 2011.godini iznosio je oko 347 milijardi eura (Slika 3). Značajnija izdvajanja za obrazovanje zabilježena su na Islandu (7,9 % BDP-a), Danskoj (7,8 % BDP-a) i Ciparu (7,2 % BDP-a), dok se najviše izdvajalo u u Estoniji (16,9 % ukupnih rashoda države), i Islandu (16,7 % ukupnih rashoda). Najniži omjer državnih rashoda za obrazovanje u BDP-u zabilježen je u Bugarskoj (3,6 % BDP-a), Slovačkoj (4,0 % BDP-a), Grčkoj i Rumunjskoj (obje na 4,1% BDP-a). U 2013. godini u Hrvatskoj su izdvajanja za obrazovanje iznosila 3,2% BDP-a s tendencijom daljnjeg smanjena (Graph 1).

B. Hrvatsko gospodarstvo u uvjetima globalne ekonomske krize

Prema [7] u 2013.godini u Hrvatskoj se nastavlja pad gospodarske aktivnosti koji se ogleda u smanjenju vrijednosti proizvedenih roba i usluga te zaposlenosti. Prema podacima Državnog zavoda za statistiku, u prva tri kvartala 2013. godine realni BDP je smanjen za 0,9%, pri čemu je najmanji pad (-0,6%), pod utjecajem dobre turističke sezone, zabilježen u trećem kvartalu. Struktura BDP-a za prvih devet mjeseci 2013.godine pokazuje da je i u Hrvatskoj najveći problem daljnji nastavak pada domaće potražnje, realno smanjena ukupna vrijednost



Slika 4. Godišnja stopa inflacije u periodu od 2011.-2013.godine
(premaHICP)

izvoza roba i usluga. U usporedbi s drugim članicama EU takvo kretanje BDP-a stavlja Hrvatsku među najneuspješnije članice, odnosno samo je Grčka (-4,3%) u razdoblju od 2008. do 2013. godine imala veću prosječnu stopu pada BDP-a od Hrvatske (-1,7%). Hrvatska i Grčka su ujedno jedine dvije zemlje kod kojih se u razdoblju od 2009. do 2013.godine konstantno bilježio pad BDP-a. Prema [7] sam oporavak od krize u 2013.godini u Hrvatskoj je među najsporijima, pa je svrstana među deset članica EU čiji je BDP realno manji nego u prethodnoj godini. U kontekstu rekordne nezaposlenosti i jakog eura, inflacija je slabila u 2013.godini na području cijele skupine zemalja EU-28 (Slika 4). Pritom je primjetno rastao broj zemalja koje su bilježile deflaciju (Grčka, Bugarska, Cipar, Irska) ili su na njenom samom rubu s razinom inflacije od 0,0% (Španjolska, Latvija, Portugal).

C. Tržište rada u Hrvatskoj

Evidencija Hrvatskog zavoda za zapošljavanje, u prvih deset mjeseci 2013. godine bilježi prosječno 342.107 nezaposlene osobe. Toliki broj nezaposlenih zabilježen je jedino 2002. godine, kad je evidentirano preko 389 tisuća nezaposlenih. Tada već, s podacima za nepunu 2013. godinu, nezaposlenih je osoba bilo otprilike 17% više nego prethodnih kriznih godina (računavši prosjek od 2008. do 2012.), sa očekivanjem daljnjeg rasta. Od ukupno 344.444 nezaposlene osobe registrirane na Zavodu krajem listopada 2013.godine, najviše njih (63%) imalo je završenu trogodišnju ili četverogodišnju srednju školu. Najmanja razina nezaposlenosti evidentirana je među onima s najvišim stupnjem obrazovanja (4,5%), dok je nešto više od četvrtine ukupno nezaposlenih osoba bez škole ili imaju završenu samo osnovnu školu. Podaci ukazuju na gotovo 30% nezaposlenih osoba između 20 i 29 godina dok je najmanje onih između 15 i 19 godina (školovanje), kao i osoba koje imaju 60 godina i više.

II. GOSPODARSTVO NA LOKALNOJ RAZINI:BRODSKO-POSAVSKA ŽUPANIJA

Želeći dati svoj doprinos u ublažavanju posljedica gospodarske krize odnosno poticanju gospodarskog

razvoja na svom području, bez pretenzija i velikih očekivanja obzirom na vrlo ograničena sredstva i utjecaj regionalne i lokalne samouprave na makroekonomska kretanja, Brodsko-posavska županija samostalno i u suradnji s drugim institucijama provodi niz aktivnosti i projekata u cilju poticanja gospodarstva odnosno poduzetništva te stvaranja preduvjeta za povećanje konkurentnosti županije kao regije (poticanje konkurentnosti poduzetnika i županije kao regije, kreditiranje poduzetnika uz subvenciju kamata, efikasno korištenje državnih potpora, poticanje obrtništva, županijska razvojna agencija, informiranje i edukacija poduzetnika, korištenje EU fondova, te poduzetničke zone u Brodsko-posavskoj županiji). Krajem prosinca 2012.godine u Brodsko-posavskoj županiji stopa aktivnosti (omjer aktivnog stanovništva-zbroj zaposlenih i nezaposlenih i radno sposobnog stanovništva) iznosila je 51,4%. U isto vrijeme stopa aktivnosti u Republici Hrvatskoj iznosila je 62,1%. U istom razdoblju stopa zaposlenosti (omjer zaposlenog stanovništva i radno sposobnog stanovništva) iznosila je 33,5% u Brodsko-posavskoj županiji, a evidentirana stopa zaposlenosti u Republici Hrvatskoj 49,6%. Broj zaposlenih izračunat je na temelju podataka o broju aktivnih osiguranika obveznog mirovinskog osiguranja pri Hrvatskom zavodu za mirovinsko osiguranje. Niske stope aktivnosti i zaposlenosti čine veći problem od visoke stope nezaposlenosti na području Brodsko-posavske županije.

A. Tržište rada na području Brodsko-posavske županije

Krajem prosinca 2013. u Hrvatskom zavodu za zapošljavanje, Područnom uredu Slavonski Brod registrirano je 18619 nezaposlenih osoba, što je za 560 osoba više nego u prethodnom mjesecu i za 71 osobu više nego u prosincu 2012. godine. Dakle, u prosincu 2013. godine nezaposlenost je povećana u usporedbi s prethodnim mjesecom za 3,1%, dok je u usporedbi s istim mjesecom 2012. godine povećana za 0,4%. U obrazovnoj strukturi nezaposlenih osoba najbrojnije su osobe sa završenom srednjom školom za zanimanja u trajanju do 3 godine i školom za KV i VKV radnike (7184 ili 38,6%), osobe sa završenom srednjom školom za zanimanja u trajanju od 4 i više godina (4258 ili 22,9%), slijede osobe sa završenom osnovnom školom (3835 ili 20,6%), osobe bez škole i s nezavršenom osnovnom školom (1699 ili 9,1%), osobe s višom školom (659 ili 3,5%), osobe fakultetskog obrazovanja (532 ili 2,9%), te osobe sa završenom gimnazijom (452 ili 2,4%). stopa nezaposlenosti u Brodsko-posavskoj županiji u 2013.godini iznosila je 32,1%.

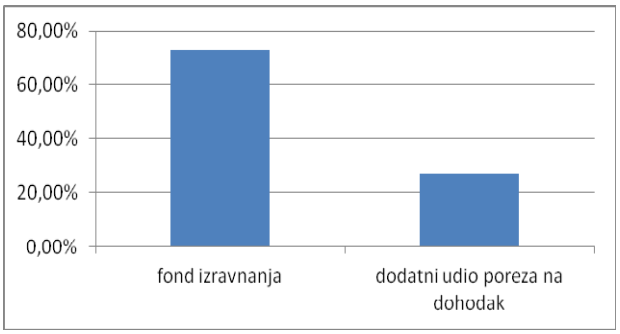
B. Važnost ulaganja u ljudske potencijale,reforma strukovnog obrazovanja

U razvijenim zemljama temelj gospodarskog i društvenog napretka čine ljudski potencijali. Ulaganje u ljudske potencijale nužno je u Hrvatskoj posebno kroz obrazovni sustav u kojem će se smišljeno otkrivati, njegovati, usmjeravati te poticati individualni potencijali i sposobnosti. Jedan od načina je cjeloživotno učenje i obrazovanje kroz koji je moguće učinkovito i pravovremeno, a nadasve smišljeno pomagati pojedincima u razvijanju njihovih talenata, vještina i znanja, bez obzira na dob. U sustave cjeloživotnog učenja u Hrvatskoj uključen je mali broj građana, iako je velik broj nezaposlenih i obrazovanih s nedovoljnim osnovnim kompetencijama za opstanak na tržištu rada ili za stalno prilagođavanje društvenim i gospodarskim promjenama i izazovima. Uvođenje učinkovitog sustava strukovnog obrazovanja važno je za transformaciju postojećeg gospodarstva u gospodarstvo temeljeno na znanju. Učinkovitost podrazumijeva reformu strukovnog obrazovanja koja treba ići u pravcu modernizacije strukovnog obrazovanja, prohodnosti, partnerstva s gospodarstvom, usklađenosti s potrebama tržišta rada i potrebama i sposobnostima pojedinca. Resorno ministarstvo donijelo je Strategiju razvoja strukovnog obrazovanja i Zakon o strukovnom obrazovanju s ciljem prilagodljivosti strukovnog obrazovanja potrebama polaznika i tržišta rada kroz modernizaciju sadržaja, metoda strukovnog obrazovanja i osiguranja mogućnosti nastavka obrazovanja. Agencija za strukovno obrazovanje i obrazovanje odraslih je od svoga osnutka aktivno uključena u provedbu brojnih međunarodnih i bilateralnih projekata u području strukovnog obrazovanja i osposobljavanja kako bi dodatnim izvorima financiranja unaprijedila kvalitetu strukovnog obrazovanja u Republici Hrvatskoj.Projekcija daljnjih proračunskih smanjenja izdvajanja za obrazovanje nameće nužnost traženja drugih izvora financiranja. Obzirom na mala proračunska izdvajanja za obrazovanje u Hrvatskoj obrazovne ustanove koriste mogućnosti koje im pružaju sredstva EU fondova.

C. Sredstva fondova EU – dodatno ulaganje u unapređenje obrazovnog sustava

Dobivena sredstva iz EU fondova srednjoškolskim obrazovnim institucijama osiguravaju uvođenje suvremenih i inovativnih sadržaja u nastavu, prijeko potrebnih strukovnim školama, unutar postojećeg sustava strukovnog obrazovanja i osposobljavanja, a u skladu sa Zakonom o strukovnom obrazovanju i Zakonom o odgoju i obrazovanju u osnovnoj i srednjoj školi. Istraživanje tržišta rada i potreba poslodavaca ukazuje na problem pronalaženja mladih ljudi s potrebnim tehničkim znanjima i vještinama koje gospodarstvo u određenom trenutku treba. Najizraženije su potrebe za mladim ljudima koji su završili srednje strukovno obrazovanje. Njihova postojeća znanja i vještine potrebno je dodatno unaprijediti kako bi mogli odgovoriti zahtjevima tržišta rada. Usklađivanje postojećeg sustava odgoja i obrazovanja sa potrebama gospodarstva treba biti prvi korak ka reformi obrazovnog sustava. Važnost strukovnog obrazovanja značajno varira od zemlje do zemlje, kao i sama organizacija, financijska izdvajanja za obrazovanje koja su uglavnom proračunska i o kojima

ovisi kvaliteta obrazovnog sustava, a time i strukovnog obrazovanja.

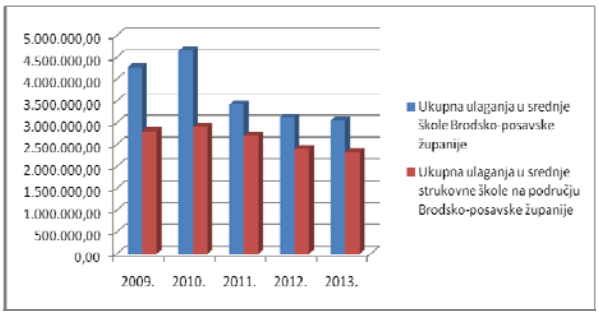


Graph 2.Decentralizirane funkcije srednjeg školstva Brodsko-posavske županije u 2013.godini-struktura prihoda
Izvor: Ministarstvo znanosti, obrazovanja i športa, 2013

D. Ulaganja u strukovne škole na području Brodsko-posavske županije

Brodsko-posavska županija uložila je značajne napore u izgradnju institucionalnih kapaciteta neophodnih za učinkovitu apsorpciju EU sredstava, a osobitu pozornost posvetila je upravo podizanju razine kvalitete strukovnog obrazovanja. Tako su u razdoblju od 2007. do 2013. kroz programe CARDS i IPA, srednje strukovne škole na području Brodsko-posavske županije realizirale EU projekte ukupne vrijednosti 28.883.000,00 kn, što je izuzetno značajan iznos, osobito ako ga usporedimo s ukupnim ulaganjem iz proračuna same županije.
Najznačajniji rezultati tih projekata su izgrađena infrastruktura, opremljeni praktikumi i laboratoriji, podignuta razina znanja strukovnih i drugih profesora, kao i razina znanja učenika, a sve s ciljem povećanja konkurentnosti učenika strukovnih škola i usklađivanja strukovnog obrazovanja s potrebama tržišta rada.
Kvaliteta strukovnog obrazovanja na području Brodsko-posavske županije određena je prvenstveno proporcionalnim izdvajanjima za srednje obrazovanja u Proračunu Republike Hrvatske, koja se potom raspoređuju jedinicama lokalne i područne (regionalne) samouprave (Brodsko-posavska županija)koje su osnivači srednjih škola na svom području. Osnivači srednjih škola utvrđuju prioritete ulaganja koji su u pravilu primarni.

U 2013.godini Brodsko-posavska županija zbog vrlo niskog BDP po stanovniku i fiskalnog kapaciteta, 72,9% prihoda za osiguranje minimalnog financijskog standarda srednjih škola ostvarila je prihodima iz sredstava poravnanja (9.075.637,00 kn), a tek 27,1% od prihoda od dodatnog udjela poreza na dohodak (3.373.027,52 kn)(Graph 2). Dakle izvorni prihodi Proračuna Brodsko-posavske županije ne doprinose unapređenju obrazovnog sustava na svom području niti u jednom segmentu, nego jedino održavanju minimalnog standarda.



Graph 3.Ulaganja u srednje obrazovanje na području Brodsko-posavske županije u periodu od 2009.-2013.godine
Izvor:Izračun autora,Proračun Brodsko-posavske županije

Iako realno Brodsko-posavska županija iz izvornog Proračuna nema sredstava kojim bi unaprijedila rad srednjih škola na svom području, gotovo 50% ukupnih sredstava koje dobiva iz Fonda poravnanja (Državni proračun) ulaže u srednje strukovne škole na svom području (Graph 3). Međutim sva ta sredstva nisu dovoljna za učinkovitije strukovno obrazovanje na području Brodsko-posavske županije i sve zahtjevnije potrebe tržišta rada. Mogućnosti za unapređenje strukovnog obrazovanja obrazovne institucije nalaze u korištenju sredstva EU.

E. Primjeri dobre prakse

- *Srednja škola „M.A.Reljković" Slavonski Brod*

Kako je riječ o županiji u kojoj je glavna gospodarska grana poljoprivreda, projekt IPA IIIc koji je provela srednja škola „M.A.Reljković" Slavonski Brod uz stručnu pomoć Brodsko-posavske županije i županijske razvojne agencije imao je dodatnu težinu. Radi se o izgradnji regionalnog centra za biotehnološka istraživanja i razvoj Brodsko-posavske županije kako bi se potaknuo gospodarski rast u županiji, temeljen na istraživačkim i razvojnim aktivnostima i transferu novih tehnologija u poljoprivredi. Regionalni centar čini jezgru inovativnog načina proizvodnje bezvirusnih sadnica (u Županiji i potencijalno u široj regiji) i pokretač je partnerskih aktivnosti vezanih uz istraživanje, edukaciju i razvoj u poljoprivredi. Na taj način se poljoprivrednim proizvođačima omogućava besplatna edukacija i savjetovanje u području agrotehnologije i razvoja poduzetničkih znanja i vještina kako bi uspješno proizvodili visoko profitabilne kulture, uveli bezvirusni materijal u proizvodnju te primjenjivali principe inovativne dinamičke proizvodnje. Ujedno projekt snažno podupire umrežavanje istraživačkih institucija i institucija koje pružaju potporu poslovanju poduzetnika s privatnim sektorom.

- *Industrijsko- obrtnička škola, Nova Gradiška*

U okviru programa IPA IV komponenta (razvoj ljudskih resursa) škola je provela projekt pod nazivom „Nova metodologija poučavanja i komunikacije u izvođenju praktične nastave u području CNC tehnologije". Navedenim projektom pojačalo se uvođenje suvremenih i inovativnih sadržaja u nastavu u strukovnim školama unutar postojećeg sustava strukovnog obrazovanja i osposobljavanja, s ciljem osiguranja

njihove prilagodljivosti na tržištu rada, povećanja kapaciteta za pružanje suvremene školske praktične nastave te ispunjavanja potreba gospodarstva utemeljenog na znanju. Realizacijom projekta dana je potpora strukovnim školama u razvoju pristupa „bottom –up“ (odozdo prema gore) kako bi svojim učenicima mogle osigurati kompetencije i pristup najnovijim tehnologijama podižući tako njihovu relevantnost na tržištu rada, a istovremeno pridonijeti razvoju obrazovnog sustava u Brodsko-posavskoj županiji u skladu s potrebama gospodarstva.

- *Tehnička škola, Slavonski Brod*

Opći cilj projekta ove škole temeljio se na razvoju i modernizaciji strukovnih obrazovnih sustava u Republici Hrvatskoj. Osim razvoja i modernizacije, specifični cilj projekta bio je unaprijediti profesionalne, materijalne i tehničke uvjete u strukovnoj školi, kao i provedbu novog kurikuluma (tehničar solarnih sustava) kroz razvoj demonstracijskog centra za solarne sustave. Naznačeni projekt osigurao je dva aspekta modernizacije obrazovnog sustava unutar Tehničke škole Slavonski Brod:

1. Razvoj kurikuluma koji prati gospodarske trendove (razvoj i primjenu solarne energije kao alternativni izvor)
2. Najefikasnije metode i tehnike prijenosa znanja unutar struke.

Tehnička škola u solarnom demonstracijskom centaru danas ima:

1. Proizvodnju električne energije - solarna centrala od 10kW (jedinstvena kao takva na području regije)
2. Praktikum za učenike s 15 radnih mjesta
3. Videokonferencijsku salu (udaljeno učenje).

IV. ZAKLJUČAK

Obrazovne institucije prepoznale su potrebu reforme obrazovnog sustava u Hrvatskoj, posebno strukovnog obrazovanja i svojim uključenjem u provedbu projekata EU na lokalnoj razini doprinijele su provedbi reforme na nacionalnoj razini. Dodatna razvojna sredstva s EU razine predstavljaju način kojim se područja i stanovništvo žele osnažiti kako bi što uspješnije koristili sve prednosti jedinstvenog tržišta roba, kapitala, radne snage i usluga. To uključuje veće tržište, veću mobilnost u radu i obrazovanju, veći izbor usluga i mnoge druge prednosti.

U pretpristupnom razdoblju Hrvatska je na raspolaganju imala oko 150 milijuna eura godišnje. Od 1.srpnja 2013.godine taj se iznos povećao na 450 milijuna, a u 2014. godini na 1,08 milijardi eura. To su doista značajna sredstva. Stoga treba pozdraviti sve aktivnosti obrazovnih institucija u provedbi natječaja za dodjelu sredstava iz EU fondova kako bi se reforma obrazovnog sustava što prije provela i unatoč skromnim izdvajanjima iz državnog proračuna Republike Hrvatske. Reformom strukovnog obrazovanja postepeno bi se smanjila stopa nezaposlenosti posebice mladih ljudi, te pokrenule gospodarske aktivnosti u zemlji. Primjeri dobre prakse strukovnih škola na području Brodsko-posavske županije omogućiti će mladim ljudima lakše i brže uključivanje u svijet rada jer će posjedovati potrebna znanja, vještine i kompetencije koje lokalno tržište rada traži i koje su važne za razvoj gospodarstva na tom području.

LITERATURA

[1] George S. Mouzakitis,”The role of vocational education and training curricula in economic development” Procedia Social and Behavioral Sciences 2, 3914–3920 Elsevier Ltd 2010

[2] Kevin R. McClure, Education for economic growth or human development? The capabilities approach and the World Bank’s Basic Education,Project in Turkey, Compare: A Journal of Comparative and International Education,London, Routledge Jan 2013.

[3] E.Pema,S.Mehay, Career effects of occupation-related vocational education: Evidence from the military’s internal labor market,Economics of Education Review 31,680-693, 2012.

[4] H.Tamer Havaa, R.Erturgutb,“ An evaluation of education relations together with technology,employement and economic development components, Procedia Social and Behavioral Sciences 2, 1771–1775, 2010

[5] Zakon o odgoju i obrazovanju u osnovnoj i srednjoj školi, Narodne novine 87/2008.

[6] Zako n o strukovnom obrazovanju,Narodne novine 30/2009,24/2010.

[7] Informacija o stanju u hrvatskom gospodarstvu,Hrvatska gospodarska komora,Centar za makroekonomske analize, prosinac 2013.

Izrada programskih komponenti u TIA Portal programskom okruženju

Danijel Maršić, Goran Malčić i Ivica Vlašić
Tehničko veleučilište u Zagrebu/Elektrotehnički odjel, Zagreb, Hrvatska
danijel.marsic@tvz.hr

Sažetak - Kod izrade programske podrške u sustavima upravljanja koji su realizirani uporabom PLC uređaja kao upravljačkog dijela i SCADA računala, odnosno HMI panela kao nadzornog dijela, uobičajena je praksa kreiranje gotovih programskih modula koji odgovaraju određenoj vrsti procesne opreme kako nalaže S88 norma. Izrada programske podrške upravljačkog i nadzornog dijela za PLC uređaje tvrtke Siemens unutar programskog paketa TIA Portal V12 dio je obavezne nastave na specijalističkom studiju elektrotehnike Tehničkog veleučilišta u Zagrebu. U okviru laboratorijskih vježbi studenti izrađuju programske dijelove upravljačkih komponenti i povezuju ih sa nadzornom HMI aplikacijom za koju također izrađuju vizualno sučelje. Ovakav pristup izvođenju laboratorijskih vježbi rezultira ishodima učenja koji su direktno primjenjivi u praksi.

I. UVOD

Sustavi automatizacije u proizvodnim procesima mogu biti jednostavni (mogu upravljati samo jednim strojem u proizvodnoj liniji) do vrlo složenih (koji upravljaju procesom proizvodnje u cijeloj tvornici). U tu svrhu razvijaju se univerzalna i specijalistička rješenja sustava automatizacije proizvodnih procesa. Cijena univerzalnih rješenja velikih tvrtki u području automatizacije često je prevelika za investitora jer specijalistička rješenja mogu koristiti samo za određeni proizvodni proces. Zbog toga se javlja potreba za razvojem vlastitog univerzalnog sustava automatizacije koji je moguće koristiti u jako širokom spektru proizvodnih procesa. Tvrtke koje izrađuju programsku podršku za sustave automatizacije, ukoliko žele biti prisutne na svjetskom tržištu, moraju to raditi prema naputcima koji su navedeni u S88 norme za šaržne procese. U šaržnim procesima izlazni učinak procesa se javlja u količinama materijala. Šaržni proces ima početak i kraj te se obično neprestano ponavlja. Šaržni procesi posjeduju neke karakteristike diskretnih i kontinuiranih procesa no zasebna su vrsta industrijskih procesa. Primjere šaržnih procesa najčešće nalazimo u industrijama pića, prehrambenoj, cementnoj, farmaceutskoj industriji i slično.

Kontrolni modul (CM) je najmanji dio sustava upravljanja koji izvršava neku osnovnu funkciju koji se prema primijenjenom S88 standardu naziva komponenta. Kontrolni modul (eng. *Control Module, CM*) je najosnovniji element fizičkog modela kojeg S88 standard ga definira kao: "*Kontrolni modul je u pravilu jedan ili grupa senzora, aktuatora, ostalih kontrolnih modula, i povezane procesne opreme, koji sa stajališta upravljanja djeluje kao zaseban uređaj*". [1] Dakle, svaki kontrolni modul osigurava izravnu "poveznicu" sa procesom preko aktuatora i senzora te tako izvršava neke osnovne upravljačke funkcije. Komponente programske podrške standardizirane su prema najčešćim tipovima uređaja korištenih u industrijskim postrojenjima. Komponente mogu biti mjerni uređaji, osjetnici, aktuatori, ali i programske komponente kao što je PID regulator koji se izvodi u programirljivom logičkom kontroleru (PLC). Koristeći ove standardne komponente moguće je lako sve kontrolne module upravljačkog sustava uklopiti u upravljački program PLC uređaja koji je povezan sa SCADA sustavom.

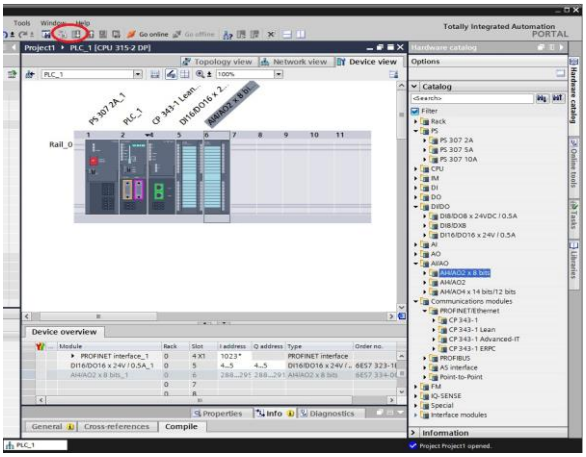
U sklopu laboratorijskih vježbi na kolegiju Sustavi upravljanja i nadzora postrojenja obrađuju se tematike vezane za izradu kontrolnih modula kroz više vježbi.

II. ALAT TIA PORTAL

Prvi dio kroz koji studenti svakako moraju proći je upoznavanje i rad sa razvojnom okolinom odnosno programskim editorom. *Totally Integrated Automation Portal* (*TIA Portal*) objedinjuje *SIMATIC Totally Integrated Automation* (*TIA*) proizvode u jedinstvenu programsku aplikaciju. Svi *TIA* proizvodi rade zajedno unutar istog programskog okruženja i pružaju podršku korisniku u svim segmentima potrebnim za stvaranje rješenja automatizacije. *TIA Portal* se koristi za konfiguraciju i programiranje PLC uređaja, ali i za vizualizaciju procesa u jedinstvenoj razvojnoj okolini. Svi podaci se spremaju u zajedničku projektnu datoteku, koristi se zajedničko korisničko sučelje za sve zadaće preko kojeg je moguće pristupiti svim programskim i vizualizacijskim funkcijama u svakom trenutku.

Prvi korak na vježbama je postupak sklopovske konfiguracije jer bez poznavanja tipa procesora nije moguće pisati programski kod. Ovim postupkom potrebno je unutar programske podrške definirati raspored i vrstu modula PLC uređaja koji odgovaraju konfiguraciji stvarnog uređaja. Kod dodavanja sklopovlja, odnosno modula treba biti obraćena pažnja da se odabere ispravan modul i njegova verzija. Zato je potrebno obavezno pročitati kataloški broj modula na njegovoj prednjoj strani i na temelju toga odabrati taj modul iz kataloga opreme u programskoj podršci. U protivnom će PLC nakon učitavanja hardverske konfiguracije javljati grešku.

[1] ISA Philadelphia Section: *ANSI/ISA S88 Batch standard – A General Overview*, 2002., www.isa.org



Slika 1. Prozor za izradu sklopovske konfiguracije

U TIA portalu spomenuti kontrolni moduli mogu biti izvedeni kao funkcije ili funkcijski blokovi koji se programiraju i pozivaju u glavnom programu OB1. Nedostatak korištenja funkcijskih blokova u odnosu na funkcije je to što se za svaki poziv funkcijskog bloka treba kreirati i pripadajući podatkovni blok, čak i ako se više puta poziva jedan te isti blok. Kod korištenja funkcija dovoljno je kreirati jedan globalni podatkovni blok u koji se spremaju podaci svih funkcija. Budući da funkcija sama po sebi ne može imati pridruženi vlastiti podatkovni blok, unutar nje se treba nalaziti neki podatkovni tip koji sadrži sve varijable koje se koriste u procesu. Pomoću tog podatkovnog tipa sve te varijable u globalnom podatkovnom bloku mogu dobiti svoju adresu u memoriji. Na taj način omogućeno je da se funkcija koristi za više istovrsnih procesa, jer će za svaki proces varijable biti na drugoj adresi.

U vježbama je dano i opisano kreiranje funkcija, njihovo programiranje, pozivanje u glavni program te rad s podacima. Radi lakšeg razumijevanja osim općenitih opisa koristi se nekoliko tipičnih primjera iz industrije.

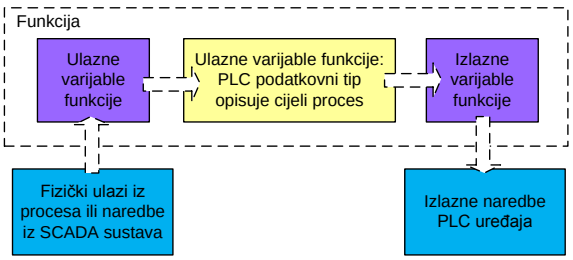
III. IZRADA UPRAVLJAČKE LOGIKE

Funkcije omogućuju izradu univerzalnih blokova koji se pozivaju u OB1 i tamo im se pridjeljuju fizičke adrese. Sva logika je sadržana u bloku koji kada pozovemo u glavnom programu ima vidljive izvode za fizičke ulaze i izlaze. Kao što je već spomenuto, nemaju vlastitu memoriju, nego se podaci spremaju preko PLC podatkovnog tipa koji sadrži varijable cijelog procesa. Ulazne varijable se upisuju pod *Input*, izlazne pod *Output*, ulazno - izlazne pod *InOut*, dok se privremene upisuju pod *Temporary* (Slika 2.). U donjem dijelu programa piše se program u jeziku ljestvičasti dijagram.

	Name	Data type	Offset	Comment
1	Input			
2	Ulaz	Bool		Ulaz
3	Output			
4	Izlaz	Bool		Izlaz
5	InOut			
6	Proces	"Proces_1"		
7	Ulaz	Bool		Ulazni uvjet
8	Izlaz	Bool		Izlaz
9	Alarm	"ALARM"		Alarm
10	Temp			
11	<Add new>			
12	Return			
13	Block_1	Void		

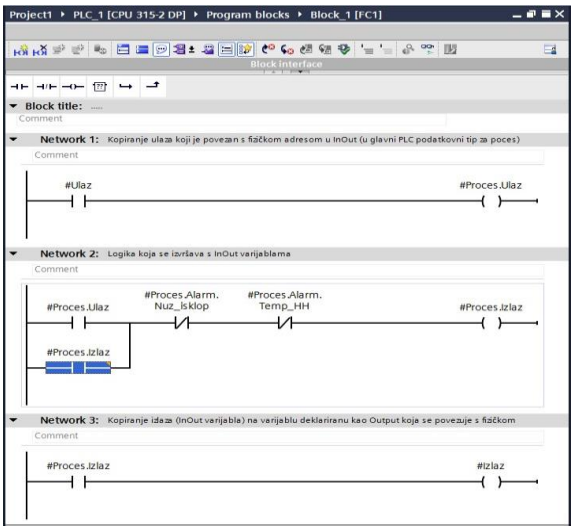
Slika 2. Primjer procesnih varijabli

Odnose između tipova varijabli deklariranih u zaglavlju najbolje prikazuje Slika 3. Sadržaj varijabli deklariranih kao *Input* treba se kopirati u odgovarajuće varijable u *InOut*. Isto tako sadržaj odgovarajućih varijabli iz *InOut* treba se kopirati u varijable u *Output* funkcije.



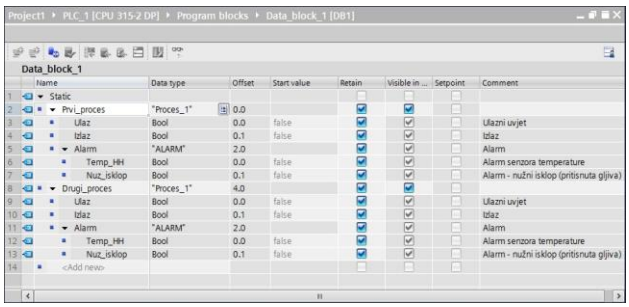
Slika 3. Blokovska shema toka podataka u funkciji

Slika 4. prikazuje logiku rada funkcije "Block_1" koja je isprogramirana u ljestvičastom dijagramu. Kod pisanja programskog koda u funkcijama ne koriste se fizičke adrese nego adrese u memoriji. Fizičke adrese dodjeljuju se bloku nakon poziva u OB1. U prikazanom primjeru bloku će se moći dodijeliti dvije fizičke adrese, na "Ulaz" i "Izlaz" te jedna adresa iz podatkovnog bloka na "Proces" koja će biti adresa varijable čiji će tip biti "Proces_1". Može se reći da se obrada podataka u bloku (funkciji) vrši pomoću varijabli deklariranih kao *InOut*. Varijable koje su deklarirane kao *Input* i *Output* služe za ulaz podataka u blok, odnosno za izlaz iz bloka i nemaju definiranu adresu u memoriji.



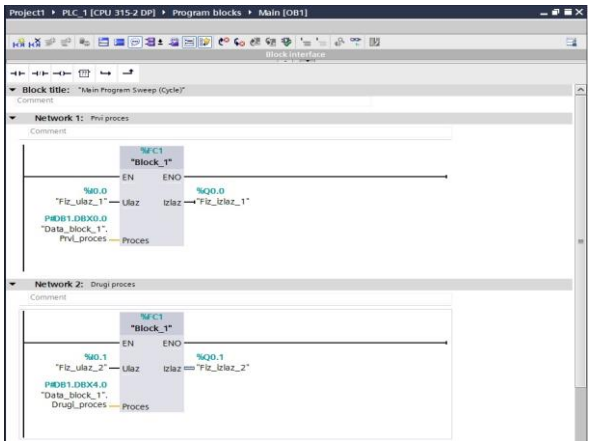
Slika 4. Logika rada funkcije "Block_1"

Varijable koje se nalaze u funkciji moraju imate definirane adrese u memoriji kako bi se podaci imali gdje spremati. To omogućuju globalni podatkovni blokovi. Slika 5. prikazuje izgled globalnog podatkovnog bloka gdje je vidljivo da su adresirane memorijske lokacije za dva procesa. Adresiranje se vrši tako da svaka varijabla koja je tipa "Proces_1", dobije početnu (referentnu) adresu, a varijable koje pripadaju toj varijabli dobiju neku vrijednost za koju su udaljene od referentne adrese. Procesi iz primjera međusobno su jednaki ali se mogu odvijati neovisno jedan o drugome. To znači da se blok sa logikom procesa mora pozvati u OB1 dva puta, za svaki proces posebno.



Slika 5. Globalni podatkovni blok

Slika 6. prikazuje izgled blokova kad se pozovu u OB1. U ovom slučaju pozvana je funkcija "Block_1" (otprije kreirana). Varijablama "Ulaz" i "Izlaz" pridružene su fizičke adrese, a pošto te varijable same po sebi nemaju definiranu adresu u memoriji, neće moći spremati podatke. No, kao što je vidljivo sa slike 4. to nije ni potrebno zbog toga jer je u logici bloka određeno da se ulazni i izlazni podaci pridružuju odgovarajućim varijablama koje su iz podatkovnog tipa "Proces_1", a te varijable imaju definiranu adresu u memoriji (Slika 5.)



Slika 6. Poziv funkcije Block_1 i pridruživanje adresa

Kako bi se funkcija izvršila mora biti pozvana iz OB1, u protivnome se ona neće nikad izvršiti. Nakon što je funkcija pozvana njenim ulazima su pridružene fizičke adrese iz procesa ili SCADA aplikacije. Podaci se sa ulaza šalju u odgovarajuće ulazno - izlazne varijable čije memorijske adrese su definirane u globalnom podatkovnom bloku. Unutar funkcije se tada odrađuje programirana logika (sve sa varijablama čije su memorijske adrese definirane u globalnom podatkovnom bloku) i rezultati (izlazni podaci) se šalju na izlaze funkcije kojima su pridružene fizičke adrese.

Na ovakav način se izvode tri vježbe za zasebne kontrolne module:

- direktno pokretanje motora,
- digitalni ulaz za diskretne senzore,
- analogni ulaz za analogne senzore.

Prolazeći kroz ove tri vježbe studenti nauče osnove rada u programskom okruženju TIA Portal V12. Najdetaljnije je objašnjena uporaba PLC podatkovnih tipova, funkcija i podatkovnih blokova. Vježbe se temelje na izradi modula koji su univerzalni za određeni tip procesa. Takvi moduli mogu se koristiti za programiranje bilo kakvog sustava ako on sadrži takve procese. To je dobra stvar za programere jer mogu puno brže napraviti programsko rješenje za zadani sustav, a izrađeni su moduli su izravno primjenjivi u praksi. Modul za direktno pokretanje i nadzor motora može se koristiti za upravljanje nekim vratima ili u nekom sličnom procesu gdje ne treba regulirati brzinu vrtnje. Modul za digitalni ulaz može se koristiti za provjeravanje stanja neke sklopke. Modul za analogni izlaz može se upotrijebiti za čitanje naponskog signala pretvarača neke fizikalne veličine i pretvaranje tog signala u stvarnu vrijednost mjerene veličine (npr. mjerenje temperature). S obzirom da je logika rada detaljno objašnjena u pisanim materijalima koji se dobiju na vježbama, student je može jednostavno razumjeti.

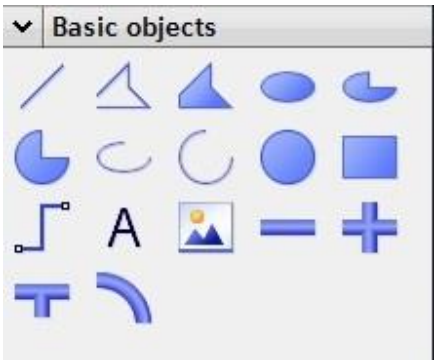
IV. IZRADA NADZORNE APLIKACIJE

SCADA (eng. ***S**upervisory **C**ontrol **A**nd **D**ata **A**cquisition*) je tehnologija koja omogućuje prikupljanje odnosno praćenje svih signala i veličina iz jednog ili više udaljenih postrojenja u kojemu se odvijaju neki procesi, te omogućuje slanje upravljačkih signala u postrojenja čime se dobiva stalni nadzor i upravljanje nad procesom.

Programska podrška *TIA Portal* koristi za kreiranje slika vizualizacije procesa na ekranu HMI (eng. *Human Machine Interface*) uređaja koji služi za upravljanje i nadgledanje strojeva i postrojenja. Prethodno definirani objekti i elementi dostupni unutar programa olakšavaju stvaranje tih ekrana. Grafički objekti i elementi su svi oni elementi koji mogu biti korišteni za vizualizaciju projekta u HMI sustavu. To uključuje tekst, tipke, dijagrame ili grafike za vizualizaciju dijelova procesa (senzori, aktuatori, uređaji). Grafički objekti mogu biti statično vizualizirani ili korišteni kao dinamički objekti uz pomoć oznaka.

U vježbama je objašnjeno uređivanje svojstava grafičkih objekata i elemenata korištenih prilikom vizualizacije procesa. Prvi korak vježbi je umetanje HMI uređaja i otvaranje njegove konfiguracije. Unutar projektnog stabla potrebno je odabrati naredbu *'Add new device'* i odrediti za koji tip odnosno veličinu ekrana će se raditi nadzorna aplikacija.

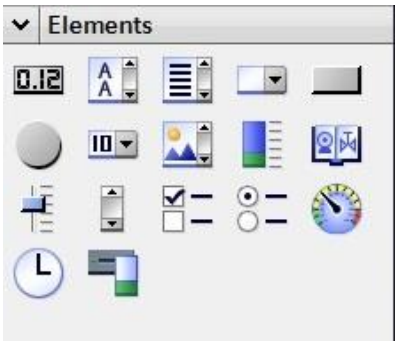
Osnovni grafički objekti nalaze se u kartici '*Toolbox*' pod nazivom 'Basic objects'. Pod osnovne objekte spadaju prethodno definirani objekti prikazani na slici 7.



Slika 7. Osnovni grafički objekti

Osnovni grafički objekti su jednostavni objekti koji mogu poslužiti za izradu odnosno dizajniranje jednostavnih rješenja vizualizacije procesa, ali se od njih također mogu stvoriti kompleksni objekti potrebni za vizualizaciju složenih procesa. Svojstva osnovnih objekata su prikazana po kategorijama i mogu se pronaći u prozoru namijenjenom za prikaz dodatnih informacija o odabranom objektu. Osim uređivanja svojstava osnovnih objekata moguće im je postaviti animacije što isto vrijedi i za naprednije grafičke objekte elemente.

Također se može raditi i sa predefiniranim elementima koji se nalaze se u kartici *'Toolbox'* pod nazivom *'Elements'*. Pod osnovne objekte spadaju prethodno definirani objekti prikazani na Slici 8. Svojstva elemenata prikazana su po kategorijama i mogu se pronaći u prozoru namijenjenom za prikaz dodatnih informacija o odabranom objektu.



Slika 8. Predefinirani elementi

U prvoj vježbi je prikazan postupak izrade jednostavne aplikacije za upravljanje motorom sa svim elementima koji se koriste prilikom vizualizacije procesa. U projektnom stablu, unutar mape *'Screens'* potrebno je odabrati opciju *'Screen'* kako bi se prikazao ekran na kojem će se dizajnirati SCADA aplikacija. Nakon što je kreiran prozor aplikacije, unutar njega se razmještaju ostali objekti i elementi koji su potrebni za funkcionalnost aplikacije.

Sljedeći korak, nakon dizajna aplikacije, je kreiranje tablice oznaka HMI uređaja (eng. *HMI Tag Table*). Tablica oznaka kreira se kako bi varijable definirane unutar tablice povezali sa objektima i elementima unutar SCADA aplikacije kako bi oni dobili funkcionalnost. Svaku od oznaka HMI uređaja potrebno je povezati sa stvarnim oznakama PLC uređaja kako bi mogli upravljati procesom putem SCADA aplikacije na HMI uređaju. Tablica oznaka kreira se unutar mape 'HMI tags' u projektnom stablu.

Objekte i elemente unutar SCADA aplikacije potrebno je povezati sa HMI oznakama kako bi aplikacija bila funkcionalna. Ukoliko bilo koji od objekata nema pridjeljenu pripadajuću HMI oznaku, objekt neće imati nikakvu funkciju animacije unutar aplikacije osim dekorativne.

Radi lakšeg snalaženja prilikom uređivanja aplikacije objektima i elementima je potrebno dodati naziv. Svi korišteni elementi će se kasnije grupirati unutar jedne skupine i ukoliko im nisu promijenjeni nazivi vrlo je teško raspoznati koji element čemu služi (npr. korištene su dvije tipke čiji su nazivi *'Button_1'* i *'Button_2'*). Kada na ekranu postoji više aplikacija potrebno je korištene objekte i elemente unutar svake aplikacije grupirati u skupine. To se radi prvenstveno iz razloga lakšeg razmještanja prozora aplikacija, ali i radi dodavanja efekata pojedinim prozorima.

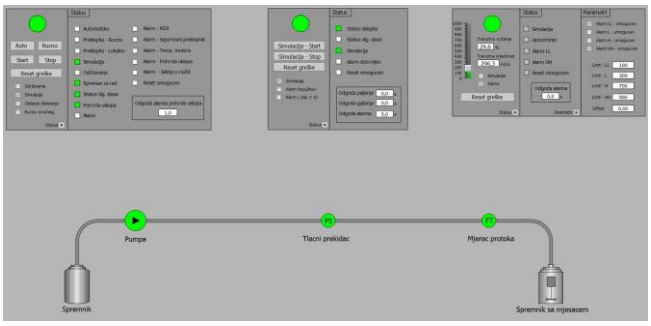
Završna faza prve vježbe je simulacija SCADA aplikacije na osobnom računalu koju studenti pokreću pomoću tipke *'Start simulation'* na alatnoj traci programske podrške *TIA Portal*. Na slici 9. prikazan je primjer izrađene aplikacije koja sadrži dva pravokutnika, dvije tipke, krug, potvrdni okvir, tekstualno polje i U/I polje upravljanju radom direktno pokretanog elektromotora. Moguće je pokrenuti i zaustaviti motor, zadati vremensku odgodu uključenja alarma nakon pojave greške te pratiti stanje rada motora.



Slika 9. Primjer aplikacije za upravljanje motorom

Druga i treća vježba su temeljene na istom pristupu izrade jednostavne nadzorne aplikacije za komponente digitalnog ulaza i analognog ulaza za koje su studenti prethodno izradili upravljački program na PLC uređaju. Nakon izrade svih komponenti pristupa se zadnjoj vježbi koja integrira sva prethodno stečena znanja na vježbama.

Vježba integracije svih komponenti je izrada sustava koji se sastoji od pumpe, tlačnog prekidača (eng. *Pressure switch*), mjerača protoka (eng. *Flow transmitter*) i dva spremnika. Spremnici su grafički objekti koji nemaju funkciju. Slika 10. prikazuje takav sustav u radu. Zelena boja indicira da sve komponente sustava rade ispravno. Pritiskom lijeve tipke miša na bilo koju od komponenata sustava (osim spremnika) otvara se dodatni prozor sa osnovnim dijelom aplikacije. Ponovnim pritiskom lijeve tipke miša na bilo koju od komponenata sustava zatvara se dodatni prozor komponente. Kako bi prikazali kartice 'Status' pojedinih komponenata potrebno je lijevom tipkom miša pritisnuti na tipku 'Status' koja se nalazi u donjem desnom uglu osnovnog dijela aplikacije.

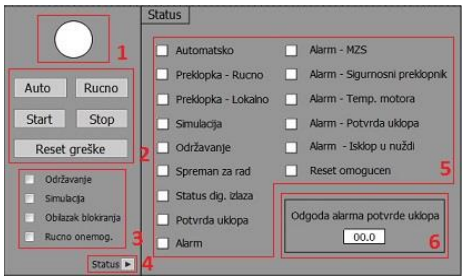


Slika 10. Zaslon vizualiziranog procesa

Studenti u završnom dijelu vježbe testiraju rad kako kompletne SCADA aplikacije tako i zasebnih prozora aplikacije. Zbog toga su na sljedećim slikama su prikazani i ukratko objašnjeni prozori aplikacije za svaku pojedinu komponentu sustava.

Slika 11. prikazuje prozor aplikacije za nadzor i upravljanje radom motora pumpe koji je sastavljen od sljedećih osnovnih dijelova sa pripadajućim funkcijama:

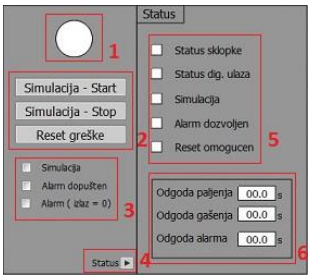
1) Indikator stanja pumpe.
2) Prostor sa tipkama za automatsko pokretanje motora, ručno pokretanje motora, simulaciju pokretanja motora i reset greške odnosno alarma.
3) Prostor unutar kojeg se aktivira način rada 'Simulacija' ili 'Održavanje'.
4) Tipka 'Status' kojom se prikazuje kartica 'Status'.
5) Prostor sa indikatorima stanja.
6) Prostor sa U/I poljem za zadavanje odgode alarma potvrde uklopa sklopnika.



Slika 11. Prozor aplikacije za nadzor i upravljanje motorom pumpe

Slika 12. prikazuje prozor aplikacije za upravljanje tlačnim prekidačem koji je sastavljen sljedećih funkcija koje su na raspolaganju operateru sustava:

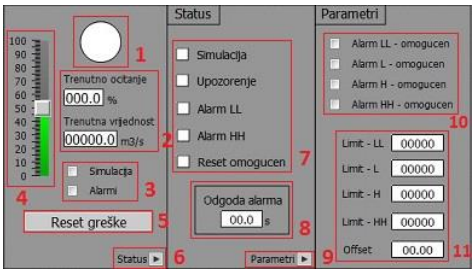
1) Indikator stanja tlačnog prekidača
2) Prostor sa tipkama za simuliranje aktivacije i deaktivacije tlačnog prekidača i reset alarma.
3) Prostor unutar kojeg se aktivira režim rada 'Simulacija' i dopuštaju alarmi.
4) Tipka 'Status' kojom se prikazuje kartica 'Status'.
5) Prostor sa indikatorima stanja (bijela boja – stanje nije aktivno, zelena boja – stanje aktivno).
6) Prostor za zadavanje odgode



Slika 12. Prozor aplikacije za upravljanje tlačnim prekidačem

Slika 13. prikazuje prozor aplikacije za upravljanje mjeračem protoka sastavljen od sljedećih funkcija:

1) Indikator stanja mjerača protoka.
2) Prostor za prikaz mjerene veličine.
3) Prostor unutar kojeg se aktivira režim rada 'Simulacija' i omogućavaju alarmi.
4) Klizač za ručno zadavanje vrijednosti mjerene veličine u režimu rada 'Simulacija'.
5) Tipka za reset greške odnosno alarma.
6) Tipka 'Status' kojom se prikazuje kartica 'Status'.
7) Prostor sa indikatorima stanja
8) Prostor za zadavanje odgode uključenja alarma.
9) Odabir se prikaza kartice 'Parametri'.
10) Prostor unutar kojeg se dopuštaju željeni alarmi.
11) Prostor za zadavanje granica uključuja alarma.



Slika 13. Prozor aplikacije za upravljanje mjeračem protoka

V. ZAKLJUČAK

Izrada upravljačke logike pojedinih komponenti procesa, vizualizacija procesa te mogućnost reakcije na proces u stvarnom vremenu postala je nezaobilazna potreba u industriji od početaka pa sve do danas. U tu svrhu najčešće se koristi PLC računalo i HMI sustav preko kojega čovjek nadzire i upravlja strojem. Radi boljeg razumijevanja procesa te lakšeg nadziranja i upravljanja procesom, izrada vizualizacije na HMI uređajima od velike je važnosti.

U ovom radu je prikazano kako studenti savladavaju problematiku izrade upravljačkog programa izradom pojedinih programskih komponenti te za iste njihovo vizualno sučelje. Korištene komponente su neizbježne u samoj praksi te znanja stečena na vježbama su sigurno primjenjiva u budućem radu. Završni dio svega je vježba koja integrira i upravljački i nadzorni dio u jednu cjelinu.

Student kroz ove vježbe lagano i postepeno gradi svoj prvi sustav automatizacije procesa, kroz čiju izgradnju dobije uvid u svu problematiku i rad sa sustavom. Vježbe traju pola semestra, no zadovoljstvo studenata se očituje kroz provedene ankete koje su više nego pozitivne. Naravo, dio njih smatra da je tematika preteška za studij, ali naše je nastojanje da budemo u korak sa potrebama tržišta rada.

LITERATURA

[1] Mario Belošević: *Izrada upravljačkog programa za module opreme unutar programske podrške TIA Portal v12*, završni rad Elektrotehnički odjel Tehničkog veleučilišta u Zagrebu, 10.07. 2013

[2] Dario Krklec: *Izrada HMI sučelja za module opreme unutar programske podrške TIA Portal v12*, završni rad, Elektrotehnički odjel Tehničkog veleučilišta u Zagrebu, 10.07. 2013

[3] Siemens; SIMATIC – STEP 7 Professional v12; Priručnik, izdanje 01/2013

[4] Siemens; *SIMATIC HMI – WinCC v12 Runtime Professional*, Priručnik, izdanje 05/2013

[5] Goran Malčić, Danijel Maršić: *Programirljivi logički kontroleri*, skripta za kolegij Procesna računala, Tehničko veleučilište u Zagrebu, Elektrotehnički odjel, Zagreb, 2009

Primjena industrijskih računalnih mreža u nastavi sa energetskim pretvaračima

Danijel Maršić, Goran Malčić i Veselko Tomljenović
Tehničko veleučilište u Zagrebu/Elektrotehnički odjel, Zagreb, Hrvatska
danijel.marsic@tvz.hr

Sažetak - Energetski pretvarači napona i frekvencije za upravljanje elektromotornim pogonima temeljenim na asinkronim motorima dio su obavezne nastave na stručnim studijima elektrotehnike. Njihovo povezivanje na više tipova industrijskih računalnih mreža je nadogradnja koja je nužna za primjenu u industriji. U radu je prikazana konfiguracija i primjena takvih sklopova koja se koristi na laboratorijskim vježbama iz područja industrijskih računalnih mreža. Prikazan je energetski pretvarač tvrtke ABB u pogonu s mrežama Profibus, EtherNet/IP i PROFINET I/O kako bi se studentima pokazale mogućnosti konfiguracije uređaja različitih proizvođača na različitim tipovima mreže s različitom programskom podrškom.

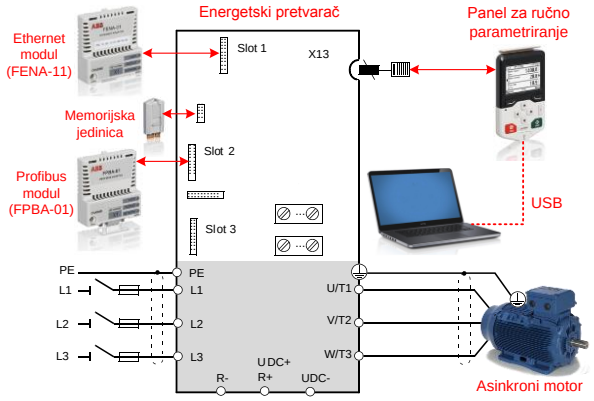
I. UVOD

U okviru kolegija Industrijske računalne mreže koji se predaje na Tehničkom veleučilištu u Zagrebu izvode se laboratorijske vježbe na kojima se pokazuje konfiguracija i rad sa više tipova mreža koje su najčešće u upotrebi. Kao krajnji objekt upravljanja na mreži koristi se energetski pretvarač napona i frekvencije koji je spojen na trofazni asinkroni motor čijim radom upravljamo. Prilikom svake vježbe, krajnji je cilj, naučiti studenta konfigurirati energetski pretvarač za rad na mreži, konfigurirati upravljačku mrežu preko osobnog računala te isprogramirati PLC računalo kao glavni upravljački element. Student se kroz ove vježbe susreće s više programskh i konfiguracijskih editora, kao i s problemom njihovog međusobnog povezivanja.

Na početku svake vježbe konfigurira se pretvarač napona i frekvencije i to ili preko ručnih komandi na samom pretvaraču ili preko programske podrške tvrtke ABB. Nakon konfiguracije uređaja, podešava se rad odabrane mreže i testira komunikacija između pretvarača i PLC uređaja. Kao zadnja stavka ostaje izrada upravljačkog programa za PLC računalo gdje se i testira rad sustava. Ovim vježbama se pokazuje koliko široko stručno znanje treba imati inženjer koji se bavi industrijskom automatizacijom, kao i koji mogući problemi mogu nastupiti kod upotrebe raznih tehnologija. Cilj vježbi je osim samoga rada na mrežama i njihovom sklopovskom opremom i programskom podrškom, pokazati i kako je poštivanjem mrežnih specifikacija moguće upogoniti opremu raznih proizvođača na istoj mreži što je ovdje slučaj. Za izvođenje vježbi su pripremljeni pisani materijali i izrađeni posebni laboratorijski modeli na kojima moraju zajedno raditi dvojica studenata da bi vježba bila uspješno izvedena.

II. KONFIGURACIJA ENERGETSKOG PRETVARAČA

Početni korak koji treba napraviti na vježbi je konfiguracija objekta upravljanja. Na ovim vježbama to je energetski pretvarač napona i frekvencije tvrtke ABB tip ACS880 za asinhrone motore snage do 2,5 kW na 400 VAC. Pretvarači su opremljeni ručnim upravljačkim komandama preko jednostavnog panela s tipkama i zaslonom na kojemu se vide parametri pretvarača u radu. Pregled energetskih i upravljačkih priključaka energetskog pretvarača prikazan je na sljedećoj slici.



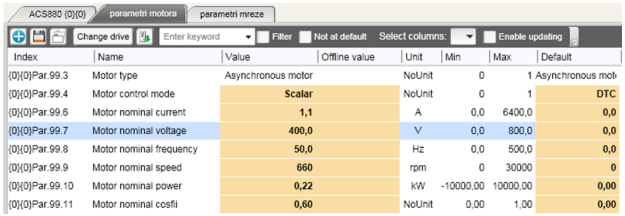
Slika 1. Osnovni dijelovi energetskog pretvarača ACS880 (ABB)

Studentima se na vježbama tumači kako se pretvarač može konfigurirati ručno, ali se konfiguracija ipak pokazuje preko namjenske programske podrške *Drive Composer Entry* tvrtke ABB koja se izvršava na osobnom računalu spojenom na pretvarač preko USB sučelja. Preko istoga student mora:

- upravljati pogonom: start, stop, smjer, referentna brzina / moment / frekvencija referenca,
- pratiti rad i stanja pogona,
- pregledati i podešavati pogonske parametre,
- pratiti signale u numeričkom i grafičkom obliku,
- stvoriti korisnički specifične radne prostore za prilagodbu parametara,
- rukovati radnim prostorom.

Na početku je potrebno u pretvarač unijeti parametre motora kako bi uređaj znao koje su njegove nazivne vrijednosti te s kojim modom upravljati njime, kako ne bi došlo do nedozvoljenih prekoračenja nazivnih vrijednosti.

Sve vrijednosti koje su potrebne za podešenje pretvarača nalaze se na natpisnoj pločici elektromotora i opisu tehnološkog procesa gdje se ugrađuje kao pogonski stroj.



Slika 2. Izgled programske podrške za podešavanje parametara

Drugi dio koji se mora podesiti na energetskom pretvaraču je mrežna komunikacija. Kako bi se spojio pretvarač na mrežu potreban je odgovarajući mrežni adapter. Svaki pretvarač opremljen je s dva mrežna adaptera i to Profibus adapter te ethernet adapter koji se upotrebljava za EtherNet/IP i PROFINET protokole. Njih je potrebno konfigurirati preko programske podrške Drive Composer Entry ili ručno preko panela da bi energetski pretvarač bio vidljiv i upravljiv preko mreže.

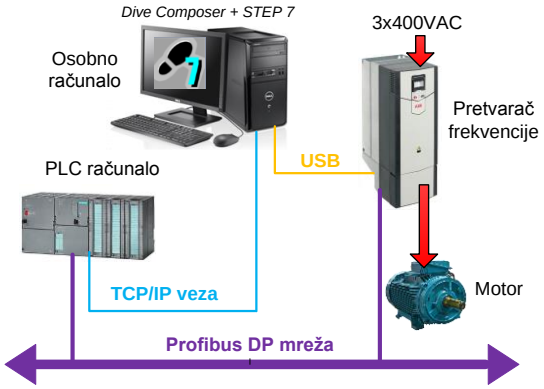
FPBA-01 Profibus DP adapterski modul je opcijski uređaj za ABB-ove pretvarače koji omogućuje spajanje pretvarača na Profibus DP mrežu. Pretvarač se u tom slučaju na Profibus mreži ponaša kao *slave* tj. kao podređeni uređaj. FENA-11 modul je kompatibilan sa svim trenutno važećim ethernet mrežnim standardima i podržava tri komunikacijska protokola: Modbus/TCP, EtherNet/IP i PROFINET IO. FENA-11 modul podržava 10 Mb/s i 100 Mb/s brzine prijenosa podataka te automatski detektira korištenu brzinu u mreži.

Preko adaptera tj. mrežnog modula moguće je upravljati pretvaračem dajući mu naredbe za npr. start, stop. Preko modula je moguće uspostaviti referencu za brzinu i moment elektromotora, te način zaustavljanja. Adapter također šalje stvarne vrijednosti stanja energetskog pretvarača, tako da se može upravljati povratnom vezom, mijenjati vrijednosti parametara pretvarača te prema potrebi resetirati grešku.

III. KONFIGURACIJA PREKO PROFIBUS MREŽE

Profibus je serijski komunikacijski standard koji omogućuje razmjenu podataka između komponenti na mrežnoj sabirnici. Profibus komunikacija se ostvaruje na razini polja gdje distribuirane jedinice kao što su ulazno/izlazni moduli, pretvarači napona i frekvencije, uređaji za analizu, mjerni sustavi, nadzorni terminali mogu komunicirati sa sustavom automatizacije u realnom vremenu [1]. Pogramirljivi logički kontroler (eng. Programmable Logic Controller, skraćeno PLC), koji radi kao nadređeni uređaj, čita ulazne podatke s podređenog uređaja te natrag šalje izlazne upravljačke veličine [5]. Vrijeme ciklusa Profibus sabirnice ovisi o broju čvorova na mreži i to vrijeme mora biti barem dva puta kraće od vremena ciklusa programa PLC računala što je važno i neophodno za vremenski kritičan prijenos podataka.

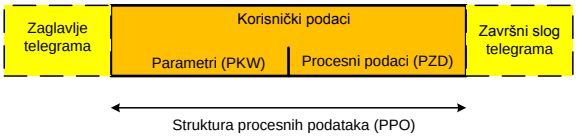
U okviru praktičnog dijela laboratorijskih vježbi uspostavlja se komunikacija između PLC uređaja serije S7-300 (Siemens) i pretvarača frekvencije serije ACS880 (ABB) spojenog preko FPBA-01 adapterskog modula na Profibus DP mrežu. Osobno računalo je spojeno preko ethernet mreže na komunikacijsku procesorsku jedinicu PLC računala i preko njega se konfigurira Profibus mreža i programira PLC računalo. Također, preko osobnog računala se konfiguriraju postavke mreže na energetskom pretvaraču tako da ga se spoji preko USB kabela (Slika 3.)



Slika 3. Principijelna blok shema sustava – Profibus DP

Preko programske podrške Drive Composer Entry potrebno je podestiti tzv. komunikacijske profile i adresu pretvarača na Profibus mreži. Komunikacijski profili su načini prenošenja upravljačkih naredbi tj. upravljačke riječi, statusne riječi, reference i stvarne vrijednosti između nadređenog uređaja i pretvarača. Za ove vježbe koristi se PROFIdrive profil koji se odnosi na strukturu procesnih podataka tj. PPO (eng. *Parameter Process data Object*). On definira strukturu podataka kojom *master* (hrv. nadređena stanica) može pristupiti *slave* (hrv. podređenoj stanici) uređaju, tj. pretvaraču koristeći metodu cikličkog prijenosa podataka. Podatkova struktura PPO za ciklički prijenos podataka podijeljena je na dva dijela:

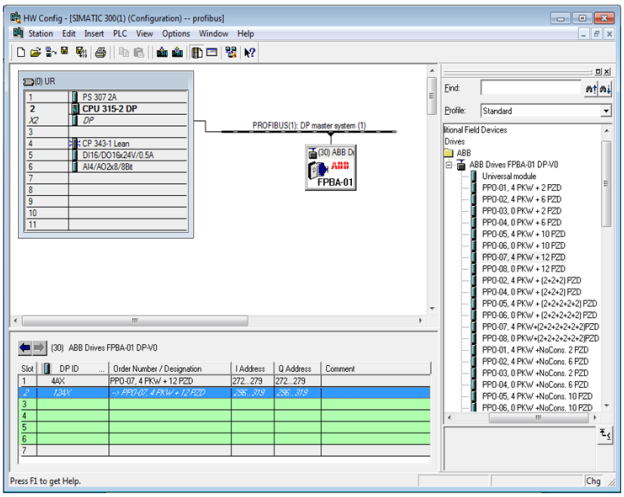
- PKW = Parametarsko područje za čitanje i mijenjanje parametara u pretvaraču. Parametarsko područje osigurava korisniku pristup i mijenjanje svih parametara pretvarača frekvencije. Može se koristiti i za čitanje dijagnostičkih poruka te poruka o greškama.
- PZD = Prikazuje procesne podake, upravljačke riječi, informacije o statusu pretvarača i aktualne vrijednosti. Procesni podaci se stalno prenose i to najvišim prioritetom prijenosa. Koriste se za upravljanje pretvaračem frekvencije na DP mreži, njegovo pokretanje i zaustavljanje, postavljanje referentnih vrijednosti i slično.



Slika 4. Podatkovna struktura PPO za ciklički prijenos podataka

U nastavku vježbi studenti u programskom paketu SIMATIC Manger konfiguriraju i programiraju i PLC uređaj, kako bi omogućili komunikaciju s energetskim pretvaračem. Prvo je potrebno u alatu za izradu sklopovske konfiguracije definirati virtualnu Profibus DP mrežu koja je u stvarnosti jednim krajem spojena na DP

sučelje procesorske jedinice PLC uređaja, a s druge strane na adapterski modul za komunikaciju pretvarača frekvencije. Isto tako, u sklopovsku konfiguraciju je potrebno upisati adresu prevarača na mreži koja je već prethodno odabrana u parametrima pretvarača. Profibus DP mreža se definira unutar sklopovske konfiguracije dvostrukim klikom na X2 koji predstavlja DP sučelje. Nakon uspješnog definiranja Profibus DP mreže i odabira adrese, na radnoj površini student vidi konfiguraciju PLC računala s pripadujućom DP mrežom koje je prethodno kreirana (Slika 5.). Za konfiguraciju nadređene stanice potrebna je GSD vrsta datoteka. Za komunikaciju sa DP-V0 Profibus protokolom postoji GSD datoteka imena proizvođača ABB. Kada je PPO struktura procesnih podataka jednom ubačena u komunikacijski modul tada je u konfiguracijskoj tablici moguće uočiti i zapamtiti adrese procesnih podataka (PZD) koje je potrebno znati prilikom izrade ljestvičastog dijagrama za nadzor i upravljanje pretvaračem frekvencije.



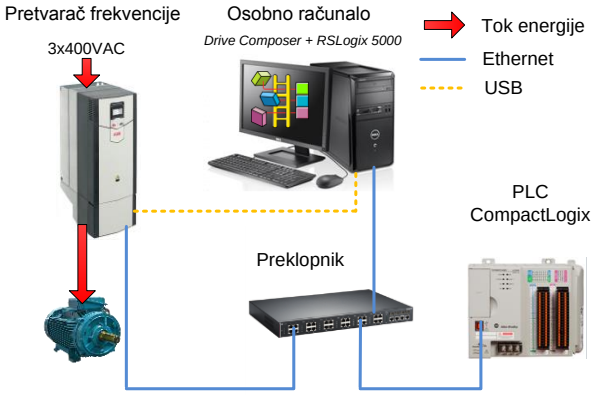
Slika 5. Podešena sklopovska konfiguracija u programu STEP 7

Zadnji dio vježbe je izrada upravljačkog programa upotrebom pripadajućih adresa procesnih riječi. Tek nakon što je konfiguracija, kako energetskog prevarača tako i PLC uređaja, provjerena od strane voditelja vježbi studenti mogu energetski pretvarač pustiti u pogon. Studenti koji prije predviđenog vremena prođu kroz obavezni dio vježbi, u okviru dodatnog zadatka provjere i Profibus komunikaciju upotrebom naprednijih struktura procesnih podataka (PPO).

IV. KONFIGURACIJA PREKO ETHERNET/IP MREŽE

EtherNet/IP (eng. IP – Industrial Protocol) je jedna od tehnologija ethernet mrežnih rješenja u industriji bazirana na otvorenoj tehnologiji, koristeći aplikacijski nivo – Common Industrial Protocol (CIP). EtherNet/IP se koristi s TCP/IP za slanje eksplicitnih poruka te se u svakom tom paketu osim aplikacijskog podatka nalazi i značenje podatka i servis odnosno zadatak koji se treba obaviti nad njime [4]. EtherNet/IP također zadovoljava i standard UDP/IP (User DatagramProtocol/Internet Protocol) koji je zbog svojih performansi i višekorisničke funkcionalnosti pogodan za potrebe rada s podacima u realnom vremenu – još poznato kao slanje implicitnih poruka (slanje poruka od jednog uređaja na mreži prema jednom ili više korisnika). Zbog toga što zadovoljava TCP/IP i UDP/IP protokole, EtherNet/IP proširuje mogućnosti rada na mreži i korisnici ga mogu upotrebljavati kako za prijenos podataka tako i za nadzor sustava (SCADA).

Laboratorijske vježbe se temelje na uspostavi komunikacije između PLC uređaja serije CompactLogix (Allen Bradley) i pretvarača frekvencije serije ACS880 (ABB) opremljenog FENA-11 adapterskim modulom. Na taj način su osobno računalo (ethernet kartica), PLC uređaj (integrirani ethernet priključak) i energetski pretvarač (FENA -11 adapter) međusobno povezani preko ethernet mrežnog preklopnika (Slika 6.).



Slika 6. Principijelna blok shema sustava – EtherNet/IP

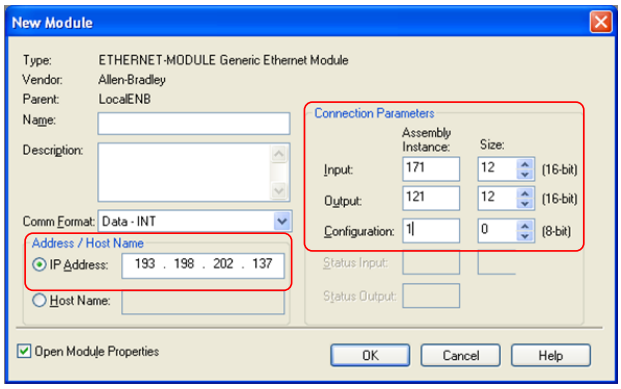
U prvom dijelu vježbi se preko USB kabela podešavaju postavke pretvarača za komunikaciju preko EtherNet/IP komunikacijskog protokola. Potrebno je unijeti IP adresu pretvarača na ethernet mreži i podesiti jedan od više različitih načina komunikacije koje podržava EtherNet/IP protokol. U ovom slučaju se koriste eksplicitne poruke kojima se zapisuju podaci na pretvarač i čitaju informacije s pretvarača. PLC uređaj obavlja ovu funkciju samostalno i direktno upisuje podatke i naredbe u parametre pretvarača i čita povratne informacije s pretvarača. Za upravljanje pretvaračem putem EtherNet/IP pretvarač mora zaprimiti upravljačku riječ koja određuje status pretvarača i reference kojima se pojedini parametri pretvarača mijenjaju. Povratne informacije koje se čitaju s pretvarača su trenutna vrijednost upravljačke riječi koja se naziva statusna riječ i trenutne vrijednosti pojedinih referenci. Za bolje i lakše upravljanje pretvaračem koriste se različite vrste poruka, tzv. Instance od kojih je svaka namijenjena za određeni tip upravljanja [2].

U ovoj vježbi studenti odabiru izlaznu instancu 121 koja omogućuje poboljšane mogućnosti upravljanja pretvaračem i podešavanje parametara pretvarača posredstvom programa iz PLC računala. Isto tako, koristi se ulazna instanca 171 u kojoj PLC računalo prima statusne informacije o smjeru vrtnje, greške i upozorenje, radna stanja, trenutnu brzinu pretvarača i sl. (Slika 7.).

Bajt	Bit 7	Bit 6	Bit 5		Bit 3	Bit 2	Bit 1	Bit 0
0	Dostignuta referenca			Pogon spreman	Vrtnja unazad	Vrtnja unaprijed	Upozorenje	Greška
1								
2	Trenutna brzina (niži bajt)							
3	Trenutna brzina (viši bajt)							
4...24	Ulazni podaci 1...10							

Slika 7. Struktura ulazne komunikacijske instance 171

U drugom dijelu vježbe studenti koriste programsku podršku RSLogix 5000 namijenjenu konfiguriranju, programiranju i praćenju rada PLC računala tvrtke Allen Bradley. Prvo je potrebno u postavkama ethernet sučelja PLC uređaja unijeti IP adresu pretvarača i ulaznu (171) odnosno izlaznu (121) instancu komunikacijskih poruka (Slika 8.).



Slika 8. Konfiguracija ethernet modula u RSLogix 5000

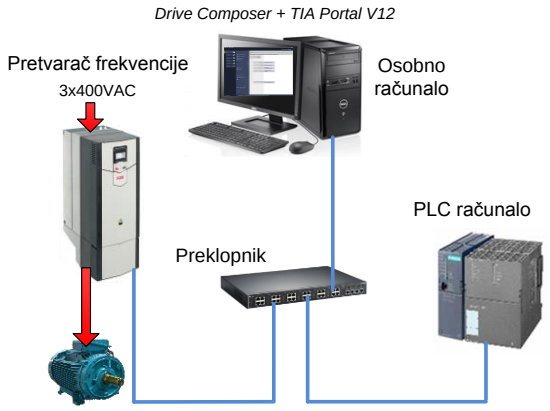
Nakon toga studenti upotrebom pojedinih adresa navedenih instanci izrađuju ljestvičastu logiku (eng. Ladder logic) za nadzor i upravljanje radom pretvarača frekvencije preko Ethernet/IP mreže. Zadnji dio vježbe je puštanje energetskog pretvarača u pogon i provjera ostalih komunikacijskih instanci za naprednije studente.

V. KONFIGURACIJA PREKO PROFINET IO MREŽE

PROFINET (*Process Field NETwork*) je otvoreni standard tvrtke Siemens za industrijsku komunikaciju temeljenu na Ethernet tehnologiji. PROFINET pruža dva različita komunikacijska protokola: PROFINET CBA i PROFINET IO. Za komunikaciju PLC računala s energetskim pretvaračem uređaja na laboratorijskim vježbama, koristimo PROFINET IO protokol koji općenito omogućava komunikaciju u realnom vremenu između distribuiranih jedinica na industrijskoj ethernet mreži. Uz standardne mrežne topologije (zvijezda, stablo, prsten). Profinet I/O podržava linijsku topologiju karakterističnu za sabirničke industrijske mreže, a to je ostvareno preko preklopnika integriranog u UI uređaj [3].

U ovoj laboratorijskoj vježbi koristi se nadređeno PLC računalo serije S7-300 (Siemens) sa zasebnim komunikacijskim modulom (CP 343-1 Advanced) koji upravlja radom pretvarača napona i frekvencije ACS880 (ABB) preko industrijske računalne mreže PROFINET I/O. Slika 9. prikazuje idejno rješenje konfiguracije sustava.

Kako bi PLC uređaj mogao komunicirati preko zasebnog komunikacijskog ethernet modula s pretvaračem frekvencije u upravljačkom programu potrebno je koristiti komunikacijske blokove FC11 „PNIO_SEND“ (za slanje podataka) i FC12 „PNIO_RECV“ (za primanje podataka). Zadaća komunikacijskoga bloka FC11 je prenositi podatke iz korisničkog programa odnosno centralne procesorske jedinice u ethernet modul CP343-1 Advanced. Funkcijski blok FC12 prima podatke iz modula CP343-1 Advanced i upisuje ih u korisnički program.

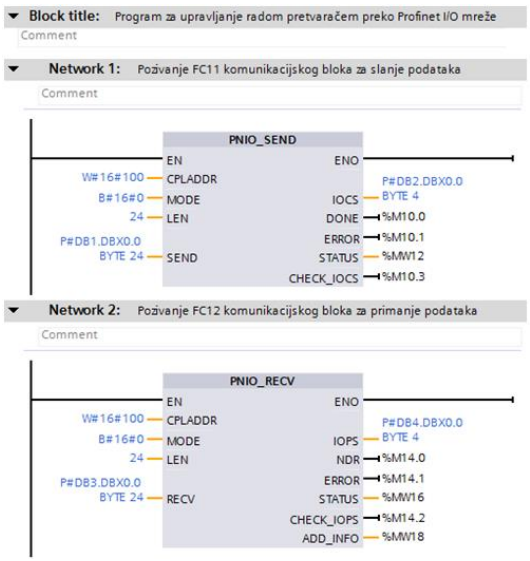


Slika 9. Principijelna blok shema sustava – PROFINET IO

Nastava se provodi korištenjem programskog paketa TIA Portal V12, tvrtke Siemens, pa su u sklopu vježbi objašnjene i osnove rada u tom paketu. TIA Portal V12 je novija generacija programske podrške za rad s PLC računalima tvrtke Siemens, koja je uvedena s namjerom da u potpunosti zamjeni programsku podršku STEP 7, koja se koristi već dugi niz godina.

U prvom dijelu vježbi se preko USB kabela podešavaju postavke pretvarača za komunikaciju preko PROFINET IO komunikacijskog protokola. Odabire se IP adresa pretvarača na mreži i strukturu procesnih podataka (u ovom slučaju PPO 7), s obzirom da PROFINET IO tehnologija koristi jednake komunikacijske profile kao i Profibus DP. U drugom dijelu vježbe studenti unutar programskog okruženja TIA Portal V12 definiraju virtualnu PROFINET IO mrežu te unose IP adresu i komunikacijske profile koji su već prethodno podešeni u parametrima pretvarača.

U zadnjem dijelu vježbe studenti izrađuju upravljački program upotrebnom komunikacijskih blokova FC 11 i FC12, te puštaju energetski pretvarač u pogon nakon provjere ispravnosti sustava od strane voditelja vježbi.



Slika 10. Prikaz primjene FC11 i FC12 komunikacijskih blokova

VI. ZAKLJUČAK

Razlog za uvođenje upravljanja energetskim pretvaračima unutar kolegija Industrijske računalne mreže je potreba za ovakvim specifičnim stručnim znanjima ne samo studenata, već i zaposlenih ljudi u industriji. Procesna industrija ne može opstati na tržištu bez kontinuiranog uvođenja novih tehnologija, a kako je razvoj računalnih mreža u industriji jedno od područja koje stalno napreduje, potreba za ovom nastavom se nametnula sama po sebi.

Kako bi se nastava izvodila što kvalitetnije, zbog obima i kompleksnosti vježbi, studenti rade u paru na jednom laboratorijskom modelu. Konstruktivistički pristup učenju, gdje studenti zajedničkim radom izrađuje mrežnu strukturu uređaja, podešavaju i programiraju uređaje na njoj, pokazao se kao najbolji za proučavanje takve tematike.

Kod ove vrste laboratorijskih vježbi, nastavnik je samo voditelj, a nikako predavač. Naime student mora naučiti da će se u svom budućem poslu na ovakvim sustavima svakodnevno susretati s potrebom čitanja tehničkih priručnika za rad s određenom opremom te uputstva za njihovo programiranje i konfiguraciju. Tako i u ovim vježbama studenti sami moraju pročitati detaljne upute i rukovoditi se njima da bi uspješno završili vježbu te ujedno se naučili koristiti tehnologijom. U radu nije navedeno da uz nastavne materijale, studenti na stolu i u računalu imaju originalne priručnike opreme na kojoj rade. Ispitivanje naučenog gradiva ne postoji, već samo to da su nakon vježbi koje traju tri školska sata uspjeli podesiti i isprogramirati uređaje smatra se prolaznom ocjenom.

Nastavni materijali su radi jednostavnijeg praćenja prepuni slika, ali uz ograničenu količinu teksta tako da studenti sami moraju shvatiti i povezati kako vježbu izvesti uspješno. Kako bi se na slikama mogli raspoznati detalji, potrebno je nastavne materijale tiskati u boji.

Problem na koji ćemo sigurno naići je činjenica da je zbog brzog razvoja tehnologije potrebno stalno nadopunjavati i osuvremenjivati nastavne materijale. Drugi problem je taj što su tehničke inovacije često popraćene novim stranim riječima koje je teško točno prevesti pa se je lakše služiti originalnom riječi.

U ovom radu, koji obiluje tehničkim detaljima, a koji su samo dio problema koje student mora riješiti na ovim vježbama željelo se pokazati koliko je ovo kompleksna tematika i zašto se takve vrste vježbi rijetko izvode na visokoškolskim institucijama. Naime često se potcjenjuje rad na laboratorijskim vježbama u odnosu na predavanja, no ovim se pokazuje da to često neopravdano.

LITERATURA

[1] Danijel Uremović: *Primjena industrijske računalne mreže Profibus DP u radu sa ABB pretvaračem serije ACS880*, završni rad - preddiplomski studij, Elektrotehnički odjel Tehničkog veleučilišta u Zagrebu, 12.07. 2012

[2] Roman Mešić: *Upravljanje pretvaračem ACS 880 sa PLC uređajem CompactLogix preko Ethernet/IP mreže*, završni rad - preddiplomski studij, Elektrotehnički odjel Tehničkog veleučilišta u Zagrebu, 03.12. 2012

[3] Julije Čučuk: *Upravljanje pretvaračem ACS880 preko ProfiNet I/O mreže sa PLC uređajima tvrtke Siemens*, završni rad - diplomski/integralni studij, Elektrotehnički odjel Tehničkog veleučilišta u Zagrebu, 18.12. 2013

[4] Goran Malčić: *Ethernet/IP - industrijski Ethernet.*, časopis Elektro. 8 (2004) , 4; 48-51 (članak, stručni).

[5] Goran Malčić, Danijel Maršić: *Programirljivi logički kontroleri*, skripta za kolegij Procesna računala, Tehničko veleučilište u Zagrebu, Elektrotehnički odjel, Zagreb, 2009

E-učenje: koncept i primjene u europskom vojnom školstvu

Vlasta Zekulić i Darko Možnik
Ministarstvo obrane Republike Hrvatske
Hrvatsko vojno učilište „Petar Zrinski“
Ilica 256B, Zagreb, Republika Hrvatska
Telefon: 01-3786 840 Fax: 01-3786 466 E-mail: vzekulic@morh.hr
Telefon: 01-3784 613 Fax: 01-3784 747 E-mail: dmoznik@morh.hr

Sažetak - Rad opisuje i objašnjava teorijski koncept i praktične primjene odnosno primjere e-učenja u europskom vojnom školstvu. Navode se motivi, prednosti i nedostaci, razlozi i mogućnosti od teorijskog koncepta do smjernica i konkretnih primjera primjene uspješnih procesa e-učenja u vojnom školstvu u nekim zemljama članicama EU.

Opisani su elementi e-učenja, vrste tehnologija e-učenja (združeno napredno distribuirano učenje /engl. JADL/ i napredno distribuirano učenje /engl. ADL/, CBT - Computer Based Training, immersive learning tj. virtualno učenje, m-learning tj. mobilno učenje, blending learning tj. hibridno učenje i kolaborativno učenje i alati /chat soba, forumi, wiki, webinar - web based seminar i webcast/).

Prikazani su razvoj i primjeri tečajeva e-učenja u europskom vojnom školstvu, kao i u primjeri primjene e-učenja u vojnom školstvu u MORH-u i OS RH na Hrvatskom vojnom učilištu (HVU) „Petar Zrinski“.

Kroz pogled u budućnost prikazuje se mjesto, uloga i značenje e-učenja u transformaciji HVU-a u visoko obrazovnu i znanstveno-istraživačku vojnu ustanovu u OS RH i MORH-u i sastavnicu sveučilišne zajednice u RH.

Sadašnje i buduće sposobnosti e-učenja u europskom vojnom školstvu. Zaključna razmatranja.

I. UVOD

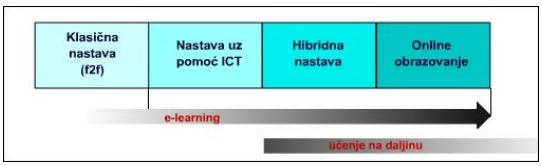
E-učenje i *učenje na daljinu* primjenjuju se u obrambenim i vojnim sustavima zemalja članica NATO saveza i EU. U obrambenim i vojnim sustavima zemalja članica NATO saveza *e-učenje* se koristi i primjenjuje pod nazivima: JADL (engl. *Joint Advanced Distributed Learning*) i ADL (engl. *Advanced Distributed Learning*).

O*pen source* sustavi za upravljanje učenjem tj. LMS-ovi (engl. *Learning Management Systems*) **MOODLE** (engl. *Modular Object-Oriented Dynamic Learning Environment*) najviše koriste akademske zajednice, dok se **ILIAS** (njem. *Integriertes Lern Informations und Arbeitsko-operations System*) koristi u obrambenim i vojnim sustavima zemalja članica NATO saveza i EU.

U MORH-u i OS RH u sustavu vojnog obrazovanja za *e-učenje* se koriste LMS sustavi MOODLE i ILIAS kao platforme i skupovi alata. Također, provodi se projekt "*Uvođenje novih načina usvajanja i održavanja znanja*" odnosno *e-obrazovanje u OS RH i MORH-u.* Određeni pripadnici OS RH i djelatnici MORH-a, a poglavito nastavnici Hrvatskog vojnog učilišta "Petar Zrinski" osposobljeni su vlastitim resursima, učinkovitim korištenjem ILIAS-a, razvijati i primjenjivati *online* ADL tečajeve i module za obrambene i vojne potrebe.

II. E-UČENJE I UČENJE NA DALJINU

E-učenje kao oblik obrazovanja postoji na više razina: kao potpuno samostalan oblik, ali i kao sastavni dio ili nadopuna klasičnog obrazovanja. Proširena klasifikacija ili "vremenska crta" *e-obrazovanja* prikazuje obrazovanje kao kontinuum e-učenja.



Slika 1. Prikaz "kontinuum"a *e-učenja*

Na slici 2. prikazan je odnos *e-učenja, učenja na daljinu i online obrazovanja*. Iako se *e-learning* i *učenje na daljinu* često izjednačuju, nije riječ o istim oblicima obrazovanja: postoje vrste *e-obrazovanja* koje se ne odvijaju online, a isto tako postoje i oblici *učenja na daljinu* koji ne koriste ICT.

A. E-učenje

E-učenje je vrsta komunikacijskog kanala kroz koji se odvija učenje. Postoje dvije skupine definicija *e-učenja* i to: 1. skupina "tehničkih" definicija stavlja naglasak na tehnologiju (na "*e*" dio u imenu). Primjer jedne od "tehničkih" definicija glasi: *"E-obrazovanje* je bilo koji oblik učenja, poučavanja ili obrazovanja koji je potpomognut uporabom računalnih tehnologija, a posebno računalnih mreža temeljenih na Internet tehnologijama.*"* i 2. skupina "pedagoških" definicija ima naglasak na obrazovanju, tj. na učenju i poučavanju ("*learning*" dio u imenu). Jedna od takvih definicija je: *"E-learning* je interaktivan ili dvosmjeran proces između nastavnika i učenika uz pomoć elektroničkih medija, pri čemu je naglasak na procesu učenja, dok su mediji samo pomoćno sredstvo koje upotpunjuje taj proces [1].*"*

Među glavnim prednostima *e-učenja* su: a) vremenska i prostorna fleksibilnost; b) interakcija (komunikacija) između polaznika i nastavnika, koja se odvija putem računala (na primjer: e-mail, forumi) te je često neposrednija i intenzivnija nego komunikacija u razredu; c) komunikacija i grupni rad na zajedničkim projektima između polaznika međusobno; d) programi *e-obrazovanja* su dostupni putem Interneta nevezano za fizičku lokaciju na kojoj se polaznici nalaze; e) programe *e-obrazovanja* mogu putem Interneta pohađati polaznici koji zbog svog posla ili drugih obveza ne mogu nazočiti tradicionalnoj nastavi u vrijeme u koje se ona odvija; f) materijali pripremljeni za *e-obrazovanje* relativno se lako obnavljaju, prilagođavaju i ponovo koriste za različite grupe polaznika; g) korištenje *interaktivnih sadržaja za učenje* i različitih *medija* za prezentiranje sadržaja te *dostupnost* sadržaja 24 sata *online*.

B. Učenje na daljinu

Učenje na daljinu (engl. distance learning, distance education) je obrazovanje ili obuka koja se nudi učenicima na različitim mjestima odnosno fizički udaljenima od učitelja ili izvora informacija. U praksi je učenje na daljinu puno složenije od ove definicije jer uključuje korištenje novih tehnologija i novih interaktivnih nastavnih metoda. Termin učenje na daljinu u novije vrijeme se zamjenjuje terminom online učenje, čime se naglašava da je riječ o posebnom obliku e-learninga ili e-obrazovanja.

Tipičan sustav za e-učenje u osnovi je web aplikacija kojoj korisnici pristupaju putem web preglednika na svojim računalima. Neke od najbitnijih prednosti e-učenja su uklanjanje prostornih i vremenskih ograničenja, olakšan pristup sadržajima za učenje te neposredno povezivanje polaznika za e-učenje. Sustavi za e-učenje omogućavaju samostalno učenje, ali i komunikaciju i suradnju, pri čemu omogućavaju i veću kreativnost nastavnika, mobilnost i fleksibilnost nastavnika i polaznika (u smislu vremena i prostora), personalizaciju, veću dostupnost materijala za učenje, nove komunikacijske kanale za suradnju nastavnika i polaznika, nove metode procjene znanja, potporu promjenama, partnerstvima i transformaciji procesa učenja i uštedu vremena, jednostavnu evaluaciju, trenutne povratne informacije i neposredno praćenje napretka pri *online* procjeni znanja.

III. KONCEPT E-UČENJA

Danas je sveprisutna potreba, ali i želja da se unaprijedi kvaliteta i dostupnost različitih obrazovnih programa. Pri tome je obrazovanje postalo *cjeloživotnim procesom* i potrebom, pa se briga o dostupnosti i kvaliteti odnosi na sveukupni ciklus *cjeloživotnog obrazovanja*. Postalo je jasno da se navedeno ne može ostvariti bez odgovarajuće i učinkovite uporabe informacijskih i komunikacijskih tehnologija u obrazovnom procesu. Upravo kvalitetna implementacija tehnologija e-učenja donosi niz prednosti u obrazovni proces i omogućava željeno novo, moderno i kvalitetno obrazovanje.

E-učenje omogućava: učenje prilagođeno individualnim potrebama studenta (bilo kada, bilo gdje), aktivno sudjelovanje i aktivnu ulogu studenata, aktualnije i ažurnije obra-zovne materijale, globalni pristup različitim obrazovnim sadržajima, smanjenje „pritiska“ na fizičke resurse (omo-gućavajući virtualne učionice, virtualne laboratorije), više komunikacije nastavnik-student i student-student, timski rad, projektni rad nastavnika i studenta, bolju iskorištenost vremena (za kvalitetnije sadržaje) pri kontaktima nastav-nik-student, samostalnu provjeru znanja i kontinuirano mjerenje i praćenje napretka studenta [5].

To je učenje i poučavanje koje je usmjereno na studente.

IV. UPRAVLJANJE PROCESOM E-UČENJA

Model upravljanja procesom e-učenja se sastoji od 3 faze: *planiranje, organiziranje i kontroliranje* procesa e-učenja. Proces upravljanja e-učenjem napreduje kroz slijedeće faze: *planiranje* procesa e-učenja (izrada operativnog plana odnosno scenarija e-učenja), *organiziranje* odnosno implementacija procesa e-učenja (uz provedbu scenarija e-učenja u praksi), *kontroliranje* procesa e-uče-nja (tj. vrednovanje različitih aspekata procesa) i *pobolj-šanje procesa i platforme*. Osnovni princip ovog procesa je *ponavljanje* - jednom kada se potvrde ili odbace početne pretpostavke kroz fazu vrednovanja, ciklus se ponavlja uz primjenu poboljšanja [6].

PLANIRANJE je iznimno bitan za cjelokupni uspjeh e-učenja, a najčešće obuhvaća aktivnosti od analize suvremenih trendova u obrazovanju i identificiranje karakteristika polaznika tečajeva za e-učenje; zatim osmišljavanje okoline za učenje koja odgovara sadržaju, ciljanoj publici i tehničkoj platformi te odabira odgovarajuće platforme; pa sve do formuliranja očekivanih ishoda i izrade operativnog plana odnosno scenarija e-učenja. Cilj faze planiranja procesa je osmišljavanje ishoda i scenarija prema određenom kontekstu. Ishodi učenja predstavljaju instrument razvoja plana i programa, te provjere njegove konzistentnosti. Na razini programa preporučeno je definirati 15 do 30 ishoda učenja, na razini predmeta 5 do 10, a na razini nastavne cjeline 5 do 15 ishoda učenja [12].

ORGANIZIRANJE osnovnog sadržaja e-učenja najčešće se radi pomoću LMS sustava koji omogućuju polaznicima da kolegije istražuju nelinearno, a isto tako i nastavnicima omogućuju izradu tečajeva na fleksibilan način, nelinearno. Redoslijed dodavanja materijala i funkcionalnosti u sam kolegij ovisi isključivo o nastavniku. Ako nastavnik i ne zna gdje započeti, npr. nije siguran treba li organizirati kolegij oko tjednog ili tematskog oblika, treba ipak odabrati jedan oblik i početi s dodavanjem materijala. U slučaju da primijeti kako je ipak drugačiji oblik prikladan za danu situaciju, nastavnik, jednostavno može promijeniti format i rasporediti sadržaj drugačije. U tradicionalnom okružju za učenje, polaznici i nastavnici koriste zajedničke prostore poput učionica, hodnika i dvorišta koja potiču društvene veze i neformalnu komunikaciju koje pojačavaju ostvareno pedagoško iskustvo. U obrazovnom okruženju gdje se koriste LMS-ovi nastav se može odvijati i u potpunosti *online*. Može se organizirati procjena i samoprocjena znanja kroz: zadaće, kvizove, lekcije i radionice.

KONTROLIRANJE napretka i ponašanja polaznika, polazi od postavljanja standarda mjerenja, mjerenja performanse, njene usporedbe sa standardima te poduzimanja potrebne korektivne akcije u svrhu poboljšanja cjelokupnog procesa e-učenja. Svrha faze kontroliranja može se najbolje izraziti kao promatranje i procjena cjelokupnog procesa e-učenja i očekivanih rezultata u odnosu na postavljene ciljeve. Aktivnosti kontroliranja napretka i ponašanja polaznika su: korištenje numeričkih skala procjene napretka polaznika, bilježenje te uvoz/izvoz ocjena na kolegiju, prikupljanje statistika o aktivnostima polaznika, otkrivanje plagijarizma u predanim radovima, prikupljanje log datoteka, prikupljanje povratnih informacija od polaznika raznim upitnicima itd.

V. TEHNOLOGIJE E-UČENJA

Postoji više tehnologija obrazovanja i obuke primjenom e-učenja, a prema NATO-voj podjeli, to su:

1. združeno napredno distribuirano učenje - JADL (engl. *Joint Advanced Distributed Learning*) i

napredno distribuirano učenje - ADL (engl. *Advanced Distributed Learning*),
2. učenje pomoću ICT-a odnosno CBT (engl. *Computer-Based Training*),
3. „uronjeno“ učenje (engl. *Immersive learning*),
4. mobilno učenje (engl. *M-learning*),
5. suradničko ili kolaborativno učenje (engl. *Collaborative learning*) i
6. kombinirano ili hibridno učenje (engl. *Blended learning*) [5].

5.1. JADL i ADL

ADL (engl. *Advanced Distributed Learning*) proizvod je *ADL inicijative* uspostavljene 1997. godine u Ministarstvu obrane SAD-a za standardizaciju i modernizaciju upravljanja „obukama“ i „edukacijama“ te načinima poučavanja. Vizija navedene inicijative je osigurati pristup učenju najviše kvalitete koja može biti „skrojena“ tj. prilagođena pojedinačnim potrebama i isplativa u pravo vrijeme i na pravom mjestu. ADL koristi strukturne i suradničke (engl. *collaboration*) metode iz industrije, akademskih zajednica i vlada koji razvijaju standarde, alate i sadržaje učenja.

Zemlje članice NATO saveza i EU mogu koristiti izrađene obrazovne on-line tečajeve za različita područja u obrambenim i vojnim sustavima.

5.2. CBT (engl. *Computer Based Training*)

CBT je obrazovni ili obučni tečaj dostavljen direktno na studentsko računalo putem CD-ROM-a ili nekog drugog uređaja za pohranu. On se razlikuje od ADL-a po tome što se ne isporučuje preko weba ili putem sustava za upravljanje učenjem i ne zahtijeva usklađenost sa SCORM (engl. *Sharable Content Object Reference Model*) standardom. Studenti se ne prijavljuju na sustav (engl. *login*) CBT, a praćenje napretka polaznika može biti praćeno lo-kalno ili po završetku tečaja.

5.3. „URONJENO“ UČENJE (engl. *Immersive learning*)

„Uronjeno“ (engl. *Immersive*) učenje koristi računalno simulirano okruženje koje se događa u stvarnom životu ili hipotetsku situaciju u grafički bogatom i dinamičnom okruženju. Studenti su stavljeni u situaciju i uključeni su u proces učenja kroz interaktivne simulacije i igre temeljene na *game-based* aplikacijama. *Immersive* učenje podupire jednog učenika ili više malih timova da rade zajedno kako bi riješili problem, uvježbali tehnike ili poboljšali svoje vještine. Ova vrsta učenja može biti web orijentirana, preuzeta na samostalna računala ili mobilne uređaje.

5.4. MOBILNO UČENJE (engl. *m-learning*)

Mobilno učenje (engl. *m-learning*) je uporaba prijenosnih računalnih uređaja, kao što su različiti iPad-i i drugi tableti, prijenosna računala, osobni digitalni pomoćnici (PDA) i pametni telefoni povezani s bežičnim mrežama. To omogućuje da se podučavanje i učenje proširi i na prostore izvan tradicionalne učionice te daje nastavnicima i polaznicima povećanu fleksibilnost i nove mogućnosti za interakcije koje su prikladne i dostupne virtualno s bilo kojeg mjesta.

5.5. SURADNIČKO ILI KOLABORATIVNO UČENJE

Suradničko ili kolaborativno učenje (engl. *Collaborative learning*) podupire nastavu kroz interaktivne metode ili online komunikacije koristeći društvene medijske alate kao što su chat sobe, tribine i video alate kao što su webi-nari i web-emitiranja. Ovi alati se mogu implementirani u učionici i online okruženju kako bi podržali tijek nastave. Oni daju mogućnost velikom broju polaznika da rade zaje-dno s nastavnikom-instruktorom. Kolaborativno učenje može biti sinkrono, kada se učenje događa u stvarnom vremenu ili asinkrono učenje, kada se ne odvija u stvarnom vremenu. Kolaborativni alati za učenje uključuju: chat sobu, forume, wiki, webinar-e/webcast-ove.



Slika 2. Suradničko ili kolaborativno učenje

5.6. KOMBINIRANO ILI HIBRIDNO UČENJE (engl. *blended learning*)

Ponekad je e-learning prisutan u klasičnom obrazovanju u školama i na fakultetima, tako da se računalo i Internet koriste za vrijeme nastave ili u izradi domaćih zadaća. U drugim sluča-jevima se obrazovni programi djelomično izvode online, tj. uz pomoć Interneta ili uz pomoć materijala distribuiranih na CD ROM-u, a djelomično na tradicionalan način. Taj se oblik zove «blended learning», «hybrid» ili «mixed mode». U zemljama članicama NATO saveza i EU koriste se LMS-ovi MOODLE i ILIAS.

VI. PRIMJENA I KONCEPTI E-UČENJA U EUROPSKOM VOJNOM ŠKOLSTVU

Vojni se sustavi oslanjaju na *obuku* i *izobrazbu* pojedinaca i postrojbi kako bi izveli iznimno teške zadaće na visokoj razini, u visoko stresnim uvjetima. I obuka i izobrazba su jednako potrebne: *obuka* jer se njom podiže razina znanja i vještina nužna za izvođenje vojnih zadaća, a *izobrazba* jer omogućava vojnom osoblju na svim razinama donošenje odluka kako, kada i gdje uporabiti vještine usvojene obukom. Dodatno, vojskama su jednako važni i učinkovitost programa (količina usvojenog znanja), kao i isplativost (potrebno vrijeme i uloženi resursi) [9].

Vojne organizacije diljem Europe imaju gotovo apsolutnu kontrolu nad obrazovanjem svojih pripadnika od dana kada pristupe u vojsku do umirovljenja. Međutim zbog suvremenih sigurnosnih trendova te raspršenosti pripadnika oružanih snaga diljem svijeta, takav centralizirani sustav izobrazbe postao im je i značajan teret. Naime, vojnici moraju kontinuirano učiti, simulirati,

uvježbavati bilo da su u matičnoj bazi, u tranzitnim zonama ili u području operacije. Mogućnost obučavanja **bilo kada** i **bilo gdje**, postao je ključni preduvjet održavanja vojne spremnosti i učinkovitosti. Stoga su oružane snage uložile u istraživanje, razvoj i implementaciju tehnologija, načela i programa kojima mogu dio odgovornosti za vojno obrazovanje prenijeti na same pripadnike oružanih snaga. Okretanjem tehnologiji, povećana je dostupnost, učinkovitost, ali i isplativost vojne izobrazbe i obuke.

Dva ključna modela koju većina vojnih organizacija primjenjuje danas su učenje na daljinu i simulatori. Danas je najveći dio nastavnih materijala za vojne potrebe mobilan i značajno smanjuje troškove jer se svi nastavni i obučni materijali pohranjuju na webu i postaju dostupni kroz računalne sustave. To omogućava kontinuirano dorađivanje nastavnih materijala, njihovu relativno laku distribuciju preko raznih računalnih platformi pa i preko „pametnih telefona“ (engl. *Smart phones*) do krajnjih korisnika. Ipak ono što je najvažnije je mogućnost prilagodbe sadržaja svakom vojniku - on sam određuje tempo, mjesto i vrijeme učenja, što je iznimno važno u području mirovnih misija i operacija, čime samostalno održava kontinuitet rasta znanja, a bez pritiska skupine ili potrebe da se prati njezin tempo. Vojnici se dodatno usmjeravaju na nove sadržaje koje trebaju usvojiti, a ne na ono što već znaju.

Uvođenjem e-učenja s jedne strane smanjuje se trošak i vrijeme potrebno za uvježbavanje, a s druge povećava spremnost, učinkovitost i sposobnost vojnika da odgovore na suvremene prijetnje. Dodatno, ADL sustavi i strategija njegove uporabe prikladna je i za distribuciju naučenih lekcija s jednog dijela područja operacija u drugo, što zahtijeva stalnu suradnju softverskih programera sa obučavateljima i instruktorima, te usklađivanje funkcionalnih i tehničkih mogućnosti vojnog sustava.

VII. UPORABA E-UČENJA U VOJNOM ŠKOLSTVU EUROPSKE UNIJE

Iako individualne zemlje članice, ranije navedene procese tehnološkog unaprjeđivanja vojne izobrazbe, provode samostalno, Europska unija je kao zajednica otišla i korak dalje. Naime, sustav europskog vojnog obrazovanja prepoznao je 2008. godine potrebu da se europsko vojno školstvo na poseban način sinkronizira, uskladi i usmjeri prema ispunjavanju europske zajedničke sigurnosne politike (engl. *Common Security and Defense Policy - CSDP*). Razlog tome su tri ključna elementa:

1. Promjene u suvremenom sigurnosnom okruženju su iznimno brze, nepredvidive i njima se teško upravlja. Pojava pojedinačnih ili usložnjenih hibridnih prijetnji u jednom kriznom području, gotovo će se sigurno pojaviti i u drugima, te vojno obrazovanje mora imati sposobnost kontinuiranog informiranja postrojbi na terenu kako bi se naučene lekcije i razvijene taktike što prije dostavile na „prvu crtu“;
2. Suvremene, međunarodne operacije prije svega provode male taktičke postrojbe kojima su zapovjednici niži, mlađi časnici čina poručnik do satnik. Sve češće se događa da taktičke postrojbe ostvaruju strateške rezultate. Ovaj element implicira da je potrebno što je moguće ranije započeti s interdisciplinarnim, međunarodnim obučavanjem kadeta i mladih časnika, kako bi već po stupanju na prvu časničku dužnost mogli odgovoriti suvremenim sigurnosnim zahtjevima zapovijedanja na terenu i
3. Sigurnosne politike Europske unije inzistiraju na rješavanju svjetskih kriza na europski način (nasuprot NATO politikama kojima upravljaju sigurnosni i politički interesi SAD-a). Krizno promišljanje u skladu s načelima i principima europske kulture i sigurnosne politike zahtijeva posebnu obojanost i usredotočenost obuke u EU vojnim akademijama.

Oblikovanje vojne izobrazbe s ovim ključnim smjernicama, dovelo je do osnivanja Fakulteta za europsku sigurnost i obranu (engl. *European Security and Defense College – ESDC*) u Brusselesu (osnovano je aktom Vijeća Europe broj 2008/550/CFSO 23.06.2008.), razvoja programa *ERASMUS militaire*, ili EMILYO (engl. *Exchange of Military Young Officers*) i mnogih drugih. Tu je posebna važnost ESDC jer on sa svojom specifičnom strukturom predstavlja „čvornu mrežu“ (engl. *hub network*) svih vojnih akademija u EU, te je kao takav ključna polaznica za distribuciju informacija, diskusija, analiza i iskustava.

U listopadu 2008. godine Instutut za strateške studije EU (engl. *EU Institute for Startegic Studies*) objavio je izvješće u kojem elaborira sve manjkavosti postojećeg neujednačenog europskog vojnog školstva te navodi nužne korake za poboljšanje. Svi uključuju web aplikacije i tehnološki pristup, a među tri najvažnija su:

- Europska obrambena agencija (engl. *European Defense Agency*) mora „razviti virtualni obučni centar koristeći svoje postojeće i buduće studije kao i druge ključne tekstove o europskoj kooperaciji. Ta web stranica se također mora koristiti kao sredstvo dizajniranja tečajeva za obuku nacionalnih voditelja programa (engl. *Progra-mme Managers*);
- Mora se stvoriti baza podataka nacionalnih obučavatelja i tečajeva dostupna svima u EU vojnom školskom sustavu;
- ESDC mora razviti klasičnu ili tradicionalnu nastavu (engl. f2f – *face-to-face*) i online tečajeve o CSDP-u, kooperaciji i kooperativnim programima koji će se provoditi kod njih, ali i u svim drugim nacionalnim vojnim akademijama [10].



Slika 3. Hijerarhija polaznika i aktivnosti u ESDC-u

Iako su diljem Europe postojali razni virtualni obučni centri poput *S-cube, Elia, European patent office* i drugih, u skladu s ovim preporukama, razvijeni su i oni specijalizirani za vojne i sigurnosne teme i trendove i to

pod okriljem ESDC-a. Nastavni materijali, knjige, izvješća, ali i čitavi ADL tečajevi dostupni su kroz sustav LMS-a pod nazivom ILIAS koji za ESDC administrira rumunjsko ministarstvo obrane.

Iz slike 1. vidljivo je kako je cijeli sustav ESDC obuke, od temeljne do najviše razine polaznika i zahtjevnosti materijala, praćen IDL (engl. *Internet based Distance Learning*) sadržajima, tečajevima i programima [11]. Uz to, razvijeno je nekoliko zajedničkih, akreditiranih i verificiranih tečajeva iz raznih područja CSDP-a koji se provode za potrebe svih zemalja članica EU-a, a jasno su i u cijelosti usredotočeni na poboljšavanje prepoznatih kritičkih vrijednosti EU-a. Ti, takozvani „zajednički moduli" (engl. *Common Modules),* su: (1) *Common module on CSDP* (sadržaj obuhvaća sve teme nužne za upoznavanje europske sigurnosne politike)*;* (2) *Common module on media training* (razvijen zbog važnosti poznavanja načina rada s medijima u području operacija); (3) *Common module on Crisis Management and Peace Support Operations* (provodi se u četiri pod-modula s ciljem obučavanja kadeta i mladih časnika svim fazama stožernog i terenskog rada u mirovnim misijama); (4) *Common module on Law of Armed Conflict* (poznava-nje ratnog prava krucijalni je preduvjet legitimnosti rada u mirovnim misijama i operacijama); (5) *Common module on Maritime Leadership* i *Common module on Maritime Navigation* (tečajevi za kadete i mlade časnike pomorske struke kako bi mogli učinkovito djelovati u multinacionalnim pomorskim operacijama poput anti-terorističke operacije *Active Endeavour* na Mediteranu ili protu-piratske operacije *Atalanta* u Somaliji) i (6) *Common module on Basic Military English* razvijen je s ciljem savladavanja i uvježbavanje engleske vojne termi-nologije.

Svaki od njih obuhvaća dva krucijalna dijela: IDL fazu i f2f fazu. U skladu sa smjernicama o potrebi lake dostupnosti informacija s jedne strane, te ograničenja troškova i vremena provedbe obuke s druge, svaki tečaj se prvo mora završiti on-line preko IDL sustava, a tek nakon toga se kadeti i mladi časnici mogu uputiti na ciljane akademije u trajanju do tjedan dana. Ova kombinacija završenih složenijeg dijela tečaja on-line (otvara se u pravilu tri mjeseca prije planiranog početka obuke) uz finalizaciju i uvezivanje svih naučenih sadržaja kroz jednotjednu f2f fazu, dovodi do mogućnosti vrednovanja ovih programa kroz ECTS sustav s 3-5 bodova.

Kroz IDL platformu ESDC-a moguće je pristupiti i drugim relevantnim vojnim institucijama i znanstvenim skupina-ma, pregledati postojeće priručnike, izvješća, pronaći postojeće pravilnike i standarde usvojene na vojnom odboru EU i sl. [15]. Također je dostupan i forum na kojem polaznici mogu razmjenjivati informacije i komentare o tehničkim ili sadržajnim aspektima platforme.

Upravo iz tih komentara i interakcije s korisnicima, uvedene su promijene 2012. godine jer je prijašnji software omogućavao ostanak na mreži samo do završetka upisanog obučnog modula. Međutim nakon prigovora polaznika da su im informacije dostupne na mreži potrebne i korisne i u kasnijem radu, a ne samo tijekom on-line tečaja, sustav je prilagođen te danas omogućava pristup svim zainteresiranim, registriranim korisnicima u matičnim akademijama, ali i u području operacija.

Danas gotovo sve vojne akademije u EU, uključujući i Hrvatsko vojno učilište „Petar Zrinski" imaju svoje IDL platforme na kojima pohranjuju i distribuiraju obučne materijale. Većina ih inzistira na završavaju određenog broja ADL tečajeva prije dolaska na neku od njihovih obuka ili seminara. Također, vojna obučna središta poput Obučnog središta za međunarodne vojne operacije i mirovne misije OS RH, koji imaju veliki broj stranih polaznika i instruktora na svojim obukama, imaju preko svojih web stranica pristup ILIAS-u te je obveza svih polaznika prije dolaska završiti određeni broj ADL programa kako bi se ujednačila inicijalna znanja većine polaznika i omogućio maksimalni učinak f2f faze obuke.

7.1. PRIMJER PRIMJENE E-UČENJA U VOJNOM OBRAZOVANJU NORDIJSKIH ZEMALJA

Nordijske zemlje Europe (Finska, Švedska, Norveška i Danska) otišle su najdalje u uporabi IT tehnologija u vojnom školstvu. Razvijenost, dostupnost i iskoristivost interneta dosegla je vrhunac 2010. godine kada je u Finskoj pravo na pristup internetu uvršteno u popis ljudskih prava [14].

2009. godine nordijske su zemlje razvile projekt „*Learning online – LONAF*" utemeljen na prepoznatim, zajedničkim izazovima u implementaciji *on-line* učenja u vojnim organizacijama. Iako *on-line* učenje i društveni mediji privlače značajnu pažnju, iznimno se slabo izučila mogućnost uporabe ovih sustava za potrebe vojne obuke i izobrazbe. Stoga su stručnjaci iz svih vojnih akademija NORDCAP-sa zajedno razvili program „*Learning space on Facebook*" na kojem su postavljeni edukacijski tekstualni i video materijali koji objašnjavaju kako učinkovito koristiti mehanizme e-učenja, s ciljem pružanja dobrih primjera praktične uporabe *on-line* alata za učenje kako bi dodatno ohrabrili daljnju uporabu ovih sredstva u vojnoj obuci.

Kako bi osigurali kontinuirani razvoj tehnologija i za e-učenje, Norveška posljednjih 8 godina, za potrebe NORDCAP-sa organizira ADL konferenciju s temom *„Kako ADL i nove tehnologije podržavaju vojno obrazovanje i obuku?"* [13] Raspon tema na ovim konferencijama ide od tehničkih postavki i načina razvoja ADL tečajeva primjerice uporabom interaktivnih PDF aplikacija (*Adobe Captivate and Adobe Connect)*, ili načina izrade modula e-učenja za iPAD i Android aplikacije, pa do predstavljanja civilne organizacija i kompanija koje vojnom obrazovanju mogu ponuditi neka od već razvijenih rješenja ili otvoriti prostor za zajednički rad iz područja ADL tehnologija poput: kognitivnog izučavanja sustava za učenje s ciljem povećavanja učinka učenja, utjecaja nastavnih tehnologija u specifičnim okruženjima kako bi se odredile specifičnosti e-učenja u raznim okruženjima (od učionice do terena u Afganistanu), razvoja elektronskog sustava mentorstva utemeljen na prirodnim jezičnim načelima gdje sustav koristi pitanja, napomene i sažetke kako bi „vodio" polaznika do konačnog zaključka kao što bi to radio i pravi nastavnik, razvoja i izrade video materijala za potrebe obuke i sl.

Također se na konferencijama prezentiraju uspješni primjeri primjene ADL-a poput onih na *Nordic Defense College* gdje se prvi modul Zapovjedno-stožerne škole u cijelosti završava kroz ADL ili ADL rješenja u vojnim školama diljem Europe poput Njemačke (sustav *SuTBw*), Belgije, Srbije i dr. Dodatni primjer je bio rješavanje problema obuke pilota na najsuvremenijim nadzvučnim

zrakoplovima (poput F-35, Grippena, Eurofightera i sl.), s obzirom da je ta obuka dugotrajna, skupa i sigurnosno klasificirana kao tajna, te se ne mogu koristiti vanjski suradnici i stručnjaci za izradu simulatora već je bilo nužno razviti vlastite IT sustave, nastavne module i simulatore.

Na njima se mogu prezentirati i uspješni vojni projekti. 2012. godine je tako predstavljen *MoLE (*engl. *The Mobile Learning Environment)* kojeg su razvili *Coalition Warfare Program NATO-a* usmjeren na rješavanje problema ograničenja koje donosi mala propusnost, ograničen pristup internetu i ograničena infrastruktura, pružajući rješenja kroz globalnu mobilnu umreženost, mobilne, satelitske tehnologije i sl. U projektu MoLE sudjeluju 24 države, od kojih je 13 iz EU. 2013. godine je predstavljen NORDCAP-sov projekt CAMO *(Cultural Awareness in Military Operations)* koji je odgovor na EU smjernicu u potrebi kulturološke senzibilizacije i obuke mladih časnika za rad u međunarodnom okruženju. Projektom je izrađen *game-based simulation*, 3D virtualni svijet u kojem se provode međunarodne operacije, a koji omogućava stavljanje mladog časnika u razne situacije i testiranje njegovih reakcija u njemu.

VIII. POGLED U BUDUĆNOST

HVU je u procesu transformacije u visoko obrazovnu i znanstveno-istraživačku vojnu ustanovu odnosno instituciju tj. sastavnicu sveučilišne zajednicekoji od ove jeseni provodi dva nova preddiplomska sveučilišna studijska programa za potrebe OS RH. Vizija je da se uvođenjem e-učenja želi u procesu obrazovanja ostvariti aktivan partnerski odnos nastavnika i polaznika, prilagođen osobinama i potrebama svakog polaznika, s jasno postavljenim obrazovnim ciljevima i jasno dodijeljenim odgovornostima za postizanje tih ciljeva. Misija je sustavnim uvođenjem e-učenja doprinositi kvaliteti sadašnjeg i budućeg sveučilišnog obrazovanja kroz stvaranje okruženja u kojem se potiče i omogućava aktivna suradnja nastavnika i polaznika. Uvođenjem i aktivnom primjenom e-učenja HVU namje-rava ostvariti četiri strateška cilja:

- unaprijediti kvalitetu obrazovanja,
- omogućiti nastavnicima i studentima ostvarivanje novih uloga u obrazovnom procesu,
- povećati konkurentnost i dostupnost učilišta i učilišnih programa i
- osposobiti studente za uporabu tehnologija cjeloživotnog učenja.

I u europskom vojnom školstvu ponuđen je određeni broj tečajeva i zajedničkih modula u području e-učenja. Svaki od njih obuhvaća IDL i f2f fazu. Sadašnje sposobnosti su na zavidnoj razini, a njihovim daljnjim razvojem, koje su precizirane Smjernicama za daljnje unapređenje sustava e-učenja osnažit će se novim sposobnostima, a polaznici će sistematizirati postojeća i usvojiti nova znanja, vještine i sposobnosti kojima će biti osposobljeni za još učinkovitiju njihovu primjenu u svakodnevnim zadaćama.

X. ZAKLJUČAK

U informacijskom društvu i vremenu *e-učenje* kao i *distribuirano* odnosno *učenje na daljinu* dobivaju sve više na značenju poglavito u europskom vojnom školstvu. Može se zaključiti kako je i u europskom vojnom školstvu napravljen značajan pomak u razvoju i uporabi sustava e-učenja, ali zbog važnosti stalnog izlaganja međunarodnom okruženju tijekom obuke i izobrazbe, većina ih ipak uvezuje ADL i f2f module u vojnom školovanju svojih pripadnika. Razvijen je i ponuđen veliki broj „gotovih" *online* tečajeva za potrebe obrambenog i vojnog sustava. Stvoreni su potrebni preduvjeti za implementaciju *e-učenja* i u obrambenom i vojnom sustavu RH.

„Znanjem u budućnost" moto je sadašnjih naraštaja i ostat će moto i naraštajima koji će doći na vojno obrazovanje u škole Hrvatskog vojnog učilišta „Petar Zrinski" u Zagrebu.

LITERATURA

[1] A. W. (Tony) Bates, „Upravljanje tehnološkim promjenama - Strategije za voditelje visokih učilišta", *CARNet* - Zagreb i *Benja* - Lokve, 2004. (prijevod), Izvornik: „Managing Techological Change: Strategies for College and University Leaders", *Jossey-Bass Publishers*, San Francisco, USA, 2000.

[2] B. Shackelford, „Project Managing E-Learning", *American Society for Training & Development Press*, Alexandria, USA, 2002.

[3] M. M. Lynch & J. Roecker, „Project Managing E-Learning", A handbook for successful design, delivery and management, *MPG Books Ltd*, Bodmin, Great Britain, 2007.

[4] J. Greene & A. Stellman,"Head First PMP" A Learner's Companion to Passing the Project Management Professional Exam, A Brain-Friendly Guide, *Oreilly Media, Inc*., CA, USA, 2007.

[5]Grupa autora, „E-Lerning Concept", NATO unclassified, *NATO, Bruxelles, Belgium*, 8.2.2012.

[6] M. Ćukušić & M. Jadrić, „E-učenje: koncept i primjena", *Školska knjiga d.d.*, Zagreb, 2012.

[7] M. Jadrić, M. Ćukušić & M. Lenkić, „E-učenje: Moodle u praksi", *Ekonomski fakultet Sveučilišta u Splitu*, Split, 2012.

[8] P. Jandrić & D. Boras, „Kritičko e-obrazovanje Borba za moć i značenje u umreženom društvu", *Tehničko veleučilišta u Zagrebu i Filozofski fakultet Sveučilišta u Zagrebu*, Zagreb, 2012.

[9] J.D. Fletcher, „Education and Training Technology in the Military", *Science magazine*, vol. 323, 2 January 2009., str. 72

[10] D. Keohane, S. deVaucorbeil, „Education and Training for EU defense equipment programmes", *EU ISS report* no. 2, Brusseles, October 2008, str. 5

[11] Ed. J. Rehl, H-B. Weisserth, „CSDP Handbook" – The Common security and defense policy of the EU, *Armed Forces of Austria*, Vienna, 2010., str. 84.

[12] B. Divjak, „Ishodi učenja u visokom školstvu", TIVA, FOI, Varaždin, 2008.

Internet adrese:

[13] http://www.adlnet.gov/partnerships/norway-adl-partnership-lab/

[14] http://www.vecernji.hr/techno/finska-prva-u-svijetu-brzi-internet-proglasila-ljudskim-pravom-161997

[15]https://esdc.adlunap.ro/repository.php?reloadpublic=1&cmd=frameset&ref_id=1

Model ubrzanog usvajanja znanja putem moodle Akademije

N. Bačun Lebarović*, V. Šebek ** i D. Kučak***
* Visoko učilište Algebra – visoka škola za primijenjeno računarstvo
Ilica 242, Zagreb, Hrvatska
e-mail: nives.bacun@racunarstvo.hr
** Visoko učilište Algebra – visoka škola za primijenjeno računarstvo
Ilica 242, Zagreb, Hrvatska
e-mail: sebek.vanja@gmail.com
*** * Visoko učilište Algebra – visoka škola za primijenjeno računarstvo
Ilica 242, Zagreb, Hrvatska
e-mail: danijel.kucak@racunarstvo.hr

Sažetak - Pojava Interneta i e- učenja promijenila je zahtjeve i pristupe edukaciji unutar samih tvrtki i edukaciji korisnika, te tako značajno utjecala na razvijanje novih pristupa, alata, modela i metoda učenja. Ciljevi ovog članka su opisati edukaciju unutar privatne tvrtke Carpio d.o.o koja se bavi prodajom ERP vlastitog softvera.

Ovim člankom će se opisati potrebe privatne tvrtke Carpio koji su doveli do uvođenja e- učenja. Zatim će se opisati Carpio Web Akademiju kao jedinstveno i višedimenzionalno internet mjesto edukacije u kojoj je tvrtka objedinila sve materijale potrebne za edukaciju svojih klijenata. Također, opisati će se model ubrzanog usvajanja znanja kao način na koji se provodi edukacija na Carpio web Akademiji te načini na koji tvrtka koristi e-učenje u svom poslovanju.

Time se želi dati uvid u jedan novi način primjene e-učenja u poslovnom svijetu koji ne zahtijeva izradu novih alata nego uspješnu primjenu postojećih na nov i inovativan način.

I. UVOD

Nagli razvoj Interneta i ekonomskog tržišta ostavilo je trag na brojne tvrtke i njihovo poslovanje jer su uvidjele da samo stalnom edukacijom i školovanjem zaposlenika mogu konkurirati na globalnom tržištu.

Čovjek koji je sposoban samostalno obavljati svoj posao u što kraćem roku cilj je svake tvrtke. Velika količina dostupnih informacija i brzina njihove razmjene dovela je do problema kako najjednostavnije i najbrže educirati korisnika. Potreba da sve informacije koje mogu biti korisne uvijek budu dostupne na jednom mjestu utjecale su na tvrtke koje educiraju da ih postave u edukacije.

Tvrtka Carpio je prepoznala da klasični načini educiranja u smislu seminarskih predavanja koje je provodila do tada, teško mogu zadovoljiti sve ove potrebe. Okrenuvši se novim tehnologijama i okruženjima kao što je Internet i mogućnostima koje on pruža osmislila je nov pristup tehnologijama e – učenja, Carpio Web Akademiju na kojoj je provodila edukaciju svojih korisnika.

Tvrtka Carpio se bavila prodajem vlastitog ERP[1] softvera posebno prilagođenom za praćenje poslovanja građevinskih tvrtki. Glavni autor članka aktivno je sudjelovao u razvoju projekta Carpio Web Akademije još 2007.g. Projekt je započeo dopunom edukacijske dokumentacije novim oblicima resursa. Standardni priručnici prevedeni su na oblik prihvatljiv za e – učenje, a video animacije i animacije rađene sa Camtasia Studiom[2] pratile su proizvode tvrtke zaokružile su ponudu resursa. Tvrtka je prešla na edukaciju putem Internet Akademije 2010.g i koristila ju je za za educiranje i certifikaciju svojih klijenata, partnera i vlastitih zaposlenika sve do 2014.g..

II. CARPIO WEB AKADEMIJA

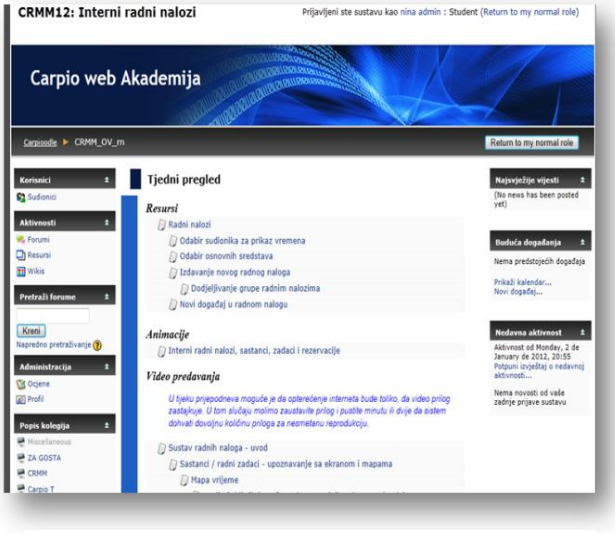
Carpio web Akademija je internet mjesto napravljenom sa Moodle 1.9x[3] alatom za učenje te ima sve karakteristike Moodle internet stranice. Akademiju ćemo opisati na dvije razine.

Prvo ćemo opisati njezino sučelje, odnosno vanjski izgled onakav kakav je vidljiv polaznicima gdje ćemo

[1] skraćeno od Enterprise Resource Planning

[2] Camtasia Studio – alat za snimanje aktivnosti na ekranu, dostupno na http://www.techsmith.com/

[3] Moodle (skraćenica za Modular Object-Oriented Dynamic Learning Environment) je besplatan softver za e-učenje u obliku platforme, također poznat kao sustav za upravljanje učenjem, ili Virtual Learning Environment (VLE), dostupan na https://moodle.org/ .



Slika 1. Izgled kolegija Carpio web Akademije

opisati izgled početne stranice te kolegija. Nakon toga ćemo opisati unutrašnju strukturu Akademije koja nije dostupna polaznicima a u kojem ćemo obuhvatiti pitanja, resurse i administriranje. korisnika.

A. *Izgled sučelja Carpio web Akademije*

Početna stranica Akademije organizirana četiri grupe kolegija. Jedna od njih je otvorena za pregled neregistriranim korisnicima, a ostale tri prate module softvera tvrtke. Na početnoj stranici je vidljivi su kolegiji koji vide neregistrirani korisnici te nazivi kolegij čiji je pregled omogućen samo registriranim korisnicima.

Izgled kolegija registriranih korisnika Akademije prikazan je na Slici 1. Kolegij je vizualno podijeljen u tri stupca. Kolegiji u centralnom dijelu se sastoje blokova resursa, animacija, video animacija, aktivnosti te drugih blokova kojima se nećemo baviti u ovom članku.

Resursi su uglavnom dijelovi priručnika vezani za temu kolegija objavljeni u htm[4] formatu, animacije su animirano objašnjenje određenih postupaka ili problema vezanih za temu kolegija.

Video animacije su, za razliku od animacija, dijelovi seminarskih predavanja koje je tvrtka snimila te prilagodila potrebu e – učenja. Video animacije su različite dužine trajanja ovisno o temi koju obuhvaćaju i sadrže strukturu koja pomaže korisniku lakše snalaženje.

Aktivnosti su kolaboracijski dio kolegija koji se sastoji od dva foruma i wiki[5] kolegija. Dio Certifikacija je vezan za certificiranje korisnika, a za određeni kolegij certifikacija se provodi putem testova koji su dostupni u određeno vrijeme i zaštićeni lozinkom pristupa.

Na Slici 1. također je vidljivo da se u lijevom gornjem uglu nalazi ime kolegija. Sa lijeve strane nalaze blokovi vezani za korisnike, aktivnosti, ocjene te kolegiji na koje je korisnik upisan pod nazivom „Popis kolegija. U desnom gornjem uglu vidljivo je ime polaznika koji se nalazi na akademiji. Blokovi vijesti, budućih događanja i nedavnih aktivnosti nalaze se sa desne strane.

B. *Unutrašnja struktura Carpio web Akademije*

Kao pravi Moodle sustav i Akademija ima određenu unutrašnju građu koja omogućuje administratorima da upravljaju s njom i njenim kolegijima. Akademija se također sastoji od:

- Banke pitanja (na različitim nivoima)
- Baze resursa
- Korisnika
- Ostale administracije vezane za pravilan rad Akademije (administracija web servera i PHPa te same baze podataka)

Pri kreiranju baze pitanja tvrtka je koristila samo tri tipa pitanja (Točno / Netočno (eng. False /True), Višestruki odabir (eng. Multiple choice), Spajanje parova (eng. Matching)). Banka pitanja svakog od kolegija Akademije je organizirana tako da je svakom pitanju određena njegova težina neovisno o tipu pitanja, nego ovisno njegovom sadržaju i zahtijevanom odgovoru.

Testovi na Akademiji sadrže slučajno odabrana pitanja iz banke pitanja odabranog kolegija. Test sadrži 20 pitanja, a njegova težina je regulirana omjerom količine jednostavnih pitanja u odnosu na one složene. Testovi na Carpio web Akademiji imaju sva pitanja na istoj stranici, te su zaštićeni pristupnim lozinkama tako da određenom testu može pristupiti samo onaj polaznik koji ima pristupnu lozinku testa. Kriterije ocjenjivanja testa određuje voditelj edukacije.

Sadržajno na Carpio web Akademiji su korištena tri tipa resursa. To su:

- Priručnici u htm formatu (standardni priručnici tvrtke koji su prilagođeni sadržaju svakog od kolegija).
- Animacije (animirani postupci različitih redoslijeda upisa podataka u sustav ili naglasci rješenja na različite sadržaje ili problem.).
- Video predavanja (snimljeni seminari koje je tvrtka održala. Sadržajno i strukturno su prilagođeni sadržaju svakog kolegija).

Akademija ima definirano ukupno15 uloga od kojih su glavne: administrator (ima prava nad kompletnim sustavom), nastavnik (ima pravo upravljati kolegijem ocjenjivati polaznike i mijenjati aktivnosti), nastavnik mentor (ima pravo dodavati ili mijenjati predavanja i ocjenjivati polaznike kolegija), kreator kolegija (ima pravo kreiranja kolegija, ali ne i interakcije sa polaznicima), student (ima prava pregledavanja i sudjelovanja u kolegiju u koji je upisan), uloge sa pravima studenta (uloge koje su dodijeljene tvrtkama korisnicima edukacije na Akademiji, a koje nećemo navesti radi zaštite njihove privatnosti), gost (imaju prava pregledavati samo

[4]Datoteka Htm formata je Hypertext Markup Language (html) datoteka.

[5]Wiki je kolekcija više Internet dokumenata nastalih suradnjom različitih autora.



Slika 3. Struktura carpio web Akademije

one javne dokumente i sadržaja) i posjetitelj (prava gosta sa dozvolom pristupa na forume, testove na onim kolegijima koji su javni).

C. Struktura Carpio web Akademije

Struktura Akademije prikazana je na Slici 3. Vidljivo je da svi korisnici, koji to imaju pravo, pristupaju cijelom sustavu preko jedinstvenog sučelja, Carpio Web Akademije i na taj način koriste akademiju. Pristup sustav upravljanja bazom podataka i upravljanju modelima omogućen preko bloka Administracija na sučelju Akademije, koji je vidljiv samo administratorima. Pristup ustavu upravljanja dokumentima je kroz kolegije i glavnu stranicu Akademije ovisno o tome gdje se dokument želi koristiti.

Blok sustav upravljanja dokumentima sadrži različite vrste objave testova. Testovi mogu biti dostupni svima, grupi vezanoj za kategoriju, potkategoriju, kolegij ili posebno namijenjenoj grupi ovisno o svrsi testa. Određeni test objavljen je na 24 sata te je na taj način dostupan korisnicima u različitim vremenskim zonama. Moguće je otvaranje foruma raznih namjena, ćaskanje (eng. chat) te zadavanje raznih zadaća i sličnih zadataka.

Sustav upravljanja bazom modela se svodi na administriranje kolegija, potkategorija odnosno samih kategorija brisanje ili dodavanjem novih kolegija, potkategorija ili samih kategorija, sigurnosnih kopija (eng. backup) svih podataka, dodjeljivanje raznih prava korisnicima, njihovo organiziranje u grupe raznih svrha, kreiranje profila i drugo.

Sustav upravljanja bazom podataka sadrži mehanizme za upravljanje i kreiranje same baze pitanja. Obuhvaća kreiranje tri različite vrste pitanja, administraciju dokumenata, slika, izvještaja, pisanih dokumenata te mogućnost kreiranja analiza testova te samu fizičku sigurnosnu kopiju baze pitanja i resursa.

Ovaj blok još sadrži i mehanizme za dodavanje ili brisanje novih korisnika, njihovih podataka, te administraciju grupa, same MySQL baze[6] i HTTP poslužitelja[7].

Moduli softvera organizirani su u kategorije sa potkategorijama unutar kojih se nalaze kolegiji. Unutar svakog od kolegija nalaze se testovi, resursi (materijali), forumi i ostali elementi baze dokumenata. S druge strane baza pitanja, statistike i rezultati testova i zadaća te razne analize kao elementi baze podataka zaokružuju svaku od potkategorija.

Tijekom rada sustava modeli mogu sami dohvatiti potrebne podatke iz baza podataka ili baze dokumenata, a nakon izvođenja rezultati pohranjuju u bazu podataka ili dokumenata.

Bazu podataka čine baza korisnika sa svim njihovim upisanim podacima i uloge koje su im dodijeljene, baza uloga sa definiranim pravima pristupa za svaku od uloga, baza pitanja, web stranice, htm dokumenti, animacije i sl. Baza dokumenata su priručnici, forumi, testovi na svim nivoima Akademije, mail korespondencija, ćaskanje, između svih korisnika, zadaće i wiki za sve kolegije kao novi oblik dokumenata koje korisnik sam može dopunjavati.

Kategorije se sastoje od nekoliko potkategorija. Potkategorije se sastoje od nekoliko kolegija, a svaki kolegij se sastoji od svoje baze podataka, baze pitanja, baze resursa i baze osobnih statistika s jedne strane.

S druge strane, postoji mogućnost administriranja aktivnosti, resursa te osobnih podataka ,uloga i grupa te izrade raznih statistika za svaku aktivnost, resurs te za samog korisnika.

Kolegij se sastoji od korisnika koji mogu imati razne uloge od učitelja raznih nivoa, administratora do studenata, gosta ili nekih korisnički definiranih uloga. Mogu imati pristup raznim resursima, wikijima,

[6]MySQL je besplatan, open source sustav za upravljanje bazom podataka.

[7]HTTP poslužitelj je web poslužitelj za različite platforme.

forumima, ispitima, zadacima, svojim osobnim podacima, statistikama ili podacima drugih korisnika, raznim statistikama ispita, pitanja, resursa ili zadaća ovisno o dodijeljenoj ulozi.

III. MODEL UBRZANOG USVAJANJA ZNANJA

Model ubrzanog usvajanja znanja sastoji se od dvije razine koje su međusobno isprepletene:

- *Metode ubrzanog usvajanja znanja* (metode pomoću koje se svakom polazniku edukacije dodjeljuje uloga edukacije na Carpio web Akademiji).
- *Trokuta komunikacije* (trokut odnosa glavnih elemenata nekog procesa i komunikacije između tih elemenata).

A. *Metoda ubrzanog usvajanja znanja*

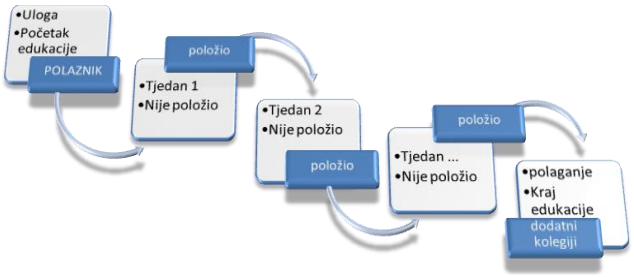
Metoda ubrzanog usvajanja znanja sastoji se u tome da se svakom polazniku Carpio web Akademije prije početka edukacije dodijeli uloga po kojoj će se ta edukacija odvijati.

Dodjeljivanje uloge edukacije ovisi o zadacima polaznika, obujmu njegovog posla te odgovornosti koju ima unutar organizacije u kojoj radi. Sastoji se određenom redoslijedu kolegija kroz koje polaznik edukacije prolazi.

Nakon što mu je dodijeljena uloga polaznik počinje sa edukacijom. Napredovanje sa kolegija na kolegij unutar uloge uvjetovan je prolaskom na testovima i aktivnosti kolegiju.

Jedan kolegij traje 7 dana. Polaznik nakon što položi kolegij na kojem se trenutno nalazi odmah kreće na sljedeći kolegij koji mu je dodijeljen ulogom. Nakon što završi sve kolegije predviđene ulogom polaznik ima mogućnost polaganja dodatnih specifičnih kolegija za posao koji obavlja ili mjesto u tvrtki.

U slučaju da polaznik ne zadovolji uvjete i ne položi kolegij on ostaje na kolegiju kojeg nije položio sve dok ne zadovolji uvjete za prolazak na sljedeći kolegij.



Slika 4. Diagram metode ubrzanog usvajanja znanja

Polaganje testa

Prvi od uvjeta za prolazak kolegija. Prolaz na testovima obuhvaća izlazak i zadovoljavanje minimalnih uvjeta za prolazak testa. Po kolegiju se polažu 2 testa jednake dubine zahtijevanog znanja i opsega znanja. Testovi se rade slučajnim odabirom pitanja iz banke pitanja kolegija, pa edukator nema utjecaja na to koje će se pitanje pojavit na kojem testu, osim u slučaju da tvrtka koja se educira ne traži suprotno.

Postoji nekoliko težina testova ovisno o dubini znanja koje je tražena. Jednom izabrana težina testa uloge ona se ne mijenja do završetka edukacije.

Pregledavanje ponuđenih materijala na Akademiji

Drugi uvjet za prolazak kolegija, Pregledavanje ponuđenih materijala na Akademiji pretpostavlja da polaznik provede neko vrijeme na kolegiju pregledavajući ponuđene materijale. To vrijeme se mjeri i automatski zbraja pregledom logova svakog polaznika kolegija posebno.

Poslovne funkcije kao uloge

Poslovna funkcija je skup povezanih poslova kojima se najsvrsihodnije obavlja određeni zadatak tvrtke sa ciljem stvaranja proizvoda odnosno pružanja usluge. Funkcijski dio organizacijskih struktura u svojoj osnovici ima grupiranje istorodnih, sličnih, neposredno zavisnih ili međusobno uvjetovanih poslova u tvrtki. Poslovne funkcije su predstavljene ulogama i grupirane po logici koja prati potrebna znanja za obavljanje poslova u tvrtki.

Imamo tri skupine uloga:

- RAC (računovodstvo)
- CT (građevinski dio)
- CRM

Duljina edukacijske uloge ovisi o njenoj složenosti i traje od 4 do 10 tjedana. Početni kolegiji su općeniti i kreću se prema sve složenijima i sve više specijaliziranima za određenu poslovnu funkciju. Neki kolegiji su zajednički svim ulogama, a neki su vrlo usko specijalizirani i jedinstveni za određenu ulogu.

B. *Funkcija trokuta komunikacije modela ubrzanog usvajanja znanja*

Trokut odnosa je trokut glavnih elemenata nekog procesa. Kako je komunikacija proces razmjene informacija na različite načine zaključujemo da je trokut komunikacije tada trokut u kojem je moguće na jednostavan način opisati odnose glavnih elemenata komunikacije nekog procesa.

U ovom članku govorimo o procesu razmjene znanja unutar modela ubrzanog usvajanja znanja. Razmjena znanja je složen proces čiji je važan dio komunikacija, no ima glavne elemente koji se tradicionalno mogu

predstaviti trokutom odnosa. Mjesto odvijanja razmjene znanja je Internet Akademija.

Model ubrzanog usvajanja znanja imati će dva glavna trokuta ili razine na kojoj će se odvijati. Prvi je vezan za komunikaciju a drugi za razmjenu znanja ili informacija između glavnih sudionika procesa.

1) Razina komunikacije

Razina komunikacije može se predstaviti trokutom komunikacije koji pokazuje ključne elemente tog procesa, znanje, izvještaje i unutrašnju komunikaciju.

Trokut komunikacije je u slučaju metode ubrzanog usvajanja znanja pomaknut. Temeljni dio je znanje, literaturu u klasičnom trokutu komunikacije predstavljaju izvješća, a unutarnja komunikacija predstavlja jezik kao treći kut tog trokuta.

Polaznik na Akademiji stječe znanje na osnovu kojeg rade se izvještaji. Izvještaji dovode do unutrašnje komunikacije u tvrtki čiji se zaposlenici educiraju. Iz trokuta razmjene prikazanog na slici 5. vidljivo je da unutrašnja komunikacija utječe na stjecanje i vrstu znanja. Svi elementi ovog trokuta utječu jedan na drugoga te na taj način poboljšavaju komunikaciju tijekom edukacije.

2) Razina razmjene znanja

Razmjena znanja kod metode ubrzanog usvajanja znanja je složeni proces čiji se glavni elementi mogu predstaviti trokutom.

U centru trokuta je Akademija kao mjesto na kojem se odvija razmjena znanja. Edukator odnosno tvrtka koja educira predstavlja izvor znanja, na odredištu su polaznici Akademije odnosno zaposlenici tvrtke koje se educira. Na vrhu trokuta je uprava odnosno rukovodstvo tvrtke koja se educira.

Akademija je mjesto na kojem se nalaze resursi koji predstavljaju znanje kako koristiti softver. Njih je tamo postavio edukator. Polaznici Akademije se koriste tim resursima i na taj način predstavljaju jednu stranicu trokuta Akademije.



Slika 5. Trokut komunikacije Modela ubrzanog usvajanja znanja

Razmjena znanja se događa u momentu kada se polaznici uče pomoću resursa kolegija Akademije. Dubina i količina znanja potrebna za neku poslovnu funkciju određuje uprava odnosno rukovodstvo. Povratnu informaciju predstavljaju uspjeh na testovima i vrijeme provedeno na kolegiju.

3) Trokut komunikacije

Trokut komunikacije kod modela ubrzanog usvajanja znanja ima elemente trokuta razmjene znanja te ih spaja sa procesom komunikacije kao što se i vidi na Slici 5. Vrhovi trokuta su:

- Edukator,
- Polaznik odnosno zaposlenik,
- Uprava odnosno rukovodstvo.

U centru modela je Akademija pa se ona nalazi i u centru trokuta modela ubrzanog usvajanja znanja. Jednu stranicu tog trokuta čine edukator i polaznik. Edukator putem resursa Akademije polaznicima predaje znanje, dok polaznici edukatoru daju povratnu informaciju o usvojenom znanju putem aktivnosti. Također polaznici mogu edukatoru ukazati na probleme u edukaciji i na taj način mu omogućiti da ju poboljša.

Druga stranica trokuta opisuje odnos edukatora i uprave odnosno rukovodstva tvrtke. Taj odnos se zasniva na raznim izvješćima koje edukator daje upravi o usvojenom znanju i raznim aktivnostima zaposlenika na Akademiji i o tijeku edukacije. Na osnovi tih izvješća uprava edukatoru daje povratnu informaciju o slijedećim koracima i načinu nastavka edukacije.

Treća stranica trokuta je ona između zaposlenika i uprave tvrtke koja se educira. Taj odnos odlikuje unutrašnja komunikacija tvrtke koja se odnosi na Akademiju i e – učenje.

4) Praćenje napretka edukacije

Način praćenja napretka je vrlo važan faktor u edukaciji te određuje njene daljnje korake i tok ovisno o željama tvrtke čiji se zaposlenici educiraju. Njen napredak prati se na dva nivoa:

- *Praćenje edukacije i ponašanja polaznika*
- *Izvještavanje o napretku i tijeku edukacije*

Ponašanje korisnika na Akademiji se analizira prateći količinu vremena i mjesto na kojem je polaznik bio na kolegiju. To se radi pomoću korisničkih logova, dok se tijek edukacije prati uspjehom na ispitima kolegija te na osnovi komunikacije koju je ostvario sa edukatorom.

Izvještavanje o napretku edukacije polaznika daje se pomoću pisanih izvješća koja se šalju svakom polazniku i upravi odnosno rukovodstvu putem e – poruka ili na neki drugi način specificiran od tvrtke čiji se zaposlenici nalaze na Akademiji.

Izvješća upravi odnosno rukovodstvu su grupna izvješća koja sadrže vrijeme svih polaznika edukacije te

tvrtke povedeno na Akademiji pojedinačno i ukupno po grupi korisnika ili na neki drugi način specificiran od strane tvrtke koja se educira. Također, ona sadrže i prolaznost na kolegijima, uspjehe svih sudionika edukacije koji su zaposlenici te tvrtke ili specificirani kao njeni članovi i plan edukacije za sljedeći period. Izvješća mogu sadržavati i dodatne informacije koje tvrtka koja se educira smatra važnima.

IV. ZAKLJUČAK

Pojava Interneta značajno je utjecala na nove pristupe poslovanju a samim time i na nove načine odnosno procese usvajanja znanja. Danas je e – učenje sastavni dio svakodnevnog života ljudi i tvrtki.

Mogućnosti učenja putem kolaboracije i razmjene informacija u što kraćem vremenskom roku velikom broju ljudi postavila je nove potrebe i trendove e – učenja.

Potrebe tvrtki koje žele educirati svoje zaposlenike su da sve informacije koje mogu biti korisne uvijek budu dostupne na jednom mjestu. Te potrebe utjecale su na tvrtke koje educiraju da ih postave i u zahtjeve svojih edukacija.

Carpio web Akademija tvrtke Carpio napravljena Moodle alatom. Predstavlja jedinstveno višedimenzionalno mjesto na kojem se nalaze svi edukacijski materijali vezani uz ERP softver tvrtke i odgovori na većinu pitanja korisnika koji rade ili počinju raditi softverom tvrtke. Sadrži dodatne informacije i iskustva vezana uz određene funkcije proizvoda tvrtke u obliku wikija i foruma. Akademija je dostupna klijentima od 0 do 24 sata na dan bez obzira na gdje se on nalazi u svijetu.

Unutar nje se svakom od polaznika kroz metodu ubrzanog usvajanja znanja dodjeljuje grupa (ovisno o kojem se korisniku (tvrtki) radi) ali i uloga (ovisno o poslovnoj funkciji koju polaznik obavlja unutar tvrtke). Uloga određuje kroz koje kolegije će se edukacija odvijati i traje od 4 do 10 tjedana.

Neki kolegiji su zajednički za više uloga, pa će se na njima nalaziti polaznici sa različitim ulogama a neki su vrlo usko specijalizirani i namijenjeni su polaznicima koji su prošli dio zajedničkih kolegija. Kolegiji unutar uloge su vezani a kroz njih polaznik prolazi individualno ovisno o tome da li je zadovoljio uvijete za prolaz kolegija ili nije. Uvjeti za prelazak na novi kolegij su prolazak na testovima i zadovoljavanje minimum aktivnosti na kolegiju. Oni su unaprijed određeni i jednaki za sve kolegije, grupe i uloge. Na takav način se korisnik 'vodi' kroz edukaciju sve do onog trenutka kada je sposoban obavljati svoj posao sa novim softverom.

Kako uloge ne ovise o korisniku nego o poslovnoj funkciji koju polaznici imaju unutar tvrtke moguće je dodjeljivanje uloga i grupa za više korisnika istodobno, neovisno o njihovoj veličini ili broju polaznika koje žele educirati. Završavanje edukacije nastaje u onom trenutku kada polaznik zadovolji sve uvjete iz svih kolegija, dakle unutar 4 do 10 tjedana.

S druge strane se polaznik, preko trokuta komunikacije, prati kroz edukaciju. Testovi koje rješava i aktivnost na kolegiju kojeg pohađa daju uvid u njegovo znanje, zalaganje ali i u onaj dio znanja u kojem bi mu trebala pomoć pri savladavanju sadržaja. Na taj način polaznik može proći kroz kolegije brže i upotpuniti svoje znanje.

Iako nam se ovaj pristup edukaciji čini prirodan i logičan, moramo uzeti u obzir da se radi o edukaciji u poslovnom okruženju gdje još uvijek vrijedi pravilo e-seminara. Vođena edukacija ovakve vrste vrlo se rijetko susreće jer je još uvijek 'primjerena školskom pristupu'.

Ipak tvrtka Carpio je uspjela primijeniti ovakav model edukacije. Kroz njega je pratila implementaciju svog softvera kod klijenata raznih veličina i profila. Kratko vrijeme edukacije ovom metodom od samo 4 do 10 tjedana znači i kratko vrijeme implementacije koja inače traje godinama a katkada nikada ne uspije.

Vidljivo je da je model ubrzanog usvajanja znanja primjenjiv, brz i efikasan. On zahvaća sve nivoe i elemente tvrtke koje se educira i stvara nove kanale komunikacije sa edukatorom ali i unutar same tvrtke.

Ovakav nov pristup postojećim tehnologijama e – učenja omogućio je inovativan način edukacije i dao tvrtki Carpio i njenim klijentima prednost na tržištu koji su tražile.

V. LITERATURA

- N. Bačun Lebarović, Organizacija edukacije u informatičkoj tvrtci primjenom modela ubrzanog usvajanja znanja, poslijediplomski specijalistički rad,Ekonomski fakultet u Zagrebu, Zagreb, 2012.
- Boggs, David, "E – Learning Benefits and ROI Comparison of E – Learning vs. Traditional Training", SyberWorks, Inc.
- O'Leonard, Karen, "Best Practeces in online Customer Training", Berstin&Associates, Listopad,2004.
- Bosnić, Ivana, "Moodle priručnik za seminar v1.0" , HrOpen – udruga za otvorene sustave i internet, Listopad 2006.
- Rice, William H., Rice, William, "Moodle 1.9: E – Learning Course Development: A complete guide to successful learning using Moodle 1.9", Packt Publishing, 2008
- Rice, William H., Rice, William, "Moodle teaching techniques: creative ways to use Moodle for constructing online learning solutions", Packt Publishing, 2007.
- Riordan, Matt, "Moodle - An electronic classroom", Cuyahoga Valley Christian Academy,
- Davis, Beth, & Carmean, Colleen & Wagner, Ellen"Moodle™ Moves To the Front of the LMS Adoption Pack", Learning Solutions Magazine, dostupno na http://www.learningsolutionsmag.com/articles/111/ , (20.01.2012.))

Masovnost, uspješnost i prolaznost u masovnim otvorenim *online* kolegijima

M. Melinščak
Veleučilište u Karlovcu, Trg J. J. Strossmayera 9, 47000 Karlovac
e-mail adresa: martina.melinscak@gmail.com

Sažetak – Masovni otvoreni online kolegiji (engl. *massive open online course* – MOOC) relativno su nov fenomen otvorenog učenja. Mogu se promatrati kao proširenje otvorenog učenja, odnosno kao nadogradnja klasičnog učenja na daljinu. Ključna je razlika u tome što su takvi kolegiji dizajnirani za neodređen broj polaznika. S obzirom na njihovo brzo širenje vode se mnoge diskusije i postavljaju pitanja o daljnjim trendovima njihova razvoja. Obilježava ih masovnost, ali je zabilježen i malen postotak polaznika koji završavaju upisane kolegije. Svrha je ovog rada u stjecanju uvida u razloge i značenje fenomena niskog postotka prolaznosti. Analizirane su ankete, dostupni izvještaji i znanstveni radovi te se daje prikaz rezultata takve analize. Izvedeni su zaključci o uzrocima značajnoga pada broja polaznika tijekom kolegija i obrazlaže se eventualni doprinos porastu postotka polaznika koji bi uspješno završavali kolegije.

I. UVOD

Masovni otvoreni *online* kolegiji (engl. *massive open online course* – MOOC) su besplatni kolegiji otvoreno dostupni neograničenom broju polaznika. I ranije su postojali otvoreni *online* kolegiji i razne varijante učenja na daljinu. Ono što MOOC-ove bitno određuje je masovnost [1-3].

Takvi kolegiji su objavljeni na Internetu, besplatni, dostupni svugdje na svijetu svakome tko posjeduje računalo i internetsku vezu, a omogućuju ih platforme kao što su *Coursera*, *edX*, *Khan Academy*, *Udacity* i druge. Pružatelji *online* kolegija su uglavnom sveučilišni konzorciji ili su blisko povezani sa sveučilištima. Štoviše, povezani su s elitnim sveučilištima poput Stanforda, Berkeleya, Harvarda, i Columbije [4].

Neki ih smatraju revolucionarnim preokretom koji nagovještava novu eru u visokom obrazovanju. U ovom ranom stadiju teško je predvidjeti daljnji tijek razvoja. *Online* obrazovanje zasigurno nije novi pristup učenju i podučavanju. To je pristup koji se razvijao godinama. MOOC-ovi su logična posljedica te evolucije. *Online* kolegiji su među najučinkovitijim i najefikasnijim metodama dijeljenja sadržaja i vještina na globalnoj razini. Mnoge uobičajeno navođene prednosti *online* kolegija su raznolikost studijskih programa i nastave koje se nude, mogućnost fleksibilnog raspolaganja vremenom, i mogućnost da polaznici balansiraju između zaposlenja i obrazovanja. Uočljivo je da *online* pokret dramatično raste i to je dovelo do kreativnijih modela poput MOOC-ova. Njihova pojava promiče razmjenu informacija širom svijeta i stvara mnogo prilika za učenje i podučavanje u različitim disciplinama [5-7].

Svaka rasprava o MOOC-ovima vodi do mnogih izazovnih i značajnih pitanja kao što je primjerice utjecaj na tradicionalno obrazovanje. Važno je pitanje ekonomske održivosti. Vrlo je upitno mogu li takvi masovni otvoreni *online* kolegiji s nekoliko stotina tisuća upisanih polaznika osigurati dovoljno kvalitetno znanje i kvalifikacije potrebne za osobni ili profesionalni razvoj i usavršavanje. Nezanemarivo je pitanje mogu li takvi kolegiji davati certifikate koji poslodavcima jamče savladano znanje i vještine. Pitanje koje je od velikog značaja za sadašnji oblik visokog obrazovanja jest tko će plaćati tradicionalno obrazovanje i koji će se zahtjevi nametnuti ako postoji mogućnost besplatnog slušanja kolegija s prestižnih sveučilišta.

Jedan od velikih izazova MOOC-ova je visoka stopa odustajanja. Kolegiji uglavnom imaju stopu završavanja od 10 % do 20 %, nekoliko kolegija bilježi čak i postotak od 2 % [7]. Glavni cilj rada je istražiti fenomen niske stope prolaznosti i kao rezultat analize dati uvid u bolje i dublje razumijevanje MOOC-ova te naći alternativne pristupe koji bi MOOC-ove činili efikasnijim. Dodatni razlog je što nam analiza *online* učenja daje dublje uvide o samom učenju i podučavanju. Prikupljanjem i razumijevanjem razloga odustajanja polaznika možemo poboljšati i *online* kolegije i tradicionalne kolegije. I konačno, analiza visokog postotka odustajanja važna je zbog ulaganja sredstava u organiziranje MOOC-ova, potrošenog vremena profesora i ostalog osoblja u organiziranju kolegija.

U radu su analizirane dostupne ankete, izvještaji, znanstveni radovi koji daju uvid u statistički relevantne značajke masovnih otvorenih *online* kolegija, njihovih polaznika i prolaznosti. Za takvu analizu su relevantni pokazatelji koji prikazuju širok raspon polaznika, ne samo studente pojedinačnog sveučilišta, nego je važna informacija struktura polaznika (studenti, studenti na doktorskom studiju, profesori, zaposleni kojima treba dodatno usavršavanje, umirovljenici…), kolika je zastupljenost u pojedinim zemljama, koji su motivi za upisivanje kolegija, koja su očekivanja, kolegiji iz kojih područja znanosti su zastupljeniji itd.

II. ANALIZA NISKE STOPE PROLAZNOSTI MASOVNIH OTVORENIH *ONLINE* KOLEGIJA

A. Opis masovnih otvorenih kolegija

MOOC-ovi su relativno nov fenomen *online* učenja. Razvili su se od ranih primjera prije pet godina a danas izazivaju značajnu medijsku pažnju i značajan interes visokoškolskih ustanova i tvrtki koje u njima vide poslovnu priliku. Mogu se smatrati proširenjem postojećeg *online* učenja, u smislu otvorenosti pristupa i neograničenosti broja polaznika [8].

Kolegiji su uglavnom organizirani kao zasebne cjeline i polaznici mogu dobiti potvrdu o završavanju (engl. c*ertification of completion*). Iako su u nekoj mjeri slični konvencionalnim predavanjima, ono što je bitno različito je što su MOOC-ovi besplatni, traže jedino pristup Internetu te interes i volju. Tipičan kolegij sastoji se od tjednih lekcija, nekoliko tema razbijeno je u kraće cjeline. Većina videa sadrži nekoliko ugrađenih kvizova (ili odabir jednog ili nekoliko od ponuđenih odgovora ili odgovaranjem u samostalnoj formi). Zatim postoje tjedni problemski zadaci koji se ovisno o vrsti kolegija sastoje također ili od pitanja ili problemskih zadataka (programiranje) ili pisanja eseja (uglavnom kolegiji iz humanističkih znanosti). Često postoji ispit na sredini kolegija i na kraju, ili samo na kraju kolegija (završni ispit). Tjedne zadaće, kvizovi i završni ispiti imaju rok nakon kojeg predaja više nije moguća (engl. *hard deadline*).

Treba napomenuti da postoje i MOOC-ovi koji se ne održavaju u određenim terminima, već je njihov materijal konstantno dostupan – tzv. *self-paced* MOOC-ovi. U radu je analiza ograničena na MOOC-ove u ograničenim terminima jer kod njih je prolaznost bolje definirana.

Moderna sveučilišta su daleko više nego samo sakupljeni kolegiji. Između ostaloga, velika vrijednost za studente je što su u bliskom kontaktu s kolegama, što sudjeluju u raznim aktivnostima unutar kampusa i u radu istraživačkih laboratorija. Možda je ova nova vrsta kolegija i načina obrazovanja dobra zamjena za prosječan kolegij s velikim brojem polaznika, no to je još uvijek na vrlo niskoj razini [5].

Jedna od ključnih značajki koja razlikuje MOOC-ove od ranijih oblika *online* učenja je što su besplatni. Polaznici ne plaćaju nikakvu naknadu. Posljedica toga je masovnost.

Pedagogija na kojoj se baziraju MOOC-ovi je također različita od tradicionalnog *online* učenja. Lekcije su drukčije strukturirane. Umjesto 45 do 60 minutnih lekcija u tradicionalnom obrazovanju i dosadašnjem stavljanju sadržaja *online*, za MOOC-ove je specifično da su predavanja uglavnom razlomljena na module čvrsto usmjerene na pojedine teme u prosječnom trajanju od 12 do 15 minuta. Glavni cilj ovakvih razlomljenih video lekcija u kraćem trajanju je omogućiti polaznicima da ih što lakše uklope u svakodnevne obveze, čime se prilagođava polaznicima raznih profila.

MOOC-ovi su inicijalno bili orijentirani na tzv. STEM discipline (engl. *Science, Technology, Engineering and Mathematics* – prirodne znanosti, tehnologija, inženjerstvo i matematika) koje dopuštaju kvantitativne procjene. Međutim MOOC-ovi se šire i na ostala područja znanosti (primjerice humanističke znanosti) time što platforme omogućuju ocjenjivanje putem istorazinske prosudbe (engl. *peer review*).

Kod klasične nastave studenti imaju redovito priliku izravno komunicirati s nastavnikom. Također imaju korist od podrške kolega koji slušaju isti kolegij. No kod MOOC-ova, unatoč nekoj vrsti podrške putem foruma i osoblja koje se povremeno uključuje u rasprave i daje odgovore i pojašnjenja, od studenta se ipak očekuje velika razina samostalnosti. MOOC-ovi daju mogućnost kvalitetnog i interaktivnog učenja, no organizirani su tako da je interakcija u virtualnom obliku, koji možda još nije dovoljno zaživio (putem blic pitanja tijekom lekcija, kvizova, zadaća), unatoč popularnosti društvenih mreža. Iskustvo je dosad pokazalo da je stvaranje snažne *online* zajednice ključna komponenta za dobro organiziran MOOC. Za učenje je nužna ljudska interakcija, makar ona bila i u virtualnom obliku. Današnja tehnologija je dovoljno razvijena da pruži medij za takvu interakciju. Unatoč važnosti tehnologije često se naglašava da se ne radi o IT revoluciji, već o revoluciji u načinu obrazovanja.

Iskustva nastavnika koji su organizirali MOOC svjedoče da treba uložiti više vremena u pripremu, bez obzira koliko dugo je nastavnik već predavao taj kolegij na sveučilišnoj razini. Činjenica da nećete biti u istoj prostoriji s polaznicima gdje je u tijeku stalna interakcija i gdje se istovremeno dobivaju povratne informacije, znači da cijeli tijek mora biti planiran do najsitnijih detalja i to prije početka kolegija. Osim samog planiranja kolegija, lekcije moraju biti snimljene, zabilježene u pisanom obliku, interaktivni kvizovi i zadaće moraju biti unaprijed pripremljeni [9].

B. Struktura polaznika i motivacija za upis

Što motivira polaznika MOOC-a? Motivacija za sudjelovanje u MOOC-ovima je značajno područje interesa za mnoge visokoškolske institucije. Mnogi su čimbenici koji utječu na motivaciju polaznika, primjerice buduća ekonomska korist, razvoj osobnih i profesionalnih vještina i znanja, izazov i postignuće, uživanje i zabava.

Prema anketi provedenoj na Sveučilištu Duke [8] polaznike svrstavaju u četiri kategorije kada je u pitanju motivacija. To su:

- Pomoć u cjeloživotnom učenju ili stjecanju i razumijevanju gradiva, bez posebnih očekivanja i potrebe za potvrdom o završavanju
- Zabava, razonoda, socijalno iskustvo i intelektualna stimulacija
- Praktičnost, često u kombinaciji s preprekama za tradicionalne mogućnosti obrazovanja
- Želja za doživljajem i istraživanjem *online* obrazovanja

Na anketi koja je prethodila kolegiju, zabava i užitak su odabrani kao važni razlozi za upis većine studenata (95 %), dok je na anketi nakon završenog kolegija većina odabrala opći interes za temu (87 %). Neki studenti upisuju *online* kolegij kako bi im to iskustvo pomoglo u odluci koji fakultet ili koje kolegije žele upisati (15 %). Značajna

manjina je tvrdila da si nije mogla priuštiti formalno obrazovanje (10 %).

Daljnja istraživanja su potrebna kako bi se razumjela motivacija polaznika na početku, ali i ono što održava motivaciju polaznika tijekom trajanja kolegija.

Uočeno je da su većina polaznika MOOC-ova ljudi koji već imaju diplomu. U tom slučaju je pitanje valjane potvrde o završetku kolegija manje važno, pošto im nije potrebno kao dokaz da su prisustvovali programu učenja i time usvojili određena znanja i vještine.

Trenutno, većina polaznika koja upisuje MOOC-ove su internacionalnog podrijetla i češće profesionalci nego studenti koji pohađaju neko sveučilište.

Paradigma MOOC-ova vjerojatno najbolje odgovara polaznicima koji su vješti u samoorganizaciji i koji su samodisciplinirani.

Prema rezultatima na *Courseri* [10] većina polaznika nisu učitelji ili profesori, samo 9 % radi na sveučilištima, 4 % u srednjim školama i 2 % u vrtićima. Velika većina polaznika bi bila jako ili u nekoj mjeri zainteresirana da uvodni materijal bude dostupan prije početka kolegija. Studenti čine 27 % polaznika, 28 % su stručnjaci i 43 % „cjeloživotni učenici“. Na pitanje o mogućnosti direktne komunikacije s instruktorom 42 % polaznika odgovara da je jako zainteresirano, 29 % odgovara da je zainteresirano, 23 % da je donekle zainteresirano i samo 5 % da nije zainteresirano. Velika većina polaznika (86 %) prati MOOC-ove od kuće, samo 8 % njih na radnom mjestu. U interakciji s ostalim polaznicima uživa 34 %, 60 % uglavnom ne komunicira s ostalim polaznicima i 6 % ne uživa u komunikaciji s ostalim polaznicima. Broj polaznika koji odgovara da im je vrlo važno dobiti certifikat o završavanju je 32 %, 25 % da im je važno, 25 % da im je donekle važno i 18 % da im uopće nije važno.

Rezultati koji se mogu istaknuti kao važni i uočljivi su vrlo visok postotak polaznika koji je zainteresiran za komunikaciju s nastavnikom. No s druge strane većina ne komunicira s ostalim polaznicima što bi bila alternativa komunikaciji s nastavnikom.

Preko 80 % polaznika na *Courseri* [9] su barem prvostupnici, a više od 40 % su magistri ili doktori znanosti, uz značajan broj upisanih koji su studenti na visokoškolskim institucijama. Mnogi polaznici upisuju kolegije iz profesionalnih ciljeva i interesa, no oni kojima takvo učenje predstavlja „hobi“ su ipak značajna većina. Na kolegijima s nižom razinom potrebnog predznanja zabilježeni su polaznici ispod 18 godina koji vjerojatno upisuju kolegije kako bi lakše odredili što ih zanima i koje sveučilište i kolegije će odabrati.

Analizom ankete na *Courseri* [9] slijedi da je najveći broj polaznika iz Sjeverne Amerike (35,2 %), zatim slijede Europa (28,2 %), Azija (21,4 %), Južna Amerika (8,8 %), Afrika (3,6 %) i Oceanija (2,8 %).

C. Postotak prolaznosti

Masovni otvoreni *online* kolegiji imaju potencijal omogućiti besplatno sveučilišno obrazovanje neodređenom broju polaznika. Zabrinutost se obično javlja zbog toga što od tisuća koje upišu kolegije samo mali dio ih uistinu završava. Broj polaznika koji odustanu od upisanih MOOC-ova je ogroman. Taj je fenomen prema nekima zanemariv jer i ako 5 % od 100000 upisanih završi kolegij to je još uvijek 5000 polaznika, odnosno daleko više nego što pohađaju studenti fizičkom prisutnošću neki kolegij u sklopu tradicionalne nastave. No ipak, ta činjenica dovodi u pitanje uspješnost MOOC-ova. Niska stopa prolaznosti i dalje se smatra jednom od većih prepreka za njihovo širenje. Godine 2007. 56,1 % akademskog osoblja je to smatralo važnom ili vrlo važnom preprekom, već sljedeće godine postotak je narastao do 61,9 %, a 2012. godine se čak 73,5 %, odnosno gotovo tri četvrtine slagalo s time [1].

O razlozima odustajanja u tradicionalnom učenju na daljinu postoji dovoljan broj studija, samo je nekoliko istraživanja provedeno za *online* obrazovanje, a za MOOC-ove koji su inačica *online* obrazovanja gotovo da i ne postoje. Vode se mnoge debate, ali nije sasvim jasno koji su to razlozi koji utječu na tako veliku stopu odustajanja, a pogotovo koji su razlozi zastupljeni u kojoj mjeri i na koje se može utjecati i kako.

Postotak prolaznosti je jedna od mjera efikasnosti *online* programa. Uz to, gubitak polaznika je trošak u smislu vremena, znanja i novca (pitanje financijske održivosti MOOC-ova je zasebno pitanje koje neće biti razmatrano u radu). Kvaliteta programa se dijelom može odrediti računanjem postotka završavanja [11] i ona je značajno niža kod *online* kolegija. Prema tomu bi se moglo zaključiti da su *online* programi manje kvalitetni i popularni no treba ipak analizom niske stope prolaznosti uvidjeti je li zaista tako. Vjerojatno prvi razlog veće stope odustajanja u slučaju *online* kolegija koji se nameće jest što su oni besplatni, neobvezni i takvo odustajanje ne snosi nikakve posljedice, što nije slučaj u tradicionalnom obrazovanju. No to zasigurno nije jedini razlog i treba sagledati koja je struktura polaznika, koji su njihovi motivi i očekivanja, uslijed čega dolazi do nezadovoljstva da bi se dobio cjelovit pogled i mogla provesti detaljnija analiza i dobiti točniji uvid u nizak postotak prolaznosti.

Trenutno ne postoje službene statistike koje obuhvaćaju sve kolegije, odnosno informacije su uvelike nepotpune. Analiza istraživačice Katy Jordan [7] pokazuje kako je najveći broj upisanih polaznika zabilježen za pojedini MOOC 180000. Još uvijek malen broj kolegija prelazi 100000 upisanih polaznika, a najčešće se kreće između 20000 i 70000. Ukupan broj polaznika koji završavaju kolegij zasad doseže najviše 20 %, no to je rijetkost i većina kolegija ima stopu završavanja ispod 10 %. Prema [8] stope odustajanja polaznika MOOC-ova koje nudi Standford, MIT i UC Berkley su od 80 do 95 %.

Treba istaknuti da se MOOC-ovi razlikuju od tradicionalnih *online* načina davanja pristupa *online* sadržaju. Mnogi stručnjaci i istraživači koji se bave analizom ovog modela učenja su naglasili velike izazove koji predstoje MOOC-ovima koje moraju savladati kako bi postali konkurentni klasičnom modelu *online* obrazovanja [6].

Problem je u tomu što kad polaznici odustanu od *online* kolegija, svaka veza s njima je izgubljena i teško je dobiti bilo kakvu povratnu informaciju. Platforme koje pružaju MOOC-ove redovito daju polaznicima ankete koje prethode početku kolegija i ankete nakon završetka, no polaznici koji su odustali od programa najvjerojatnije

ne ispunjavaju te ankete i time je nemoguće dobiti direktan odgovor o razlogu odustajanja.

Ono što također još nije jasno jest mogu li MOOC-ovi utjecati na povećanje broja diplomiranih studenata, pošto su studentima ista predavanja koja su im obvezna u sklopu programa tradicionalnog obrazovanja dostupna i kao MOOC. Zabilježeni su slučajevi u kojima, nakon što profesori organiziraju MOOC iz kolegija kojeg predaju na sveučilištu, velik dio studenata prestane pohađati nastavu i upiše MOOC. Upravitelji akademskih ustanova su i dalje sumnjičavi predstavljaju li MOOC-ovi održivu metodu nuđenja *online* kolegija, ali vjeruju da pružaju značajne spoznaje o pedagogiji *online* učenja. Manjina (23 %) upravitelja akademskih ustanova smatra da su rezultati *online* učenja inferiorni u odnosu na klasično obrazovanje [1]. Značajno je da upravitelji s institucija koje i same nude MOOC-ove imaju bolje mišljenje o rezultatima *online* obrazovanja.

Prema izvještajima posljednjih nekoliko godina pada postotak godišnjeg porasta ukupnih upisa studenata u SAD-u (primjerice sa 6.9 % 2009. godine na -0.1 % 2011. godine), no udio *online* upisa u ukupnom broju upisa raste i to kontinuirano kroz sve godine (primjerice s 9,6 % 2002. godine na 32,0 % 2011. godine). Nedavni rezultati pokazuju da 44,6 % akademskog osoblja smatra da treba više vremena i vještina za organizaciju online kolegija nego tradicionalnog, dok se samo 9,7 % ne slaže s time [1].

Zanimljivo je da veći postotak smatra *online* kolegije većim izazovom u institucijama koje nude takvu vrstu obrazovanja. Mišljenje osoblja s institucija koje ne nude niti jednu vrstu takvog obrazovanja mogu se u nekoj mjeri smatrati nerealnom prosudbom koja nije temeljena na iskustvu. U statistikama je ponekad teško razlučiti stvarne pokazatelje jer one nisu dostupne zasebno za institucije koje nude MOOC-ove i one koje ih ne nude, a izgleda da to značajno određuje njihovo stajalište.

Ono oko čega se vode rasprave jest je li *online* obrazovanje prikladno za sve studente. Godine 2007. preko 80 % akademskog osoblja je izrazilo da „studenti trebaju imati više discipline kako bi bili uspješni u *online* kolegijima“ i disciplina se smatra jednom od većih prepreka za široku primjenu *online* obrazovanja. Iskustvo s *online* obrazovanjem je samo učvrstilo taj stav i 2012. godine se 88,8 % akademskog osoblja izjasnilo o potrebi veće discipline studenata za uspješnost u *online* kolegijima [1].

III. RASPRAVA

Prema nekim autorima [8] jednaki su razlozi odustajanja i u tradicionalnom obrazovanju i u *online* obrazovanju. Iako neki razlozi mogu biti identični, sigurno se fenomen odustajanja od *online* kolegija ne može u potpunosti poistovjetiti s razlozima odustajanja u klasičnom obrazovanju. Svakako je korisno sagledati poznate uzroke odustajanja u klasičnom obrazovanju i razmotriti koji od njih su valjani i za *online* obrazovanje.

Razlozi odustajanja u klasičnom obrazovanju se dijele na akademske, osobne i financijske [8]. S obzirom da su MOOC-ovi besplatni teško da financijski razlozi imaju znatniji utjecaj. Mogu biti razlog kod *online* obrazovanja koje se plaća i koje nude neke tvrtke.

Slično kao i u pohađanju klasičnih kolegija do odustajanja najviše dolazi tijekom prva dva tjedna. Neki od razloga odustajanja su primjerice nezadovoljstvo *online* okruženjem, razlike između profesionalnih i osobnih interesa, struktura tečaja, niska razina povjerenja u učenje na daljinu, dvojbe oko uspješne *online* komunikacije, nesposobnost efikasne uporabe softvera potrebnog za učenje na daljinu, nedovoljno predznanje i preopterećenje informacijama te konačno depresonalizirano okruženje [8]. Poglavito treba razmotriti manjak centralne uloge koju ima nastavnik u klasičnom obrazovanju i koje su posljedice izostanka njegove uloge.

Gotovo svaki instruktor učenja na daljinu kao i svaki student imaju različita objašnjenja zašto studenti odustaju od pohađanja kolegija. Ipak, sva se objašnjenja mogu svrstati u dvije kategorije [12]:

1. studenti odustaju od učenja na daljinu iz istih razloga kao i kod klasičnih kolegija
2. razlozi za odustajanje su povezani s fundamentalnim razlikama između dva modusa učenja

Razlozi zbog kojih polaznici upisuju *online* programe su uglavnom: fleksibilnost rasporeda (polaznik sam određuje vrijeme kad će ispunjavati obveze predviđene programom kolegija), udobnost i učinkovitost koje pruža rad od kuće ili s bilo kojeg mjesta (gdje postoji veza s Internetom i računalo), vjerovanje da će se program kolegija dobro uklopiti u njihove profesionalne ciljeve, profesionalno usavršavanje, postizanja višeg stupnja znanja iz pojedinog područja i povjerenje u uglavnom prestižna sveučilišta koja pružaju usluge masovnih otvorenih *online* kolegija.

Razlozi odustajanja mogu biti uvelike raznoliki. Nije zabilježen razlog koji bi se pokazao dominantnim. Razloge se može podijeliti u osobne, razloge povezane sa zaposlenjem ili razloge povezane sa samim programom i organizacijom kolegija [8]. Osobni razlozi mogu biti primjerice pomanjkanje vremena za završavanje zadataka (iako je za svaki MOOC navedeno prosječno i očekivano tjedno opterećenje u satima), teško usklađivanje s rasporedom kolegija, obiteljski problemi i drugo. S obzirom na visok postotak polaznika koji su zaposleni razlozi odustajanja mogu biti poslovne obveze koje se mijenjaju tijekom trajanja kolegija i nisu predvidljive pri samom upisu, manjak podrške od strane poslodavca, preveliko opterećenje kod punog radnog vremena i istovremenog pohađanja *online* kolegija itd. Kada su u pitanju razlozi vezani za sam program kolegija, oni mogu biti: preteško ili prelagano gradivo, zadaci, kvizovi; preveliko ili premalo predznanje (kod svakog MOOC-a je navedeno potrebno predznanje, no nekad polaznici možda ne procjenjuju realno svoje znanje); problem može nastati kod istorazinske prosudbe kojom polaznici mogu biti nezadovoljni (pogotovo kada se radi o humanističkim znanostima jezik može biti problem – s jedne strane manjak znanja jezika polaznika kojima engleski nije materinji otežava pisanje i istovremeno nezadovoljstvo polaznika engleskog materinjeg jezika koji dobivaju procjene svojih radova od polaznika s nedovoljnim znanjem jezika); manjak komunikacije jedan na jedan koja je svojstvena klasičnom obrazovanju bilo da

su konzultacije studenta i nastavnika ili studenata-kolega međusobno (postoje forumi, ali veći dio polaznika ne sudjeluje na njima, barem ne aktivno, unatoč popularnosti društvenih mreža i takve vrste komunikacije); manjak interesa za program kolegija i neispunjena očekivanja. Ponekad razlozi mogu biti i tehničke prirode, primjerice manjak podrške od strane instruktora, težak pristup Internetu u nekim zemljama, hotelima itd.

Treba li se zabrinuti zbog stope polaznika koji odustaju, ili ne, jest stalni predmet rasprava. To uvelike ovisi o pretpostavljenoj namjeni MOOC-ova. Ako je cilj dati priliku otvorenim i besplatnim visoko kvalitetnim kolegijima s elitnih sveučilišta tada nas visoka stopa odustajanja ne treba brinuti. Međutim, većina se slaže da bi bilo korisno povećati stopu prolaznosti i uvidjeti razloge zašto i u kojim fazama polaznici odustaju od pohađanja kolegija. Pitanje je žele li sveučilišta pridobiti najbolje studente ili doprinijeti društvenom dobru ili pojačati i povećati vlastitu važnost. Želimo li razviti dublje razumijevanje ljudskog kognitivnog rasta ili usavršiti metode podučavanja za specifičan sadržaj [13].

Proslavljeni fizičar, nobelovac, poznat po svojim predavanjima, Richard Feynman, u predgovoru svojoj knjizi *Six Easy Pieces* [14] napisao je da je dobro podučavanje moguće jedino ako postoji izravni individualni odnos između studenata i dobrog nastavnika – situacija u kojoj student raspravlja o idejama, misli o stvarima, govori o stvarima. Nemoguće je naučiti mnogo samo sjedeći na predavanjima, pa čak ni rješavanjem zadataka koji su dodijeljeni. Nažalost njegova vizija podučavanja često nije prisutna ni u tradicionalnoj nastavi. No kada profesor organizira kolegij u obliku MOOC-a, mora biti svjestan da ta vrsta osobne interakcije zasigurno nestaje. Kod MOOC-ova koje obilježava prvenstveno ono „M“, to jest masovnost mogućnost da se razgovara sa svakim pojedinim studentom nestaje, nestaje prave interakcije, tijek je asinkron i uglavnom u jednom smjeru – od profesora prema svim studentima odjednom. To znači da polaznik postaje potpuno odgovoran za svoje svladavanje gradiva, ne postoji jedan na jedan interakcija, no može biti student na student, organiziranim u male studijske grupe. Zadaća organizatora MOOC-a je omogućiti tu interakciju.

IV. ZAKLJUČAK

MOOC-ovi se uglavnom razlikuju od konvencionalnog *online* učenja na daljinu nedostatkom ili vrlo limitiranom akademskom podrškom.

Zaključno, istraživanja pokazuju da su razlozi za odustajanje kod MOOC-ova kompleksni, višestruki i međusobno povezani. Stjecanjem uvida u razloge odustajanja polaznika institucije koje pružaju takve programe mogu profitirati razvijanjem strategije za smanjivanjem odljeva polaznika i održavanja što više stope upisa u svojim programima. Na neke razloge se može utjecati, na neke ne jer su posljedica same prirode MOOC-ova.

Mnogo je različitih čimbenika koji mogu uzrokovati da polaznici odustanu od upisanih kolegija. Zasigurno je manjak osobne podrške nastavnika i način na koje MOOC-ovi daju povratne informacije jedan od važnijih razloga. Slijede loša organizacija kolegija, nedostatak vremena ili motivacije da se završi kolegij. Zasad ne postoji zabilježena korelacija između trajanja kolegija i stope završavanja kolegija. Jedan od uzroka su teškoće uslijed pohađanja kolegija na nematerinjem jeziku. Postoje neki dokazi da istorazinska prosudba može utjecati na veću stopu odustajanja, vjerojatno vezano uz višu razinu potrebnog angažiranja koje se traži od polaznika, kao i problemi koji nastaju uslijed nedovoljnog poznavanja jezika. Ostali faktori mogu uključivati razinu i kompleksnost gradiva koje se izlaže u kolegiju.

Neki organizatori MOOC-ova počinju istraživati nove načine pružanja više razine osobne podrške. No, pružanja osobne podrške kod tako velikog broja polaznika, u kolegijima čije je glavno obilježje masovnost, ostaje velik izazov. Organizatori kolegija i nastavnici zasad pružaju podršku u vidu foruma. Pružanje osobne podrške jedan na jedan, odnosno nastavnik-polaznik bi značajno povećalo troškove i oni bi rasli s brojem polaznika. Neki profesori pokušavaju prebroditi taj problem potičući studente koji su diplomirali ili završili kolegij da učestvuju u izvođenju MOOC-ova i pomognu u pružanju dodatne podrške i usmjeravanju polaznika.

LITERATURA

[1] I. E. Allen and J. Seaman, Changing Course: Ten Years of Tracking Online Education in the United States. Sloan Consortium, 2013.

[2] O. Simpson, Supporting Students in Online, Open & Distance Learning. New York: Routledge, 2013.

[3] T. Anderson, The Theory and Practice of Online Learning. Athabasca University Press, 2008.

[4] "MOOCtalk | Let's teach the world." [Online]. Available: http://mooctalk.org/. [Accessed: 06-Feb-2014].

[5] F. G. Martin, "Will massive open online courses change how we teach?," Commun. ACM, vol. 55, no. 8, pp. 26-28, August 2012.

[6] S. M. North, R. Richardson, and M. M. North, "To adapt MOOCs, or not? That is no longer the question," Univ. J. Educat. Res., vol. 2, no. 1, pp. 69-72, 2014.

[7] K. Jordan, "MOOC Completion Rates: The Data.," MOOC Completion Rates: The Data. [Online]. Available: https://www.katyjordan.com/MOOCproject.html. [Accessed: 06-Feb-2014].

[8] L. Yuan and S. Powell, MOOCs and Open Education: Implications for Higher Education. 2013. [Online]. Available: http://publications.cetis.ac.uk/2013/667. [Accessed: 06-Feb-2014].

[9] Universities UK - Massive open online courses: higher education's digital moment?. London: Universities UK, 2013.

[10] "Coursera," Coursera. [Online]. Available: https://www.coursera.org/. [Accessed: 06-Feb-2014].

[11] P. A. Willging and S. D. Johnson, "Factors that influence students' decision to dropout of online courses," J. Asynchronous Learn. Netw., vol. 8, no. 4, pp. 105–118, 2004.

[12] S. Carr, "As distance education comes of age, the challenge is keeping the students," Chron. High. Educ., February 2000.

[13] B. D. Voss, Massive Open Online Courses (MOOCs): A Primer for University and College Board Members. Washington: AGB Association of Governing Boards of Universities and Colleges, 2013.

[14] R. P. Feynman, R. B. Leighton, and M. L. Sands, Six Easy Pieces: Essentials of Physics Explained by its Most Brilliant Teacher. New York: Basic Books, 2011.

CTS

International Conference on

COMPUTERS IN TECHNICAL SYSTEMS

Steering Committee

Chairs: Vlado Sruk, University of Zagreb, Croatia
Leonardo Jelenković, University of Zagreb, Croatia
Dejan Škvorc, University of Zagreb, Croatia
Jadranko Novak, Nomen, Rijeka, Croatia
Nikola Mišković, University of Zagreb, Croatia

Members: Matjaž Colnarič, University of Maribor, Slovenia
Željko Hocenski, University of Osijek, Croatia
Niko Guid, University of Maribor, Slovenia

INVITED PAPERS

Learning From Success Cases in Steel Industry Electrical and Automation Applications

Paolo Paganini
* A.I.S. – Automation Integrated Solutions S.p.A., Sovizzo (Vicenza), Italy
p.paganini@ais-automazione.it

Abstract - This paper presents an overview on the evolution of the application of electrical and automation technologies in the industrial processes, with more details in the hot rolling mills for metals.

Different aspects are taken into consideration, like electrical motors, power electronics, process control and information management.

Finally, a proposal of a target framework for the future possible applications matching the present needs of industries, like improvement of production procedures and energy saving.

Rather than going into deep details, the aim is to bring to the surface a certain amount of topics in order to support future actions requested by industries managers as well by electrical and automation companies.

I. INTRODUCTION

The overview begins with an analysis of the features of some typical industrial processes. Although the results can be applied to quite a wide range of industries, we focus on some specific applications, which can be considered as good example and, at the same time, are well present in the geographical area of South-Eastern Europe and Mediterran.

It is useful to first define the industrial processes considered in this paper in order to set the limits of this analysis.

The focus is on those processes where production is powered by medium to big electrical motors, each rated hundreds of kW, where production flow must be coordinated by a centralised control and where processes are repetitive.

According to this definition it comes that the the focus is mostly on metals production and processing, cement, material handling, oil & gas, mining.

It is worth to highlight that, even though with some difference among countries, all these kinds of processes are well established in Europe.

However, most of those industries are now facing well known problems of energy tariffs and availability, environmental issues and, last but not least, cost and quality of finished products.

All such matters must necessarily play a major role in the design of modern applications.

II. THE TECHNOLOGY BEHIND THE PROCESSES



Fig.1: Old Style “Cross Country” rolling mill

When considering the “heavy”industrial processes, it comes logical to consider only the evolution of the process design and layout. In other words, the focus is on the process technology itself, a glimpse of whose is easily achievable when viewing the process.

Just as an example, if we consider old style hot rolling mills we would come across the so called “cross country” design like the one shown in Fig.1:

Here only one electrical motor, usually operating at fixed speed, was powering several machines (stands). In most cases, such motors was a slip ring motor (or wounded rotor type), whose power and speed and consequently torque could not be controlled. (Fig.2)

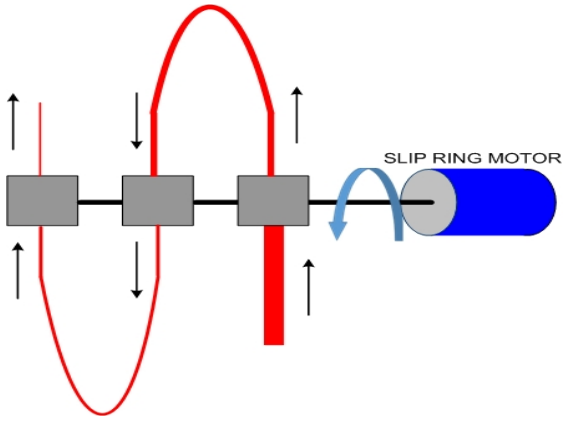


Fig.2: Schematic of “Cross Country” layout



Fig.3: Modern “In Line” rolling mill

The major step forward in this process evolution was the upgrading to continuous in line rolling mills, like in the next picture, where each machine (stand) is powered by an independent electrical motor. (Fig.3)

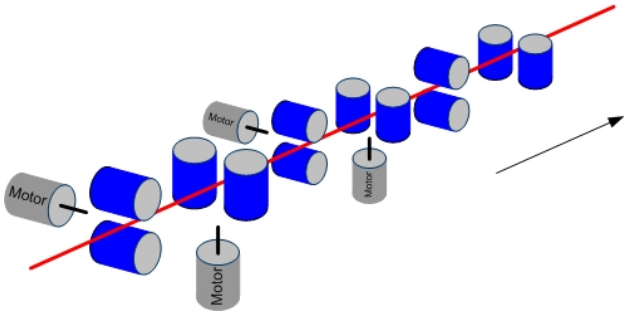


Fig.4: Schematic of “In Line” rolling mill

This layout was an outstanding breakthrough in the rolling mills technology, thanks to innovative and future oriented approach of many companies and workshops supplying mechanical equipment for rolling mills, several of those were located in Italy.

III. The Role of Electricals and Automation

The previous examples is interesting because its shows a well known technology, but what it has been the role of electrical equipment and automation?

It may not be immediately noticed that the evolution of rolling mill design has been possible because there was the availability of electrical motors, whose power, speed and torque could be controlled with accuracy.

The slip ring motors of obsolete “cross country” mills were replaced by DC (direct current) motor; the control of the voltage and current of a DC motor allows the regulation of the delivered speed and power by means of DC power converters.

DC motors have been on the market since almost 150 years but only after semiconductors technology has been broadly available, it has been possible to have accurate and not expensive DC power converters built with semiconductor thyristors.

IV. Unexpensive High Technology

This is an intersting pattern: we have old and well established technology (rolling of steel bars), we have DC motors (150 years old technology) and both of them have been brought to a new life by an innovative and unexpensive technology.

The semiconductor technology was developed by engineers and researchers in the Universities in cooperation with big Companies, but after only a few years the application market was almost completely ruled by small and medium companies.

In Italy, the country I know better, between late ’60 and ’90 there has been a period were a significative portion of the production of the DC power converters, one of the pillar of the modern industrial processes, was located in only one industrial district around the city of Vicenza, that someone called the “Italian Silicon Valley” (by the way, it is not by chance that Mr.Federico Faggin [1], one of the inventors of the one of the very first industrial microprocessor in California (the Zilog Z80) was born and received first years of education in Vicenza)

Only a few years before the availability of semiconductors, none of these small and medium companies could have had access to the supply of electricals and automation for the steel plants.

The new technology and an excellent background of education provided by technical high schools and nearby Universities boosted an economical “miracle”, whose effects are still there today, even though the peak period has been reached 15 years ago.[2], [3]

The example of rolling mill is interesting not only for the intrinsic technological properties but also because it helps to highlight a common condition that should never be neglected: the electrical and automation suppliers are not generally the owners of the process technology, which is in the hands of the process and mechanical suppliers. It means that the success of a new electrical and automation technology is based either on the cooperation with a design supplier or, on the opposite, on a solution that is not relying on the process.

It is a matter of fact that several big design companies have acquired in house capacity in electricals and automation or some big electrical decided to move toward the process design.

V. Looking at the Future

So the question today is: how can we adopt the same “DC motors and DC converters approach” approach using state of art technology and apply it to industrial processes, without being a big Corporation?

To get a good answer it is first necessary, however, to finish this quick survey of the effects of electrical and electronic equipment on the industrial processes.

(By the way, just as a mention, today DC motors technology is being replaced by AC motors with inverters, allowing higher efficiency and accuracy)

A. *The Control Systems*

The next step upwards is toward the control systems: the availability of big electrical motors controlled by semiconductor power electronics was not fully effective without an automatic control system, delivering set points and operating control loops. The computer technology was requested and MIPRO audience is the most appropriate to appreciate this technology.

B. *The Returning Pattern*

The technical evolution of power electronics and industrial control systems has been quite similar and moving along parallel paths: after a first research and development stage, a period has followed where many companies proposed their own products with proprietary technology. They have been quite successful because they fulfilled the industrial companies demand, no matter how they were designing their products.

C. *The lesson to be learned*

Of course, there were (and there are) also some dangerous situations to be avoided in order to keep a company safe.

The model of the successful small/medium company, selling its high tech and unexpensive products all over the world collapsed when the products became naturally aged after some years of operation, but the original supplier was not existing anymore or located too far away from the plant or the documentation lost.

So only those companies featuring good products, a solid structure, a distributed geographical presence and compliance with standards have been able to survive.

In addition to this, the market is now dominated by electrical and automation products manufactured with excellent quality and low prices by companies much bigger than the average european typical size.

So now it is not requested to know how to manufacture most of the electrical and automation equipment, but rather how to use it at best and the lowest cost.

VI. THE CONCLUSIONS

Going back to the previous question, where do we have to look for interesting opportunities in the application electrical and automation technology in the heavy industry processes, without being a big Corporation and without being a process design company?

If we look at the conditions fostering the success of the electrical suppliers back in the age of “DC motors and first computers”, we can see which were the common grounds:

- there was availability of a basic technology
- the technology was well tested
- the technology was not expensive
- there were the request of the heavy industry to improve production.

So, today, which electrical and automation technology is well tested, technologically stable, not expensive and easily available?

Here we propose an answer: information management and processing has never been so powerful as today, thanks to the merging of low cost and really portable devices with distributed data transmission and processing.

Moreover, data processing capacity can be installed anywhere, not only in the “traditional” computers. It means that we can have huge processing power also inside a power converter controlling a motor but also thousands kilometers far away from the industrial process.

Moreover, which are the needs of the industry today, with emphasis on Europe? Some are here proposed:

- Energy savings
- Structured maintenance scheduling
- Detailed collection of data on production
- Improve production procedures: merging the analysis of process parameters with production data can improve the fine tuning of production procedure, resulting in cost reduction.
- Support to technical staff: easier training, troubleshooting and access to documentation

These technologies can be developed also by small/medium sized companies, this is the good new.

Anyway, the good old workhorse of the heavy industries, the DC motor, will not disappear. The upgrading of power converters towards full integration with distributed data processing is still at the first step.

[1] http://en.wikipedia.org/wiki/Federico_Faggin

[2] I http://www.itisrossi.vi.it/

[3] http://www.unipd.it/

The multimedia devices portable automated test system

Vukota Pekovic*, Ivan Kastelan **, Gordana Velikic* and Nikola Teslic*
* RT-RK Institute for Computer Based Systems, Novi Sad, Serbia
** University of Novi Sad, Faculty of Technical Sciences, Novi Sad, Serbia
vukota.pekovic@rt-rk.com

Abstract - Rapid development of new technologies adds to an architectural complexity of multimedia devices. The last generations of multimedia devices combine features of multiple devices into a single device. Complex devices require more testing and verification methods, and more time to find a fault. Very often inoperative device has to be transported to operator's facilities for complete diagnostics. Transportation of inoperative devices to repair/service centers prolongs time from a problem reporting to problem correction, increases costs, and adds to the customer's discontent. Field test system comprises various input interfaces which cover multimedia device's audio/video output, and allows functional verification, wide A/V analysis and picture quality measurement. Field test system is ideal for verification of STB/DVD/Blu-ray player's devices at customer's homes or retailers offices and also for monitoring of head-end signal quality in broadcast systems.

I. INTRODUCTION

Rapid development of new technologies combined with higher demands for user interaction contributes to architectural complexities of multimedia devices. Latest developmental trends merge features of different devices into a single multipurpose multimedia device. The integrated device requires additional tests and verification methods, which increases time needed to find a fault. Often an inoperative device is transported to operator's facilities for comprehensive diagnostic. Transportation of inoperative devices to either repair or service centers prolongs time from a problem reporting to the problem correction, and thus increases repair costs and user's discontent.

It does not surprise that a prompt problem solution is highly evaluated in customers' satisfaction surveys [1]. A need exists to relocate the device comprehensive health diagnostics to customers' homes. In many cases, a device that the customer reports as inoperative is fully functional and the 'fault' is incorrect use of the device. Such cases can be easily eliminated and solved with troubleshooting steps that can be navigated with the help of remote customer support. However, if the device is inoperative, many faults can be easily solved at the customer's home with a portable test system that can detect nature of the fault at the location. Some benefits of such diagnostic system are: fast and reliable device health diagnostics, no transportation, supplement unit rental or storage costs, saved personal data, and increased user's satisfaction. Although the system would decrease service costs significantly – the estimated annual saving is over $35,000 [2], envisioned system is not commercially available. Current state-of the-art is in the phase of research and prototypes, reported by S3Group [3] and Regenersis [4] who are still facing problems with power and connector blocks. In this paper we describe successfully deployed field test system prototype, which has powerful features and compact size.

II. THE FIELD TEST SYSTEM

Field Test system (Fig. 1) is a portable, stand alone, automated multimedia device test system intended to be used in customers' homes for verification of unit's functionality. Typical deployments are in a field service vehicle. The system equips field technicians with a powerful tool which performs fundamental functional tests in order to diagnose possible malfunctions of customer's device. The system enables greater flexibility of test environment, wide range of tested devices, rapid adaption and easier applicability to new device releases. Carefully designed with optimized parts guarantees good overall retail system cost. The system adds to greater reliability of testing during the test and verification process, and facilitates and speeds up the field work of a technician.



Figure 1. Prototype of the Field Test System

This work was partially supported by Ministry of Education, Science and Technological Development of the Republic of Serbia, under grant No. TR32034, year 2011-2015.

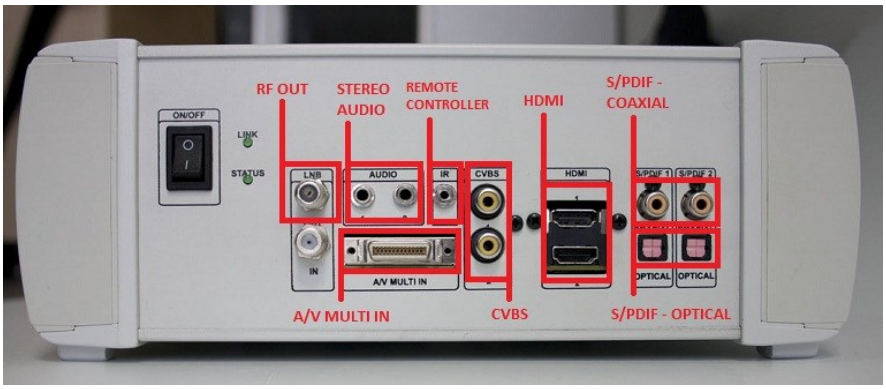


Figure 2. Front-view of the Field Test System prototype

Integrated software modules enable testing of devices that allows black box testing approach. The black box testing approach [5]-[6] is a testing in which system can influence only input and output signals of a device under test (DUT), while internal structure of the DUT remains hidden and unknown to the testing system.

The system prototype comprises various input interfaces which cover any DUT's audio/video output and enables wide audio-video analysis and picture quality measurement. The prototype is particularly good for verification of set top boxes (STB), DVD and Blu-ray players. Complete infrastructure for field diagnostics allows technician to reliably decide whether to replace inoperative device or to correct minor defect. Field test system includes protocols for transfer of collected test execution information to central service center database and generates reports necessary for repair process.

Field test system device supports real time capture and streaming of SD, ED and HD video content up to 1080i60 from wide variety of digital or analog A/V sources, captures digital or analog audio up to 192 kHz with 24-bit resolution. The prototype has built-in VHF/UHF Modulator, grabber device, internal power supply, and control unit. It has a 10-inch touch screen display and Windows 8 Pro Operating System, and can be used with specially developed Executor Home Screen application.

Available connectors for the prototype are shown on Fig. 2 and summarized in Table 1.

TABLE I. FIELD TEST SYSTEM CONNECTORS

Connectors	
2x Analog audio input	Stereo audio on 3.5 mm TRS jack connector
2x HDMI	Type A receptacle HDMI
1x Multi A/V input	26 pin Mini Delta Ribbon (MDR) connector - Component, S-Video, SCART
2x Composite NTSC/PAL input	RCA connector
2x LNB in/out	F type connector
2x S/PDIF coaxial	Coaxial RCA connector
2x S/PDIF optical	TOSLINK
Power supply	Integrated in platform cover with passive cooling, 20-240V 10W
IR	Remote control emulator output

Interface specification for the Field Test System prototype is listed in Table 2.

TABLE II. FIELD TEST SYSTEM INTERFACES

Interfaces	
VHF/UHF Modulator output	RF, support for most QAM-, OFDM- and VSB-based modulation standard
Supported Modulator Output	Support for most QAM-, OFDM- and VSB-based modulation standard Direct digital conversion to 47..1000MHz Play out of I/Q sample files up to 9.14Msamp/s Programmable attenuator
Digital Video Input	HDMI, SDI
Analog Video Input	Composite, Component, S-Video, SCART
Supported HD Format	720p50, 720p59.94, 720p60, 1080PsF23.98, 1080p23.98, 1080PsF24, 1080p24, 1080PsF25, 1080p25, 1080PsF29.97, 1080p29.97, 1080PsF30, 1080p30, 1080i50, 1080i59.94 and 1080i60
Supported SDI formats	HD and SD SDI Receiver, with complete SMPTE Video and Audio Processing, as per SMPTE 425M, 292M and SMPTE 259M-C
Supported SD formats	625/25 PAL, 525/29.97 NTSC and 525/23.98 NTSC
HDMI Compliance	HDMI 1.3
Multiple Rate Support	SDI, HDMI and component analog video connections are switchable between SD and HD. SDI switches between 270 Mb/s standard definition SDI and 1.5 Gb/s HD-SDI

Rear view of the prototype is shown on Fig. 3, on the next page. Rear interfaces include:

- USB – connector used for barcode scanner,
- Serial port – used for barcode printer,
- Power supply – for the device.

Field test system has live preview through LAN interface. This feature enables extra help from external source center, such as remote service operator. Field test system has included automatic video signal and standard detection.



Figure 3. Rear-view of the Field Test System prototype

The prototype system captures and stores images or video sequences to internal memory for on device image processing, transfers grabbed images/video to host device (Tablet) for future analyzes, allows easy image capture or video streaming of multiple A/V signals into single computing server which can be spatially dislocated from video source, sampling of analog video, audio and control signals up to 32MBPS for measurement of amplitude and temporal characteristics of the signals, DSP based platform which is capable to execute desired processing or analysis algorithm on captured audio and video content, voltage and frequency measurement on A/V inputs (CVBS, RGB, Line In, LNB etc.), embedded remote control emulator, support and compliance with consumer and professional A/V standards.

Test environment has to be configured before the Executor Home Screen application first run. The list of configuration files which need to be set is shown on Fig. 4, on the next page.

TestEnvironmentConfiguration.ini file saves a list of all devices used in testing process and files with macros related to specific devices.

TestDeviceConfiguration.ini saves relations between DUT and created test plans. These relations are based on barcodes of DUTs and IDs of the test plans.

During initialization the user chooses server and database to connect. This will be saved for next log in (Fig. 5).

The cursor of the application's main page is positioned on the Barcode field, as shown on Fig. 6.

The barcode can be scanned or entered manually. Next, the test plan is selected and during execution the test cases are displayed with test case results (fail/pass), as shown on Fig. 7.

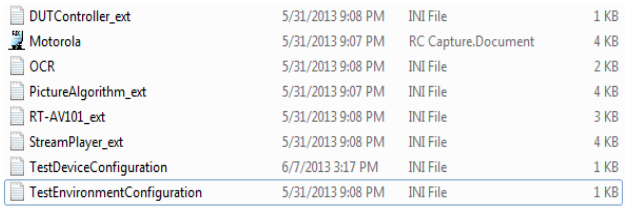


Figure 4. Configuration files for Executor Home Screen application

Single test steps can be checked during the execution at the bottom of the page. System Check displays the list of devices used for testing.

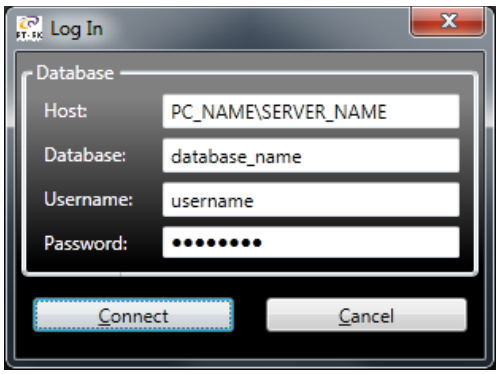


Figure 5. Login screen for Executor Home Screen application

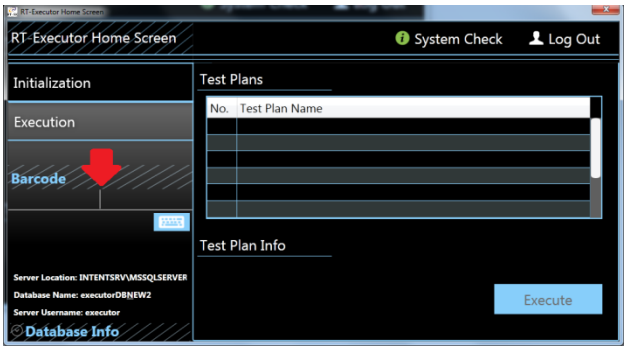


Figure 6. Main page of the Executor Home Screen application

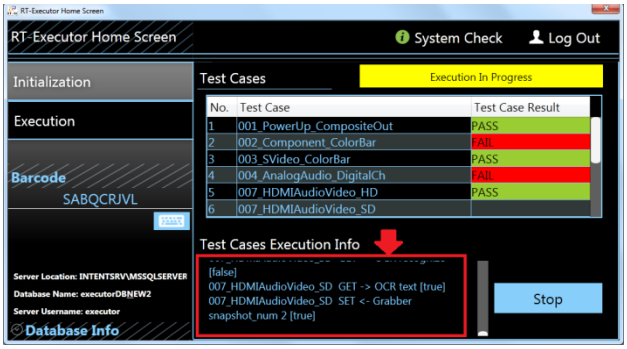


Figure 7. Test execution

III. CONCLUSION

Technology development has changed roles of multimedia devices from relatively simple tasks and limited interactions with users to multifunctional devices that cut across various dimensions. The role changes have brought different user habits and added to complexity of devices. Customers demand high quality of service and rapid problem solving response time. This leads to demands to automate the test process to cut the test time and to reduce costs. The core issue is to develop reliable tests and test hardware. Mobility of such test stations is an issue due to device's size and robustness. Current trend is to take reported device to the service facility. This adds additional costs and does not address "no fault found" issue. However, what has been missing is a powerful portable tool that would enable to move the test from repair facilities to home. We described a test field system, compact in size, loaded with testing features that will enable field technician to conduct tests previously reserved for testing locations, at user's home. Although the system is intended for use at customer's home, it can be deployed at retailers' offices or used by service providers to monitor head-end signal quality in broadcast systems.

REFERENCES

[1] A. Cuadra, M. Cutanda, A. Aurelius, K. Brunnstrom, J.E. Lopez de Vergara, M. Varela, J-P. Laulajainen, A. Morais, A. Cavalli, A. Mellouk, B. Augustin, I. Perez-Mateos: "Ecosystem for customer experience assurance", *International Conference on Smart Communications in Network Technologies (SaCoNeT)*, 2013, pp. 1-5

[2] C. Radford: "How home test technology is revolutionizing the reverse logistics for set top boxes", *A Technical White Paper from Regenersis*, 2011.

[3] S3Group, http://www.s3group.com/

[4] Regenersis, http://www.regenersis.com/

[5] D. Marijan, V. Zlokolica, N. Teslic, V. Pekovic, T. Tekcan: "Automatic Functional TV Set Failure Detection System", *IEEE Transactions on Consumer Electronics*, 2010, pp. 125-133

[6] I. Kastelan, V. Pekovic, N. Teslic, T. Tekcan, D. Marijan: "Automatic Black Box Testing of television systems on the final production line", *IEEE International Conference on Consumer Electronics*, 2011, pp. 869-870

PAPERS

Static characteristic and describing function of polynomial compensated nonlinearities

O. Kuljača, K. Horvat and Z. Eškinja
Brodarski Institute, Zagreb, Croatia
ognjen.kuljaca@hrbi.hr, krunoslav.horvat@hrbi.hr, zdravko.eskinja@hrbi.hr

Abstract - Paper deals with polynomial compensation nonlinearities. Most actuators are nonlinear sporting nonlinearities such as deadzone, saturation and others. One way of dealing with such actuators is to try to compensate nonlinearities by polynomial representation of inverse nonlinearity. Compensated actuators can improve behavior of the system, but then we have a new nonlinearity in the system, polynomial compensated nonlinearity. Paper will show how to obtain static characteristic and describing function of such nonlinearities.

I. INTRODUCTION

Every control system contains at least one actuator. Virtually all actuators have nonlinearities. Depending on the control system these nonlinearities can be neglected or one has to deal with them in some way. Nonlinear control theory is rich with different methods of compensation or stability analysis of systems with nonlinear actuators ([9],[11],[12],[13],[14]).

Many actuators have odd nonlinearity followed by linear filter type element. Examples are some electrical motors, hydraulic cylinders and solenoid actuators. There are different methods for dealing with such systems (introduction of ditter signal, different anti wind-up schemes etc,). This paper deals with static characteristic and describing function of such nonlinearities compensated by polynomial representation of inverse actuator nonlinearity. Examples are given for compensated nonlinearities of type dead zone, saturation, deadzone with saturation and friction.

II. DESCRIBING FUNCTION

Many actuators with odd nonlinearity can be represented by structure shown in Figure 1.

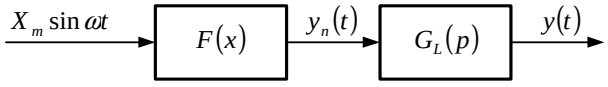


Figure 1. Actuator represented with decoupled nonlinear F(x) and linear parts $G_L(p)$, p=d/dt

Describing function is an equivalent gain of nonlinear element, defined by the harmonic linearization method of nonlinear static characteristic ([2], [9], [10], [11], [12] and many others). It is a known method of analysis and synthesis when nonlinear system can be decoupled into linear and nonlinear parts. If the linear part of the system has the characteristics of low-pass filter and if we apply periodical signal to the system, output signal will have the same base frequency as input signal with damped higher frequencies.

If the amplitudes of higher harmonics are relatively small when compared to the amplitude of the first harmonic, i.e. if filter hypothesis holds, output signal can be approximated by it's first harmonic. Mathematically, first harmonic of the output signal, for the sinusoidal input signal $X_m \sin \omega t$, can be expressed by Fourier expressions:

$$y_N(t) \approx Y_{P1} \sin \omega t + Y_{Q1} \cos \omega t \tag{1}$$

$$Y_{P1} = \frac{1}{\pi} \int_0^{2\pi} F(X_m \sin \omega t) \sin \omega t \, d(\omega t) \tag{2}$$

$$Y_{Q1} = \frac{1}{\pi} \int_0^{2\pi} F(X_m \sin \omega t) \cos \omega t \, d(\omega t) \tag{3}$$

where Y_{P1} and Y_{Q1} are first Fourier coefficients. Describing function is the ratio between first harmonic of the output signal and input signal in complex form:

$$G_N(X_m) = P(X_m) + jQ(X_m) = \frac{Y_{P1}}{X_m} + j\frac{Y_{Q1}}{X_m} \tag{4}$$

where $P(X_m)$ and $Q(X_m)$ are coefficients of the harmonic linearization ([2], [9], [10], [11], [12] and many others).

Determination of describing function boils down to the determination of integral expressions for the known static characteristic of the nonlinear part of the system.

There are many conventional nonlinearities for which static characteristics and describing functions are theoretically derived and given in analytical form.

The problem arises when the static characteristic of nonlinear system cannot be analytically expressed or integral expressions cannot be solved. In that case describing function can be determined by experiment or simulation ([3], [4], [5], [6], [7], [8]) or some method of numerical integration [10].

In this work we will use numerical integration in Simulink as given in [10]. The approach is very simple, yet effective. In essence, equations (2) and (3) are directly

Research was partially supported by HZZZ grant 09/193 Fuzzy Logic Adaptive Elastic Controller.

applied in Simulink on nonlinear element outputs. More details on this approach to obtaining describing function are also given in [13] and [14].

III. Polynomial Compensation of Odd Symmetric Nonlinearity

Ideally, if one could completely compensate a nonlinearity, a compensated nonlinearity would become a linear gain. This, in general, is not possible for many reasons. However, some level of compensation could be achieved. In this paper we deal with polynomial compensation of odd symmetric nonlinearities. Two main characteristics of such nonlinearities is that they are odd symmetric and time invariant.

A static neural network compensation scheme that yields polynomial compensation is described in more details in [13] and [14].

We deal with actuator with architecture shown in Figure 1. It is important that in practice $y_n(t)$ is measurable. This is often the case because many motor drives and hydraulic drives have nonlinearity in part of their electronic driver and this value can be measured. We then connect an input with changeable amplitude and measure and save inputs and outputs. Once this is done we will use recorded data to train simple neural network in such way that we try to achieve neural network that will have opposite, i.e. inverse action to nonlinearity. Of course, it is impossible to achieve exact opposite action, but in many cases it is possible to achieve good enough representation to improve nonlinear characteristic. Schematic of data generation and recording is shown in Figure 2.

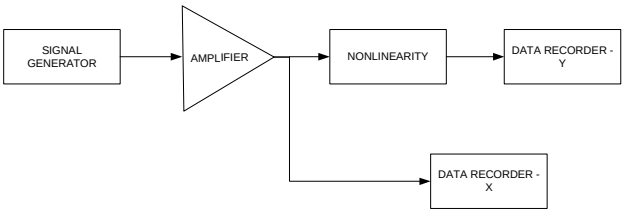


Figure 2. Data generation and recording

Then, we can use neural network to compensate for nonlinearity as shown in Figure 3 ([13], [14]).

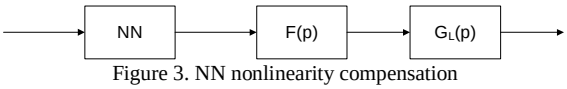


Figure 3. NN nonlinearity compensation

where NN stands for neural network, F(p) is nonlinearity and $G_L(p)$ is linear part of the system.

The idea is that neural network compensates for some effects of nonlinearity since neural network is trained to act as inversion of nonlinear function.

However, realizing neural network sometimes can be complicated. One possibility is to replace the neural network with its describing function. Describing function can be graphically obtained by using procedure given in [10] and described in more details in [13] and [14].

Realization of neural network or graphical representation of describing function can be complicated, with number of calculations included in the process or with need to realize detailed look-up tables for describing function. So, the next step is to obtain polynomial form describing function that could be used in final compensation of given nonlinearity.

So, we represent graphic describing function of compensating neural network by polynomial $G_{NN}(x)$:

$$G_{NN}(x) = a_n x^n + a_{n-1} x^{n-1} + \ldots + a_1 x + a_0 \qquad (5)$$

Polynomial approximation is easier to realize than neural network, i.e. we can use simpler hardware to do so.

The procedure is experimental and it works as follows. First, we decide on domain of input signal (X, Figure 2). Then, we generate and record the signal to use in training neural network.

In our case datasets X and Y were generated using chirp signal generator with starting frequency 0.1 Hz, final frequency 20 Hz and simulation end time 1000s with fixed step size 0.001. Odd symmetric nonlinearities are static nonlinearities so frequency really does not have an effect on them but we used chirp signal in order to generate a number of different data points.

After data was recorded, one can use dataset Y as network input and dataset X as network output for training. There are many tools available for neural network training. Here, Matlab neural networks toolbox was used. Feed forward network had two layers, first layer with seven nodes and sigmoid tangent activation function and one node in output layer with linear gain activation. Network was trained for seven hundred epochs with gradient descent with adaptive learning rule backpropagation. The same network architecture was used for all nonlinearities described in this paper. Choice of number of nodes and activation functions is heuristic and depends on required precision.

IV. Static Characteristic And Describing Function Of Polynomial Compensated Nonlinearities

As explained before, polynomial is obtained from describing function of obtained neural network. The degree of polynomial is determined heuristically. Then, one should inspect static characteristic of compensated nonlinearity to obtain the picture of behavior of polynomial compensated nonlinearity. In this Section we will deal with examples for deadzone, saturation, deadzone with saturation and friction.

Compensation with polynomial is shown in Figure 4.

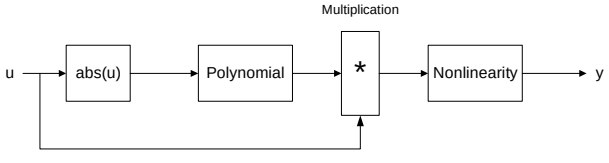


Figure 4. Compensation with polynomial

NN training data for deadzone with upper and lower limits equal to 1 was obtained using scheme from Figure 2 with amplification of 5. For describing function we used amplitudes from 0 to 3.53. Frequency is of no importance since we deal odd symmetrical nonlinearity so imaginary part odd describing function is 0.

Static characteristic of deadzone with upper and lower limits equal to 1 is shown in Figure 5.

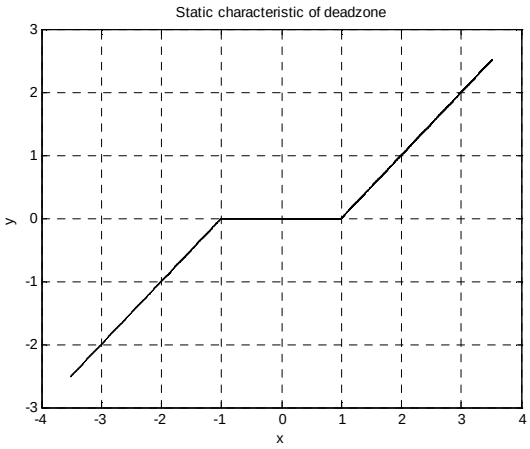


Figure 5. Static characteristic of deadzone with upper and lower limits equal to 1

With given data and described procedure deadzone shown in Figure 5 was compensated with third order polynomial function:

$y=-0.3024x^3+2.2337x^2-5.4734x+5.9814$

Static characteristic of compensated deadzone is shown in Figure 6.

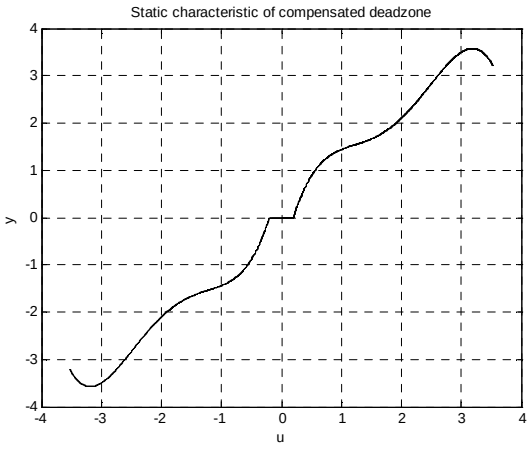


Figure 6. Static characteristic of compensated deadzone

NN training data for saturation with upper and lower limits equal to 0.5 was obtained using scheme from Figure 2 with amplification of 3.53. For describing function we used amplitudes from 0 to 3.53. Frequency is of no importance since we deal odd symmetrical nonlinearity so imaginary part odd describing function is 0.

Static characteristic of saturation with upper and lower limits equal to 0.5 is shown in Figure 7.

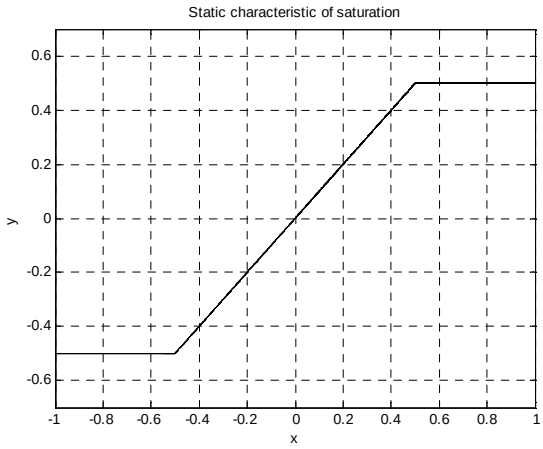


Figure 7. Static characteristic of saturation with upper and lower limits equal to 0.5

With given data and described procedure saturation shown in Figure 7 was compensated with third order polynomial function:

$Y=0.1929\ \ x^3-1.0843x^2+1.2508\ x+1.4042$

Static characteristic of compensated saturation is shown in Figure 8.

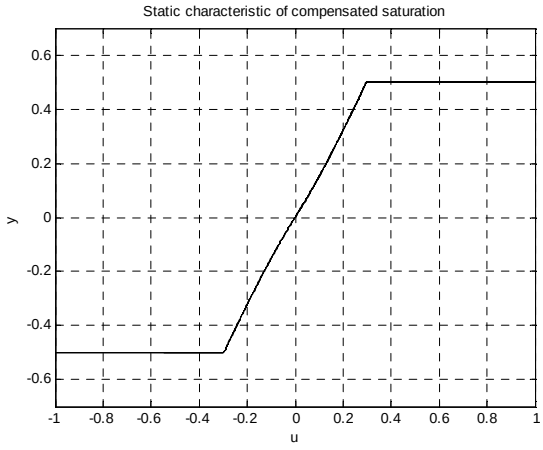


Figure 8. Static characteristic of compensated saturation

Static characteristic of dead zone with saturation with upper and lower limit equal 1 with start and end of dead zone equal to 0.3 is shown in Figure 9.

NN training data for saturation with upper and lower limits equal to 0.5 was obtained using scheme from Figure 2 with amplification of 1.5. For describing function we used amplitudes from 0 to 1.5. Frequency is of no importance since we deal odd symmetrical nonlinearity so imaginary part odd describing function is 0.

With given data and described procedure deadzone with saturation shown in Figure 9 was compensated with fifth order polynomial function:

$y=-4.4035x^5+20.8821x^4-39.8614x^3+38.8677x^2-19.7612x+5.7324$

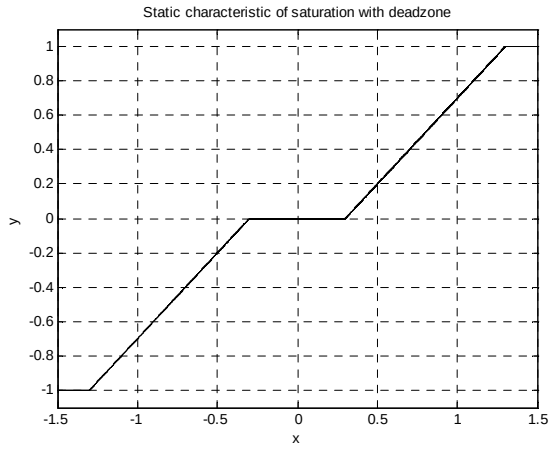


Figure 9. Static characteristic dead zone with saturation with upper and lower limit equal 1 with start and end of dead zone equal to 0.3

Static characteristic of compensated dead zone with saturation is shown in Figure 10.

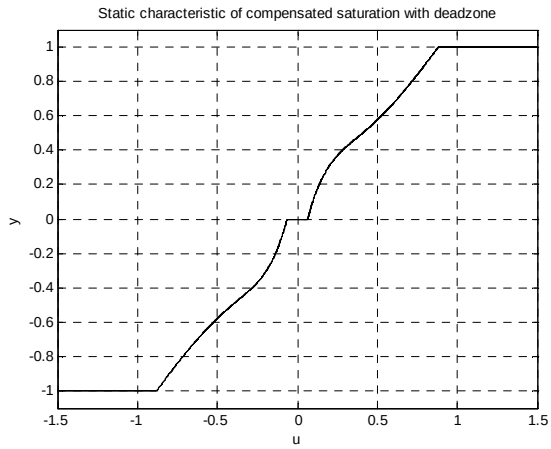


Figure 10. Static characteristic of compensated dead zone with saturation

Static characteristic of friction function y=sign(x)[abs(x)+0.5] is shown in Figure 11.

With given data and described procedure friction shown in Figure 11 was compensated with fourth order polynomial function:

$y= 0.0461x^4-0.3810x^3+1.0264x^2-0.8100x+0.5233$

Static characteristic of compensated friction is shown in Figure 12.

It can be seen that for case of friction, as with the case of saturation effects of compensation are not so clearly seen. This is due to hard limits nonlienarities which are very hard to approximate in general. Thus, it is hard to compensate them as well.

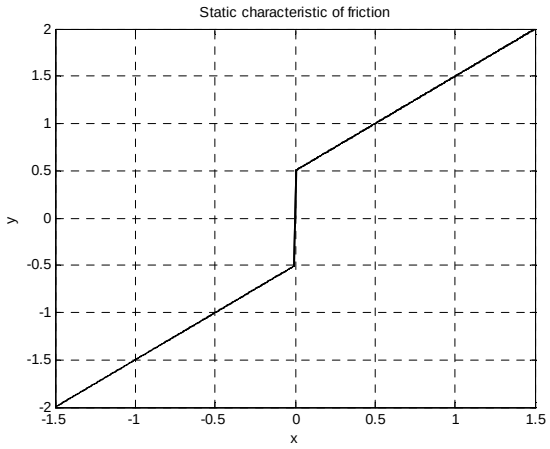


Figure 11. Static characteristic of friction

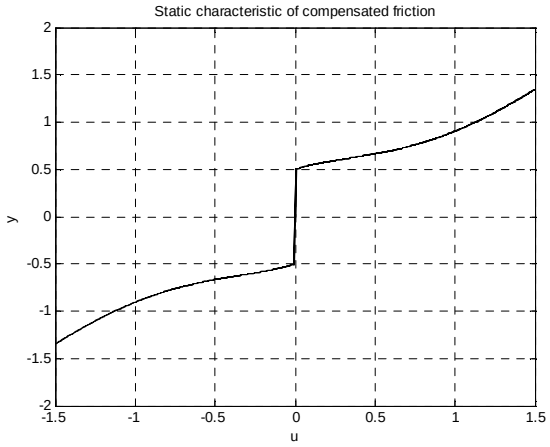


Figure 12. Static characteristic of compensated friction

The case involving the deadzone show much better compensation results. This is due to fact that deadzone can be actually inverted.

Since neural networks are universal approximators with respect to smooth functions it is expected that in case of smooth functions compensation would show better results than for hard limit nonlinearities. This is the goal of future research.

It can be seen that all static characteristics of compensated nonlinearities are still odd symmetric. Thus, their describing function will depend only on gain. Describing functions of polynomial compensated nonlinearities are obtained using method described in [10], [13] and [14] are shown in Figure 13-Figure 16.

All describing functions are, as expected, positive and not dependent on frequency. Thus, a stability analysis of a system with compensated nonlinearity can be performed using linear stability analysis methods ([9], [12]).

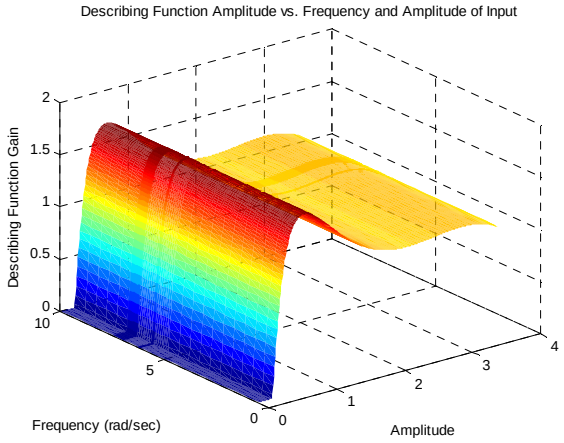


Figure 13. Describing function – compensated dead zone

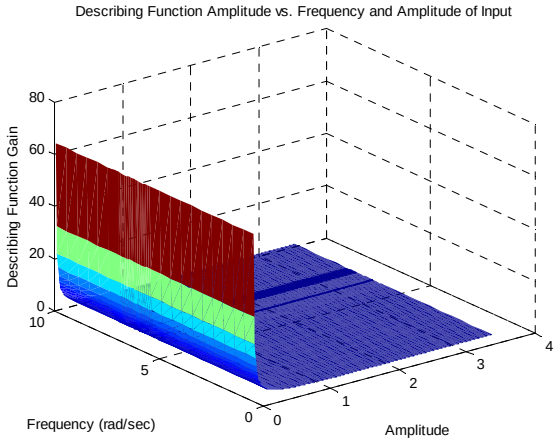


Figure 16. Describing function – compensated friction

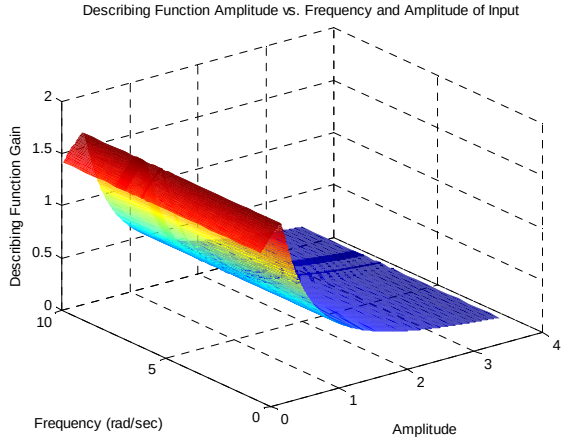


Figure 14. Describing function – compensated saturation

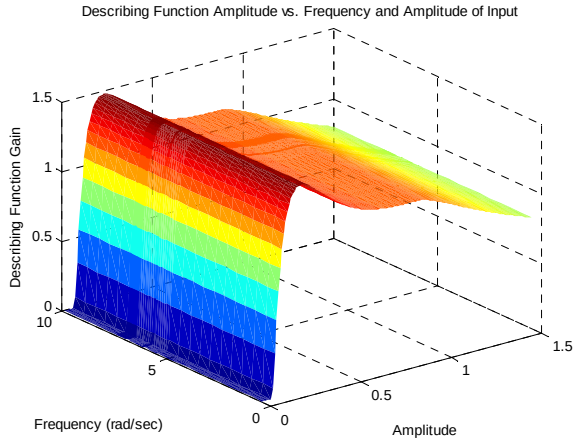


Figure 15. Describing function – compensated dead zone with saturation

V. CONLUSION

The paper dealt with static characteristics and describing functions of polynomial compensated odd symmetric nonlinearities. It was shown that polynomial compensated nonlinearity will be odd symmetric if the original nonlinearity is odd symmetric. That fact results in positive frequency invariant describing functions of compensated nonlinearities that allow for use of linear stability analysis methods.

REFERENCES

[1] Kreyszig, E. (1993). *Advanced Engineering Mathematics (7)*, John Wiley and Sons, New York, USA

[2] Novogranov, B.N. (1986). Determination of Frequency Domain Characteristics of Nonlinear Systems (in Russian), Masinostroenie, Moskow, Russia

[3] Kuljaca, O., Tesnjak, S., Vukić, Z. (1999). Describing Function of Mamdani Type Fuzzy Controller with Input Signals Derived From Single System Input and Singleton Ouput Membership Functions, *Proceedings of the 1999 IEEE Hong Kong Symposium on Robotics and Control*, Volume I, HKSRC '99, Hong Kong, 1999.

[4] Sijak, T., Kuljaca, O., Antonic, A., Kuljaca, Lj., (2005). Analytical Determination of Describing Function of Nonlinear Element with Fuzzy Logic, *Proceedings of the IEEE International Symposium on Industrial Electronics 2005*, ISBN 0-7803-8739-2, Dubrovnik, Croatia, June 2005

[5] Sijak, T., Kuljaca, O., Kuljaca, Lj. (2004). Describing function of generalized static characteristic of nonlinear element, *Proceedings of REDISCOVER 2004, Southeastern Europe, USA, Japan and European Community Workshop on Research and Education in Control and Signal Processing*, ISBN 953-184-077-6, Cavtat, Croatia, June 2004

[6] Sijak, T., Kuljaca, O., Kuljaca, Lj., (2007). Engineering Procedure for Analysis of Nonlinear Structure Consisting of Fuzzy Element and Typical Nonlinear Element, *Proceedings of the 15th Mediterranean Conference on Control and Automation*, Athens, Greece, July 2007

[7] Sijak, T., Kuljaca, O., Kuljaca, Lj., (2007). Computer Aided Harmonic Linearization of SISO Systems Using Linearly Approximated Static Characteristic. *Proceedings of EUROCON 2007 The International Conference on Computer as a Tool*. ISBN 1-4244-0813-X, Warsaw, Poland, September 2007.

[8] Sijak, T., Kuljaca, O., Tesnjak, S. (2001). Stability analysis of fuzzy control system using describing function method, *Proceedings of 9th Mediterranean Conference on Control and Automation*, ISBN 953-6037-35-1, Dubrovnik Croatia, June2001

[9] Slotine, J.J.E., Li, W. (1991). *Applied Nonlinear Control*, Prentice Hall, ISBN 0-13-040890-5, Engelwood Cliffs, New Jersey, USA

[10] Schwartz, C., Gran, R. (2001). Describing Function Analysis Using MATLAB and Simulink. *IEEE Control Systems Magazine*, Vol. 21, No. 4, (August 2001), pp. 19-26, ISSN 1066-033X

[11] Vukic, Z., Kuljaca, Lj., Donlagic, D., Tesnjak, S. (2003). *Nonlinear Control Systems*, Marcel Dekker, ISBN 0-8247-4112-9, New York, USA

[12] Netushil at al. *Theory of Automatic Control*, Mir, Moscow 1978

[13] Kuljaca, O., Horvat, K., Gadewadikar, J., "Describing function analysis of systems with NN compensated odd symmetric nonlinear actuators", *Proceedings of the 35th International Convention MIPRO 2012*, pp. 788-793, Opatija, Croatia, May 2012

[14] Kuljaca, O., Horvat, K., Gadewadikar, J., Tare, B., "Polynomial Compensation of Odd Symmetric Nonlinear Actuators via Neural Network Modelling and Neural Network Describing Function ", Engineering/Computing and Systems Research E-Conference CISSE 2012, International Conference on Industrial Electronics, Technology & Automation IETA 12, 2012

A new approach for maximum utilization of the mother steel plates for rectangular steel plates orders

Lasko Kasapinov*, Mile Jovanov** and Dimitar Trajanov**
* Open Mind Solutions LTD
St. Kliment Ohridski 20A-3
Skopje, Macedonia
** Faculty of Computer Science and Engineering
University Ss. Cyril and Methodius
st. Rugjer Boshkovikj 16
Skopje, Macedonia
lasko.kasapinov@gmail.com, mile.jovanov@finki.ukim.mk, dimitar.trajanov@finki.ukim.mk

Abstract – A solution to the steel slab design problem provides huge savings in the steel industry. Generally, this is a combinatorial optimization problem and it is NP-complete. In this paper a dynamic programming model is presented in order to provide an elegant optimal solution for one variation of the problem. The heuristic used in the model is presented as well. The computational results obtained on real-life instances of the problem are reported and compared with the results obtained on the same instances with the legacy optimization model, and they show a considerable improvement.

I. INTRODUCTION

There are many types of metal products produced in the steel industry. One of the produced types is rectangular steel plates. The process of production planning usually begins with the Book of Orders, i.e. list of orders that need to be met with the appropriate priority. The orders specify the dimensions and desired steel (quality) grade of the steel plates. On the other side, objects with cuboid shape and different dimensions called slabs are used as raw material in the process of production. The slabs are obtained by casting the molten steel and have their own chemical and mechanical structure. The slabs are processed further and subsequently rolled into large steel plates (mother-plates) with the same volume and steel grade as the slabs. Mother-plates are finally cut across their length into plates with dimensions and steel grade as required in the corresponding orders.

A major problem in the steel industry is combining of ordered steel pieces for maximum utilization of the material from the mother-plate, and maximum utilization of required resources. Resolving this issue is important because minimal improvements in the manufacturing process can result in huge savings. The combination of the assortment problem and the trim loss problem is known as the cutting stock problem (CSP).

The general problems associated with the standard one-dimensional cutting are known as NP-complete problems. In many cases this kind of problems can be modeled using mathematical programming and solution can be found using approximate methods or some heuristics. The goal is to design a plan for one-dimensional cutting of necessary pieces of material into many pieces of ordered products in such manner that the overall loss will be minimized while various requirements that may occur in practice will be satisfied.

There are several known attempts to resolve the problem of cutting pattern generation in the literature. The beginnings and the first definitions are set by Kantorovich since 1960. Scientific research has led to several formulations and solutions. The most famous formulation has been composed by Dyckhoff [4] in 1990. According to his theory, the problem can be classified based on four characteristics: size, type of task, the semi-product assortment (raw material) and product range. There is a strong connection between the cutting and packaging, so the same classification scheme applies to both categories of problems.

Many of the models and solution techniques for CSP have been used as useful starting points for this research. This paper continues with a literature study covering few different formulations and solution methods of the CSP in the steel industries, in section II and III. In section IV we present a formulation for the specific CSP and the legacy implemented solution of the problem in Section V. Section VI presents the new implemented algorithm for the problem. In section VII, the performances of the proposed algorithm are compared to the performances of the initially implemented one by analyzing the results obtained on real-life instances of a local large scale steel production company. Finally, the conclusions are presented in section VIII.

II. BASIC CONCEPTS OF CSP

The CSP is a type of matching problem on a sparse graph, allocating set of materials to a given set of orders. Each order contains information on the required quality,

the dimensions of the plates and the required number of pieces (these are the most important parameters for simpler modeling of our problem). Through the quality and some of its other attributes, each order limits its conceivable material i.e. slabs. Several orders can be produced from the same slab and there are no limitations on the number of slabs that can be requested. There are many other constraints that make the problem hard to solve, and the solution to such a problem depends on the size and diversity of the products and the technology applied in the production process.

The optimal grouping of orders as well as defining the optimal production should give answers to the following questions:

1. Which orders will be produced?
2. Which slabs should be rolled into necessary mother-plates to be further cut into ordered plates?

It is necessary to meet all the production constraints to answer the above questions and to address the following issues:

1. reducing the waste material i.e. achieving a minimal rest;
2. reducing the number of created "free" slabs (slabs that have no defined order);
3. increasing the capacity utilization (slabs strive to be as many as possible within the limits);
4. producing all the ordered products as scheduled;
5. increasing the number of emergency orders produced.

III. REVIEW OF THE RELATED WORK

Basically, three different classes of solution methods for solving CSP exist. Algorithmic methods guarantee to find the optimal solution for the problem. The drawback of the algorithmic methods is the computational complexity which often results into high running times, especially on larger instances. For this reason, purely algorithmic approaches for solving CSP have been rare until the past few years. On the other hand, heuristic methods hardly ever find the optimal solution, but usually generate an acceptable solution faster. A heuristic is considered acceptable if the found solutions are close enough to the known or believed optimal solution. Heuristic methods are also highly domain-dependent, in sense that they use information about the particular problems for which they are developed. Therefore, these methods usually cannot be used in solving even similar problems. Meta-heuristic methods are solution methods that orchestrate an interaction between local improvement procedures and higher level strategies to create a process capable of escaping from local optima and performing a robust search through the solution space (Hiller, 2010). These methods include procedures that employ strategies for overcoming the trap of local optimality in complex solution space and efficiently explore the search space to find optimal solution.

A. Linear programming

A solution of the problem can be obtained by use of Linear Programming (LP), and there are several such methods for finding solutions of CSP already. For example, Gilmore and Gomory [2] have suggested a "column generation" method which shows excellent results over a small number of subjects. However, since the problem relies on combinatorics, this method cannot be effectively applied to the steel plate production process where there exist wide range of products and many limiting factors. This model has to generate all possible solutions to ultimately choose the best solution, but it prolongs execution time significantly. Heinz et al. [5] have introduced an integer program solver to solve the problem by removing the naturally arising symmetries. Further, Gargani and Refalo [3] have stated that the models used for constraint programming approaches to this problem are basically linear models over binary variables and these models are not well suited to a constraint programming approach because of the limited domain reduction they produce. Integer programming solvers can tighten the formulation by adding cutting planes and they use the relaxed optimal solution to guide the search. Therefore, the authors, based on the structure of the problem, have introduced logical global constraints. Additionally, with their particular strategy for variables and their values, they have defined a particular heuristic for searching the results around specific combination i.e. Large Neighborhood Search (LNS). Moreover, by using certain static constraints in their model, they try to avoid creation of symmetrical solutions. Finally, they have got pretty fast solution, but a solution that is suitable for smaller instances. The solution that uses their model and limitations, leads to poor performance when searching for results in case of larger instances.

B. Delayed Column Generation

The problem can be solved more efficiently with certain modifications, i.e. the desired result can be achieved without having to create every possible combination. The idea is to use the Simplex method and the process of determining the next cutting pattern to be based on a basic pattern obtained by LP solution complemented with global and logical constraints. This is achieved exactly with the delayed column generation (DCG) technique which provides greater and more efficient propagation in the search for solutions. Before using DCG methods, it is necessary to drop the limitation on variables to be integers only. When the non-integer solution is found, the approximation (rounding) to integer can be achieved in few ways. For example, the proposed solution can be rounded down to the nearest integer. But if the solution is rounded to the lower integer, then a residual material will be created and that material must be filled with appropriate orders separately by some ad-hoc methods, otherwise the combination would not be acceptable and would not be an allowed solution.

According to Belov et al. [7], two points are important for the practical solvability of the column generation subproblem: the quality of cutting planes and a robust rounding heuristic. The heuristic helps to find an optimal solution long before cutting planes have forced integrality.

In 2006, Bettinelli et al. [6] have introduced linear relaxation of their model and have created an effective solution which is optimized with column generation and the lower bound found in this way is used in a branch-and-price algorithm.

C. *Sequential heuristic procedures*

The basics of heuristic approach are laid by Haessler in 1971, who is particularly concerned with problems arising in the paper industry. In his papers he has proposed that sequences of setting the product should be produced and generated using heuristics, and this would give fast but not exactly the most optimal solution. Therefore the speed is achieved but the proposal is not an effective one. With the aim to improve the solution, Haessler and Sweeney [10] have proposed a hybrid solution that basically is a delayed column generation method complemented with heuristics.

The prominent example of the "local heuristic search" class of procedures, we have analyzed the Sequential Heuristic Procedures (SHP) method. The SHP procedures outperform the capabilities of linear programming and can be used for solving more complex problems. These procedures create the decisions sequentially (one by one) by defining templates until all the necessary requirements are met and thus the smallest material waste is achieved. The SHP procedures not only allow control of the waste, but also control of other factors for which the standard LP procedure does not provide it.

Eshghi [8] has designed an Ant Colony Optimization (ACO) algorithm for solving the one-dimensional cutting problem. This algorithm is based on designed probabilistic laws and it is inspired from the ants' social behavior when they are looking for the shortest way. Eshghi stated that "meta-heuristics for solving the one-dimensional cutting problem will result in much desired answers with small solving time that could be a good basis for researchers' future research".

IV. MATHEMATICAL MODEL OF CSP

The definition of CSP represents correlation between orders, mother-plate and the corresponding slab. The width of the mother-plate is determined by the widest order, and before you allocate all orders in a mother-plate it is impossible to calculate its width and weight. On the other hand, there are many eligible slabs with different geometries for each order. So, for a given weight of a

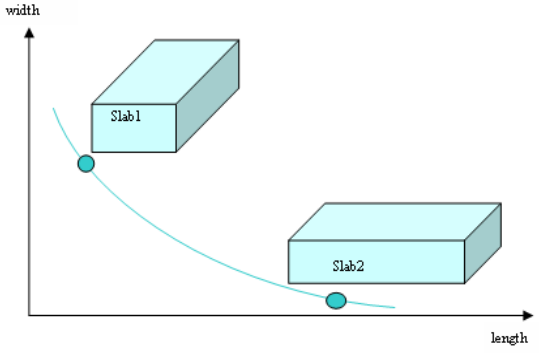


Figure 1. Eligible slabs with different geometries for each proposal

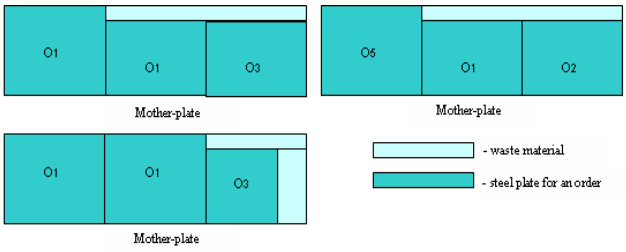


Figure 2. Correlation between mother-plate and orders

mother-plate, the length of the slab is strongly dependent on the width of the slab, as shown in Fig.1.

The trim loss strongly depends on the cutting plan of the mother-plate and the mother-plate can be cut into plates for suitable orders with or without trim loss. Examples of cutting plans with trim loss are shown in Fig.2.

The specific CSP that is being solved in this paper can be described in terms of inputs, outputs, and constraints for production design. The inputs to the problem are the following:

- List of orders with information about size of the rectangle, number of commissioned plates, quality grade, and constraints defined with specific chemical and mechanical properties;
- List of slabs with size of the rectangle, quality grade, and chemical and mechanical structure;
- Table of quality grades and their interdependence;
- Table of coefficients for material transformation from slabs to steel plates;
- List of values of approved dimension ranges for steel plates, values of several different ratios between dimensions of slabs and steel plates.

The output is a list of slabs and corresponding mother-plates along with the locations of the orders on each mother-plate.

There are few constraints in the real production process and they are described as follows:

- The weight of the mother-plate is equal to the weight of the used slab transformed by appropriate coefficient depending on the size of the mother-plate;
- The quality grade of the mother-plate is equal to the quality grade of the used slab;
- An order can be produced from slab with higher quality than the order's quality;
- The dimension of the mother-plate must be in the defined dimension range;
- Mother-plate length-to-width ratio must be in a predefined range.

We define the model as follows. Assume that there are n orders to be produced by a certain set of slabs. Orders o_1, o_2... o_n (*further referred directly with 1, 2 ..., n*) have a measured weight w_o. For a simpler presentation of

our model and in order to show the use of constraints, we will make the assumption that all constraints are replaced with just one "color constraint", and the new rule that should be satisfied is that one slab can be used for production of orders with up to two different colors. This modification can be perceived as a slight generalization of CSP and it is known as the knapsack problem with color constraint in the integer programming community [11]. Therefore, for each order o we define color c_o, where $o \in \{1, 2 \ldots, n\}$, and let's suppose that there are d different colors i.e. $c \in \{1, 2 \ldots, d\}$.

Let's suppose that we have m slabs $s_1, s_2 \ldots, s_m$ with u_s, $s \in \{1, 2 \ldots, m\}$, where u_s represents the measured weight (capacity) of slabs.

In a solution of CSP:

1. The total weight of the orders allocated in a slab must not exceed the weight of the slab.
2. The orders allocated in a slab can be of two different colors only.

Let the matrix X_{nm} represent the affiliation of orders to a slab, where each element $x_{os} \in \{0, 1\}$ and $x_{os} = 1$ iff the order o can be produced by the slab s.

The second matrix Y_{nm} with elements $y_{cs} \in \{0, 1\}$ represents the color of the orders, where $y_{cs} = 1$ iff the slab s contains at least one order with color c. According to the restrictions we can define that an order may be produced only by one slab. Therefore we define the matrix X'_{nm} which represents the final affiliation of orders to a slab, where each element $x'_{os} \in \{0, 1\}$ and $x'_{os} = 1$ iff the order o is proposed to be produced by the slab s. Finally, we have:

$$\sum_{s=1}^{m} x'_{os} = 1, \forall o \in \{1,2,..n\} \quad (1)$$

The weight of the orders must not exceed the weight of the slab (the coefficient for material transformation from slabs to steel plates is presumed to be one):

$$\sum_{o=1}^{n} w_o x'_{os} \le u_s, \forall s \in \{1,2,..m\} \quad (2)$$

If an order is proposed to be produced by a slab then it is obvious that the slab contains that order's color, and $x_{os} \le y_{cs}$, for every $o \in \{1, 2 \ldots, n\}$. The color constraint could be expressed by $\sum y_{cs} \le 2$, where $c \in \{1, 2 \ldots, d\}$, $s \in \{1, 2 \ldots, m\}$. The goal is to achieve the required production orders using less slab weight or in other words to produce orders from slabs while waste material is minimal:

$$\min(u_s - \sum_{o \in \{1,2,..n\}} w_o x'_{os}), \quad s \in \{1,2,..m\} \quad (3)$$

The solution we defined above is quite different from the basic model of CSP. For a clearer picture of the goal we seek, we have defined the rest of the slab and we are trying to minimize that unused weight of slabs. It is obvious that the lower limit of this minimization is 0 (zero) which means that we have achieved proposal for production of orders with the same weight as the weight of the allocated slab.

V. INITIALLY IMPLEMENTED SOLUTION

The current system in the company that we have initially implemented uses algorithm implemented in the Oracle PLSQL programming language and it is integrated in an ERP environment.

The implemented algorithm is based on LP procedures complemented with branch and bound and constraint satisfaction algorithms of constraint programming [9]. It starts by defining the relation between slabs and orders in such manner that provides information about which slab can be used for production of each order according to the rules. This selection prevents the algorithm from irregular combinations because in the next step the algorithm creates possible combinations of previously selected orders for each slab. Employed symmetry breaking constraints and some heuristics exclude undesirable combinations. The process of combining the orders uses the weight of orders to fill the slab until the full capacity of slab is reached or exceeded. Additionally, orders are combined starting with the biggest order – a heuristic that provides us with better propagation in searching. The algorithm uses greedy approach [12] and tries to fill the slab with one order first. In the next step, results are sorted and the best combination is selected for each slab. This approach finds the patterns and possible solutions in a sequential and iterative way.

The implemented algorithm generates solutions with high quality and provides large savings in the company [13]. But, it is a time-consuming process and the execution time is significantly prolonged when it is used on large instances. Therefore, we implemented additional heuristics that provide acceptable solution without the need for creating all possible combinations. In that way, we obtained the optimal solution to our CSP.

Our further research was aimed at improving the execution time and brought us to a better algorithm developed by using dynamic programming paradigm [14], as presented in the next section.

VI. THE PROPOSED ALGORITHM

The analyzed problem is a problem in combinatorial optimization. It can be decomposed into overlapping subproblems and it has an optimal substructure i.e. optimal solution to the given problem can be obtained by using optimal solutions to its subproblems. Therefore we can use methods of dynamic programming paradigm to find an optimal solution to the problem. At first, we simplify the problem by breaking it down into simpler subproblems and we try to find solutions to the subproblems. We use bottom-up fashion to search the possible combinations of orders [15]. Equation (3) makes us sure that the combination with minimal trim loss will be selected. In the next step, we construct the solution to the main problem using the solutions to all the smaller subproblems. The algorithm is implemented in an integrated information system using the Oracle PLSQL programming language.

We classify orders into groups of orders $g \in \{1, 2..., i\}$, and each group consists of orders which may be combined together to be produced from one slab. This means that each group consists of orders with the same quality, thickness and width. Let's define the array C_g of possible combinations of orders for each group of orders $g \in \{1, 2 ..., i\}$, and $C_g(l) \in \{0, 1\}$, where $l \in \{1, 2 ..., k\}$. Each element l of the array represents the total length of orders i.e. the length of the constructed mother-plate, and the capacity or weight of the mother-plate can be calculated easily because the other dimensions are already defined by the group where the orders belong to. Further, $C_g(l) = 1$ iff there is a possible combination of orders with total length equal to l. Initially, each array will have the value 1 for the zero element: $C_g(0) = 1$. This default value will be used later in the algorithm to create a combination for the order being considered. Next, each order from the appropriate group will search the array in reverse order from the highest element with value 1 in the array to the first element. For each element l that has value 1 in the array, we define a new element l' in the array and the value of l' will be set to one, i.e. $C_g(l') = 1$. The newest element in the array represents the total length of orders i.e. $l' = l + l_o$, where l_o is the length of the current order. We have at least one new element which is the length of the current order according to the predefined value $C_g(0) = 1$. As previously mentioned, each element in the array can be easily converted into the weight of combined orders i.e. the minimal weight of the slab required for production of the combined orders.

According to the production constraints it is not necessary to expand the arrays with all possible combinations of orders, but it is sufficient to perform combining the orders until the maximum allowable length of mother-plate is reached. Analogous to the above, some calculated values in the array can be excluded if the corresponding slab (a theoretical slab with weight equal to the weight of combined orders) has dimensions smaller than the minimum allowed. These restrictions are similar to the color constraint in the entry model of CSP, and the basic algorithm does not create irregular combinations of orders using matrix Y.

The next step is to find appropriate slabs. For that purpose, we calculate the weight of the orders in any proposed combination of orders (where the element has value 1 in array C_g). Further, for each match of suitable slab and element of the array that satisfy the rule defined in (2), the unused material of slab is calculated.

It is worth to mention here that in our extended model we use a cube structure (3D array) instead of one-dimensional arrays. The first dimension of the cube is the length of the constructed mother-plate i.e. the element l of the array C_g. The second dimension represents the various groups of orders, and the third dimension of the cube is used for storing values of C_g as well as values of other attributes of orders which are used for checking whether the real-life production constraints are met.

At the end, the best results are selected starting from the slab with the smallest waste and the appropriate combination of orders, as it is defined in (3). Additional procedure ensures the validity of the final affiliation of orders to a slab, as presented in (1).

The steps of the algorithm are summarized as follows:

```
S1: create group of orders (GO) and initialize array Cg
S2: create possible combinations of orders:
    for each GO loop
      for each order O in GO take length lo and loop
        get last element l of Cg with value 1
        loop (start searching Cg from the end)
          if lo + l < max_len_mother_plate then Cg(lo+l)=1
          get element prior to l
        end loop for Cg
        take next order
      end loop for orders
      next GO
    end GO loop
S3: find appropriate candidate slab
    for each slab S loop
      for each Cg with the same grade as the slab loop
        find the last element l of Cg
        loop
          if weight_of_orders <= weight_slab and
             constraints_are_satisfied
          then Xos=1, where o represents all orders from Cg(l)
          get element prior to l
        end loop
        get next Cg
      end loop for Cg
      get next slab
    end loop for slabs
S4: select the best results – define X'
```

VII. EXPERIMENTS AND RESULTS

In this section the computational performances of the initially implemented algorithm and the proposed algorithm are presented and analyzed. We have created an environment on PC Intel Pentium 3,4GHz, 2GB of RAM, Windows XP operating system and Oracle Database 10g, for the testing purposes. Both the initially implemented algorithm and the proposed one were implemented and compared in this environment.

The test data sets were obtained from the existing ERP system and they are real-life production data. To monitor the performances of the proposed algorithm under various conditions, the test instances were grouped together into three different-sized data sets (groups): a small data set, a medium and a large data set. The small data sets used for testing consisted of 300 slabs and 500 orders, the medium sets 1000 slabs and 1500 orders, and the large data sets 2000 slabs and 3000 orders.

The main purpose of the computational tests was to validate the newly proposed algorithm with actual data given by the company, to compare the two models specified above and to find the impact of the new approach on everyday production. Table 1 summarizes the characteristics of the real cutting stock data sets and presents adequate numerical results for these examples. The data set specifications (number and weight of slabs and orders) are rounded and the results are averaged within each group of data sets. The first column gives the identification of data set groups and the second column identifies the algorithm we have used. The next columns

TABLE I. PERFORMANCE OF CURRENT AND PROPOSED ALGORITHM FOR CSP

		Slabs		Orders		Used Slabs		Satisfied orders		Performance	
Data set group	**Algo rithm**	**Pcs**	**Weight (t)**	**Pcs**	**Weight (t)**	**Pcs**	**Weight (t)**	**Pcs**	**Weight (t)**	**Execution Time (s)**	**Trim loss (%)**
Small	Init.	300	1490	500	1600	220	1150	360	1130	25	1,74
	Prop.	300	1490	500	1600	260	1250	400	1230	22	1,6
Inter-mediate	Init.	1000	4890	1500	4700	710	3700	1200	3620	130	2,16
	Prop.	1000	4890	1500	4700	850	3950	1280	3870	110	2,03
Large	Init.	2000	9650	3000	9500	1300	7900	2460	7610	360	3,67
	Prop.	2000	9650	3000	9500	1690	8070	2560	7900	250	2,11

refer to quantity information about slabs and orders that have been used for testing purposes. Further, we show the number and total weight of used slabs, the number and total weight of satisfied orders as well as the total loss produced by the algorithm. The last two columns provide us with information about the quality of the obtained solutions in terms of algorithm execution time and produced waste material. It is noticeable that the proposed algorithm ***creates solutions with better utilization of available orders*** (weight of satisfied orders over weight of orders), the proposed design of mother-plates ensures ***better usage of slabs*** i.e. less trim loss, and the proposed algorithm ***takes less time*** than the initially implemented one. The results vary in different circumstances and there are bigger improvements in propagation and quality of solutions when larger data sets are used.

Table 2 provides information about the achieved improvements on satisfied orders and execution time with the proposed algorithm over the initially implemented algorithm. If we compare the obtained results, we can see that we have achieved improvement from 10% on execution time on small data sets, up to almost 30% better execution time on large data sets. Also, there is over 17% improvement in satisfied orders when the algorithm is used on large data sets.

VIII. CONCLUSION

In this paper we presented a new approach for maximum utilization of the mother steel plates for rectangular steel plates' orders. The model we have presented is based on dynamic programming paradigm and is more effective than the linear programming model complemented with global constraints. The results show the fact that a portion of production trim loss can be reduced by better planning and those results can be obtained faster with the proposed algorithm. The use of optimization techniques reduces waste materials which results in huge savings in steel industry. Furthermore, the better production planning provides better utilization of stock material, and improves the financial performance of the company by improving the investment level of the stock material.

TABLE II. NEW ALGORITHM RESULT IMPROVEMENTS

	% of improvements	
Group of data sets	**Satisfied Orders**	**Execution Time**
Small	8	10
Intermediate	11	16
Large	17	29

ACKNOWLEDGMENT

The research presented in this paper is partly supported by the Faculty of Computer Science and Engineering, at Ss. Cyril and Methodius University in Skopje.

The authors wish to express their gratitude to paper reviewers for the valuable suggestions for the improvement of the paper.

REFERENCES

[1] G. Dutta, "A survey of mathematical programming applications in integrated steel plants," Informs Published Online, October 2001

[2] P.C. Gilmore and R.E. Gomory, "A linear programming approach to the cutting-stock problem – Part II," Operations Research, pp.863-888 December, 1963

[3] A. Gargani and P. Refalo, "An efficient model and strategy for the steel mill slab design problem," Principles and Practice of Constraint Programming – CP 2007, Volume 4741, pp.77-89, 2007

[4] H. Dyckhoff, "A typology of cutting and packing problems," European Journal of Operational Research 44, 1990

[5] S. Heinz, T. Schlechte, R. Stephan and M. Winkler, "Solving steel mill slab design problems," Constraints, Volume 17, Issue 1, pp. 39-50, January 2012

[6] A. Bettinelli, A.Ceselli and G. Righini, "A branch-and-price algorithm for the two-dimensional level strip packing problem," 4OR, Volume 6, Issue 4, pp. 361-374, September 2007

[7] G.Belov, G.Scheithawer, "A cutting plane algorithm for the one-dimensional cutting stock problem with multiple stock lengths," European Journal of Operational Research 141, pp. 274-294, 2002

[8] K. Eshghi, "An ACO algorithm for one-dimensional cutting stock problem," Journal of Industrial Engineering International, Vol.1, No.1, September 2005

[9] S. Heinz, T. Schlechte, R. Stephan , "Solving steel mill slab problems with Branch and Price," ZIB-Report 09-14, May 2010

[10] Robert W. Haessler and Paul E. Sweeney, "Cutting stock problems and solution procedures," European Journal of Operational Research, Volume 54, pp. 141-150, 1991

[11] M. Dawande and J. Kalagnanam, "The multiple knapsack problem with color constraints," Citeseer, 1998

[12] G. Scheithauer and J.Terno, "A branch-and-bound algorithm for solving one-dimensional cutting stock problems exactly," Applicationes Mathematicae 23, Volume 2, pp. 151-167, 1995

[13] L. Kasapinov, M. Jovanov and D. Trajanov, "On the maximum utilization of the mother steel plates for rectangular steel plates orders," Proc. of Eleventh CIIT Conference, Bitola, Macedonia, 2014, (in print)

[14] M.Held and Richard M. Karp, "A dynamic programming approach to sequencing problems," Journal of the Society for Industrial and Applied Mathematics, Volume 10, No.1, pp. 196-210, 1962

[15] T. Ono and T. Ikeda, "Optimization of two-dimensional guillotine cutting be Genetic Algorithms," EUFIT, September 1998

Simulation Environment for Creating Artificial Range Data in Underwater Object Reconstruction

Ivan Aleksi*, Dieter Kraus**, Željko Hocenski*
*Josip Juraj Strossmayer University of Osijek, Faculty of Electrical Engineering, Osijek, Croatia.
**Hochschule Bremen, University of Applied Sciences, Institute of Water-Acoustics, Sonar-Engineering and Signal-Theory, Neustadtswall 30, D-28199 Bremen, Germany.
Ivan.Aleksi@etfos.hr, Dieter.Kraus@hs-bremen.de, Zeljko.Hocenski@etfos.hr

Abstract: This paper proposes a simulation environment for creating simulated ranging data of high resolution sonar systems. It enables the assessment of underwater object reconstruction techniques and the verification of various methods for automated target detection. As an input, the proposed environment uses the AUV's trajectory, a model of the sonar system attached to the AUV and 3D CAD models of underwater objects. As an output one can get a point cloud of the whole underwater object or 2D images, i.e. range data, which are smaller parts of a large object's 3D model. Noise is modeled and used as additional input in order to generate outputs that are closer to the real-world. The simulated data is processed and evaluated, where well known sonar signal processing methods are used to process the simulated acoustic signals. Automated target detection is achieved by processing the simulated 2D range data with advanced image processing techniques.

Keywords: Simulation Environment, 3D Models, Object Reconstruction, Automated Target Detection, Sonar.

I. INTRODUCTION

Underwater Inspection Mission (UIM) is usually carried out by a human divers and remotely operated vehicles (ROV). In order to automate UIM, Automated Unmanned Vehicles are used recently. Usually, AUV is moving underwater in a 3D space by following a predefined 6D trajectory of AUV configurations, i.e. 3D positions and 3D orientations, c.f. Fig. 1. Generally, a trajectory is planned in a manner that the surface under inspection is in the sonar scanning range. Sonar system is attached on an AUV and it enables range data calculation. As a scanning result, a large set of 3D points is obtained, a.k.a. point cloud. Point cloud enables surface inspection on a dry end of the system, i.e. on a computer's screen.

II. SIMULATION ENVIRONMENT

Signal processing in sonar systems is computationally intensive task. Computational complexity grows with increasing the number of time samples, the number of transducer elements and the number of antenna beams [1]. Various methods for sonar signal processing and their different implementations are compared in [1]. Proposed methods are implemented in MATLAB and C++ programming languages. In order to implement mentioned signal processing methods in C++, a special programming environment is used [2]. It reduces the implementation and verification time by combining MATLAB and C++ programming languages.

When using methods with large execution time values, simulation lasts long too. In order to speed up a range finding method, a straight forward approach is used in [3]. Instead of calculating range data with time consuming signal processing methods, a range finding is simulated by 3D model-based geometry approach. Range data is calculated as an intersection point between each 3D model's triangle and each antenna's beam, and it is done for each AUV's configuration, i.e. position and orientation. Intersection points are calculated in the antenna coordinate system and are transformed to the environment coordinate system. AUV trajectory planning is done in [3] by placing convex polygons around ship hull models and by placing straight lines beside harbor wall models. With known AUV's trajectory it is possible to transform calculated intersection points from the antenna c.s. to the referent c.s. Finally, a set of intersection points between antenna beams and model's triangles forms a point cloud.

Intersection point calculation is implemented in a GPU in order to speed up the reconstruction mission simulation [4]. Total speedup is between 100 and 200. GPU-based solution is involved in order to accelerate the execution time. Measured average speed-ups were ranging from 100 to 200.

Automated target detection is done in [5]. Firstly, a 20 cm edge length cubes are placed on the model's surface as anomalies. They are representing the so called Mine Like Objects (MLO). Automated MLO detection is done in [5] by projecting 3D range data to a 2D range image. Subsequently, image processing techniques are applied. Each 2D image correspond to a 10 m long scanned model part. Images are made with orthographic projection of a point cloud on a 2D trajectory-relative backplane.

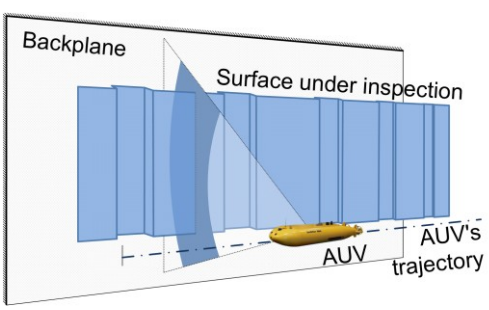


Figure 1. Underwater Inspection Mission (UIM) carried out by an Autonomous Underwater Vehicle (AUV) which is following predefined AUV's trajectory.

A. Point Cloud Reconstruction

Simulation model proposed in this paper can provide a user with a 4D point cloud, where each point consists of a 3D point $\mathbf{p}_W$ and 1D color. For the point cloud reconstruction, input parameters are: 3D environment model, 6D AUV trajectory and antenna parameters.

In this paper 3D models are made of N triangles, where each triangle has three 3D vertices. Therefore, the environment is described with the matrix of size $N \times 9$. As an example, the pyramid model is illustrated in Fig. 2.

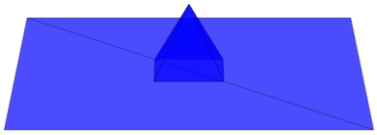


Figure 2. Environment model of a pyramid made of triangles.

Antenna parameters define the number of time samples $n = 1,\ldots,N$, the number of transducer elements $m = 1,\ldots,M$ and the number of beams $k = 1,\ldots,K$. In each of k beams, range finding method provides a time sample n_k from which the range r_k is calculated for corresponding k-th azimuth φ_k and elevation θ_k angle. From r_k, φ_k and θ_k a point cloud point is defined in polar coordinates. Transformation to Cartesian coordinates is done with eq. (1).

$$\mathrm{p_w} = \mathrm{r_k}[\cos(\theta_\mathrm{k})\sin(\varphi_\mathrm{k}) \;\; \cos(\theta_\mathrm{k})\cos(\varphi_\mathrm{k}) \;\; \sin(\theta_\mathrm{k})]^\mathrm{T} \qquad (1)$$

Parameter $\mathrm{w} = 1,\ldots,\mathrm{W}$ defines an index of a point in the point cloud ${}^2\mathbf{p}_W = [\mathrm{x_w} \;\; \mathrm{y_w} \;\; \mathrm{z_w}]^\mathrm{T}$ in the antenna's c.s.

AUV trajectory consists of AUV 3D position $\mathbf{p} = [x\ y\ z]^T$ and 3D orientation $\mathbf{a} = [\alpha\ \beta\ \gamma]^T$. Antenna c.s. can be relatively transformed concerning the AUV's c.s. It can be translated by the position $\mathbf{p}_A$ and rotated by the orientation $\mathbf{a}_A$. Let $\mathbf{R}(\mathbf{a})$ be the rotation matrix by orientation vector $\mathbf{a}$, and let ${}^2_0\mathbf{H}$ be the transformation matrix that transforms point cloud points ${}^2\mathrm{p_w}$ from the antenna c.s. S_2 to the points ${}^0\mathrm{p_w}$ in the referent c.s. S_0, c.f. eq. (2) and Fig. 3.

$$\begin{gathered} {}^0\mathrm{p_w} = {}^2_0\mathbf{H} \cdot [\,\mathrm{p_w^T} \;\; 1\,]^T \\ {}^2_0\mathbf{H} = {}^1_0\mathbf{H}\,{}^2_1\mathbf{H} = \begin{bmatrix} \mathbf{R}(\mathbf{a})\mathbf{R}(\mathbf{a}_\mathrm{A}) & \mathbf{R}(\mathbf{a})\,\mathbf{p}_A + \bar{\mathbf{p}} \\ [0\ 0\ 0] & 1 \end{bmatrix} \\ {}^1_0\mathbf{H} = \begin{bmatrix} \mathbf{R}(\mathbf{a}) & \bar{\mathbf{p}} \\ [0\ 0\ 0] & 1 \end{bmatrix}, \quad {}^2_1\mathbf{H} = \begin{bmatrix} \mathbf{R}(\mathbf{a}_\mathrm{A}) & \mathbf{p}_A \\ [0\ 0\ 0] & 1 \end{bmatrix} \end{gathered} \qquad (2)$$

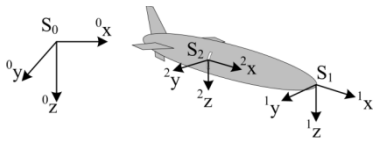
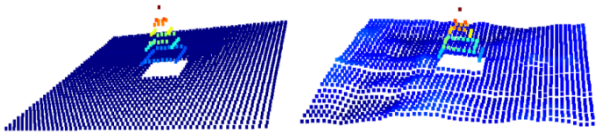


Figure 3. Referent coordinate system (c.s.) S_0, AUV's c.s. S_1 and antenna's c.s. S_2.

In this work, a noise is added to the simulation in two ways. Range noise is added to the calculated range $r_k' = \mathrm{r_k} + \xi_\mathrm{k}$. Reconstructed point cloud is illustrated in Fig. 4, in cases with/without the range noise.



(a) (b)

Figure 4. Reconstructed point cloud: without noise (a); with the range noise (b).

In order to correlate noise in neighbor points, a 2D low pass filtering operation is incorporated.

Orientation noise is added to the AUV orientation with two components: measureable and non-measureable noise. Measureable noise Ω_i is caused by the underwater environmental forces that are moving an AUV off the referent trajectory. Non-measureable noise Ω_i' is caused by the angular sensor measurement error. Finally, orientation $\mathbf{a}_i' = \mathbf{a}_i + \Omega_\mathrm{i} + \Omega_\mathrm{i}'$ is filtered and unwrapped to the interval $\pm\pi/2$. One can observe that the orientation noise cause changes in point cloud density, c.f. Fig. 5.

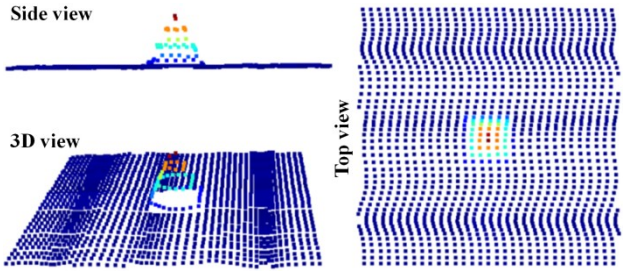


Figure 5. Reconstructed point cloud with added angular noise.

B. Range Image Reconstruction

Range image reconstruction is an orthographic projection of a 3D point cloud on a 2D image plane. It is done for a certain part of AUV's trajectory. In this work, a 10 m long trajectory is used for simulated imaging of one part of the underwater object. Main goal is to calculate a distance along a normal vector of a 2D image plane. Distance is calculated from a certain pixel position in 2D image to an intersection point with a 3D model, c.f. Fig 6.

Proposed projection is done in a trajectory-relative manner. Firstly, the centroid $\bar{\mathbf{p}}$ and mean orientation $\bar{\mathbf{a}}$ is calculated from one part of AUV's trajectory. If sonar imaging is placed in z,y-plane, with zero beam in y-direction, normalized vector $\mathbf{v}_1$ represents scanning direction and is defined with eq. (3), c.f. Fig. 7.

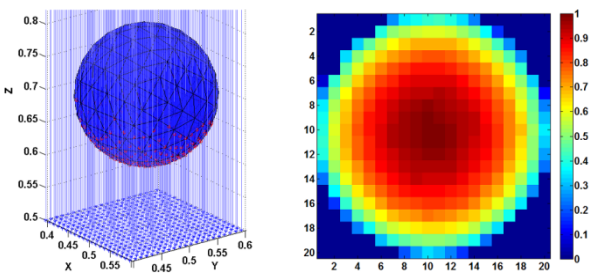


Figure 6. Orthographic projection of a 3D sphere model (a) on a 2D image (b) by calculating distance along perpendicular line between a certain pixel in 2D image and the intersection point with 3D model.

$$\mathbf{v}_1 = [\mathbf{R}(\bar{\mathrm{a}}) \cdot [0\ \ 1\ \ 0]^{\mathrm{T}}]^{\mathrm{T}} \tag{3}$$

Distance values $d_w^{'} = \mathbf{v}_1^{\mathrm{T}}({}^0\mathbf{p}_w - \bar{\mathbf{p}})$ is calculated between a point $\mathbf{p}_w$ and a plane placed in the centroid $\bar{\mathbf{p}}$ and the normal vector $\bar{\mathbf{v}}_1$. With distances $d_w^{'}$, mirrored image is represented since it is calculated from the antenna origin. In order to get the real image, distance values must be inverted $d_w = \max(d_w^{'}) - d_w^{'}$.

In order to get 2.5D height map, i.e. a range image, a point cloud rotation is required. Normal vector of a back-plane is set along the z-axis, i.e. in the $\mathbf{z} = [0\ \ 0\ \ 1]^T$ direction. Rotation is done by angle $a = \angle(\mathbf{v}_1, \mathbf{z})$, c.f. eq. (4), about the new axis $\mathbf{v}_2 = (\mathbf{v}_1 \times \mathbf{z})$, c.f. eq. (5) and Fig. 7.

$$a_{\mathbf{v},\mathbf{z}} = \angle(\mathbf{v}_1, \mathbf{z}) = \tfrac{\pi}{2} - \mathrm{atan2}(v_z, \sqrt{v_x^2 + v_y^2}) \tag{4}$$

$$\mathbf{v}_2 = (\mathbf{v}_1 \times \mathbf{z}) = \begin{bmatrix} \mathbf{x} & \mathbf{y} & \mathbf{z} \\ v_x & v_y & v_z \\ 0 & 0 & 1 \end{bmatrix} = [v_y,\ -v_x,\ 0]^T \tag{5}$$

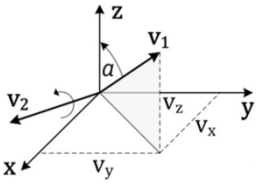
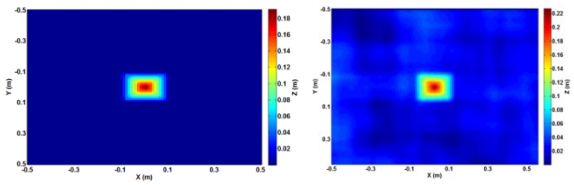


Figure 7. Rotation about vector $\mathbf{v}_2 = (\mathbf{v}_1 \times \mathbf{z})$ by angle $a = \angle(\mathbf{v}_1, \mathbf{z})$.

Proposed rotation is done with the axis-angle rotation matrix $\mathbf{R}(\mathbf{v}, a)$. In order to get the image borders in alignment with x, y-coordinates, additional rotation is required about the z-axis by the angle $\psi_b = \mathrm{atan2}(v_{1,y}, v_{1,x})$. Finally, rotated point cloud points $\mathbf{p}_{w,R}$ are defined with eq. (7).

$$\begin{aligned} \mathrm{p}^{'} &= \left(\mathrm{R_z}(\psi_\mathrm{b}) \cdot \mathrm{R}(\mathrm{v}, \mathrm{a})\right) \cdot {}^0\mathrm{p_w} \\ \mathrm{p_{w,R}} &= [\mathrm{p}^{'}(1)\ \ \mathrm{p}^{'}(2)\ \ \mathrm{d_w}]^{\mathrm{T}} \end{aligned} \tag{6}$$

Proposed techniques are used to reconstruct the 2D range image, c.f. Fig. 8, of the 3D point cloud which is illustrated in Fig. 4.



(a) (b)

Figure 8. Reconstructed range image from the point cloud illustrated in Fig 4.: without noise (a); with range noise (b).

C. Sonar Signal Simulation

Continuous wave pulse echoes are simulated using proposed 3D model-based simulation environment. In this work, carrier signal is defined with 8 periods of sinus function, carrier frequency $f_c = 1$ MHz, pulse duration $t_{pulse} = 8/f_c$ and the sampling frequency $f_s = 5$ MHz. Let $\mathbf{t}$ be time sample vector with time samples. Points ${}^0\mathbf{p}_w$ are intersection points between sonar's beams and model's triangles, while points $\mathbf{p}_m = [x_m\ \ y_m\ \ z_m]^T$ are centers of transducer elements in the antenna c.s., c.f. Fig. 9.

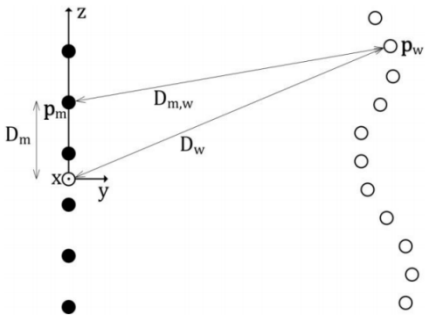


Figure 9. Transducer elements array (left) and the calculated point cloud points (right). Distances define time delays for CWP superposition.

Distance $D_w = |\mathbf{p}_w|$ is a distance between the antenna origin and an echoing point, while $D_{m,w} = |\mathbf{p}_w - \mathbf{p}_m|$ is a distance between an echoing point and a certain transducer element.

Echoes from a point cloud points are superimposed, as it is proposed in Alg. 1. Time delays $\tau_{m,w}$ and phase shifts $\varphi_{m,w}$ are calculated for M transducer elements and W echoing points (Alg. 1, Lines 1-2). If w-th echoing point is in the antenna range (Alg. 1, Line 3) and the CWP's time samples are inside the interval $[1, N]$ (Alg. 1, Lines 6-9), w-th echoing point amplitudes are superimposed to m-th channel signal (Alg. 1, Line 11). Amplitude A can be a constant value or a distance dependent function [6].

Algorithm 1. Superimposing sonar signal echoes from a point cloud.

Inputs: $f_c, f_s, \mathbf{t}, D_w, D_{m,w}, t_{pulse}, M, N, W, \vartheta_{max}, \vartheta_{min}, A, c$
Output: $\dot{x}$

1: $\tau_{m,w} = (D_{m,w} + D_w)/c$
2: $\varphi_{m,w} = \tau_{m,w} \cdot 2\pi f_c$
3: **For each** w
4: **If** $\mathrm{atan2}(z_w, y_w) \le \vartheta_{max} \ \wedge\ \mathrm{atan2}(z_w, y_w) \ge \vartheta_{min}$
5: **For each** m
6: $i_0 = [\mathbf{t}(1) \cdot f_s]$
7: $i_1 = \lfloor \tau_{m,w} \cdot f_s \rfloor - i_0$
8: $i_2 = \lfloor t_{pulse} \cdot f_s \rfloor + i_1 - i_0$
9: **If** $i_1 < 1 \ \vee\ i_2 > N$
10: **Go to** line 4
11: $\dot{x}(m, i_1{:}i_2) = \dot{x}(m, i_1{:}i_2) + A \sin\left(2\pi f_c \cdot \mathbf{t}(i_1{:}i_2) - \varphi_{m,w}\right)$

Sonar system with $N = 3378$ time samples, $M = 128$ channels, $K = 91$ beams, distance ranging from 5.5 m to 6.0 m, artificial for the 64th channel is shown in Fig. 10.

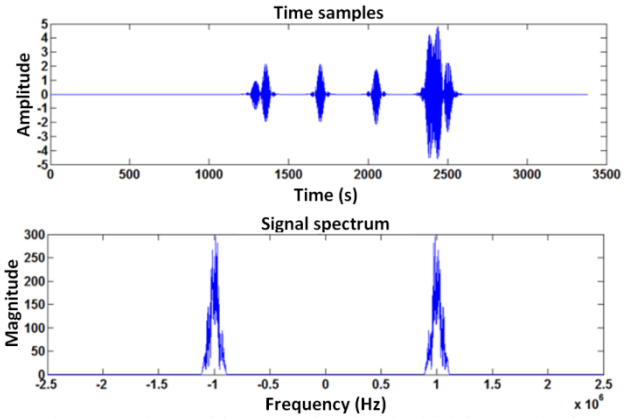
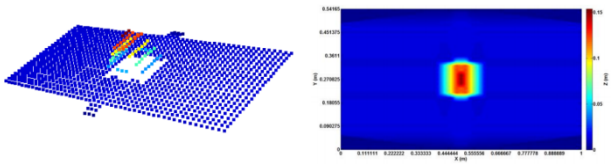


Figure 10. Time and frequency domain of artificial sonar signal.



(a) (b)

Figure 11. Reconstructed pyramid: point cloud (a); range image (b).

Simulated point cloud reconstruction of a pyramid model and the use of artificial signals is shown in Fig. 11. One can observe a different noise when compared with the ones described in Subsections II. A and II. B. Noise presented in Fig. 11 is caused by digitization errors and algorithmic imperfections.

D. Parallel GPU Implementation

CPU-based (central processing unit) simulation is time consuming and it sometimes last a few hours. Therefore, a GPU-based (graphic processing unit) simulation is used to accelerate the computationally intensive task of underwater reconstruction simulation [4]. Generally, speed-up is 50-100x without any optimization involved [7]. Proposed simulation environment has a speed-up of 150-250x. GPU implementation reduces the execution time of a simulation from 4 hours to a few minutes. Basic idea is to accelerate the intersection points calculation in parallel on a GPU, c.f. Fig. 12.

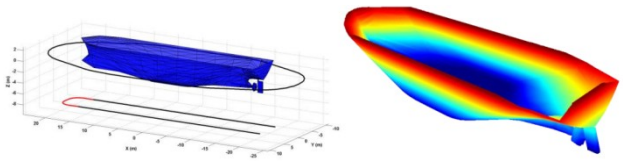


Figure 12. Ship hull model and AUV's trajectory (a); Reconstructed point cloud of the ship hull (b).

E. Automated Target detection

Main goal of an automated target detection is to detect anomalies on a ship hull during the underwater inspection mission [5]. Usually, a human technician analyses recorded images after the underwater scanning procedure ends. With the use of computers and acoustic vision it is possible to replace human vision. Automated target detection used in proposed simulation environment uses orthographic projection of a point cloud in order to create 2D images. Subsequently, 2D image processing algorithms are used to detect anomalies, e.g. MLOs (Mine Like Object). Top hat filter is used to detect MLO-like cubes with 20cm edge length. Thresholding technique is applied to segment detected MLOs from its surrounding background. For example, a harbor wall-like model with three MLOs is illustrated in Fig. 12 a, while detected MLOs are illustrated in Fig. 12 b.

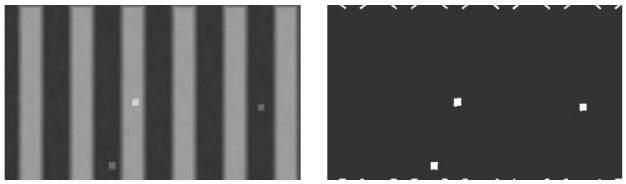


Figure 12. Ship hull model and AUV's trajectory (a); Reconstructed point cloud of the ship hull (b).

III. Conclusion and Discussion

In this work a simulation model is presented for creating artificial sonar data, 3D point clouds and 2D range images. Main goal of proposed simulation is to create the environment for automated target detection. Noise is modeled as a range noise and AUV's orientation noise, measureable and non-measurable. Range noise is simulating signal processing errors of a target range detection, while the orientation noise simulates AUV trajectory following. GPU accelerated parallel intersection point calculation is involved in order to reduce the simulation duration. With achieved speed up of 200x, a few hours long simulation of complete ship hull reconstruction is reduced to only a few minutes long simulation. Proposed simulation model is applicable for simulating other range data-based systems.

References

[1] Aleksi, Ivan; Hocenski, Željko; Dieter, Kraus: "Comparison of Acoustic Imaging Methods and their Variations with Sequential Implementations," 34th International Convention on Information and Communication Technology MIPRO2011, Opatija, Croatia, 23.-27.5.2011.

[2] Aleksi, Ivan; Kraus, Dieter; Hocenski, Željko:"Multi-Language Programming Environment for C++ Implementation of SONAR Signal Processing by Linking with MATLAB External Interface and FFTW," 53rd International Symposium ELMAR-2011, Zadar, Croatia, 12.-14.9.2011.

[3] Aleksi, Ivan; Dieter, Kraus; Hocenski, Željko: "Trajectory Planning Based on 3D CAD models for Autonomous Underwater Vehicle Inspection," 32nd Conference on Transportation Systems with International Participation, Automation in Transportation, pp. 108-113, Zagreb, Croatia 2012.

[4] Aleksi, Ivan; Kraus, Dieter; Hocenski, Željko: "CUDA-based parallel underwater target reconstruction," 31st International Conference Science in Practice SIP2013, Bremen, Germany, 27-29.10.2013.

[5] Aleksi, Ivan; Kraus, Dieter; Hocenski, Željko; Keser, Tomislav: "Simulated Surface Anomaly Detection in Underwater Object Reconstruction," Automation in Transportation 2013, Croatia, Osijek, 20-23.11.2013.

[6] Blondel, P.: The Handbook of Sidescan Sonar, Praxis Publishing Ltd, Chichester, UK, 2009. ISBN: 978-3-540-42641-7.

[7] Kirk, D. B.; Hwu, W.-M. W.: Programming Massively Parallel Processors: A Hands-on Approach (Applications of GPU Computing Series), Morgan Kaufmann, ISBN13: 978-0123814722, February, 2010.

Cooperative ramp metering simulation

Martin Gregurić, Edouard Ivanjko and Sadko Mandžuka
University of Zagreb Faculty of Transport and Traffic Sciences/Department of Intelligent Transportation Systems, Zagreb, Croatia
martin.gregurich@gmail.com, edouard.ivanjko@fpz.hr, sadko.mandzuka@fpz.hr

Abstract - The increase of vehicle numbers in recent decades resulted in road traffic congestion problems. Such congestions are a characteristic of densely populated urban areas and occur daily during morning and afternoon rush hours. Urban areas have been suffering from the lack of space needed to build new road infrastructure. The traffic congestion problem can be solved by applying new traffic control approaches from the domain of intelligent transportation systems (ITS). One of the applied methods from ITS is known as ramp metering and is used to increase the throughput of urban highways with many on- and off-ramps. Nowadays ramp metering is used in cooperation with additional control approaches like variable speed limit control (VSLC). Prior to implementation, such cooperative traffic control systems have to be tested in simulations using real world traffic data. One of the used simulators is CTMSIM which enables macroscopic simulation of highway traffic and local ramp metering approaches. In this paper the CTMSIM simulator is augmented to enable simulation of cooperative ramp metering algorithms, stand-alone VSLC, and cooperation between ramp metering and VSLC. Augmented simulator is tested using some limited available traffic data with the Zagreb bypass urban highway as a case study.

I. INTRODUCTION

Urban highways cannot fulfil originally projected high level of service (LoS) today anymore. LoS is defined as a group of qualitative measures which characterize operational conditions within traffic flow and their perception by motorists and drivers [1]. Main reason for low values of LoS on today's urban highways is high amount of traffic during increased daily (also known as rush hour) or seasonal traffic demand. In the same time urban highways are in most cases integrated into the urban road network and environment so there is no more space for infrastructural build-up.

On- and off-ramps on urban highways are common places for bottlenecks since they are the merging points of mainstream and on-ramp flow. Congestions at on-ramps become problematic if mainstream and/or on-ramps are overburdened with heavy transit traffic or/and a large number of local drivers. Local drivers prefer to use the urban highway instead of the local urban road network in time of daily migration [2]. Such a traffic scenario is common at most urban highways.

Regarding mentioned problems, new traffic control concepts are being applied to ensure desired LoS. On-ramp merging control method known as ramp metering and mainstream variable speed limit control (VSLC) are the most used control methods on urban highways [3]. Ramp metering and VSLC are important services used in intelligent transportation system (ITS) based traffic control architectures.

Ramp metering implementations are generally simulated by using various traffic simulation programs with the ability to simulate interaction of all existing traffic flows in a highway system. Traffic simulators are divided into two major categories: macroscopic and microscopic. Macroscopic simulators compute cumulative traffic flow parameters (e.g. speed, flow, and density) and their relationships to each other according to traffic flow equations. Individual vehicles are not considered. Most used macroscopic simulators are: CTMSIM, FREFLO, AUTOS, METANET and VISUM. Microscopic simulators continuously or discretely compute parameters (e.g. position, speed, maximum acceleration rate, etc.) of every individual vehicle during simulation. Most used microscopic simulators are: PARAMICS, MITSIM, CORSIM, VISSIM, AIMSUN, and TRANSIM.

This paper tackles the problem of cooperative highway traffic control systems simulation using a macro-simulation environment. CTMSIM is selected as a suitable macroscopic-highway simulator because of its open architecture based on Matlab. New functionalities and control approaches can be easily added and verified. In order to perform analysis and to validate the impact of cooperative control systems on highway traffic flows, it is necessary to adequately extend CTMSIM. In its original version it only supports local ramp metering algorithms without the possibility to simulate cooperative control approaches between several on-ramps or speed limit control (SLC) systems.

CTMSIM simulator is augmented to support SLC using a modified graphical user interface (GUI) to set a fixed speed limit value for every segment (cell) of the highway model. This value is used as the default speed limit value. Additionally, it is possible to change the speed limit value during simulation, which transforms SLC into variable SLC (VSLC). In order to enable cooperative ramp metering simulation, a hierarchical control simulation framework is developed and added to

CTMSIM. Using mentioned augmentation cooperative ramp metering algorithm, HELPER is implemented to verify the new framework. Augmented CTMSIM also enables simulation of cooperation between the implemented ramp metering algorithm HELPER and VSLC. Zagreb bypass highway is used as test case for evaluation of the developed hierarchical control framework and VSLC.

This paper is organized as follows. Second section briefly describes concept and problems which can be mitigated by application of ramp metering. Concept of cooperation between on-ramps and other highway control systems in order to achieve better results in ramp metering is presented in third section. Fourth and fifth sections describe basic features of the CTMSIM simulation environment and implemented augmentations. Sixth section contains simulation results of the comparative analysis of implemented ramp metering algorithms, VSLC and cooperative approaches. Paper ends with conclusion and future work description.

II. RAMP METERING

Common source of congestions on urban highways is merging of mainstream and on-ramp traffic flows as presented in Fig. 1. Area where these two types of traffic flows are actually interacting is known as downstream bottleneck. Consequences of highway congestions manifest through the following indicators: traffic demand exceeding road capacity, increased number of accidents and incidents, on-ramps queues that spill over into the urban arterial roads and peaks in traffic demand resulting from platooned vehicle entries from on-ramps [4]. Location of the downstream bottleneck close to the on-ramp and the spillback effect on the adjacent local urban road network is denoted in Fig. 1.

The main objective of ramp metering is to control queues at on-ramps in order to maintain maximal capacity flow on a highway. This is achieved by restricting access of on-ramp traffic to mainstream traffic with special traffic lights and appropriate signalization [5]. Visualization of complete installation of a local ramp metering system on a highway is shown in Fig. 2. Most important part of the ramp metering system is the algorithm which determines the "access rate reduction" for every on-ramp flow.

Generally it is possible to divide ramp metering algorithms in two major categories: local (or isolated) and coordinated. Local strategies take into account only traffic conditions on particular on-ramp and nearby segment of

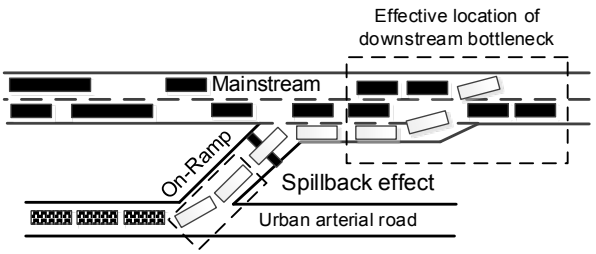


Figure 1. Graphical illustration of downstream bottleneck and spillback effect

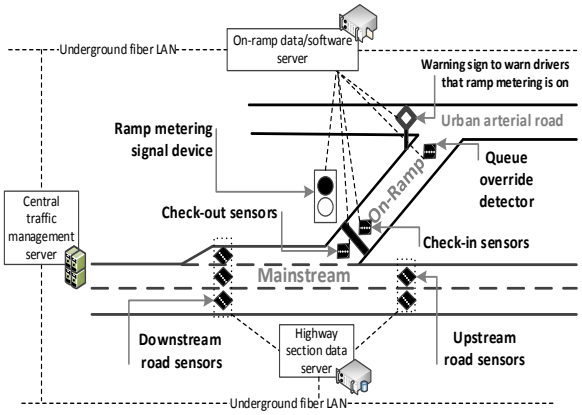


Figure 2. Basic ramp metering installation on a highway section

highway where they are applied. Most representative member of this algorithm category is ALINEA. Main working principle of ALINEA is to keep the downstream occupancy of the on-ramp at a predefined value using a closed loop control structure [6].

Coordinated algorithms consider traffic situation on the overall highway system. According to overall highway traffic state, metering rate for every on-ramp is adjusted. These coordinated ramp metering algorithms are further divided into cooperative, competitive and integrated algorithms. Simulation of cooperative algorithms is addressed in this paper and therefore they will be explained in the next section with more details. Competitive algorithms first compute two or more sets of metering rates using different algorithms types. The more restrictive one will be selected as the actually implemented rate. Typical representative of competitive ramp metering algorithms is SWARM. The SWARM algorithm has two major modules. First module performs local, while second module conducts global predictive ramp metering based on the same traffic control scenario. Metering rates for every on-ramp obtained by local and global algorithm are compared and smaller value is selected as final ramp metering rate [7]. Integrated algorithms are based on optimization of a specific LoS value while considering constraints such as maximum allowable on-ramp queue, bottleneck capacity, etc.

Some authors also make an alternative classification of real-time (traffic-responsive) and fixed-time ramp metering strategies [6]. Real-time ramp metering strategies are based on real-time traffic data obtained from highway sensors while fixed-time strategies use historical data without insight into the current traffic situation.

III. COOPERATIVE RAMP METERING

A cooperative control system is defined as a set of control entities that share information and/or tasks to accomplish a common, though perhaps not singular, objective. Cooperation in ramp metering was originally achieved by exploiting adjacent on-ramps queue capacity in order to perform effective mitigation of congestion related to a particular area with several close on-ramps. In such a highway network configuration, there exists the risk that mainstream congestion originating from an on-ramp can cause severe congestion that includes one or

more upstream on-ramps. Such congestion spread closes the access of mainstream vehicles to upstream off-ramps enlarging the congestion problem.

Latest work in cooperative ramp metering includes cooperation between classic cooperative ramp metering framework and other highway management services, driver information systems and vehicle itself [7]. In such a way, cooperation with additional highway traffic control systems can be established.

A. Cooperation between on-ramps

As mentioned, main problem with local or isolated ramp metering is a risk of fast upstream congestion propagation known as a "shock wave". This traffic scenario is common at a highway section with an on-ramp when local ramp metering cannot efficiently control the merging process between high traffic demands at on-ramp or/and in mainstream. This situation creates a mainstream "shock wave" which in relation to the motion of the vehicles travels backward or upstream. Upstream on-ramps with local ramp metering applied are not aware of downstream traffic conditions and therefore cannot provide an adequate solution to prevent the congestion spread.

Cooperative ramp metering algorithms work in two phases. In the first phase metering rate for each on-ramp is computed by its local algorithm. The second phase involves further adjustment of each local metering rate according to overall highway system information. Overall information involves traffic data from every on-ramp and its associated mainstream section traffic data. This phase is done in a cooperative unit. Main goal of mentioned cooperation is to achieve higher highway throughput by adjusting on-ramp metering rates upstream of the on-ramp affected by congestion.

HELPER was the first algorithm which used the mentioned cooperative ramp metering principle. It contains local traffic responsive metering algorithm and a cooperative unit. The cooperative unit uses data obtained from local ramp metering units to monitor the whole controlled highway section. If congestion or a long on-ramp queue is detected, it declares the belonging highway section with its on-ramp as master. Furthermore, several upstream on-ramps are declared as slave on-ramps. Slave on-ramps have the role to create virtual queues in order to reduce queue length on the congested on-ramp.

B. Cooperation with additional traffic control systems

Except cooperation between several on-ramps, additional cooperation can be established within the form of a superstructure containing several traffic control sub-systems. Such sub-systems are: VSLC, line prohibiting changes and various driver information services. In cooperation with the core ramp metering system they can provide better control over involved traffic flows. In this paper, emphasis is on cooperation between VSLC and ramp metering. VSLC is applied shortly before HELPER creates virtual on-ramp queues and ends when the congestion is reduced.

IV. SIMULATOR CTMSIM

CTMSIM is an interactive simulator based on macroscopic traffic models specifically designed for highway traffic flows simulations. Its original version supports several local and competitive ramp metering algorithms including on-ramp queue control. Highway model is based on the Asymmetric Cell Transmission Model (ACTM) and allows implementation and development of user-pluggable on-ramp flow and queue controllers [8]. In order to apply ACTM, the highway is divided into I cells. Basic ACTM architecture with marked variables is shown in Fig. 3.

Detailed ACTM used in CTMSIM is explained in [8] and in continuation basic equations are given. Simulation time is divided into K intervals with length Δt. Number of vehicles moving from cell i to cell $i+1$ (or mainstream flow) during interval k can be obtained according to (1):

$$f_{i[k]} = \min\{v_i(1-\beta_{i[k]})\left(n_{i[k]} + \gamma r_{i[k]}\right), w_{i+1}\left(\bar{n}_{i+1} - n_{i+1} - \gamma r_{i+1[k]}\right), F_{i[k]}\}, \quad (1)$$

$$F_{i[k]} \triangleq \min\{\overline{f_i}, \frac{1-\beta_{i[k]}}{\beta_{i[k]}}\bar{s}_i\}, \quad (2)$$

where $\beta_{i[k]}$ denotes split ratio for off-ramp flow, and γ represents on-ramp flow blending coefficient, both from interval [0, 1]. The blending coefficients define amount of traffic flow added or separated from mainstream traffic flow before its value is computed [9]. Furthermore, v_i denotes normalized free flow speed and $n_{i[k]}$ represents number of vehicles (or mainstream density) in cell i at time step $k\Delta t$. The parameter $r_{i[k]}$ denotes number of vehicle entering cell i, from its associated on-ramp at time step $k\Delta t$, while w_{i+1} represents normalized congestion speed in cell $i+1$. $F_{i[k]}$ denotes congested flow which leaves cell i, and $\overline{f_i}$ represents mainline capacity of cell i while $\bar{s}_i$ denotes off-ramp capacity in cell i. On-ramp flow values are now determined with constraints given by:

$$r_{i[k]} \leq l_{i[k]} + d_{i[k]}, \quad (3)$$

$$r_{i[k]} \leq \vartheta_i(\bar{n}_i - n_{i[k]}), \quad (4)$$

$$r_{i[k]} \leq c_{i[k]}, \quad (5)$$

where $l_{i[k]}$, denotes number of vehicles queuing in the on-ramp in cell i at time $k\Delta t$, and $d_{i[k]}$ represents demand for on-ramp in cell i. The parameter $c_{i[k]}$ denotes value of metering rate computed by the chosen ramp metering algorithm, while ϑ_i denotes on-ramp allocator parameter. On-ramp flow (metering rate) that can be merged with mainstream is obtained by:

$$r_{i[k]} = \min\{l_{i[k]} + d_{i[k]}, \vartheta_i\left(\bar{n}_i - n_{i[k]}\right), c_{i[k]}\}. \quad (6)$$

Equation (7) describes computation of off-ramp flow in cell i during time interval k:

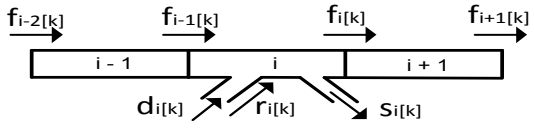


Figure 3. Basic ACTM architecture [9]

$$s_{i[k]} = \frac{\beta_{i[k]}}{1-\beta_{i[k]}} f_{i[k]}. \quad (7)$$

Mainstream density in the cell *i* during interval *k+1* can be computed from mainstream conservation equation:

$$n_{i[k+1]} = n_{i[k]} + f_{i-1[k]} + r_{i[k]} - f_{i[k]} + s_{i[k]}. \quad (8)$$

Mainstream speed in the cell *i* is obtained according to the following equation:

$$v_i^c = \min\left(\frac{f_{i[k]}/(1-\beta_{i[k]})}{n_{i[k]}+\gamma r_{i[k]}}\left(\frac{L_i}{\Delta t}\right), v_i^{ff}\right), \quad (9)$$

where v_i^{ff} represents free flow speed value for cell *i,* and L_i denotes length of cell *i*.

V. AUGMENTATION OF CTMSIM

In order to simulate cooperative ramp metering control, additional high level traffic control framework was developed as augmentation of the original CTMSIM simulator as presented in Fig. 4. The ACTM equations are modified to enable definition of a speed limit value for each cell separately.

A. Cooperative ramp metering module

In order to simulate cooperative highway control systems first step in CTMSIM augmentation was to enable cooperation in ramp metering. This is enabled through the cooperative ramp metering module. As it can be seen in the bottom part of Fig. 4 the original CTMSIM simulation sequence goes only through defined cells in a particular time step. In the proposed augmentation, additional simulation step is added at end of each time step. It is computed after all cells have been simulated. This additional simulation step is related to the cooperative ramp metering module and it has access to data from all cells so it can compute optimal local ramp metering rates and VSLC values used in the next time interval *k*.

First task of this cooperative ramp metering module is to monitor traffic conditions at every on-ramp during simulation. If critical density is exceeded in a current simulation time step, the central monitor module marks this cell's on-ramp as the "master" on-ramp. Metering rate at the "master" on-ramp is then decreased by 10 % of current traffic demand. Three upstream cells with on-ramps of the "master" on-ramp are declared as "slave" on-ramps. Metering rate at the "slave" on-ramps are decreased by 40 % of current traffic demand what creates virtual queues on them.

Second task of cooperative module is to enable override possibility for every metering rate, previously

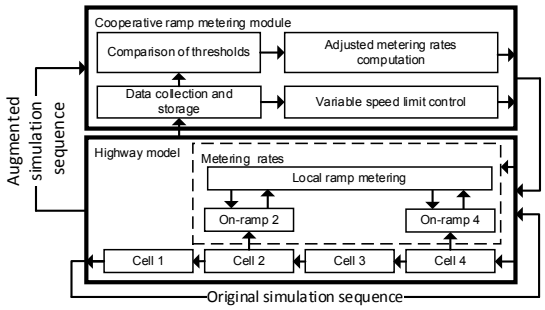


Figure 4. Augmented CTMSIM simulation structure

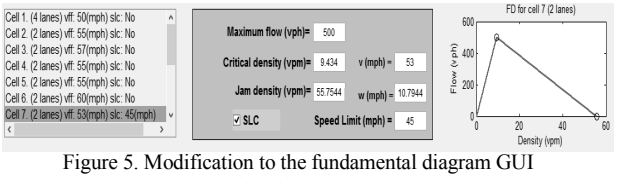


Figure 5. Modification to the fundamental diagram GUI to include the SLC option

computed by a local ramp metering algorithm. This feature is important for the HELPER algorithm which firstly computes metering rates for each on-ramp based on a local algorithm and then uses the cooperative module to obtain the final ramp metering value for each on-ramp.

B. Speed limit module

Second modification of CTMSIM involves implementation of SLC for every cell in the simulation model. SLC is implemented through modification of the cell mean speed equation given with (9) into:

$$v_i^c = \min\left(v_i^{SLC}, \frac{f_{i[k]}/(1-\beta_{i[k]})}{n_{i[k]}+\gamma r_{i[k]}}\left(\frac{L_i}{\Delta t}\right), v_i^{ff}\right), \quad (10)$$

where v_i^{SLC} denotes the current SLC value for *i*-th cell. SLC value must be lower than free flow speed value of the current cell. Original CTMSIM traffic fundamental diagram GUI is modified for definition of SLC parameters as presented in Fig. 5.

VI. SIMULATION RESULTS

To test the CTMSIM augmentations, different simulation scenarios were defined. Each traffic scenario includes increased traffic demand at the on-ramp in cell 6. Such scenario enables a simulation environment to test cooperative highway control strategies when congestion starts to build up. Obtained simulation results are divided into three groups. First group of simulation results contains comparative analysis of local ramp metering algorithms. Analysis is conducted using the following ramp metering algorithms: local (ALINEA), competitive (SWARM) and cooperative (HELPER). For comparison, simulation scenario without ramp metering control is also added to the analysis. Second group of simulation results are addressed to VSLC validation. Third simulation group presents cooperation of VSLC with the HELPER ramp metering algorithm.

A. Zagreb bypass and traffic flow data

Urban highway section between the nodes Jankomir and Lučko on the Zagreb bypass is selected to be the simulation use case model. Mentioned section contains 70 % traffic generated by the nearby town Zagreb [10]. Additionally, this section has increased traffic load during the whole day and also experiences a significant effect from daily migrations. Traffic demand for every on-ramp on this section is reconstructed according to Ljubljana bypass traffic flow data, since accurate hourly traffic flow data for the Zagreb bypass are not available. For this reconstruction Zagreb bypass average daily traffic flow from [10] and Ljubljana daily flow change characteristic were used. Graphical presentation of the ACTM model of highway section between nodes Jankomir and Lučko is presented in Fig. 6. Green colour represents on-ramps and blue colour off-ramps. First cell is denoted yellow.

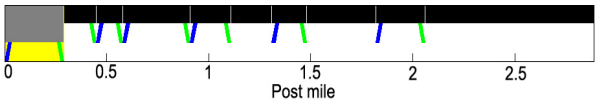


Figure 6. ACTM model of section between nodes Jankomir and Lučko

B. *Cooperation between on-ramps*

In order to compare obtained simulation results appropriate LoS measures have to be used. LoS measures used in this analysis are: travel time (TT) and delay. TT is a simple LoS measure which gives information how much time one vehicle needs to travel through observed highway segment. This measure is used to observe throughput of mainstream in the same time ignoring queues at on-ramps. TT is measured in minutes. Delay is defined as the difference between the actual time spent by all vehicles on a congested highway and the time spent in case they have travelled at free flow speed. Delay also considers vehicles which are waiting in on-ramp queues or in mainstream queues caused by the bottlenecks. It is measured in vehicle-hours.

Results of LoS analysis are shown in Tab. I. According to Tab. I. among the stand-alone ramp metering algorithms, cooperative algorithm HELPER has achieved best TT due to its restrictive nature. Lowest delay was achieved in the simulation scenario without ramp metering. In this basic scenario the blending coefficient γ was set to 1. This means that all vehicles from the on-ramps are immediately merged with mainstream under the condition that in a current cell maximal mainstream capacity is not exceeded. Such behaviour induces smaller on-ramps queues but significantly increases traffic density of mainstream which also increases average TT. These specific behaviours can be seen in Figs. 7 and 8.

HELPER maintains increased mainstream throughput by distributing vehicles and consequently the waiting time to the "slave" on-ramps queues. This behaviour causes longer queues at particular "slave" on-ramps increasing the delay, respectively. Additionally, "slave" on-ramps decrease the number of incoming vehicles from the upstream on-ramps regarding the current bottleneck location. This reduces impact of congestion back-propagation on the mainstream throughput which consequently decreases TT.

C. *Variable speed limit control*

Firstly, VSLC is analysed as a standalone highway traffic control strategy. Analysis includes comparison of simulation results obtained without and with applied VSLC at the same highway simulation model. VSLC is applied in the cells 2, 5 and 6. Main goal of VSLC is to

TABLE I. Results of cooperative analysis between different traffic control algorithms

Traffic control algorithm	TT (min)	Delay (vehicle-hour)
None	7.06	15.87
ALINEA	3.90	36.88
SWARM	3.71	41.49
HELPER	3.40	22.63
VSLC	5.59	12.24
HELPER + VSLC	3.30	21.50

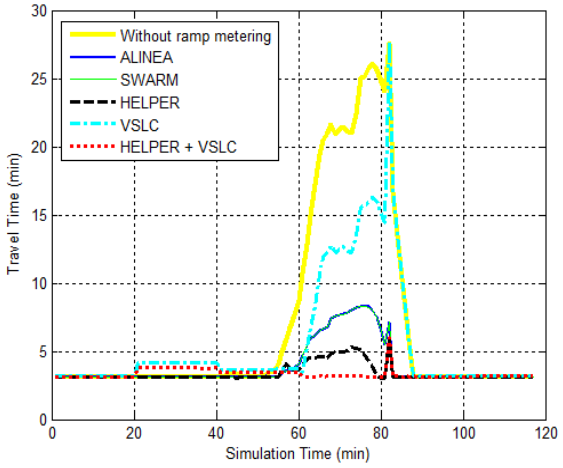


Figure 7. Results of comparative analysis according to travel time

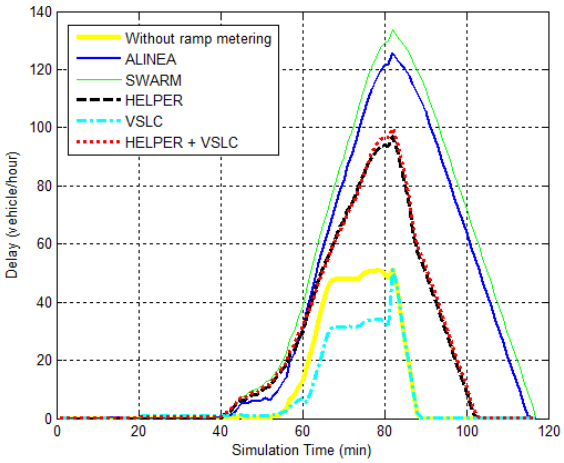


Figure 8. Results of comparative analysis according to delay

gradually decrease speed of the upstream flow before congestion starts. Such approach firstly gradually decreases mainstream speed, but enables higher mainstream speed during the congestion period unlike the scenario without VSLC.

Analysis of mainstream speeds can be seen in Fig. 9. VSLC prevented traffic standstill during most of the congested period. VSLC produces vehicle platoons which are traveling at lower speeds in comparison to the free flow speed by decreasing mainstream speed. This impacts the process of mainstream congestion back-propagation. From Fig. 10 it is possible to notice that mainstream density is lower at upstream cells of the congested cell 6

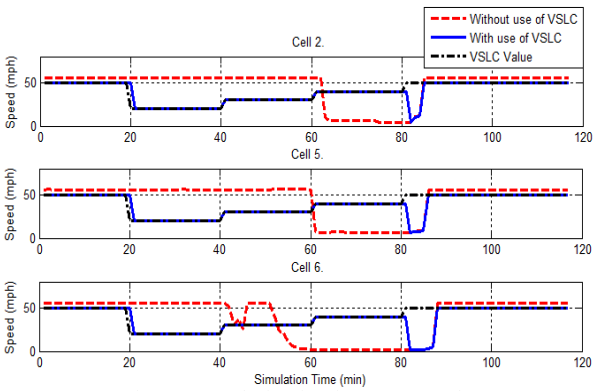


Figure 9. Analysis of mainstream speed with and without VSLC

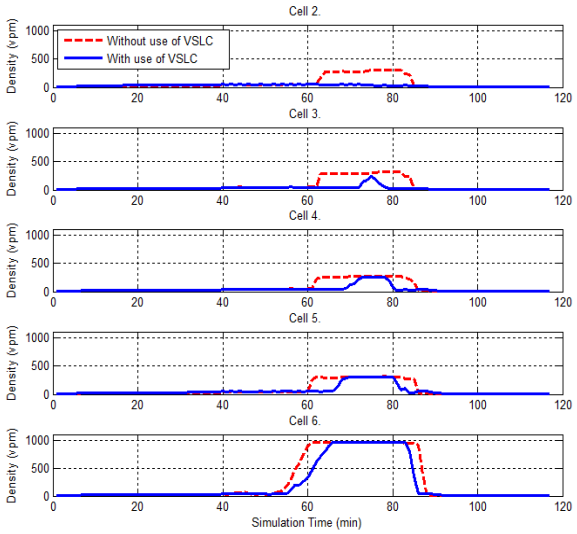


Figure 10. Analysis of mainstream density with and without the use of VSLC

with VSLC in contrast to the situation without VSLC.

D. Cooperation between on-ramps and VSLC

In this scenario cooperation between HELPER ramp metering algorithm and VSLC is simulated. Results of comparative analysis presented in Figs. 7 and 8 show that cooperation between HELPER ramp metering algorithm and VSLC produces smallest TT. VSLC decreases mainstream speed starting in the cell before the last "slave" on-ramp cell and ending in the congested cell. This decreases speed of the incoming vehicles to the area between last "slave" on-ramp and congested cell. Decreased speed of mainstream vehicles contributes to the lower values of density in the area between congested and last "slave" on-ramp. It can be concluded that virtual queues provided by cooperation between HELPER and VSLC significantly reduce traffic density upstream of the congested on-ramp. Lower upstream density of the congested on-ramp leaves additional mainstream capacity to accept vehicles which have origin from congestion back-propagation.

HELPER, as the standalone ramp metering algorithm and cooperation between HELPER and VSLC produce similar values of delay. The reason for such results is HELPER's reduction of metering rate on the "slave" on-ramps in order to reduce mainstream density upstream of the congested on-ramp. VSLC as the part of HELPER and VSLC cooperation does not have a strong influence on the delay as in the case of TT due to lower mainstream density. It can be concluded that VSLC has lower influence on delay if the mainstream density is decreased by HELPER exploitation of on-ramp queues. Standalone VSLC produces better delay results because of higher value of mainstream density. Other ramp metering algorithms produce higher values of delay due to higher number of vehicles in the congested on-ramp queue.

VII. Conclusion And Future Work

In this paper cooperative systems simulation in traffic control is tackled. Possibilities of cooperative control application in urban highway management are described. Cooperation between on-ramps in ramp metering, and between ramp metering and VSLC are chosen as most common systems. Simulation requirements are explained and framework for macroscopic highway traffic simulator CTMSIM augmentation is proposed. Proposed framework includes additional simulation step for cooperative control computations (between on-ramps and between ramp metering and VSLC) and augmentation of the ACTM highway model to include SLC. Both traffic control systems can be used as standalone or in cooperation.

To test implemented CTMSIM augmentations the ramp metering algorithm HELPER and VLSC are used. Simulations verify operational work of the proposed cooperative framework. All simulations are made using part of Zagreb bypass, the segment between nodes Lučko and Jankomir. Simulation results show that cooperation between VSLC and cooperative ramp metering provides best TT and delay values in comparison with other ramp metering algorithms involved into analysis.

Future work will include implementation of additional cooperative ramp metering algorithms and an optimization algorithm for cooperation between ramp metering and VSLC.

Acknowledgment

The research reported in this paper is partially funded by the FP7 - Collaborative Project: Intelligent Cooperative Sensing for Improved traffic efficiency - ICSI (FP7-317671) and by the EU COST action TU1102. Authors wish to thank Nikola Bakarić for his valuable comments.

References

[1] M. Treiber, and A. Kesting, "Traffic flow dynamics", Springer-Verlag Berlin Heidelberg, 2013.

[2] M. Gregurić, M. Buntić, E. Ivanjko, and S. Mandžuka, "Improvement of highway level of service using ramp metering", Proceedings of the 21st international symposium on electronics in transport – ISEP 2013, Ljubljana, Slovenia, 25. – 26. March, 2013.

[3] D. Li, and P. Ranjitkar, "Assessing Ramp Metering and Variable Speed Limits Strategies for Auckland Motorway", ITS World Congress Tokyo, Japan, 2013, Paper 4161, 2013.

[4] M. C. Esposito, J. W. Polak, R. Krishnan, and M. Pleydell, "A Global Comparison of Ramp-Metering Algorithms Optimizing Traffic Distribution on Motorways and Arterials", Road Transport Information and Control Conference and the ITS United Kingdom Members' Conference, 25. – 27. May, 2010. pp. 1 – 6.

[5] C. Lee, B. Hellinga, and K. Ozbayc, "Quantifying effects of ramp metering on freeway safety", Proceedings of 38th Conference on Accident Analysis and Prevention, Waterloo, Ontario, Canada, 2006, pp. 279–288.

[6] M. Hasan, M. Jha, and M. Ben-Akiva, "Evaluation of ramp control algorithms using microscopic traffic simulation", Transportation Research Part C 10, MIT, Cambridge, MA 02142, USA, 2010. pp. 229–256.

[7] M. Papageorgiou, C. Diakaki, V. Dinopoulou, Y. Wangand, and A. Kotsialos, "Review of Road Traffic Control Strategies", Proceedings of the IEEE, Vol. 91, No. 12, December, 2003, pp 2043–2067.

[8] A. Kurzhanskiy, and P. Varaiya, "CTMSIM Traffic Macro-Simulator for MATLAB", User Guide, TOPL Group, UC Berkeley, 2008.

[9] G. Gomes, and R. Horowitz, "Optimal freeway ramp metering using the asymmetric cell transmission model", Transportation Research Part C 14, 2006. pp. 244–262.

[10] G. Štefančić, D. Marijan, and S. Kljajić, "Capacity and Level of Service on the Zagreb Bypass", Promet - Traffic&Transportation, Zagreb, Croatia, Vol. 24, No. 3, 2012. pp. 261-267

Interpretation of divers' symbolic language by using hidden Markov models

Mario Menix, Nikola Mišković and Zoran Vukić

University of Zagreb Faculty of Electrical Engineering and Computing
Laboratory for Underwater Systems and Technologies
Unska 3, Zagreb, Croatia

E-mail: {nikola.miskovic, zoran.vukic}@fer.hr

Abstract - This paper reports the results of a system based on hidden Markov models (HMM) that is used to interpret both static and dynamic divers' hand signals using real time video feed. Two methods of collecting features that describe diver gestures are described and two types of HMMs are investigated: one based on discrete outputs variable distribution and the other based on mixture of Gaussians outputs variable distribution. For 8 basic diver gestures, 800 data samples were collected and the results were analyzed for the purpose of determining the most appropriate *i)* feature vector describing diver gestures, *ii)* HMM parameters (number of states and mixture components), and *iii)* type of the hidden Markov model. Finally, the quality of performance for recognition of each gesture is evaluated and analyzed, and necessary steps for improving the overall system are reported.

I. INTRODUCTION

Divers operate in a very challenging environment where even the slightest unexpected situation can have catastrophic results. Professional divers are usually involved in risky underwater operations some of which include surveying a part of a seabed (search and recovery missions), collecting samples (marine biology missions), documenting a site (marine archaeology missions) or performing inspections of underwater facilities (underwater inspection and maintenance missions). Having in mind the dangers that divers are exposed to during their activities, in the last couple of decades, underwater robots are being put to use more frequently. However, application of autonomous underwater vehicles or remotely operated vehicles still cannot completely replace a human diver, especially during highly sophisticated operations such as intervention, repair, etc.

The FP7 CADDY project (Cognitive Autonomous Diving Buddy) recognizes the necessity for a human diver and has a goal to develop a multicomponent robotic system, consisting of a surface autonomous marine platform and an autonomous underwater vehicle that will help divers during their underwater activities by performing a role of a diver "guide" (leading the diver through the underwater), diver "observer" (monitoring diver behaviour), and diver "slave" (helping the diver perform different set of tasks). In order to achieve these roles, establishing a communication channel between the diver and the buddy robot is of high importance. Since divers usually communicate using a predefined set of hand gestures, it is natural to use the same symbolic language as means of communication between divers and robots underwater.

Gesture recognition is an intriguing research area and different methods such as hidden Markov models (HMM), conditional random fields, particle filtering and condensation, finite-state machine as statistical modeling, optical flow, skin colour, connectionist model, etc. are being investigated, [5]. Among these methods, HMM has proved to be the most frequent tool, [2-7]. It has been successfully applied for spatial-temporal processes such as speech recognition [2], protein modeling [6] and gesture recognition [4].

This paper focuses on the application of the HMMs for interpretation of divers' hand signals using real time video feed. In this paper, we use the output mixed Gaussian distribution in HMM to improve the recognition rate.

The paper is organized as follows. The continuing part of Section II briefly describes the image processing system used to obtain gesture features. Section III focuses on the HMM theoretical background directly related to the applied approach. Section IV gives details on the specific problem of interpreting diver gestures and Section V describes experiments and provides analysis of the results. The paper is concluded with Section VI.

II. IMAGE PROCESSING SYSTEM

The image processing system has been mostly adopted from the publicly free-to-use system described in [1]. The software locates important hand features accurately and increases recognition tolerance to different hand rotations and tilt. A feature vector provided by the software as output includes three components: x and y coordinates of the palm center in the video frame, and the number of fingers that are stretched out and currently visible. All three of the mentioned features are of high importance for gesture recognition.

The research leading to these results has received funding from the European Union Seventh Framework Programme under FP7-ICT project "CADDY - Cognitive Autonomous Diving Buddy" Grant Agreement Number: 611373.

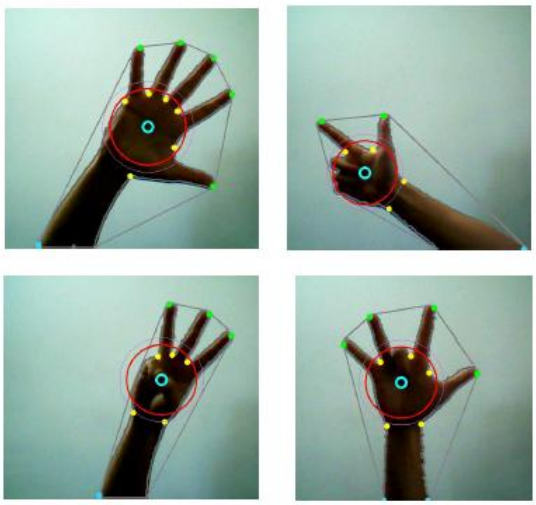


Figure 1. Snapshots from [1] showing hand images as recognized by the image processing system.

The software takes an RGB image captured by a web camera as an input and processes it to generate a binary image which provides enough information of hand contour. The binary image is used to calculate the contour and the convex hull of the contour. The palm position can be initially estimated by the information extracted from convex hull, and the fingertips' position can be detected from the hand contour. Examples of the results obtained from the image processing software are shown in Fig. 1.

III. HIDDEN MARKOV MODELS

Here we briefly review the theory of hidden Markov models (HMM) and some problems related to HMM training and human gesture recognition. Detailed description of the conventional HMM can be found in [2].

HMM is a statistical Markov model in which the system being modelled is assumed to be a Markov process with unobserved (hidden) states. HMM is characterized with *N* (number of states in the model, where individual states are denoted as *S*), *M* (the number of distinct observation symbols by state, where individual symbols are denoted as *V*) and $\lambda = (A, B, \pi)$ where,

$A = \{a_{ij}\}$ is the state transition matrix with

$$a_{ij} = P\left[q_{t+1} = S_j | q_t = S_i \right], \quad 1 \le i, j \le N , \quad (1)$$

$B = b_j(k)$ is the observation symbol probability distribution

$$b_j(k) = P\left[v_{k,t} | q_t = S_j \right], \quad 1 \le j \le N ; 1 \le k \le M , \quad (2)$$

and $\pi = \{\pi_i\}$ is the initial state distribution

$$\pi_i = P\left[q_1 = S_i \right], \quad 1 \le i \le N \quad (3)$$

where $S = \{S_1, S_2, \ldots, S_N\}$ is a set of states, q_t is a state at time *t*, and $V = \{v_1, v_2, \ldots, v_M\}$ is a set of symbols.

Given the observation sequence $\boldsymbol{O} = O_1 O_2 \ldots O_T$ and a model $\lambda = (A, B, \pi)$, the task is to efficiently compute the probability of the observation sequence $P(O | \lambda)$ with the given model (Problem 1), and to adjust the model parameters $\lambda = (A, B, \pi)$ to maximize $P(O | \lambda)$ (Problem 2).

The observation symbol probability distribution $P\left[v_{k,t} | q_t = S_j \right]$ can consist of discrete symbols or continuous variables. If the observations are discrete symbols, the observation model can be represented as a matrix in the form

$$B(i,k) = P\left[v_{k,t} | q_t = S_i \right]. \quad (4)$$

If the observations are vectors, it is common to represent $P(O_t | q_t)$ with a Gaussian:

$$P\left[v_{x,t} | q_t = S_i \right] = N\left(\boldsymbol{x} | \boldsymbol{\mu}, \boldsymbol{\Sigma} \right) \quad (5)$$

$$N\left(\boldsymbol{x} | \boldsymbol{\mu}, \boldsymbol{\Sigma} \right) = \frac{1}{\sqrt{(2\pi)^d}\sqrt{|\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})} . \quad (6)$$

A more flexible representation is a mixture of M Gaussians

$$P\left[v_{x,t} | q_t = S_i \right] = \sum_{m=1}^{M} P\left[M_t = m | q_t = S_i \right] \times N\left(x_{m,i} | \mu_{m,i}, \Sigma_{m,i} \right) \quad (7)$$

where M_t is a hidden variable that specifies which mixture component to use and $P\left[M_t = m | q_t = S_i \right] = C(i,m)$ is conditional prior weight of each mixture component. In our approach, we implement both discrete and mixture of Gaussians outputs variable distribution.

A. Solution to Problem 1

In order to calculate the probability of the observation sequence $\boldsymbol{O} = O_1 O_2 \ldots O_T$ with the given model $\lambda = (A, B, \pi)$ the most straightforward way of doing this is through enumerating every possible state sequence of length *T* (the number of observations), but fortunately there is more efficient procedure called *forward-backward* procedure. Solving this problem will provide us with the answer how to successfully recognize which trained HMM model (trained gesture) is the closest match to our observation sequence (gesture).

The general approach will be to iteratively update the weights. We do this by defining

$$\alpha_t(i) \triangleq P(\boldsymbol{o}_1 \boldsymbol{o}_2 \ldots \boldsymbol{o}_t, \; q_t = S_i) \text{ and} \quad (8)$$

$$\beta_t(i) \triangleq P(\boldsymbol{o}_{t+1} \boldsymbol{o}_{t+2} \ldots \boldsymbol{o}_T | q_t = S_i) \quad (9)$$

where $\alpha_t(i)$ is the probability that model is in state q_i at step *t* having generated the first *t* elements of sequence $\boldsymbol{O}$, and $\beta_t(i)$ is the probability that the model is in state $q_i(t)$ and will generate the remainder of the given target sequence, i.e., from $t+1 \to T$.

We can solve both for $\alpha_t(i)$ and $\beta_t(i)$ inductively (initialization – (10) and (12), induction – (11) and (13), and termination – (14)) :

$$\alpha_1(i) = \pi_i \cdot b_i(\boldsymbol{o}_1), \quad 1 \le i \le N \tag{10}$$

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i)\, a_{ij}\right] b_j(\boldsymbol{o}_{t+1}), 1 \le t \le T-1, 1 \le j \le N \tag{11}$$

$$\beta_t(i) = 1, \quad 1 \le i \le N \tag{12}$$

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(\boldsymbol{o}_{t+1})\, \beta_t(j), \; 1 \le t \le T-1, 1 \le j \le N \; . \tag{13}$$

We can also derive

$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i) = \sum_{i=1}^{N} \beta_1(i) = \sum_{i=1}^{N} \beta_t(i)\, \alpha_t(i) \; . \tag{14}$$

B. Solution to Problem 2

The most difficult task is to adjust the model parameters $\lambda = (A, B, \pi)$ to maximize $P(O \mid \lambda)$ because given any finite observation sequence as training data there is no optimal way of estimating the model parameters. This is the main issue when training specific HMM models for a specific gesture. But we can choose the model parameters so that $P(O \mid \lambda)$ is locally maximized using the *Baum-Welch* procedure (part of the expectation-maximization method).

First we define the probability of being in state S_i at time *t*, and state S_i at time *t+1* , given the model and observation sequence, i.e.,

$$\xi_t(i, j) = P(q_t = i, q_{t+1} = j | \boldsymbol{O}, \lambda) \; . \tag{15}$$

From definition of the forward-backward algorithm, we can rewrite $\xi_t(i, j)$ in the form:

$$\begin{aligned} \xi_t(i, j) &= \frac{P(q_t = i, q_{t+1} = j | \boldsymbol{O}, \lambda)}{P(\boldsymbol{O}|\lambda)} = \\ &= \frac{\alpha_t(i)\, a_{ij} b_j(\boldsymbol{o}_{t+1})\, \beta_{t+1}(j)}{P(\boldsymbol{O}|\lambda)} = \\ &= \frac{\alpha_t(i)\, a_{ij} b_j(\boldsymbol{o}_{t+1})\, \beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_t(i)\, a_{ij} b_j(\boldsymbol{o}_{t+1})\, \beta_{t+1}(j)} \end{aligned} \tag{16}$$

and the probability of being in state S_i at time *t* :

$$\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i, j) \tag{17}$$

It has been proven by Baum et al. in [11] that maximization of $Q(\lambda, \bar{\lambda})$ over $\bar{\lambda}$ leads to increased likelihood (19) what leads to the estimation of an HMM.

$$Q(\lambda, \bar{\lambda}) = \sum_{Q} P(Q|O, \lambda) \log\left[P(Q, O|\bar{\lambda})\right] \tag{18}$$

$$\max_{\bar{\lambda}}[Q(\lambda, \bar{\lambda})] \Rightarrow P(O|\bar{\lambda}) \ge P(O|\lambda) \tag{19}$$

A set of reasonable re-estimation formulas for A, B and π are:

$$\bar{a}_{ij} = \frac{\text{no. of transitions from state } S_i \text{ to } S_j}{\text{no. of transitions from state } S_i} = \frac{\sum_{t=1}^{T-1} \xi_t(i, j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \tag{20}$$

$$\bar{b}_j = \frac{\text{no. of times in state } j \text{ and symbol } \boldsymbol{v}_k}{\text{no. of times in state } j} \tag{21}$$

where $\bar{\pi}(i) = \gamma_1(i)$ is expected frequency in state S_i at time $t = 1$.

In other words if we define the reestimated model $\bar{\lambda} = (\bar{A}, \bar{B}, \bar{\pi})$, computed using (20) and (21), we have found a new model from which the observation sequence is more likely to have been produced. If we iteratively use $\bar{\lambda}$ instead of λ and repeat the estimation calculation, we can then improve the probability of the observation sequence from the model until some limiting point is reached.

IV. SYSTEM IMPLEMENTATION

In this section, we present details of our proposed model, applied to recognize 8 prime diver hand signals, as shown in Fig. 2.:

"Look at me" – static gesture with two fingers straight up pointing to eyes,

"I have a cramp" – dynamic gesture with the fingers bending to and from the palm,

"Going up" – static gesture with a thumb pointing up,

"Running out of air" – dynamic gesture with an open palm facing down and moving left and right,

"Turn around" – dynamic gesture with one finger pointing up and making a circle in the horizontal plane

"Danger" – extended arm with open palm moving left and right,

"Slow down" – dynamic gesture with an open palm facing down and moving up and down

"Stop" – static gesture with an open palm and fingers together pointing up.

A simplified scheme of the envisioned approach is given in Fig. 3. The first step is to detect the hand from the sequence of image frames. Using the image processing system the hand is recognized and the hand feature vector is obtained for the training and recognition process. The image processing system only extracts hand features when the hand is detected, therefore reduces the computation load when the hand is not detectable.

In this paper we observed HMM models that use two different feature vectors. In the first case, the feature vector, an observation in HMM, includes 3 components:

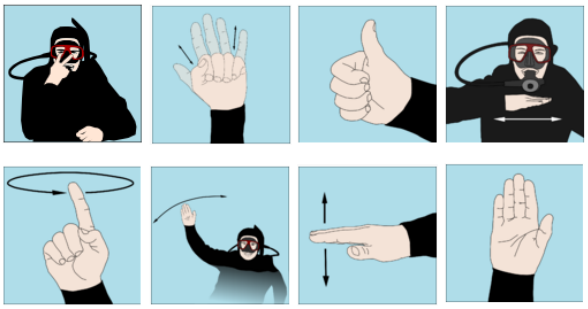


Figure 2. Diver hand signals used in the paper: "Look at me", "I have a cramp", "Going up", "Running out of air", "Turn around", "Danger", "Slow down" and "Stop" (source: Wikipedia)

$$O = (x, y, \text{number of fingers})$$

where x and y are coordinates of the palm centre in the video frame and the third component is the number of fingers stretched out and currently visible. This feature vector is given directly from the image processing system. In the second case, the feature vector includes two components:

$$O = (\text{orientation, number of fingers})$$

where the second component is the same as the third component in the first feature vector and the first component is orientation feature. The orientation is determined between two consecutive points from hand gesture path, using

$$\theta_t = \arctan\left(\frac{y_{t+1} - y_t}{x_{t+1} - x_t}\right) \quad t = 1, 2, \ldots T-1 \qquad (22)$$

where T represents the length of gesture path. The orientation θ_t is quantized in segments of 20° in order to generate codewords from 1 to 18.

As a part of research presented in this paper, we implemented three different training models, to test their efficiency and, based on the experimental results, choose the best among them. The models are chosen based on the form of the feature vector (three or two component feature vectors) and also on the type of output (and input) for the HMM (discrete or mixture of Gaussians outputs variable distribution). The models that have been taken into account are listed in Table I.

V. EXPERIMENTAL RESULTS

The experiments were conducted in such a way that for each of the 8 diver gestures reported in the previous section, 100 video samples were taken and used for the training dataset. For each sample, a total of 36 images were taken with the camera during a period of around four seconds (camera has a rate of 8.6 fps), meaning that every gesture is described using 36 feature vectors extracted from the images. All the experiments were implemented in Visual C++ (using OpenCV library) and Matlab R2009a. The used camera was a Truevision webcam (640x480 pixel image resolution).

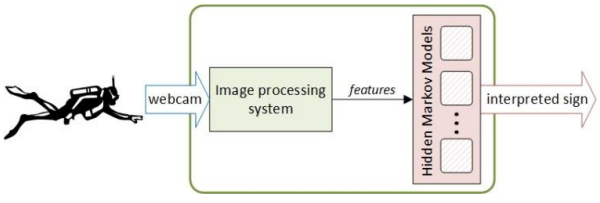


Figure 3. System scheme

TABLE I. HMM TYPES USED IN ANALYSIS

MODEL SHORT NAME	NO. OF COMPONENTS IN FEATURE VECTOR	OUTPUT TYPE
MG-2	2	mixture of Gaussians outputs variable distribution
D-3	3	discrete outputs variable distribution
D-2	2	discrete outputs variable distribution

The HMM parameters are derived through the training process. The training loop was programmed to end either when the maximum number of iterations has been achieved (5 in this particular case) or when the logarithm of likelihood does not increase.

This extensive set of collected data was used to determine the most reliable HMM model that can be used to determine diver gestures. To evaluate our approach overall, we carried out the following analysis from the obtained data:

- The first analysis aims to determine whether the recognition rate is higher using a model based on 3 or 2 component feature vector.
- The second analysis is to determine the best number of states and/or mixed Gaussians and to compare the recognition rate of the models in regards of their outputs variable distribution, whether it is mixture of Gaussians or discrete.

A. *First analysis – choosing the feature vector*

In order to determine whether the proposed feature vector containing two (orientation of motion and number of fingers) or three (hand position and number of fingers) components gives better performance, models D-2 and D-3 with discrete outputs variable distribution were observed. For the sake of generalization, multiple experiments were analyzed for the cases with different number of states Q. The results are shown in Fig. 4.

As the experiment shows, regardless of the number of states, better recognition is achieved using the D-2 model. The recognition rate using the D-2 model is around 50%,

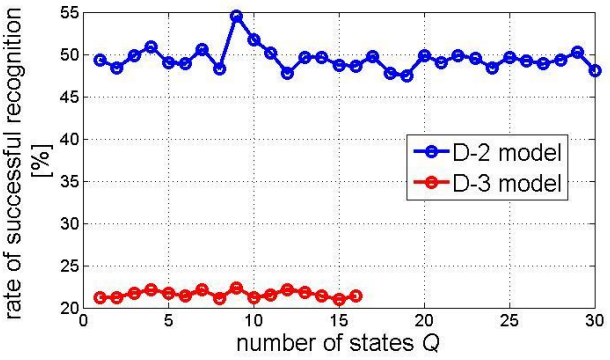


Figure 2. Rate of successful recognition for D-2 and D-3 models.

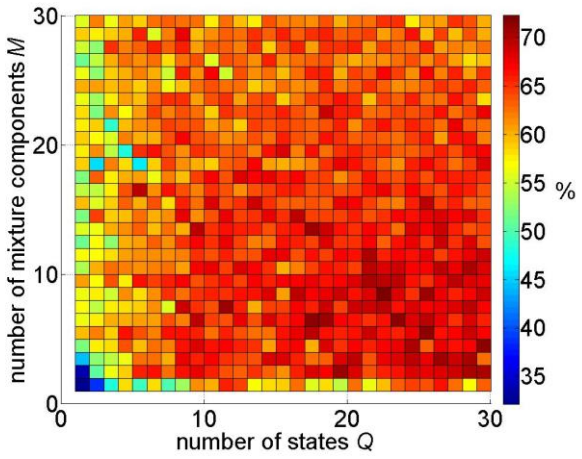


Figure 3. Rate of successful recognition for the MG-2 model.

peaking at 54% with a HMM with 9 states, where for the same number of states the recognition rate for the D-3 model is barely above 22%, with an average 21.5%.

The conclusion raised from this experiment is that models using two-component feature vector give significantly better results.

B. *Second analysis – choosing the number of states and/or number of mixture components, and outputs variable distribution*

This experiment was performed in order to determine which model has better performance relative to the output variable distribution, the one that uses discrete or mixture of Gaussians. Models D-2 and MG-2 were tested. These two models significantly depend on the model parameters.

The D-2 model depends on the number of states Q. As it was shown in the previous experiment (see Fig. 4), the best recognition rate of 54% was obtained with Q=9.

The MG-2 model, on the other hand, depends on two parameters: M – the number of mixture components and Q – the number of states. Both parameters were varied in the region from 1 to 30. The results of recognition rate are shown in Fig. 5. The MG-2 model has produced a recognition rate of 72%, improving the recognition rate of the D-2 model by almost 20%. It is clear from Fig. 5 that better results are acquired for a higher number of states and the number of mixtures being around 10.

C. *Full diver symbol verification*

The previous analysis has shown that the best results are obtained using a mixed Gaussian HMM model with M=8 and Q=22. For the sake of full verification of the system, we performed verification of the model for all 8 gestures. The comparison between the D-2 and MG-2 model is summarized in Table II and Table III where rows represent the commanded gesture, and the columns are the interpreted gesture. For the sake of clarity, we entered only the values for the dominant interpretation.

In the case of D-2 model, the gesture "Slow down" appears most often as an interpretation of the commanded gesture. The reason for this is that during this gesture the whole palm is not visible, therefore the image processing system cannot accurately calculate the position of the hand nor the number of fingers. Gestures "Cramp" and "Stop" are very similar when observing the component orientation. Their main distinguisher should be the component number of fingers, which is, as explained in the image processing system, a component that is not

TABLE III. RATE OF SUCCESSFUL RECOGNITION PER GESTURE FOR THE D-2 MODEL.

		INTERPRETED GESTURE							
		"Look"	**"Cramp"**	**"Going up"**	**"Out of air"**	**"Turn around"**	**"Danger"**	**"Slow down"**	**"Stop"**
COMMANDED GESTURE	**"Look"**							95%	
	"Cramp"		67%						
	"Going up"			33%				40%	
	"Out of air"				50%		37%		
	"Turn around"					70%	28%		
	"Danger"						90%		
	"Slow down"							80%	
	"Stop"		28%					28%	40%

TABLE II. RATE OF SUCCESSFUL RECOGNITION PER GESTURE FOR THE MG-2 MODEL.

		INTERPRETED GESTURE							
		"Look"	**"Cramp"**	**"Going up"**	**"Out of air"**	**"Turn around"**	**"Danger"**	**"Slow down"**	**"Stop"**
COMMANDED GESTURE	**"Look"**	~ 45%							
	"Cramp"		>95%						
	"Going up"			~ 60%					
	"Out of air"				>95%				
	"Turn around"					>95%			
	"Danger"						>95%		
	"Slow down"							~ 70%	
	"Stop"		~ 70%						13%

being accurately calculated and therefore in need of improvement. The gesture “Out of Air”, like the gesture “Danger”, are the only two gestures with left-right movement, and so the model has trouble distinguishing between them. The similar scenario appears for the gesture "Turn Around", since the movement of the hand in a circle trajectory is transferred to a left-right movement on the 2D camera frame. Again, the main distinguisher should be the component number of fingers.

In the case of the MG-2 model, the situation has significantly improved for all gestures except for the "Stop" gesture where the accuracy of interpretation decreased to 13% with the most cases of interpretation as the "Cramp" gesture.

To sum up, while the DP-2 gives somewhat reliable interpretation of only 4 out of 8 trained gestures, the MG-2 model gives high chances of correct recognition for 6 gestures and misinterprets only one gesture.

D. Comments on the experiments

From the presented results, it is clear that the MG-2 model gives the best results, compared to other tested models. However, a few comments should be stated.

During our research, we also covered the ergodic and LR (left-right) HMM. For the ergodic model, a connection can be made between any two different states, while in the LR model only a forward connection can be made, i.e. once a state has been passed it could not be achieved again. Our experiments showed no significant difference between these two models so all above mentioned experiment results are ones achieved using the ergodic model.

Another important parameter in the training of the HMM is initial value of the parameter matrices that have been set to random values in the experiments presented here.

It should also be mentioned that in some cases (e.g. when training the D-3 model with 17 or more states) matrices describing the HMM would become singular and model parameters could not be calculated.

The covariance matrix is another parameter that should be taken into consideration when analyzing the quality of the HMM. There are three different type of covariance matrix: diagonal, spherical and full. Experiments have shown that using the diagonal covariance matrix with more mixture components produced better results in comparison to the full covariance matrix with fewer mixture components. As a result, the diagonal covariance matrix was used and its value was kept constant during the experiments.

VI. CONCLUSIONS

This paper describes a HMM based system for recognizes diver signals from colour video image sequences. Our database contains 100 video sequences for each isolated gesture, as well as 100 test video sequences for each of the 8 prime diver signals.

We have shown that the HMM with the mixture of Gaussians outputs variable distribution and a two component feature vector gives the best performance among the tested models. The results show an average of 72% accuracy.

Although the system showed some good features there is room for improvement. The system is suitable for real-time application, however some modifications and additional research effort has to be invested in order to apply it to the underwater.

In the conducted experiments, the camera was fixed in from of the test subject. Since this methodology has to be applied on a diver, by observing her/him from a moving underwater vehicle, the approach to recognizing gestures had to be changed. Some research on compensation of the observing vehicle movements has been reported in [12]. Our future research will focus on measuring hand motion relative to the divers head, i.e. in the divers coordinate system. It is expected that this would solve the problem of a moving observation underwater vehicle.

Further on, the presented approach will be extended to the detection of dynamic gestures that are performed using both hands, thus extending the interpretation of the diver vocabulary.

In addition to that, we will investigate the possibility of using stereo-cameras and high resolution sonars to obtain the highest possible hand reconstruction quality, thus improving the image processing algorithms.

REFERENCES

[1] Chen, Wei-chao, "Real-Time Palm Tracking and Hand Gesture Estimation Based on Fore-Arm Contour." Master thesis, University of Science and Technology, Taiwan, 2011

[2] Rabiner, L. et al., *Fundamentals of speech recognition*, Prentice Hall, New Jersey, 1993

[3] Ćurković, P., Prepoznavanje govora korištenjem skrivenih Markovljevih modela. University of Zagreb, Zagreb, 1999.

[4] Elmezain, M. et al., "A hidden markov model-based continuous gesture recognition system for hand motion trajectory." Pattern Recognition, 2008. ICPR 2008. 19th International Conference on. IEEE, 2008.

[5] Nguyen-Duc-Thanh, N. et al. "Two-stage Hidden Markov Model in Gesture Recognition for Human Robot Interaction." Int J Adv Robotic Sy 9.39 (2012).

[6] Krogh, A. et al. "Hidden Markov models in computational biology: Applications to protein modeling." Journal of molecular biology 235.5 (1994): 1501-1531.

[7] Chen, Weiying, et al. "Real-time 3d hand shape estimation based on image feature analysis and inverse kinematics." Image Analysis and Processing, 2007. ICIAP 2007. 14th International Conference on. IEEE, 2007.

[8] Chang, Y.H. & C.M., Automatic Hand-Pose Trajectory Tracking System using Video Sequences., Master Thesis, Department of Information and Computer Engineering, Chung-Yuan Christian University, Chung-Li, Taiwan, 2010

[9] Yin, Xiaoming, and Ming Xie. "Finger identification and hand posture recognition for human–robot interaction." Image and Vision Computing 25.8 (2007): 1291-1300.

[10] Nam, Yanghee, and K. Wohn. "Recognition of space-time hand-gestures using hidden Markov model." ACM symposium on Virtual reality software and technology. 1996.

[11] Shotton, Jamie, et al. "Real-time human pose recognition in parts from single depth images." Communications of the ACM 56.1 (2013): 116-124.

[12] Buelow, Heiko, and Andreas Birk. "Gesture-recognition as basis for a human robot interface (hri) on a auv." OCEANS 2011. IEEE, 2011.

Cooperative Control of Heterogeneous Robotic Systems

N. Mišković, S. Bogdan, I. Petrović and Z. Vukić
Department of Control And Computer Engineering
Faculty of Electrical Engineering and Computing
University of Zagreb
Unska 3, Zagreb, Croatia
E-mail: {nikola.miskovic, stjepan.bogdan, ivan.petrovic, zoran.vukic}@fer.hr

Abstract – Establishing a cooperative heterogeneous robotic system comprising autonomous vehicles operating in air, on ground and on/under water is a challenging task that can significantly contribute solving complex problems related to environmental monitoring (e.g. prediction, prevention and recovery from natural disasters) and security (e.g. border control, fight against terrorism). This paper presents a short overview of a project that involves cooperative control of heterogeneous robotic systems at the University of Zagreb Faculty of Electrical Engineering and Computing. An overview of the vehicles to be used in the heterogeneous systems is given and the envisioned scenario of neutralization of suspicious underwater objects is presented. Tasks of each vehicle are described and emphasis is placed on the actions and communication routes that are required to ensure cooperative behavior.

I. INTRODUCTION

Heterogeneous robotic systems are systems that comprise mobile robotic units operating on land, in air and on/under water. While individual robotic systems today successfully perform complex tasks in their medium, activities through all three media are imposed by industrial needs as well as in seven societal challenges defined by the European Commission within the Horizon 2020 programme, specifically in the areas of:

1. environment: prediction, prevention and recovery from natural disasters; development of systems for global surveillance and informing about the state of the environment, and

2. security and protection: fight against crime and terrorism; strengthening security through border control; increasing resilience in cases of crisis and disasters.

The primary motivation for research in the areay of cooperative control of heterogeneous robotic systems is to enable sinergy between autonomous unmanned land, air and marine robots for the purpose of accomplishing a joint mission (e.g. surveillance of sea and land borders). The idea of the research is to unite existing robotic systems acting in different media into a unique cooperative heterogeneous system shown in Fig. 1.

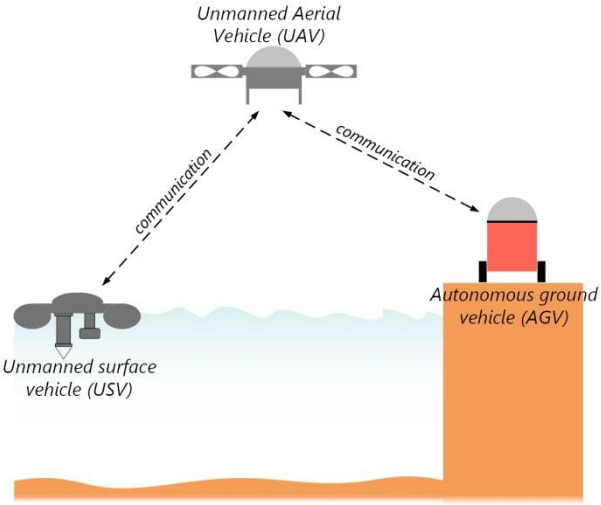


Figure 1. Concept of a cooperative heterogeneous robotic system.

The aim of the proposed research is to develop and experimentally validate cooperative control algorithms that maintain formation comprising three autonomous mobile robots (land vehicle, surface marine vessel and an autonomous aerial craft) and thus ensure their coordination. In order to accomplish this research goal, the following activities have to be performed:

1. Development and implementation of robotic cooperative control algorithms with kinematic and dynamic constraints as well as constraints in the communication channel.

2. Field trials using real robotic systems for the purpose of experimental validation of the developed algorithms and identification of necessary modifications and improvements. The developed algorithms will be tested in a scenario where autonomous robots are moving in a formation in which the aerial craft has a function of ensuring communication channel between the land vehicle and the surface marine platform.

The work described in this paper has been supported by the University of Zagreb as a short term financial support under the contract number 2013-ZUID-19.



a) *b)* *c)*

Figure 2. Photos of the vehicles comprising the heterogeneous robotic system: a) autonomous surface marine platform (*PlaDyPos* made at UNIZG-FER), b) autonomous aerial vehicle (*ArduCopter Quad* by jDrones) and c) autonomous ground vehicle (*Guardian* by Robotnik) with a manipulator.

II. AUTONOMOUS MOBILE ROBOTS

Mobile robots that will be used are available from three research laboratories from the Faculty of Electrical Engineering and Computing at the University of Zagreb: Laboratory for robotics and intelligent control systems – LARICS (unmanned aerial vehicles), Laboratory for Underwater Systems and Technologies - LABUST (unmanned underwater and surface marine vehicles) and Autonomous Mobile Robotics Group (autonomous land vehicles)

The **unmanned surface vehicle (USV)** is an overactuated platform with 4 thrusters forming the "X" configuration. This thruster configuration enables motion in the horizontal plane under any orientation. The platform has been developed at LABUST and the current version is 0.35m high, 0.707m wide and long, as it is shown in Fig. 2a), and it weighs approximately 25kg. The control computer (isolated from environmental disturbances inside the platform hull) is in charge of performing control and guidance tasks (dynamic positioning, path following, diver following) and all the data processing. Apart from the compass, batteries and CPUs, the platform is equipped with a GPS for determining the platform position, Ultra-Short Baseline (USBL) used to determine the position of the diver relative to the platform, a wireless modem used to transmit data from the platform to the surface station. The platform has been primarily developed for the purpose of diver tracking, [1]. More information on the characteristics of the platform can be found in [2].

Due to its simple mechanical design and open-source software, **unmanned aerial vehicle (UAV)** *ArduCopter Quad*, from jDrones company, shown in Fig. 2b) is one of the most popular aerial platforms with 4 propellers. Electronics, designed around Arduino processor, comprise complete set of sensors and features – 6 Degree of Freedom IMU stabilized control, gyro stabilized flight mode enabling acrobatics (loops and barrel rolls), magnetometer for heading determination, barometer for altitude hold, IR sensor integration for obstacle avoidance, and sonar sensor for automated takeoff and landing capability. Sophisticated control algorithms allow mounted camera stabilization. Ground station software modules provide automated waypoint navigation and wireless command & telemetry for long distance communication with real-time graphs of flight data. On board flight telemetry data storage can be used as well. This vehicle has recently been used to test hybrid adaptive control for aerial manipulation, [3].

The **autonomous ground vehicle (AGV)** is a *Guardian* produced by Robotnik from Spain, specially designed for applications in the domains of safety, inspection, research, and the high level of mobility allows it to move in rough and inaccessible terrain (stairs, slopes, etc.). The vehicle has an embedded computer with Linux OS that can be connected to either wirelessly or via etherner. The complete programming platform is based on ROS (Robot Operating System) that allows easy extension if the vehicles capabilities. Fig. 2c) shows the Guardian vehicle with a Powerball manipulator and a tool, both produced by a German company Schunk.

III. SCENARIO DESCRIPTION

The scenario that has been chosen for the demonstration of the cooperative heterogeneous robotic system is neutralization of unexploded underwater ordnance (UxO) which is an extremely dangerous and destructive activity since an underwater explosion has a destruction radius much higher relative to the one on land. The scenario does not deal with the neutralization process itself (defusing the UxO) but rather with handling the UxO and transporting it to a safe location. The proposed scenario can significantly contribute to expelling UxOs to the surface and taking them to a safe place on land autonomously, without human intervention or with limited intervention from a safe distance.

The proposed procedure is described in the following steps:

- Step 1: Attaching and activation of the self-inflatable balloon to the underwater object

A diver or a remotely operated underwater vehicle (ROV) attaches a balloon that can be inflated either by using acoustic activating signal from a safe distance, or by using a timer. Once the diver or ROV are at a safe distance, the balloon is inflated and the unknown dangerous object emerges to the surface. The balloon can optionally emit GPS coordinates so the location of the

UxO is known with a precision of a couple of meters radius.

- Step 2: Positioning the heterogeneous cooperative robotic system

The unmanned surface vessel (USV) and the unmanned aerial vehicle (UAV) that hovers above the USV, go to the location where the UxO is positioned.

- Step 3: Tugging the UxO to land

The USV followed by the UAV tugs the UxO to the location on land, next to the sea line, where the autonomous ground vehicle (AGV) with a robotic manipulator is located.

- Step 4: Autonomous ground vehicle takes the UxO

The AGV, equipped with a manipulator, takes the object out of the sea and carries it to a safe location where neutralization can be performed.

As a part of the proposed experiment, the following simplified mission will be executed. The mission is explained from the perspective of each individual vehicle.

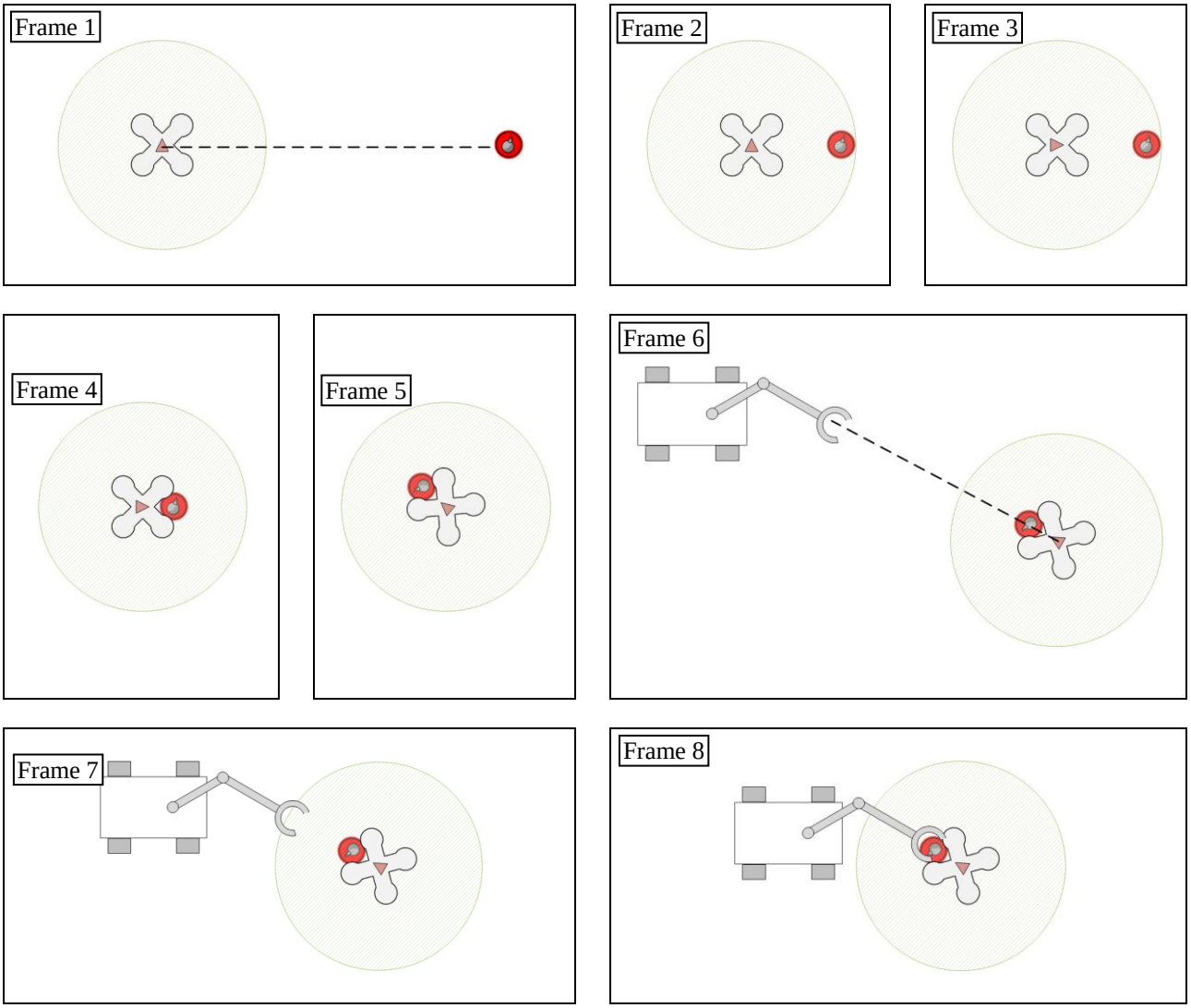


Figure 3. The scanario execution shown in time frames. The gren circle around the USV shows the dowwards field of view of the UAV hovering above the USV. Red circle represents the UxO.

A. The USV mission states

The USV mission states are comprised of the following elementary maneouvres: dynamic positioning (DP), heading control, and line following. Detailed description of the implemented algorithms used to execute these maneouvres can be found in [4], [5].

The following five states are executed sequentially.

1. Initial dynamic positioning (DP)

The USV dynamically positions around a predefined point A until the UAV receives the information about point B where the object (UxO) is located. The transition to the following mission state is executed when the USV receives position B.

2. Line following towards the object

Once the point B is known, the USV begins the line following maneuver from its initial point to point B where the UxO is located, as it is shown in Frame 1 of

Fig. 3. For this maneuver, only sensors onboard the platform are used (compass and GPS). At the moment when the UAV detects the UxO in the field of view (as shown in Frame 2 of Fig. 3), i.e. when relative distance between the platform and the object can be determined, the transition to the following state is executed.

3. "Grasping" the UxO

In this state, fine positioning of the platform relative to the UxO is possible since precise data on relative distance between the Uxo and the USV in the USV's coordinate system are available. The USV rotates towards the UxO so that it can "grasp" it between the two legs as it is shown in Frame 3 of Fig. 3. Once the desired orientation is achieved, the DP algorithm that minimizes the distance between the platform and the object is initiated, as shown in Frame 4 of Fig. 3. The mission transitions to the following state once the contact between the platform and the object is detected from the UAV field of view, and when the position of the autonomous ground vehicle (AGV) C sent by the UAV is available.

4. Line following towards the AGV

Once the contact between the USV and the UxO is achieved, and when position C is available, the USV rotates towards the ground vehicle while holding the object between the two legs as it is shown in Frame 5 of Fig 3. Once the desired orientation is achieved, line following from the current position to position C is initiated as shown in Frame 6 of Fig. 3. Only sensors onboard the USV are used (compass and GPS) for this maneuver. When the UAV detects the AGV in the field of view as shown in Frame 7 of Fig. 3, i.e. when relative distance between the USV and the AGV can be determined, the transition to the following state is executed.

5. Approaching the AGV

In this state, fine positioning between the USV and the AGV is possible since precise data on relative distance between the USV and the AGV is available. The USV dynamically positions itself to a point in front of the AGV, determined by the UAV, as shown in Frame 8 of Fig. 3. Once the desired position is obtained, the USV rotates with the UxO towards the AGV so that simple object grasping by the AGV is enabled.

B. The UAV mission states

The UAV mission states are based around hovering above the USV extending the field of view from above. The field of view of the UAV is represented as a green circle in Fig. 3.

The following five states are executed simultaneously during the whole mission.

1. Visual DP above the USV

The UAV detects the USV position within the video frame at all times and positions itself above the USV using DP algorithms based on position obtained from the video image. The UAV keeps constant altitude at all times while the orientation is arbitrary.

2. Transmitting UxO coordinates

If the UAV has information about the UxO coordinates (these coordinates will be transmitted to the UAV by an operator through the wireless link), they are transmitted via WiFi to the USV below. In this state, the UAV serves as an extended antenna for the USV.

3. UxO detection

The UAV searches for the UxO (a predefined marker on the object) within the video image all the time. If the object is found, the UAV sends the relative distance between the detected object and the USV to the USV.

4. AGV detection

The UAV searches for the AGV (a predefined marker on the AGV) within the video image all the time. If the AGV is found, the UAV sends the relative distance between the detected AGV and the USV to the USV.

5. Transmitting AGV coordinates

The UAV sends the AGV coordinates via WiFi to the USV all the time. These coordinates are received from the AGV itself.

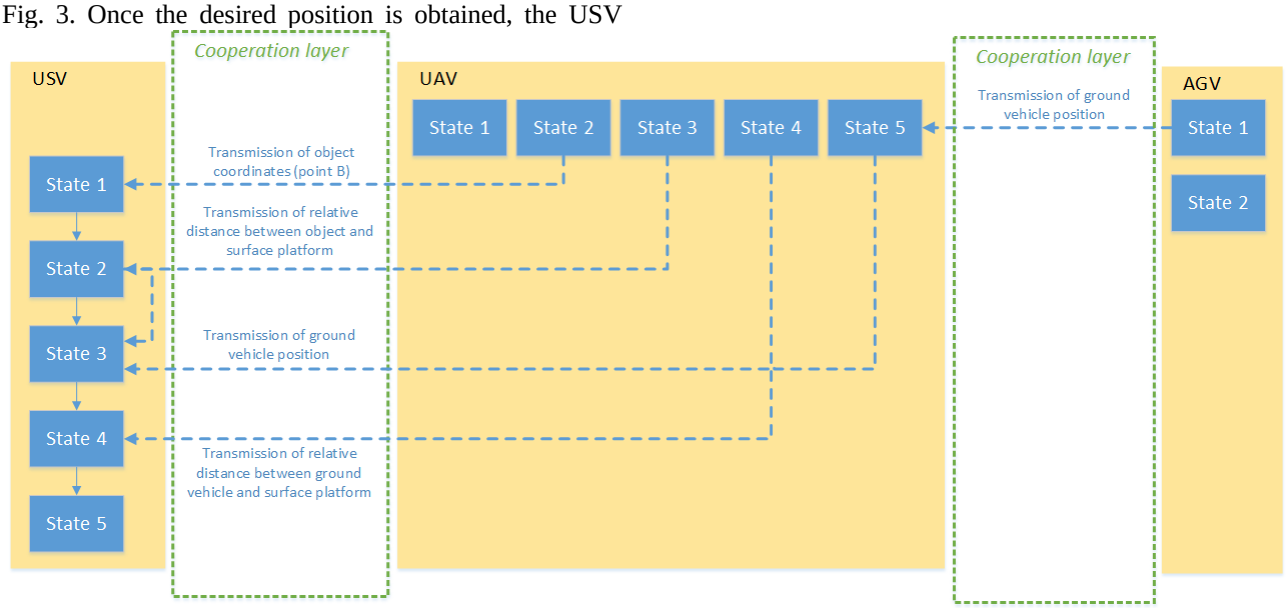


Figure 4. The conceptual representation of the mission flow from the perspective of the each unmanned vehicle.

C. The AGV mission states

1. Transmitting position coordinates

The AGV is at a location on ground next to the sea line and transmits its own GPS coordinates to the UAV through the wireless link all the time.

2. Remotely operated object grasping

Once the UxO is close enough to the AGV, the remote operator can use the manipulator onboard the AGV to grasp the UxO and carry it to a safe location for neutralization.

Fig. 4 graphically shows the mission states from the perspective of the individual unmanned vehicles. In the same figure, the cooperation layers are shown. These layers in fact represent the communication that occurs between the vehicles. The data that is being communicated is the baseline of the cooperation of the heterogeneous robotic system.

IV. CONLUSION

This paper presents an overview of the project with the task to establish a cooperative heterogeneous robotic system. The systems that will be used for experiments are described and the scenario of neutralization of the unexploded ordnance is given.

The next step is to combine already existing vehicle behaviors to support the cooperative system. The cooperation will be established by maintaining the formation between the unmanned surface vehicle and an unmanned aerial vehicle that serves as an extended antenna and a communication router. Once the algorithms are verified in a simulation, real life experiments will be conducted.

REFERENCES

[1] N. Mišković, Đ. Nađ, N. Stilinović, and Z. Vukić, "Guidance and control of an overactuated autonomous surface platform for diver tracking", Proceedings of 21st Mediterranean Conference on Control & Automation (MED), pp. 1280-1285, 2013.

[2] N. Mišković, Đ. Nađ, Z. Vukić and B. Marszalek, "Laboratory Platforms for Dynamic Positioning – Modeling and Identification", Proceedings of the 8th IFAC Conference on Control Applications in Marine Systems, pp. 161-166, 2010.

[3] M. Orsag, C. M. Korpela, S. Bogdan, and P. Y. Oh, "Hybrid Adaptive Control for Aerial Manipulation", Journal of Intelligent & Robotic Systems, Volume 73, Issue 1-4, pp 693-707, January 2014

[4] N. Mišković, Z. Triska, Đ. Nađ, and Z. Vukić, "Guidance of a Small-Scale Overactuated Marine Platform – Experimental results" Proceedings of the 34th international convention on information and communication technology, electronics and microelectronics vol. III (MIPRO/CTS 2011), pp. 39-44, 2011.

[5] V. Vladić, N. Mišković, and Z. Vukić, "Quick Identification and Dynamic Positioning Controller Design for a Small–Scale Ship Model" Proceedings of the 2012 20th Mediterranean Conference on Control & Automation (MED), pp. 1385-1390, 2012.

Program Solutions for the Complex Cooling Process via LADDER Logic

Tomislav Špoljarić, Mato Fruk, Goran Vujisić
Department of Electrical Engineering
University of Applied Sciences, Zagreb
Konavoska 2, Zagreb, Croatia
tomislav.spoljaric@tvz.hr, mato.fruk@tvz.hr, gvujisic@tvz.hr

ABSTRACT - This paper describes designing complex process control in a cooling system. Control solutions specifically designed for simple cooling system are already available and these simple control designs include temperature control limited to one cycle of operation. For more complex cooling processes with variable operation cycles it is necessary to introduce a superior process control design, which coordinates the operations of simple processes. Within each cycle temperature is controlled by temperature regulator, while cycle operation exchange of individual processes is managed by microprocessor device whose program was written in LADDER logic in this study.

I. INTRODUCTION

Today majority of small modern cooling systems used for food storage demand simple temperature control. Cold room temperature is monitored by room probe and evaporator temperature is monitored by defrost probe. Automatic temperature control is maintained in one cycle with two different periods. In first period refrigerating process is done, with compressor, condenser and expansion valve controlled by room temperature probe and pressure probes used to monitor refrigerant pressure in suction and discharge lines. Second period is used for defrosting process, meaning that the ice that was accumulated on evaporator section during refrigeration period needs to be melted. Second period is only needed in storage rooms with temperature below and around 0°C, with first period being significantly longer than the second. After one cycle, overall process repeats.

Most electronic temperature regulators used for simple cooling process control can be used for one cycle regulation with refrigeration/defrost period. More complex electronic regulators can control two different time cycles of operation.

Storage and freezing of various types of food and goods demand different levels of cooling process complexity. While with most goods simple one cycle processes are satisfying, there are situations when more complex control must be designed. This type of cooling process control can easily be designed by incorporating two devices: electronic temperature regulator for temperature control which is subjected to electronic device operated by LADDER logic and used for time cycle control.

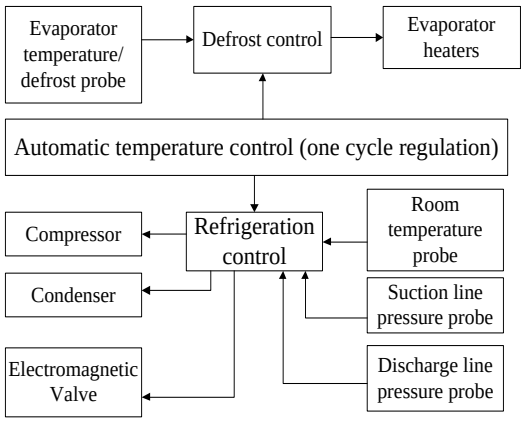


Fig. 1 Block diagram of simple cooling process

II. TIME FLOW DIAGRAM IN COMPLEX COOLING SYSTEMS

Basically, previously described simple cooling process is now expanded into complex by having variable instead of fixed refrigerating and defrosting periods and by taking into account the number of cycles. This process is being designed with consideration to nature of goods that need to be stored or frozen. In some cases individual quick freezing processes in freezing tunnels demand various time cycles of operation because the type of food or goods that is processed on that occasion exhibits different levels of humidity into local atmosphere during overall process, thereby creating additional ice in the evaporator section. This results in earlier defrost demand and shortened refrigeration period when process is in the earlier stage of operation.

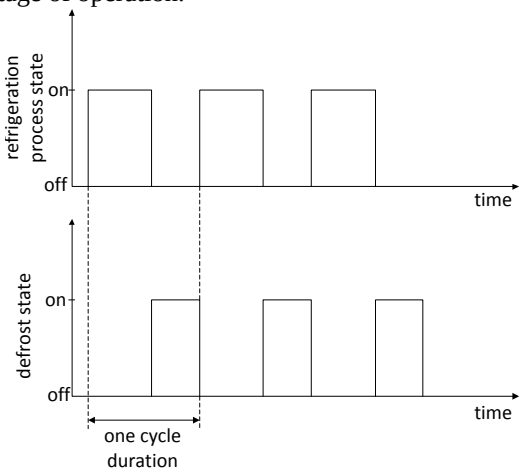


Fig. 2 Time flow diagram of simple cooling process

Refrigeration control is done by switching compressors, condenser fans and electromagnetic valves according to room temperature and lines pressure. When defrost control is in charge evaporator heaters are switched according to evaporator temperature probe, commonly known as defrost probe. Time flow diagram of described simple cooling process is shown in fig. 2.

In complex cooling process variable cycles are introduced, meaning that refrigeration and defrost periods are also variable. While electronic temperature regulators allow only one cycle of operation, time cycle control can be done by microcontroller device. Electronic temperature regulators have several digital inputs with various functions, and one of them is a defrost start signal. This type of signal can be used for signalling end of refrigeration period and start of defrosting period. Small period called pause period has to be incorporated in LADDER program to ensure that switching of refrigeration components is blocked while defrost components are still in operation and vice versa. This period also ensures that electric power supply of cooling system isn't overloaded with unnecessary load, because defrost heaters operation is not needed, even for a few seconds, while refrigeration period is in charge and may cause inappropriate tripping of electric protection devices. In pause period all system components are switched off.

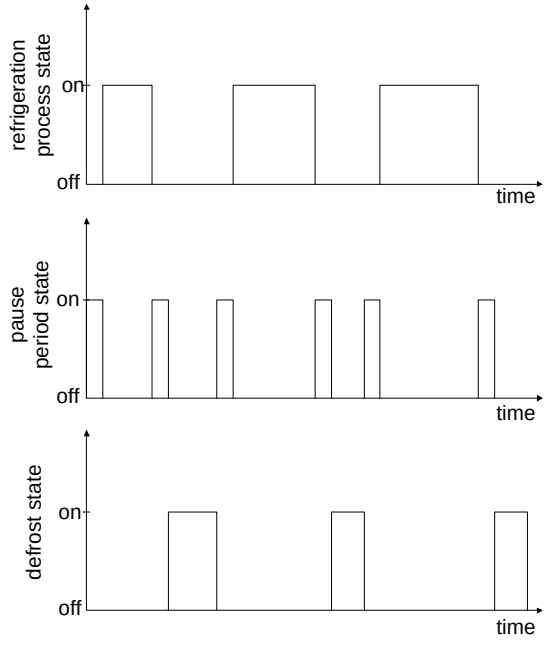


Fig. 3 Time flow diagram of complex cooling process with three cycles of operation

First step in designing complex cooling process control is time flow diagram design. One has to take into account overall time needed for goods to be frozen, number of cycles of operation and different operation period times. As mentioned before, earlier periods of cooling or freezing process demand shorter periods of refrigeration and longer defrosting periods. The freshness of goods and humidity that is transferred from goods to local atmosphere results in creating additional ice on the evaporator section. Later periods of refrigeration process can be longer with shorter defrosting periods, as goods are now drier and humidity is lower. Time flow diagram example of such a process is shown in fig. 3.

III. TEMPERATURE CONTROL IN COMPLEX COOLING SYSTEMS

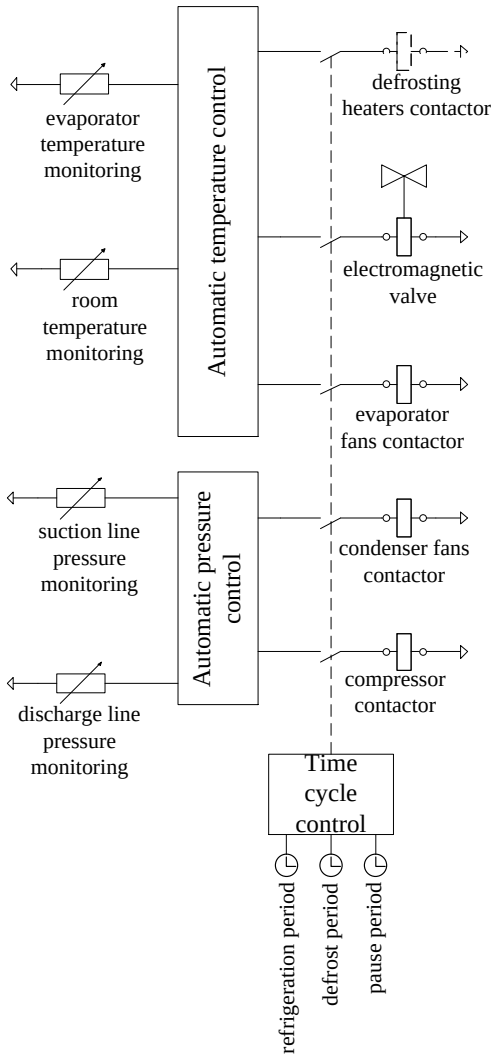


Fig. 4 Block diagram of complex cooling process control

Temperature control in complex cooling system is subjected to time cycle control. Electronic temperature regulators give signals to contactors with respect to measured temperature and pressure signals. Contactors are being switched in two stages with relay contacts. First stage operates according to the period of operation cycle in time cycle control, and the second stage operates according to measured temperature and pressure in temperature control.

Due to independent regulations of pressure and temperature in temperature control, pause

period has to be implemented. This way, safe switching of various components is ensured as mentioned before. In simple cooling process temperature regulator is working in one cycle and ensures safe switching of various elements. In complex systems operations of defrosting heaters (controlled by evaporator temperature), compressor, condenser fans (controlled by lines pressure), evaporator fans and electromagnetic valve (controlled by room temperature) are all independent operations and have to be treated in such manners, when creating a process design.

Block diagram of complex cooling process is shown in fig. 4.

There are more simple systems operated in complex manners, such as cooling systems of minor electric powers. Pressure control in these systems works only as a protection with electromechanical pressure control devices, while room temperature is main element in providing satisfying cooling in cold rooms of such systems. Regulation of these simple systems consists only of superior time monitoring device and temperature regulator, without pressure probes as monitoring elements. In this paper, a more complex approach has been selected. Temperature and pressure regulations with four independent probes are more common in cooling systems that demand complex cooling process design.

IV. TIME CYCLE CONTROL IN COMPLEX COOLING SYSTEMS

Time cycle control in complex cooling system is superior to temperature control mentioned before. In this stage of control microcontroller operated by LADDER logic program gives orders to electronic temperature regulator as a signal for defrost start/stop and gives signals to contactors according to current period of operation. All LADDER diagrams shown in this paper are based on actual diagrams written and simulated in easySoft® pro 6 software package, and are meant to be used with Eaton's EASY® relay devices.

First period of each cycle is pause period, after which refrigeration period takes control. In refrigeration period compressor contactor and condenser fans contactors are controlled by pressure regulator. Evaporator fans and electromagnetic valve are operated by electronic temperature regulator.

First pause period in each cycle starts with M01 signal and its contact set to ON stage, which sets timer T01 and begins counting time scheduled for pause period. After T01 counts up its contact closes which ensures resetting of defrost period timer and markers while setting refrigeration period marker. This is also a starting period in overall process. When refrigeration period timer T02 counts up, its contact ensures resetting first pause marker M01 and timer T01 and setting second pause period marker M03 and timer T03.

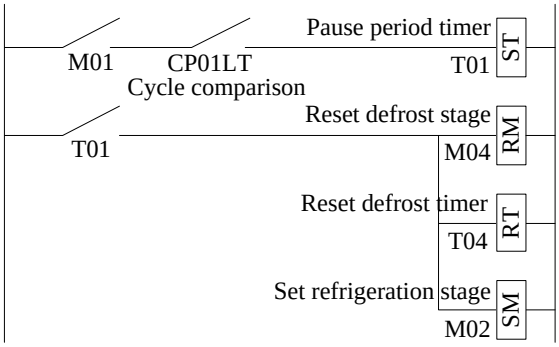


Fig. 5 LADDER diagram for pause period control, after defrost and before refrigeration period

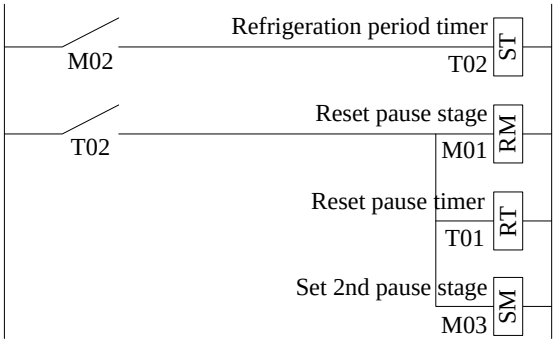


Fig. 6 LADDER diagram for control of refrigeration period

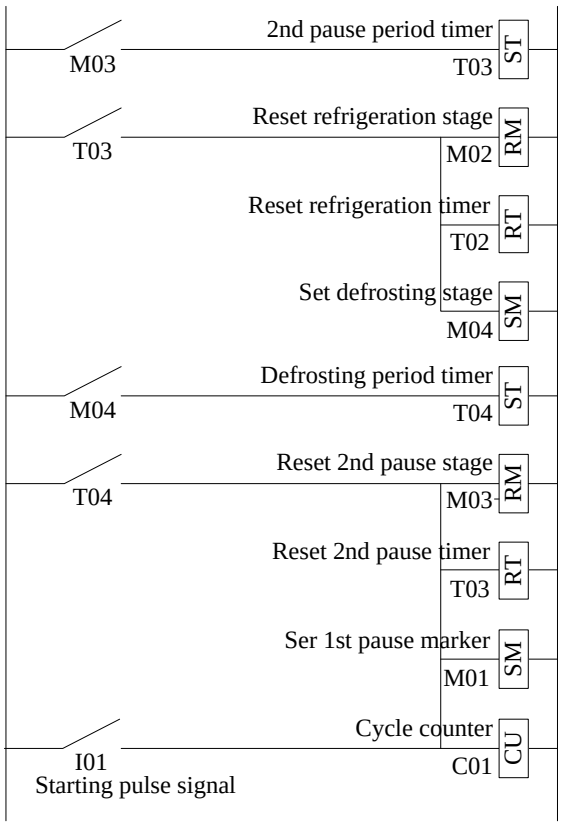


Fig. 7 LADDER diagram for pause period control before defrost and defrosting period

In defrost state pressure control works independently but contactors of condenser fans and compressor remain switched off, since time cycle control does not allow them to be switched on. Also, pressure probes in both lines do not allow compressor and condenser fans to operate, since electromagnetic valve is turned off and there is no refrigerant flow through lines.

After 2nd pause period timer T03 counts up, refrigeration marker and timers are reset, and defrosting period marker M04 and timer T04 are set. When timer T04 counts up, cycle repeats itself according to comparison CP01 of counter block C01 implemented to count number of cycles by setting marker M01 in ON stage again.

Times stored in timers may also be variable, since they are stored as data in particular memory location selected by process programmer. On fig. 5 two different types of timers are shown.

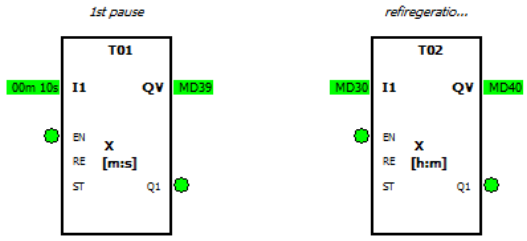


Fig. 8 On-delayed timer function blocks for fixed and variable time setpoint

First timer, used for pause period has time setpoint I1 fixed and set to 10 seconds. Second timer, used for refrigeration period, has a variable setpoint located in particular memory location in microcontroller memory (MD30). Actual inputs of both timers are set to two different memory locations (MD39, MD40) allowing program process designer or end user (optionally via SCADA or some other visualization device) to monitor time elapsed on each period of current cycle. Other inputs on timer block are enable bit (EN): digital signal used to start the timer, reset bit (RE) and stop bit (ST). Output Q1 is used as on-delayed switching contact.

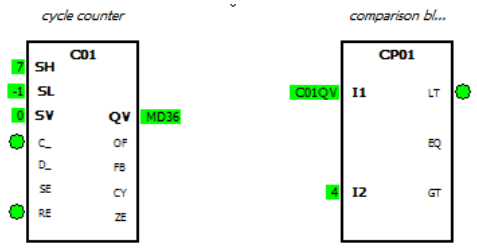


Fig. 9 Counter and comparison function blocks with fixed and variable setpoints

When setting cycle number, counter block needs to be defined. More important signals SH, SL and SV present upper, lower setpoint and preset value. Bit C_ presents count coil that counts on every rising edge. Bit D_ defines count direction. Actual value of the counter is set in QV output. Comparison block takes two values (I1, I2) and transfers their comparison as digital signal to three types of switching contacts defined as bit outputs in comparison block definition: LT (lower than), EQ (equal), GT (greater than). These function blocks are used to define number of cycles and condition if current cycle has reached that limit. Variability of all mentioned values can be easily obtained by using SCADA/HMI systems, therefore giving accessibility to end user.

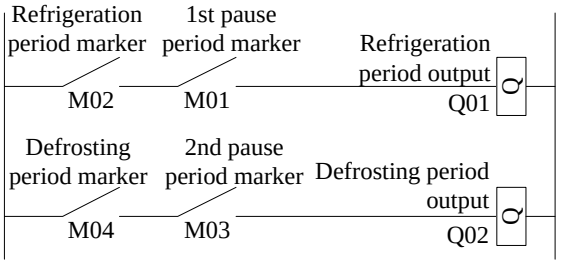


Fig. 10 LADDER diagram for refrigeration control and defrosting period outputs

Digital outputs are set to ON state when current period and previous pause period markers are set on. When proceeding to next pause period, previous pause period marker is reset, setting the output bit to zero. Logic of described output switching process is shown in fig. 10 and fig. 5, 6 and 7 respectively.

V. COMPLEX COOLING PROCESS IN FREEZING TUNNEL

Individual quick freezing process in freezing tunnel is observed as a practical approach to the problem of complex cooling. Goods that are frozen in this type of process behave in same manner as mentioned in first chapter of the paper, such as raising level of humidity in freezing tunnel's space.

Measurements were taken from the cooling system consisting of three freezing tunnels with different carrying capacities (two smaller tunnels have capacity of 15 tons of fresh fish and one large tunnel has capacity of 30 tons of fresh fish). Every tunnel works in four or six cycles, depending on the type and freshness of fish to be frozen. All cycles have fixed time presets, varying from first to last cycle. Refrigeration periods within each cycle are fixed but different in each cycle. Defrost periods are fixed and same in all cycles, as well as pause periods are too.

Automatic temperature control is realized with Carel's MPXPRO® electronic regulators which control evaporator fans in refrigeration periods, and evaporator electric heaters in defrosting periods. These regulators have one digital input reserved for starting signal of defrosting period which enables electric heating of evaporator section in coordination with evaporator temperature probe. In

refrigeration period evaporator fans are switched on according to signal from room temperature probe.

Automatic pressure control of compressor and condenser units is performed by Carel's microcontroller unit pRACK® that has one digital input reserved for starting signal from superior microcontroller device EASY 820-DC-RC® performing time cycle control. When in refrigeration period, EASY relay gives signal (by setting Q01 output to ON stage) to pRACK unit to start with pressure control. Suction line's pressure is used to control compressor, and discharge line's pressure is used to control condenser fans. In this particular approach compressor and condenser fans switching states are not observed, because temperature control and monitoring in freezing tunnels is more interesting from the aspect of HACCP (Hazard Analysis Critical Control Points) inspection, since food storing and freezing has to satisfy certain number of regulations concerning food quality.

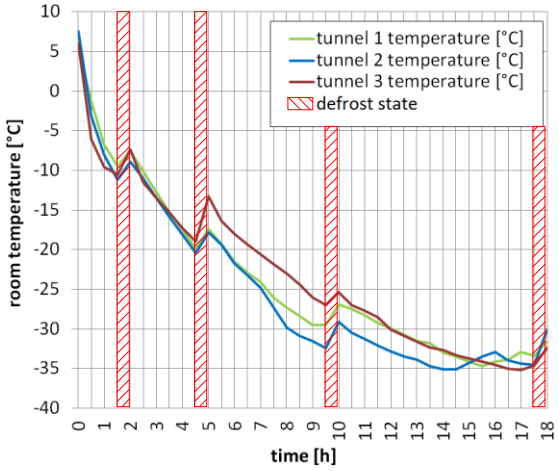


Fig. 11 Time-temperature graph of complex cooling process in three freezing tunnels loaded with 120% carrying capacity

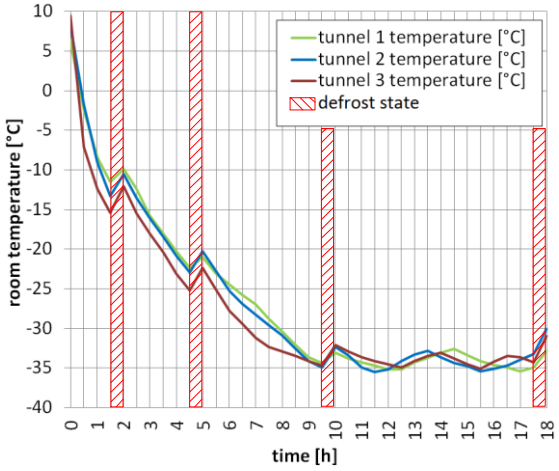


Fig. 12 Time-temperature graph of complex cooling process in three freezing tunnels loaded with 100% carrying capacity

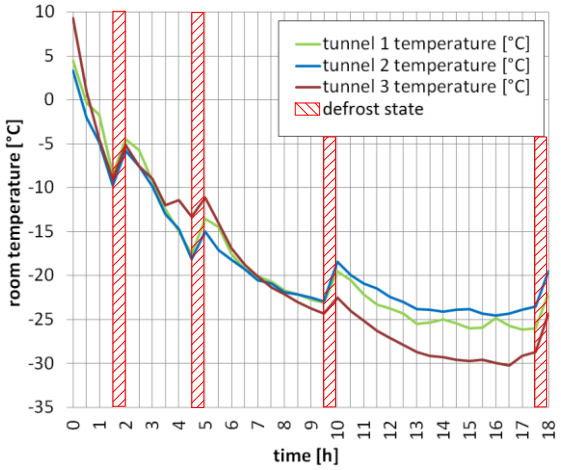


Fig. 13 Time-temperature graph of complex cooling process in three freezing tunnels loaded with 200% carrying capacity

Graphs of temperature measurements are shown on fig.11, 12 and 13. Measurements were taken from four cycle cooling processes with tunnels carrying 120%, 100% and 200% of fresh fish capacity, with 30 minutes sample time. Observed four cycle process lasts for 18 hours. Device used for generating HACCP inspection reports was Carel's PlantWatch PRO®. This device has minimum sample time of 30 minutes when observing room temperature in creating HACCP report.

Working conditions and setpoints were same for all three freezing tunnels. Temperature setpoints of all three tunnels has been set to -35°C, with temperature difference set to 2°C, meaning that in refrigeration period evaporator fans switch from -33°C to -35°C.

With process in defrosting period heaters are switched on until evaporator section temperature reaches 20°C, with temperature difference set to 2°C. Due to pretty long sample rate of data acquisition from the measuring system and short fixed period of defrost period, switching states of defrosting elements are almost impossible to monitor. On temperature/switching state graphs it can only be seen that defrosting elements are switched on when defrosting period takes place in observed process.

With process in refrigeration period, one can observe and conclude from these measurements how the effect of overload can affect directly to the electrical supply system, with all electrical components working continuously. Switching states of evaporator elements have been observed for different conditions of tunnel loading. Load equal to carrying capacity (100%) shows that in last cycle automatic temperature regulator switches fans off, as the goods have reached desired temperature, with their quality preserved (fig.15).

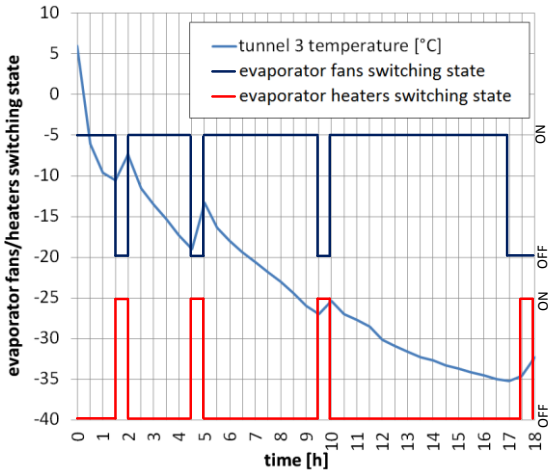


Fig. 14 Switching states during four cycle complex cooling process in largest freezing tunnel loaded with 120% carrying capacity

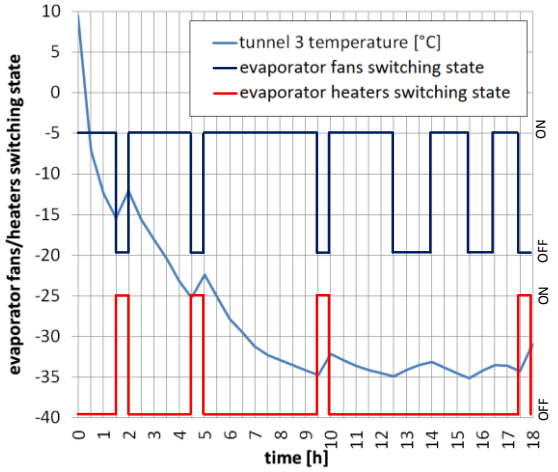


Fig. 15 Switching states during four cycle complex cooling process in largest freezing tunnel loaded with 100% carrying capacity

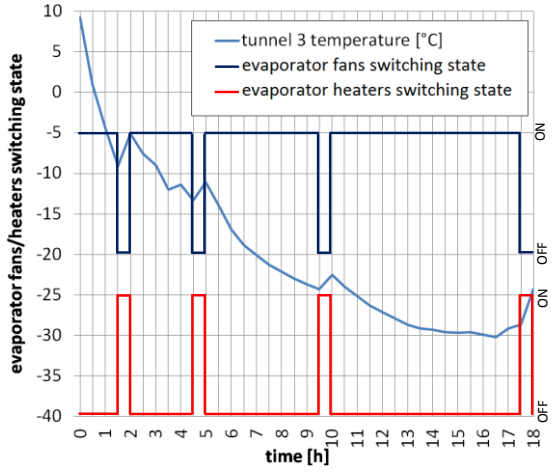


Fig. 16 Switching states during four cycle complex cooling process in largest freezing tunnel loaded with 200% carrying capacity

Switching states/temperature graph (fig. 14) for slightly larger overload above rated shows continuous operation of all elements included in refrigeration period, but also reaching the desired temperature with satisfying quality of frozen goods, while in heavily overloaded freezing tunnel room temperature never reaches default temperature setpoint, making goods' quality questionable.

VI. CONCLUSION

Temperature regulation for simple cooling process is commonly used in food storage and quick freezing, and used as one cycle cooling process. But with more demandable cooling processes of specific goods, such as various types of fish, it's complexity needs to be expanded. Various durations of cooling cycles consisted of refrigeration and optional defrost period, demanded for successful processing of goods, require more complex process design than the one described as simple process with automatic temperature/pressure regulation only.

Most affordable and optimal design lies in microcontroller coordinating with electronic temperature and pressure regulators. Possibilities for end user to control complex processes are numerous (with optional SCADA/HMI systems) and at the same time having possibility to preserve administrating of the system for service and parameter change purposes.

Described results in last chapter of this paper show that this type of control design is successful in operation and with its low-cost equipment acceptable for end user. It's simple core consisting of microcontroller driven by LADDER program and automatic regulation devices for specific purposes (such as temperature and pressure control) and possibility of control elements expansion (SCADA/HMI systems, microcontroller networking) is also an advantage of this type control design.

BIBLIOGRAPHY

[1] D. Gadže, T. Špoljarić, “Izvedbeni projekt elektroinstalacija rashladnih sustava – ELO-001-13”, University of Applied Sciences, Zagreb, February. 2014.

[2] easy820-DC-RC datasheet – Article no. 256271, Eaton online catalog, Eaton, 2013.

[3] easy800 User Manual – 4th edition, Moeller, 2005.

[4] J. Garden-Robinson J., Food Freezing Guide, revised edition, NDSU, 2011.

[5] Becker B.R., Fricke B. A., “Freezing Times of Regularly Shaped Food Items”, University of Missouri, Kansas City, February 1999.

[6] Earle R. L., Earle M. D., Unit Operations In Food Processing, web edition, The New Zealand Institute of Food Science & Technology, 2004.

The Use of the Internet and Wireless Communications in the Monitoring and Control of Industrial Processes

Viša Tasić*, Marijana Pavlov*, Darko Brodić**, Vladimir Despotović**, Dragan R. Milivojević*
*Mining and Metallurgy Institute, Bor, Serbia
** University of Belgrade, Technical Faculty in Bor, Bor, Serbia
e-mail: visa.tasic@irmbor.co.rs

Abstract - This paper describes the application of Internet and wireless communications in the monitoring and control of production processes in the metallurgical company RTB Bor, Serbia. A special attention is paid to the presentation of the software solutions. Implemented distributed control systems have shown remarkable reliability in practical work, low cost of maintenance and investment in their expansion. The described solution is applicable for monitoring and control of processes with the control system response time of about a minute or longer.

I. INTRODUCTION

Our previously presented works describes the development of the PLC based on microcontroller (Motorola MC68HC11) [1, 2] and applications for monitoring and control of industrial processes [3-6]. Such monitoring systems are implemented at all key production lines in the Copper Mining and Smelting Complex Bor (RTB Company).

The RTB Company is engaged in excavation and processing of copper ore (mines, flotation, copper smelting and refining facilities) [2]. The process equipment in almost all production units in the RTB Company was very old with the basic local process control systems. As an investor the RTB Company has always demanded low cost monitoring systems. For these purposes the specific process control systems have been designed and developed [2-5]. Monitoring and control of process parameters in many production units in the RTB Company include interaction with the process from remote locations. For that reason, the individual control systems were interconnected in a large industrial network [3].

Dynamic development of new communication technologies provides new ways for remote monitoring and control of industrial processes. For that reason, the realized monitoring and control systems are constantly upgraded and improved until nowadays, both in terms of hardware and software. In this paper we describe the use of programs for data exchange and synchronization over the Internet and wireless networks with the aim of remote monitoring and control of industrial processes in the RTB Company.

II. OVERVIEW OF REMOTE MONITORING OF INDUSTRIAL PROCESSES IN THE RTB COMPANY

Remote monitoring and control in the RTB Company in previously realized monitoring systems was achieved over remote workstations. The workstations were connected using base-band modems via leased lines with a number of PLCs [2]. Communication protocol, BSCP (Binary Synchronous Communication Protocol) was specially designed and applied in order to support real-time data transfer between all network nodes [7]. That was costly because of the large amounts of time required for development of communication software and hardware.

In the past decade, carefully planned and realized industrial computer network was realized in the RTB Company. It consists of several LAN segments connected in a single network [3, 4]. Connection of distant LAN segments is achieved either with conventional telephone copper wires and the appropriate modems, or with wireless access points (AP). Wherever it was possible, wireless connection is applied. 'Point to point' mode is used for wireless AP operation. For short distances, connection between the network nodes is provided by using the UTP cables. In some cases the connection between the nodes or LAN segments is established using the mobile phone network services with GPRS data modems, HCP SHARK Quad-band GSM/GPRS [8].

The special SCADA real-time application named Process Control Program (PCP) [9] was developed in order to support PLC functions and to perform data transfer, analyses and real-time result interpretation. The PC which is directly connected to a PLC, running the PCP, acts as data server in the network configuration. Remote workstations execute Remote Control Program (RCP), a passive version of PCP. RCP is used for distant monitoring only. RCP can access the databases stored on the server and retrieve the new data. Each production unit has the server entity that consists of the PLC and PC workstation, as shown in Figure 1.

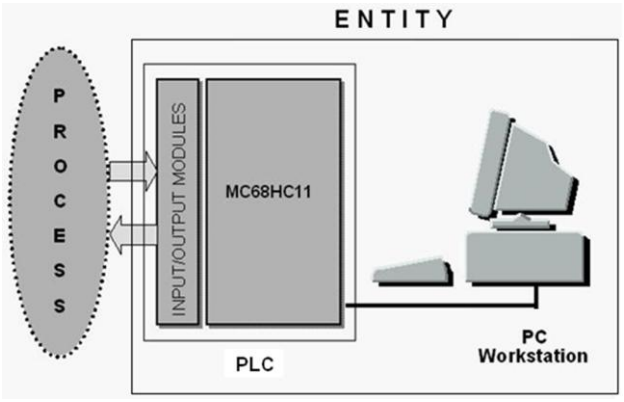


Figure 1. Diagram of a simple control system (entity)

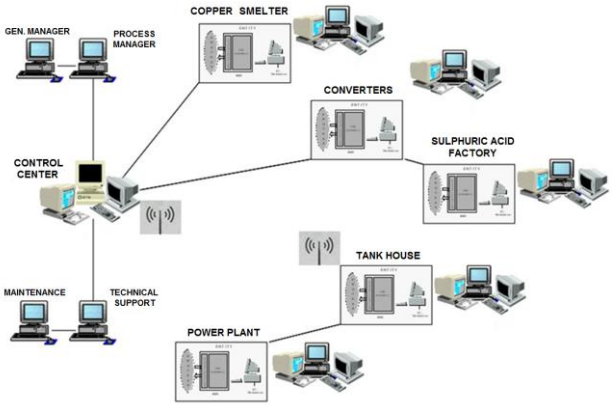


Figure 2. Block diagram of realized industrial network structure

The realized industrial network consists of several sub-networks with a number of clients for remote monitoring as shown in Figure 2. Smelting of copper concentrates is such process in which the majority of the process parameters change slowly over time (temperatures, pressures, and flow rates). Slow change of process parameters was the main reason for adoption that PCP provides data for RCP on every minute. PCP generates very short text files (ASCII), with the size of few KB, which contains actual information about monitored process parameters. The workstations that execute RCP simply download these files over LAN. From a technical point of view, this is a low-cost solution having in mind that distant LAN segments are mostly connected over conventional telephone copper wires (RTB Company owns private telephone lines inside and between the most of its facilities).

III. THE NEW WAYS OF REMOTE MONITORING AND CONTROL OF INDUSTRIAL PROCESSES

The constant evolution of the Internet and information technologies brings new methods for remote monitoring and control of the industrial processes and equipment. The Internet technologies provide low cost mechanisms for connecting computers anywhere in the world. One such technology is digital subscriber line (DSL) where computers have a permanent connection to the Internet. RTB Company uses DSL technology to lower remote monitoring costs. The most of remote monitoring implementations in the RTB Company uses simple file transfer protocols (FTP) to transfer data between two distant locations. This requires software to transfer and view data. For that purpose Gbridge software is used.

Gbridge is free software that is used to synchronize folders at remote computers with the folders on host computer [10]. Gbridge automatically forms a collaborative, encrypted VPN (Virtual Private Network) that connects computers directly and securely [11]. It allows users to remotely control a Windows host, transfer files, share folders, and automatically synchronize files between PCs. Also, Gbridge does not require any modifications to the network firewall. It creates a soft VPN between host and remote PCs using the gbridge.com website [10] and a Gmail account to login in. This VPN is fully meshed so that multiple clients can see one another and participate in file transfers, remote control, and shared file locations. Thus, Gbridge provides many-to-many connectivity instead of a one-to-one link as with most remote control utilities [12]. SecureShare is the Gbridge feature built in VPN that enables sharing files or folders on one computer, to all other participants of VPN.

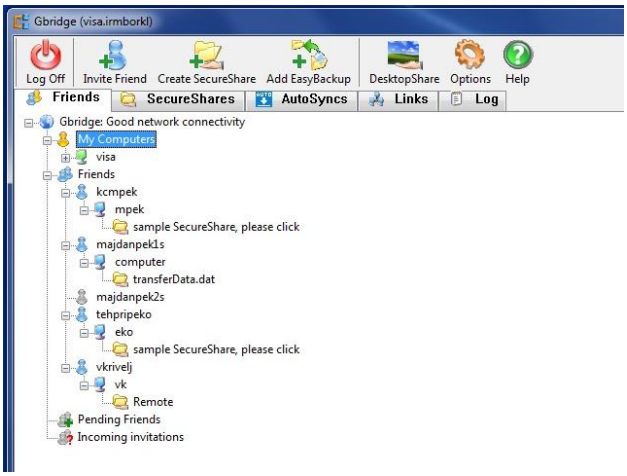


Figure 3. Gbridge Friends tab with SecureShare folders provided on 'friend' (host) computers

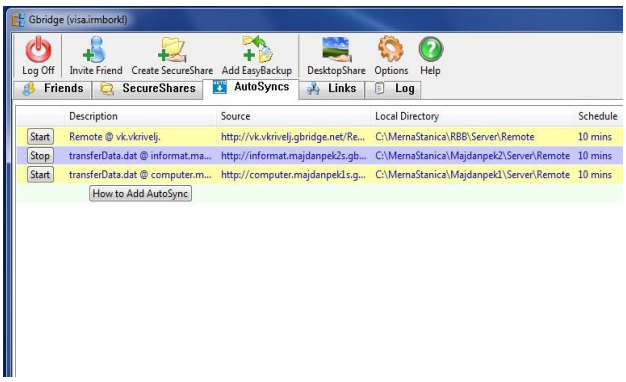


Figure 4. Gbridge AutoSyncs tab with a number of tasks running

Gbridge AutoSync feature typically involves two computers: one as the source, which provides the SecureShare, and the other is the destination which actually does the automatic synchronization copying data from SecureShare to a local folder. The feature needs to be invoked by first navigating to the "friends" tab (as shown in Figure 3) and find SecureShares created on those source computers, then one can click on these SecureShares to start the auto sync task.

For the purpose of remote monitoring of industrial processes we always need to get the latest copy of files generated by PCP. This requires creating the AutoSync tasks as shown in Figure 4. The destination computer always tries to synchronize the contents of the "Source" SecureShare according to a predefined schedule. AutoSync performs smart scan on the source directory and would only send the difference over, so it uses a negligible amount of data when the two folders are already synchronized. If AutoSync misses a scheduled synchronization due to the source or destination computer being offline, it will resume the synchronization as soon as both computers are available [10].

With Gbridge we can track parameters of industrial processes from any place in the world that has Internet access. RCP reads current information on the state of industrial process parameters from a local folder on the same computer that executes RCP. Particularly useful option is using a wireless connection on a laptop computer to monitor the industrial processes.

When the process control must be performed from a remote location (for example, when a manager wants to change the process parameters of a control loop from his home) TeamViewer software is used.

TeamViewer is a software package for remote control and file transfer between computers [13]. TeamViewer has to be running on both host and client computers. Once installed it generates a partner ID and password as shown in Figure 5. To establish a connection from a local client to a remote host, the host ID and password must be entered into the local TeamViewer window (shown in Figure 5).



Figure 5. TeamViewer common window with local ID and password (on the left) and partner ID (on the right)

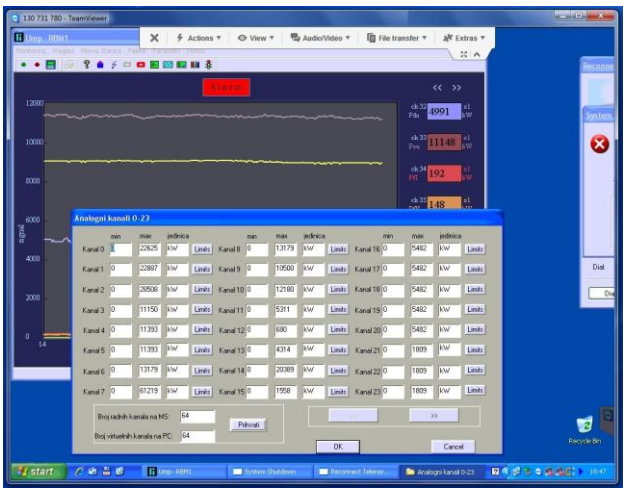


Figure 6. Example of remote control of PCP parameters using TeamViewer

Example of TeamViewer usage for changing the parameters of a PCP is shown in Figure 6. TeamViewer is also very suitable for transferring large files between the two computers because there are no file size limits as is the case when e-mail is used. Although TeamViewer can also be used just for monitoring of control computer (observing its desktop) we didn't use this option.

CONCLUSION

In this paper we describe solutions for monitoring and control of industrial processes using the Internet and wireless communications. Presented solutions were applied in practice for monitoring using Gbridge and control processes using TeamViewer in the metallurgy of copper. The main benefit of given solutions is their ease of use, low price, and its applicability in the different combinations of operating systems and hardware. The described solutions are applied in our environment for control of processes with the control system response time of about a minute or longer.

Furthermore, proposed approach has another benefit – the existing SCADA applications (PCP and RCP) didn't require any modifications at all. In this way, the new communication technologies with described software solutions provide new possibilities for monitoring and control of the production processes with minimal investment.

ACKNOWLEDGMENT

This work was funded by the Grant of the Ministry of Education, Science and Technological Development of Republic of Serbia, as a part of Project No. TR33037: "Development and Application of Distributed System for Monitoring and Control of Electrical Energy Consumption for Large Consumers."

REFERENCES

[1] D.R. Milivojevic and V.Tasic, "MMS in Real Industrial Network," Information Technology and Control, vol. 36, pp. 318-322, 2007.

[2] V.Tasić, V.Despotović, D. Brodić, M.Pavlov, D.Milivojević, "Twenty years of monitoring and control of electricity consumption in RTB Bor, Serbia," Proceedings of 36th International Convention, MIPRO 2013, pp. 1134-1138, Opatija, Croatia, 20.5.-24.5.2013.

[3] D.Milivojević, V.Tasić, M.Pavlov and V.Despotović, "Synthesis of DCS in Copper Metallurgy," Proceedings of XLII International Scientific Conference on Information, Communication and Energy Systems and Technologies ICEST 2007, Book 2, pp. 629-631, Ohrid, FYR Macedonia, June 2007.

[4] V. Tasic, D. R. Milivojevic, M. Pavlov and V. Despotovic, "Reducing the Costs of Electricity in the Copper Mining and Smelting Complex RTB Bor (Serbia) by Using of the Distributed Control System," Proceedings of 33rd International Convention on Information and Communication Techology, Electronics and Microelectronics MIPRO 2010, pp. 664-667, Opatija, Croatia, 24.05.-28.05.2010.

[5] V.Tasić, D.Milivojević, M.Pavlov, V.Despotović, D. Brodić, "Microcontroller Based Systems for Peak Load Reduction," Proceedings of 35th International Convention, MIPRO 2012, pp. 919-923, Opatija, Croatia, 21.5.-25.5.2012.

[6] D.Milivojević, M.Pavlov, V.Despotović and V.Tasić, "Half a Century of Computing in the Serbia Copper Mining and Metallurgy Industry," IEEE Annals of the History of Computing, no.3. vol. 34, pp. 34-43, July-September 2012.

[7] V.Tasić, D.Milivojević, V.Despotović, D.Brodić, M.Pavlov and I.Stojković "Communications in Realized Industrial networks," Proceedings of XLVIII International Scientific Conference on Information, Communication and Energy Systems and Technologies ICEST 2013, Volume 2, pp. 619-622, Ohrid, FYR Macedonia, 26.6.-29.6.2013.

[8] HCP SHARK Quad-band GSM/GPRS, http://www.hcp.rs/wp-content/uploads/file/en%20-%20HCP%20Shark%20usermanual.pdf (accessed 10 February 2014)

[9] D.Milivojevic, V.Despotovic, V.Tasic and M.Pavlov, "Process control program as an element of distributed control system," Information technology and control, vol. 39, pp. 152-158, 2010.

[10] Gbridge, http://www.gbridge.com (accessed 10 February 2014)

[11] R.Venkateswaran, "Virtual Private Networks," IEEE Potentials, Lucent Techno. Bell Lab. USA, pp. 11-15, 2001.

[12] PCWorld, http://www.pcworld.com/article/234062/gbridge.html (accessed 10 February 2014)

[13] TeamViewer, http://www.teamviewer.com/en/index.aspx (accessed 10 February 2014)

Community structure in networks: Girvan-Newman algorithm improvement

Ljiljana Despalatović*, Tanja Vojković**, Damir Vukičević**
*The University Department of Professional Studies, University of Split, Croatia
**Faculty of Natural Sciences and Mathematics, University of Split, Croatia
ljiljana.despalatovic@oss.unist.hr

***Abstract*—Real world networks often have community structure. It is characteristic that the groups of nodes are connected denser within themselves and rarely with each other. The Girvan-Newman method for the detection and analysis of community structure is based on the iterative elimination of edges with the highest number of the shortest paths that go through them. By eliminating edges the network breaks down into smaller networks, i.e. communities. This paper introduces improved Girvan-Newman method where multi-edge removal is allowed, and presents the results of the application of both methods to the existing real social network (Zachary karate club), the computer-generated network and the tumor genes and their mutations network. The improved algorithm in practice reduces the number of operations, but retains the same computational complexity, so it cannot be applied to networks with a very large number of nodes. The most important feature of our improvement is that the result is graph-theoretical invariant, while original algorithm depends on the vertex labeling.**

***Index Terms*—Girvan-Newman, complex networks algorithms, edge betweenness**

I. Introduction

A network is a set of nodes and links connecting them. Studying network properties belongs into graph theory area which is a part of discrete mathematics. With the development of computers that can perform operations on very large amounts of data, modeling and analysis of network performance is extended to other scientific disciplines and becomes an interdisciplinary topic.

In mathematics, network is represented with a graph, where nodes are *vertices* and links *edges*. *Graph* is ordered pair $G = (V, E)$, where $\varnothing \neq V = V(G)$ is set of vertices, $E = E(G)$ is set of edges such that every $e \in E$ connects two vertices $u, v \in V$ called *endpoints* of e. Two connected vertices are *adjacent* vertices and we say that the edge and its endpoint are *incident*. A *loop* is an edge that connects a vertex to itself. The number of edges incident to a given vertex is the *degree* of that vertex, where every loop is counted twice. A graph is *regular* if all its vertices have the same degree.

In the real world, complex systems in different disciplines form networks such as social networks, World Wide Web, transport networks, citation network, metabolic networks, protein interaction networks, and gene regulatory network. Network exist in almost every area of human existence, from neuroscience to sociology, from economy to information technology. Those networks are called *complex networks* or *real-world networks*. Although graph theory is almost 300 years old (Euler solution for Könisberg bridges problem in 1736. is considered as the birth of the graph theory), the study of complex networks is a young discipline, only a little more than a decade old.

Unlike random (Erdos-Renyi model) and regular graphs, which were the focus of the graph theory before the Erdos-Renyi model, most real world networks are not random nor regular and have common topological properties. Some of them are small world effect, scale-free property, clustering or network transitivity, and community structure.

The property of complex networks that nodes are divided into groups or communities in which the links connecting nodes within the community are dense and the connections between nodes from different communities are sparse is the focus of this study. The possibility of detecting communities in networks can have practical applications. Communities in social networks could represent people with similar preferences or interests. The epidemiology could use the results to recognize the communities and the connections between people in order to prevent the spread of disease [1]. Kininmonth et al. [2] used the community detection methods to analyze genetic information exchange among the coral reefs through transport larvae. Communities on the Internet could be pages with similar content. From these examples, it is clear that the detection of communities in complex networks has broad application in a variety of scientific disciplines.

The rest of the paper is organized as follows. In Section 2 we present methods used for community structure detection in complex networks. Section 3 describes the Girvan-Newman method and our improvement of the method. In Section 4 we introduce modularity, a qualitative measure of network decomposition. Section 5 presents and discusses results achieved with our modification, and finally, Section 6 concludes the article.

II. Methods for community structure detection in complex networks

Main methods for finding communities in a network are variants of hierarchical clustering methods used in social network analysis [3]. Hierarchical clustering could be agglomerative or divisive, depending on decision to start from

empty network and add the edges to form communities, or to start with complete network and remove edges until the communities are formed. In both cases, the process could be represented with *dendrogram*, which is a hierarchical tree with clusters as a nodes in tree, and single nodes as leaves. In agglomerative process, dendrogram is built from leaves to the root, and in divisible process from root to the leaves (Fig. 3a).

A. Agglomerative methods.

The traditional hierarchical agglomerative clustering algorithm starts with empty graph which consist of nodes of original graph without edges. In every step of the algorithm edges are added, starting from "stronger" to "weaker" links.

Edge weight can be calculated in different ways. For example, edge weight could be the number of node-independent or edge-independent paths between vertices. Two paths are node-independent if they share no other vertices than the path endpoints. Similarily, they are edge-independent if they share no edges. The number of those paths represents the number of vertices (or edges) which should be removed from the graph in order to disconnect path endpoints [4].

New communities are formed in the consecutive steps of the algorithm. Iterative algorithm process could be stopped at any iteration, but usually, it is done until all the edges are added to the graph. In the end, all the nodes and edges form one community.

B. Divisive methods.

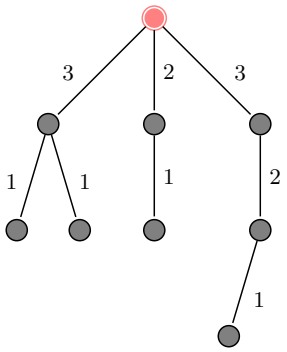
Opposite of agglomerative methods, divisive methods start from the complete graph and, in every iteration, remove the edges with the highest weight. In every step weight calculation is repeated, since the weight of remaining edges changes with edge removal. The Girvan-Newman algorithm, which will be described in detail in the next section, is an example of divisive algorithm. In that algorithm, the order of removal of the edges with the highest weight is not defined, so it could produce different results depending on implementation. Modifications of that algorithm exist in which the community structure is built for every possible order of highest weight edge removal.

III. Girvan-Newman method

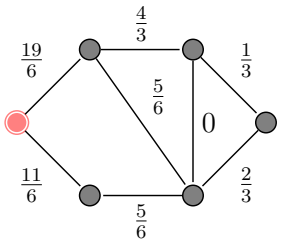
Girvan-Newman method is divisive method where edge weight is the number of shortest paths passing through the edge. That value is called *edge betweenness* and it is a generalization of central vertex betweenness which determines vertex influence on other vertices in network. Vertex betweennes is the number of shortest paths passing through the vertex, therefore, edge betweenness is the number of shortest paths passing through the endpoints of the edge.

We can describe Girvan-Newman algorithm in the following way:

1) Calculate edge betweenness for every edge in the graph.



(a) A tree



(b) Not a tree

Figure 1: The number of shortest paths in a graph: a) In a tree there is unique path from source to every vertex. b) If graph is not a tree there is more than one path to every vertex. Each edge has weight which is fraction of the paths flow through the edge.

2) Remove the edge with highest edge betweenness.
3) Calculate edge betweenness for remaining edges.
4) Repeat steps 2-4 until all edges are removed.

In order to calculate edge betweenness it is necessary to find all shortest paths in the graph. The algorithm starts with one vertex, calculates edge weights for paths going through that vertex, and then repeats it for every vertex in the graph and sums the weights for every edge.

Let us choose source vertex $s \in V$ in a graph. In the most simple case graph is a tree (Fig. 1a) and there exists only one shortest path from the source vertex to any other vertex. Starting from the leaves we assign the value 1 to the edges that connect leaves with the rest of the tree (because there is only one shortest path to s passing through that edge). Moving upwards in the tree we assign edge value as a sum of values assigned to the edges directly below it, increased by 1. The number of shortest paths in the tree from source vertex to every other vertex passing through particular edge is that edge weight value. By repeating the process for every vertex and calculating the sum of weight values for every edge, we calculate edge betweenness for every edge.

If a graph is not a tree (Fig. 1b), it is possible that more than one shortest path connects source vertex with some other vertex. In that case, value $\frac{k}{l}$ is assigned to the edge on shortest path to source vertex, where k is the number of shortest paths to source vertex from the endpoint of the edge that is closer to source and l is the number of shortest paths from source vertex to other endpoint of the edge, and multiplied by the

number of shortest paths from source that pass through edges below farther vertex incresed by one. Assignment starts from the edge that has maximum distance from source vertex.

The algorithm for calculating edge betweenness is performed in two parts. In the first part of the algorithm, using breadth-first search, distance from source is assigned for every vertex and also the number of shortest paths from source to vertices. In the second part, starting from edge incident to the vertex with maximum distance from the source vertex as endpoint (the last vertex visited in the first part on the algorithm), the numbers of shortest paths passing through edges are calculated for every edge. For every vertex $i \in V$ the triple (d_i, w_i, b_i) is calculated, where d_i is the distance from the source vertex, w_i is the number of shortest paths from source vertex to vertex i, and b_i is the number of shortest paths between source vertex to any vertex in graph that pass through vertex i.

For this algorithm we denote $Adj(v)$ as the set of all vertices adjacent to $v \in V$.

The first part of the algorithm for vertex marking:

1) For initial vertex $s \in V$ let $d_s = 0$, $w_s = 1$, $b_i = 0$.
2) Let $d_v = \inf$, $w_v = 0$, $b_v = 1$ for all $v \neq s \in V$.
3) Create queue Q, $Q \leftarrow \{s\}$. Create list L, $L \leftarrow \{s\}$.
4) While Q is not empty:
   a) Dequeue $i \leftarrow Q$.
   b) For each vertex $j \in Adj(i)$:
      i) If $d_j = \inf$ then $d_j = d_i + 1$, $w_j = w_i$. Enqueue $j \rightarrow Q$. Push $j \rightarrow L$.
      ii) If $d_j \neq \inf$ and $d_j = d_i + 1$ then $w_j += w_i$.
      iii) If $d_j \neq \inf$ and $d_j < d_i + 1$, do nothing.

Efficient implementation of this part of the algorithm could be done by using abstract data type queue.

The second part of the algorithm starts from the vertex that was last marked in the first part of the algorithm and visits vertices in reverse order than they were visited in the first part of the algorithm. Only one shortest path from source passes through the last marked vertex.

The second part of the algorithm for edge betweenness calculation:

1) While L is not empty:
   a) Pop $i \leftarrow L$.
   b) For each vertex $j \in Adj(i)$:
      i) If $d_i < d_j$ then $b_i = 1 + \sum_j \sigma_{ij}$.
      ii) If $d_i > d_j$ then $\sigma_{ij} = \frac{w_j}{w_i} * b_i$.

Both parts of the algorithm are performed for all source vertices s and edge betweeness for every edge is calculated as a sum of the edge betweennesses calculated in every step.

The computational complexity of this part of algorithm is $O(mn)$, where m is the number of edges and n is the number of vertices. After each edge betweenness calculation, the edge with highest edge betweenness is removed and the algorithm is repeated until there is no remaining edge. The algorithm complexity is therefore $O(m^2n)$.

In complex networks it is often the case that more edges have the same highest edge betweenness. Since the recalculation of edge betweenness has $O(mn)$ complexity, in order to reduce the number of calculations we could remove all edges with the highest edge betweenness in the same step. On the other side, Girvan-Newman algorithm does not say anything about the order of removal for those edges leaving that detail to implementation.

Consequently, we introduce modification of the Girvan-Newman algorithm:

1) Calculate edge betweenness for every edge in the graph.
2) Remove all edges with highest edge betweenness.
3) Recalculate edge betweennes for remaining edges.
4) Repeat 2-4 until graph becames empty.

Worst-case time complexity is still $O(m^2n)$, but in networks with strong community structure the number of calculations could be significantly reduced.

The main problem of Girvan-Newman algorithm is the fact that it is not really an algorithm that has a graph as input and community structure as output. Namely, if the labeling of the vertices of the graph is rearranged, than the result of Girvan-Newman algorithm may change. Hence, there is no unique output for a given input (which should be a property of any algorithm). Our modification amends this shortcoming of the Girvan-Newman algorithm. Namely, our community structure indeed does not depend on the labeling of the vertices of the graph.

Also, this modification may reduce the number of operations of the algorithm. However, this strongly depends on the observed graph. There are graphs for which no improvement will be obtained and there are graphs for which significant improvement will be obtained. The most drastic example may be edge-transitive graphs (i.e. graphs such that for each two edges e and f there is the automorphism that maps e to f). Standard Girvan-Newman algorithm has complexity $O(m^2n)$ on this graphs also, while our algorithm does not repeat calculation for each edge, but does everything in a single step reducing complexity to $O(mn)$.

IV. Modularity

The algorithms used for community structure detection in networks give an overview of possible communities. Simulation of decomposition done by Girvan-Newman method gives more information of network structure. Knowing network structure, it is possible to predict critical connections in the network, and therefore, control the network. For example, in the electricity distribution network detecting critical connections could be a method to prevent the collapse of the system that has occurred in the United States in 1965, 1996, and in September 2011.

However, the question is when an optimal decomposition of a network into communities is reached. To answer that ques-

tion it is not necessary to decompose network in communities of the size one. Qualitative measure of network decomposition is called **modularity** [5].

For a particular partition of a network into k communities, let us define $k \times k$ matrix E such that e_{ij} is the fraction of number of edges that connect vertices between community i to community j in the total edge number. Then on the diagonal of the matrix E is the fraction of edges that are located within the same community, so the trace of the matrix $tr(E)$ is the fraction of edges that will not be removed in process of removing edges. A good partition is the partition with high value of the trace. However, that number will be the highest in the case where all the vertices are in the same community, so it does not always give necessary information about the structure of the communities.

We define modularity:

$$Q = \sum_i (e_{ii} - a_i)^2 = tr(e) - \|E^2\|,$$

where $\|\|$ is the sum of matrix elements, as the fraction of edges that connect vertices within communities decreased with the expected fraction of edges within partition in random graph [6]. When the fraction of edges within the communities is higher than it is in random graph, the value is $Q = 0$. As Q is approaching the value 1, the community structure in network is better. In most cases, the value of Q is between 0.3 and 0.7 [5].

Value Q is calculated in every step of divisible algorithms. The maximum value q gives us the best partition of the graph.

V. Results

We tested the Girvan-Newman algorithm with the single edge removal and modified algorithm with the multiple edges removal on three different complex networks. The first, Zachary's karate club; the second, computer generated random network; and the third, network of cancer genes with co-occurring and anti-co-occurring mutations.

A. Zachary's Karate Club

Zachary's karate club is the example of the real-world network for which it is known how the network was split in reality. Fig. 2 shows the friendship network between 34 members of a karate club at a US university observed by an anthropologist Wayne Zachary for a period of three years, from 1970 to 1972 [7].

At the beginning of the study there was a conflict between the club president (node 34) and the karate instructor (node 1). During the study club was divided into two parts and has become the most frequently tested dataset in the network communities detection area. The Girvan-Newman algorithm divides the network into two communities with an accuracy of 97% (which is also Zachary's accuracy) with wrong classification of the node 3 (in Zachary's study the node 9 was wrongly classified). Fig. 3a shows the dendrogram generated using the Girvan-Newman algorithm, while Fig. 3b shows the dendrogram generated using implementation of modified method with multiple edge removal. It is clear that our method leads to faster separation of the communities i.e. reduced number of operations (Fig. 4). The algorithm modularity values are high when the network is divided into two communities which is expected concerning the nature of community and the process of the separation. However, modularity is the largest in the network partition to the three communities. On Fig. 2 we can see that there exists a community of five nodes (5, 6, 7, 11, 17) with high number of links within community and small number of links to other communities.

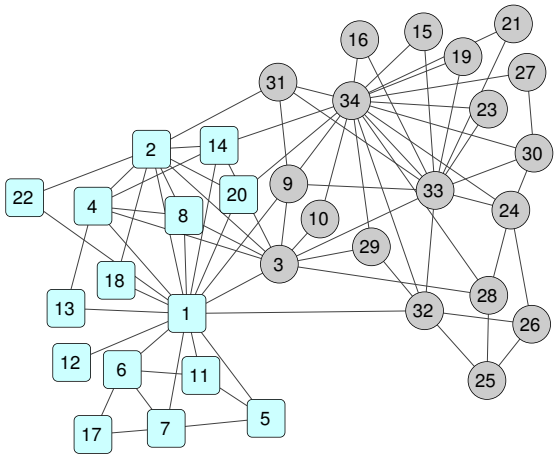


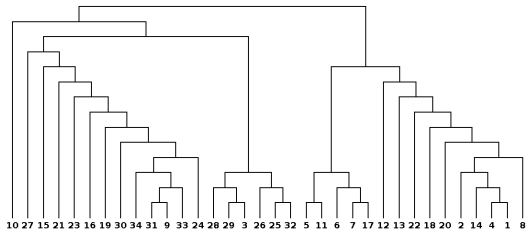
Figure 2: Zachary karate klub

B. Computer generated network

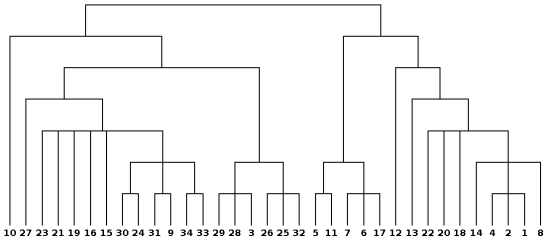
In addition to real-world examples we analysed computer-generated network with strong community structure. For this purpose, network with 48 vertices divided into communities of 16 vertices is generated with probability p_{in} that edge is within community and p_{out} that edge is between communities. Fig. 5 presents the modularity values for Girvan-Newman algorithm and for our multiple-edge-removal modification of the algorithm. The modularity function maximum is reached in both algorithms when the network is divided in three communities, but needed fewer steps to complete calculation in the latter case. Both algorithms correctly classify all nodes.

C. The network of cancer genes with co-occurring and anti-co-occurring mutations

The network of cancer genes with co-occurring and anti-co-occurring mutations (CCA network) is presented in [8]. CCA networks are studied in order to better understand the development of cancer and community detection in such networks is an important property. The cancer genes are nodes and co-occurring and anti-co-occurring mutations are links between them. The tested network has 306 nodes and 1366

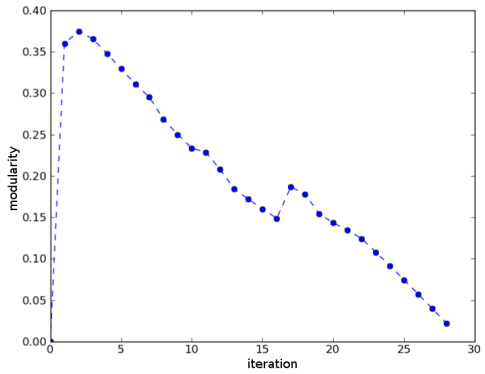


(a) Single edge removal

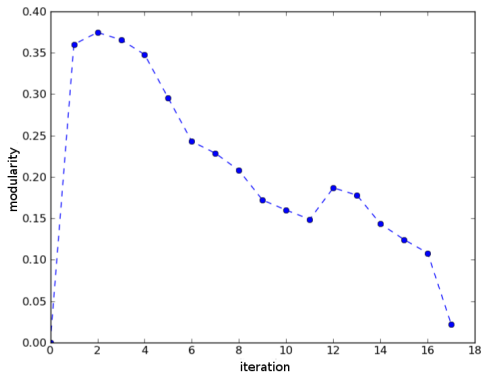


(b) Multiple edge removal

Figure 3: Zachary's karate club



(a) Single edge removal

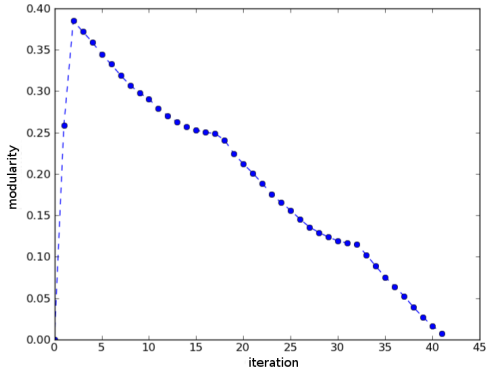


(b) Multiple edge removal

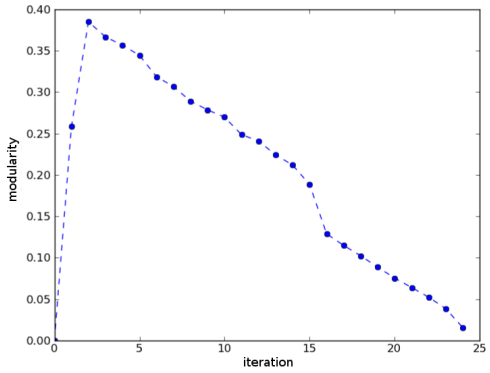
Figure 4: Zachary's karate club - modularity

edges, of which 1355 are co-occuring and 11 are anti-co-occuring.

The number of iterations in this example needed to divide



(a) Single edge removal



(b) Multiple edge removal

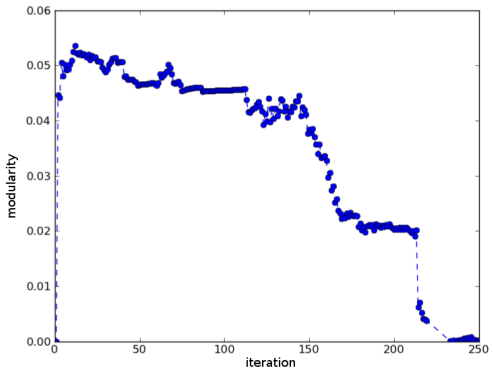
Figure 5: Computer generated network modularity

the network into communities with modularity 0 is much smaller when multiple edge removal is done. Both algorithms correctly classify all the nodes.

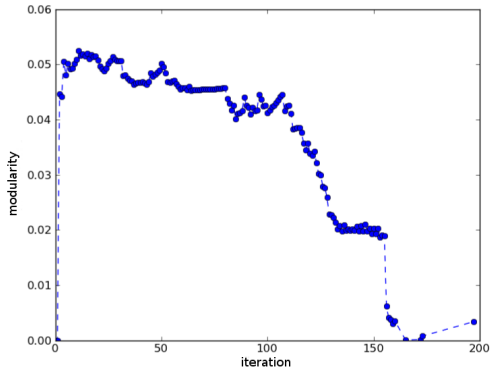
VI. Conclusion

Community detection in networks is important for different scientific and practical areas. It is not a surprise that in this article we refer to articles of physicists, sociologists, mathematicians, biologists, epidemiologists, etc. Complex network research is young discipline and with its rapid development reflects the flourishing of interdisciplinary research. Graph theory is widely used in various applications in different areas. Some of them are mentioned in this article. The robustness of a network and analysis of system collapse due to the elimination of nodes or links between nodes is of great importance in the study of networks. How can we stop the propagation of diseases? How can we determine the most important airlines or highways? Algorithms decribed in this article help us to experience the world around us.

Girvan-Newman algorithm is one of the first algorithms that deals with detecting communities in networks and as such suffers from certain "childhood diseases". Number of operations is proportional to m^2n, or n^3 for sparse networks, which limits performance on networks with up to tens of



(a) Single edge removal



(b) Multiple edge removal

Figure 6: A network of cancer genes with co-occurring and anti-co-occurring mutations modularity

thousands of nodes. Methods based on modularity, [6] [9] [10], the use of the eigenvectors of matrices [11], and greedy algorithms [12] give us smaller complexity than Girvan-Newman method. Regardless, in this article we focused on the Girvan-Newman method and its improvement, as a starting point for any subsequent methods. This small algorithm improvement reduces the number of operations, but still not enough to apply it on large complex networks. But on smaller networks it has a high percentage of correct classification of nodes and it is independent of vertex labeling, which puts it ahead of other methods.

REFERENCES

[1] S. Kitchovitch and P. Liò, "Community Structure in Social Networks: Applications for Epidemiological Modelling," *PloS one*, vol. 6, no. 7, p. e22220, 2011.

[2] S. Kininmonth, M. J. H. Van Oppen, and H. P. Possingham, "Determining the community structure of the coral seriatopora hystrix from hydrodynamic and genetic networks," *Ecological Modelling*, vol. 221, no. 24, pp. 2870–2880, 2010.

[3] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D. Hwang, "Complex networks: Structure and dynamics," *Physics Reports*, vol. 424, no. 4-5, pp. 175–308, 2006.

[4] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 12, pp. 7821–7826, 2002.

[5] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physical Review E - Statistical, Nonlinear and Soft Matter Physics*, vol. 69, no. 2 Pt 2, p. 16, 2004.

[6] A. Clauset, M. E. J. N. Newman, and C. Moore, "Finding community structure in very large networks," *Physical Review E*, vol. 70, no. 6, pp. 1–6, 2004.

[7] W. W. Zachary, "An information flow model for conflict and fission in small groups," *Journal of Anthropological Research*, vol. 33, no. 4, pp. 452–473, 1977.

[8] Q. Cui, "A network of cancer genes with co-occurring and anti-co-occurring mutations," *PLoS ONE*, vol. 5, no. 10, p. 8, 2010.

[9] M. E. J. Newman, "Fast algorithm for detecting community structure in networks," *Physical Review E*, vol. 69, no. 2, pp. 1–5, 2004.

[10] D. Chen, Y. Fu, and M. Shang, "A fast and efficient heuristic algorithm for detecting community structures in complex networks," *Physica A: Statistical Mechanics and its Applications*, vol. 388, no. 13, pp. 2741–2749, 2009.

[11] M. E. J. Newman, "Finding community structure in networks using the eigenvectors of matrices," *Phys. Rev. E*, vol. 74, p. 036104, Sep 2006.

[12] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, p. P10008.

Performance Evaluation of Websocket Protocol for Implementation of Full-Duplex Web Streams

D. Skvorc, M. Horvat, and S. Srbljic
University of Zagreb/School of Electrical Engineering and Computing, Zagreb, Croatia
dejan.skvorc@fer.hr, matijah@acm.org, sinisa.srbljic@fer.hr

Abstract – Besides traditional synchronous HTTP request-response communication paradigm that was used on the Web for decades, modern Web services require more flexible communication system that enables asynchronous messaging between clients and servers, as well as server-initiated data delivery. Among several Web-based asynchronous communication paradigms emerged recently, the *Websocket* protocol and corresponding *Websocket* API are accepted as a pivotal framework for implementation of full-duplex asynchronous Web streams.

In this paper, we evaluate the performance of the *Websocket* protocol with respect to underlying TCP protocol. We compare the two against the latency and amount of generated network traffic. The results show that, except a small overhead imposed due to initial handshaking, *Websocket*-based communication does not consume any more network traffic than plain TCP based communication. However, it is still slightly inferior in terms of latency.

I. INTRODUCTION

Since its announcement in the early 1990's, the HTTP protocol is the primary communication protocol for the World Wide Web [1, 2]. HTTP is an application-level protocol built on top of the TCP/IP protocol stack. It was designed to operate in client-server settings with strict separation of concerns between clients and servers. Client is responsible to ask for data or service from a server by sending an HTTP request. On the other hand, server responds with an HTTP response, carrying the results of execution of the action required by the client. Server hosted data and services are delivered to the clients upon their requests only, rather than being delivered autonomously when they become available on the server. Request-response paradigm of the HTTP protocol dictates a synchronous operation of the communication parties, which means that every HTTP request from a client is followed by an HTTP response sent by server. This prevents or at least diminishes the application of the protocol in modern Web-based environments that require more sophisticated and flexible communication models.

Recent research in the field of real-time Web tried to overcome the limitations of the HTTP protocol by upgrading the natively non-asynchronous communication paradigm and tuning it to operate in asynchronous settings. That resulted in various conceptual models, such as HTTP-based polling and long polling [3], and technologies, such as *Comet* [4] and *BOSH* [5]. However, each of these technologies either provides only a subset of the required functionalities or tradeoffs the system's simplicity, scalability, and performance for the sake of functionality. For example, *Comet* and *BOSH* only simulate the asynchronous communication between client and server by maintaining two TCP connections, one for upstream, and another one for downstream communication. Recently introduced *Websocket* protocol [6, 7] and *Websocket* API [6, 8], developed as part of the HTML 5 standard [9], are currently the only true asynchronous full-duplex communication framework for the Web. *Websocket* protocol and *Websocket* API bring the expressiveness and flexibility of the TCP sockets to the Web applications, where access to the raw TCP sockets is otherwise not allowed. *Websocket* technology is now natively supported in all major Web browsers and Web servers, while *Websocket* API libraries are available for almost any programming language used in Web development [10, 11, 12, 13, 14, 15].

Many discussions and written reports have been published so far comparing the performance of the *Websocket* and related technologies [16, 17, 18]. All of them highlight the *Websocket* as a breakthrough technology for implementation of asynchronous full-duplex real-time Web streams. However, little is known about the performance degradation this high-level protocol imposes over a plain TCP socket based communication, which still remains an underlying transport mechanism.

In this paper, we compare the *Websocket* and the TCP protocol against two performance-related dimensions: latency and amount of generated network traffic. Although both protocols provide similar communication capabilities, in general they should not be considered as alternatives to each other since they operate at different levels of the Internet protocol stack. While native applications may directly access the raw TCP protocol through the sockets API, until recently Web applications were limited to the HTTP protocol and its synchronous request-response communication paradigm. The *Websocket* protocol is an application-level protocol built on top of the TCP and specifically designed to expose TCP-like communication capabilities to Web applications, which do not have access to the raw TCP sockets. Thus, the goal of our research was to determine the amount of inevitable performance loss if full-duplex asynchronous communication system operates within the limitations of Web settings.

The rest of this paper is organized as follows. Section 2 outlines the most important characteristics of the *Websocket* protocol and shortly describes the *Websocket* API. We emphasize the *Websocket* handshake and *Websocket* frame as key factors used during the evaluation. In Section 3, we describe the evaluation procedure, the experiments for benchmarking the performance of the *Websocket* protocol, and the results gained from these experiments. Section 4 concludes the paper.

II. WEBSOCKET PROTOCOL AND WEBSOCKET API

The *Websocket* protocol [6, 7] is an application-level protocol built on top of the TCP. It enables bidirectional data transport in Web sessions, and is considered as a viable alternative to HTTP polling techniques [3], which are implemented as tradeoffs between functionality, efficiency, and reliability. As such, the *Websocket* protocol is suitable for Web-based near real-time traffic required in games, chats, stock exchange information systems, multimedia systems, remotely controlled systems, etc.

Although similar in functionalities to the regular TCP protocol, the *Websocket* protocol is backward compatible with the existing Web infrastructure. The *Websocket* handshake is compliant with the HTTP 1.1 specification. Furthermore, *Websocket*-based communication inherits all the benefits of the existing Web infrastructure, such as

- native support in Web browsers, including origin-based security model
- proxy and firewall traversal
- URL-based endpoints which enable multiple, and potentially unlimited number of services running on a single TCP port
- elimination of length limits imposed by plain TCP protocol

Fig. 1 shows a sequence diagram of a *Websocket* session and emphasizes the communication overhead the *Websocket* protocol requires over the plain TCP protocol. Since the *Websocket* is a TCP-based protocol, it requires a TCP connection to be established between client and server before any *Websocket*-based interaction can occur. Handshaking a TCP connection requires three messages to be exchanged between client and server. At this point, two peers that have direct access to the plain TCP protocol can start sending application-specific payload data to each other. However, for *Websocket*-based communication, a *Websocket* session should be established first.

To establish a session, client sends a *Websocket Upgrade Request* to the server, upon which server responds with a *Websocket Upgrade Response*. From this point forward, even the Web-based client and server can send data back and forth in asynchronous full-duplex mode.

A. Protocol Handshake

The *Websocket* handshake takes the advantage of the HTTP protocol in the session establishing phase. The *Websocket Upgrade Request* is a regular HTTP GET request with session endpoint specified as a request URI component. HTTP headers `Upgrade` and `Connection` are indication to the server that the client wants to switch from regular HTTP to the *Websocket* protocol for the rest of the session. Other fields enable negotiation of additional *Websocket* options, such as protocol version, subprotocol indication, and security-related parameters, such as script origin and *Websocket* key used to confirm the connection acceptance.

```
GET /chat HTTP/1.1
Host: server.example.com
Upgrade: websocket
Connection: Upgrade
Sec-WebSocket-Key: dGhlIHNhbXBsZSBub25jZQ==
Origin: http://example.com
Sec-WebSocket-Protocol: chat, superchat
Sec-WebSocket-Version: 13
```

If the server supports the *Websocket* protocol, it replies with an HTTP response with a status code set to `101 Switching Protocols`. From this point forward, the HTTP-based communication is finished and all further communication is based on *Websocket* frames.

```
HTTP/1.1 101 Switching Protocols
Upgrade: websocket
Connection: Upgrade
Sec-WebSocket-Accept: s3pPLMBiTxaQ9kYGzzhZRbK+xOo=
Sec-WebSocket-Protocol: chat
```

B. Data Framing

Units of data transfer in the *Websocket* session are called *Websocket* frames. The *Websocket* protocol supports binary data frames, textual UTF-8 encoded data frames, and control frames for protocol-level signaling. Minimal framing information is added to each payload data, as shown in Fig. 2.

Each *Websocket* frame begins with at least 2 bytes long header prepended to the payload data. Depending on the length of the payload data and the direction of the communication, the length of the header may increase up to 14 bytes.

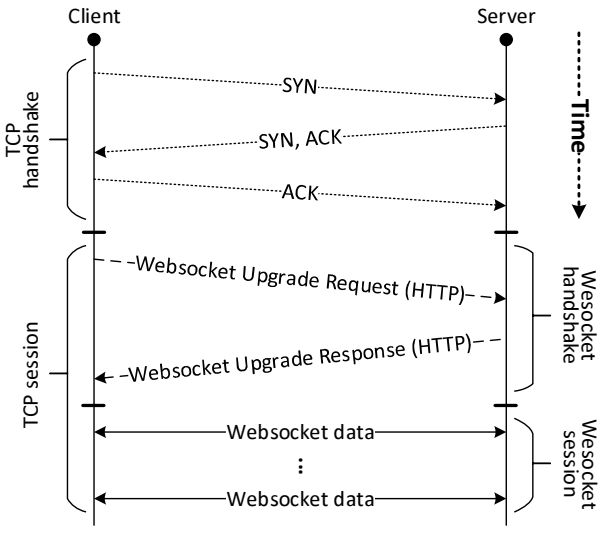


Figure 1. *Websocket*-over-TCP sequence diagram

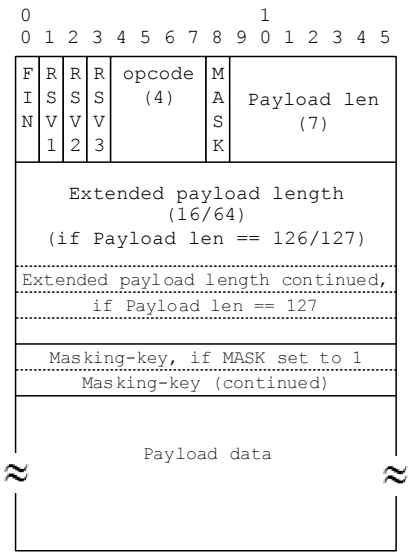


Figure 2. *Websocket* frame structure

First byte contains control bits and 4-bit operation code `opcode`, which defines whether the payload data should be interpreted as binary data, textual data, or protocol-level signaling. Next bit is a `MASK`, which if set to 0 signals that the payload data are transferred in clear form, and obfuscated otherwise. Following 7 bits are used to indicate payload data length. If payload data length is up to 125 bytes, this value is directly specified in `Payload len` field. If payload length is between 126 and 65535 bytes, the `Payload len` field is set to 126, and two additional bytes are added to specify the actual payload data length. Finally, if payload data are longer than 65535 bytes, the `Payload len` field is set to 127, and 8 additional bytes are added to specify the actual payload data length.

Client-originated frames must additionally be obfuscated with a 32-bit `Masking key`, specified in the following 4 bytes of the frame header. This is required to disable the unwanted payload processing performed by the network intermediaries that may discover patterns of HTTP traffic in *Websocket* frames. For example, if a *Websocket* frame contains a pattern of HTTP traffic, an HTTP proxy may accidentally decide to cache the data as if they were regular HTTP traffic. This is a security vulnerability known as HTTP proxy cache poisoning [19, 20]. For frames sent from server to the client, this masking is not necessary, thus 4 bytes containing a `Masking key` are omitted in server-originated frames. The *Websocket* frame ends with payload data of specified length.

Table 1. Event-driven *Websocket* API

Callback	Description
`onopen`	invoked when *Websocket* session is established, signalizes that the protocol is ready to transfer payload data
`onerror`	invoked whenever an error occurs
`onclose`	invoked when one of the peers has terminated the session
`onmessage`	invoked when an incoming message from another peer has arrived

C. *Websocket API*

Besides the protocol itself, *Websocket* specification also defines the API to interact with the protocol [6, 8]. The *Websocket* API is defined by its states of readiness, responses to a networking or messaging event, and message types available for data transfer between client and server. Table 1 summarizes the protocol events exposed to applications through an event-driven callback-based API.

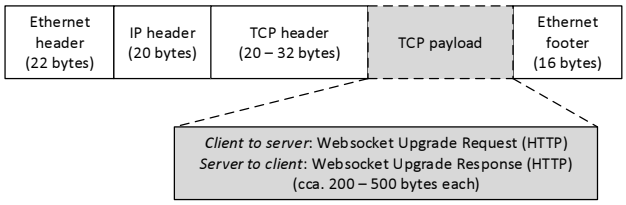
III. PERFORMANCE EVALUATION

Performance evaluation of the *Websocket* and the TCP protocol we present in this paper consists of comparison of the network traffic and data transfer time these two protocols consume to transfer given amount of data between client and server. Since the size of data frames in both protocols depends solely on the size of the payload data, we evaluate the protocols' efficiency from the network traffic perspective analytically, using the protocol specifications defined in *Websocket* RFC [7] and TCP RFC [21], respectively. On the other hand, data transfer time is evaluated experimentally in a laboratory test bed.

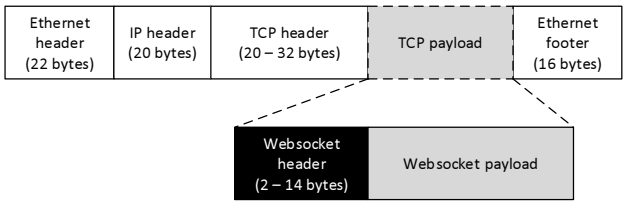
A. *Analitical Evaluation of Network Traffic*

Fig. 3 shows the embedding of the protocol fields in the Internet protocol stack. Regardless of whether the two peers communicate directly via TCP or through the *Websocket* protocol, lower level protocol fields, including the TCP header, are always present in all network packets. Thus, our analysis takes into account only the overhead the *Websocket*-based communication incurs over the plain TCP-based communication. This overhead includes two HTTP messages transferred during the *Websocket* handshake, as well as *Websocket* frame headers for each frame transferred during the session.

The overhead of the *Websocket* handshake is fixed in length and typically counts few hundreds of bytes. Since this handshake is performed once per session, its significance decreases with each new *Websocket* frame carrying payload data. Thus, the rest of our evaluation is focused on long-running sessions, where the impact of the



a) Network packet structure during the *Websocket* handshake



b) Network packet structure during the *Websocket* session

Figure 3. Protocol relationship in the Internet protocol stack

initial handshake is diminished and the only observable traffic overhead is caused by the *Websocket* data framing.

When given amount of payload data are transferred using plain TCP protocol, these data are directly embedded as TCP payload. However, when transferred as a *Websocket* frame, then the TCP payload consists of both payload data and a *Websocket* frame header, as shown in Fig. 3b. This relation is given in (1), where *data* is the length of the payload data, P_{TCP} is the length of the TCP payload when these data are transferred as a TCP payload, P_{WS} is the length of the TCP payload when the same data are transferred as a *Websocket* payload, while H_C is the length of the *Websocket* frame header for client-originated frames.

$$\begin{aligned} P_{TCP} &= data \\ P_{WS} &= data + H_C \end{aligned} \tag{1}$$

As shown in Fig. 2 and described in Section 2, the *Websocket* protocol uses variable frame header length that depends on the size of the payload data. The *Websocket* framing scheme is given in (2), where *data* is the length of the payload data, H_S is the length of the frame header for server-originated frames, while H_C is the length of the frame header for client-originated frames. As described in Section 2, masking of client-originated frames requires 4 additional header bytes in relation to server-originated frames.

$$\begin{aligned} H_S &= \begin{cases} 2, & 0 \le data \le 125 \\ 4, & 126 \le data \le 65535 \\ 10, & data \ge 65536 \end{cases} \\ H_C &= H_S + 4 \end{aligned} \tag{2}$$

The network traffic overhead O_P a *Websocket*-based communication incurs over a plain TCP-based communication is given in (3). The dependence of this overhead on the size of the payload data is given in Fig. 4. The graph shows a worse case, with a client-to-server directed communication only, where each *Websocket* frame is prepended with 4 additional bytes carrying a masking key.

$$O_P = \frac{P_{WS} - P_{TCP}}{P_{TCP}} \cdot 100\% \tag{3}$$

Significant difference in performance of the two protocols occurs only for tiny messages of just few bytes in size. In this range, the *Websocket* protocol generates almost an order of magnitude more network traffic than the TCP. However, even with this additional traffic, the total overhead is no more than six bytes, as defined in (2). For bigger messages, the *Websocket* frame size converges very fast towards the non-framed TCP payload size. For messages bigger than 1 kB, the overhead is less than 1 %. The relation shown graphically is given for small range of payload data sizes only, where the difference in the behavior of the two protocols is significant. We intentionally omitted the rest of the graph, where this relation asymptotically approaches to 0.

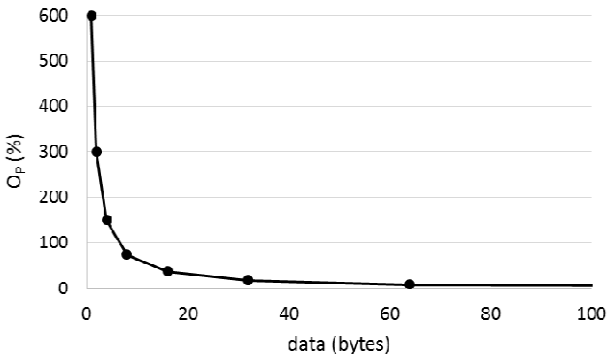


Figure 4. Network traffic overhead of the *Websocket* protocol in relation to the TCP protocol

Except the initial handshaking performed through one HTTP round trip at the beginning of the session, the amount of network traffic generated by the *Websocket* protocol is comparable to that generated by the plain TCP. With a carefully designed framing scheme, which gradually decreases the amount of framing information as the payload data decrease in size, the *Websocket* protocol generates negligibly more network traffic than plain TCP, even for the smallest amounts of data.

B. Experimental Evaluation of Data Transfer Time

Experimental environment for measuring data transfer time consists of two host machines, one playing the role of a server, while other being a client. Hardware and software configuration of these host machines are given in Table 2.

Both protocols are similar in terms that they both require particular form of negotiating the protocol parameters between communicating parties, prior to be able to transfer actual data through the network. Thus, we separately measure the time required to negotiate a connection and the time required to transfer a chunk of payload data from one side to the other.

In our first experiment, we measure the time required for client and server to establish a TCP connection and a *Websocket* session, respectively. Results are given in Fig. 5.

Table 2. Experimental environment configuration

	Client	**Server**
Hardware	CPU: AMD Turion II P520 RAM: 6 GB	CPU: AMD Athlon X2 5000+ RAM: 5 GB
OS	Windows 8 64-bit	Windows 8 64-bit
Network	1000BASE-T (Gigabit Ethernet, host machines directly connected using UTP Cat5 Ethernet cable)	
TCP implementation	*java.net.Socket* (Java JDK 1.7)	*java.net.Socket* (Java JDK 1.7)
Websocket implementation	*org.eclipse.jetty.websocket.**client**.** (Jetty 9.1.0)	*org.eclipse.jetty.websocket.**servlet**.** (Jetty 9.1.0)

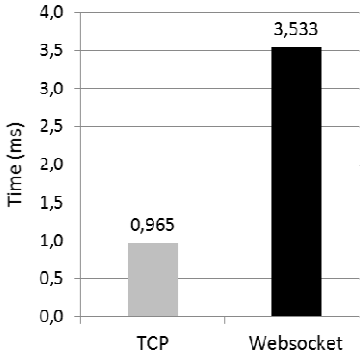


Figure 5. Comparison of TCP and *Websocket* handshake duration

In our laboratory environment with client and server being directly connected with a Gigabit Ethernet cable, establishing a *Websocket* session lasts 3.7 times longer than establishing a TCP connection. The reason for such a slow performance of the *Websocket* protocol is the fact that the *Websocket* protocol is not a real transport protocol, but rather an application-level protocol built on top of the TCP. This means that each *Websocket* session requires a TCP connection to be established first, as shown in Fig. 1. Thus, setting up a *Websocket* session includes all the actions required for setting up a TCP connection, plus an additional HTTP round trip for a *Websocket* handshake. Besides the time required to transfer two HTTP messages through the network, this additional handshake also includes the time necessary for client and server to allocate the necessary resources and initialize the *Websocket* API callbacks required to perform a *Websocket*-based communication in the rest of the session.

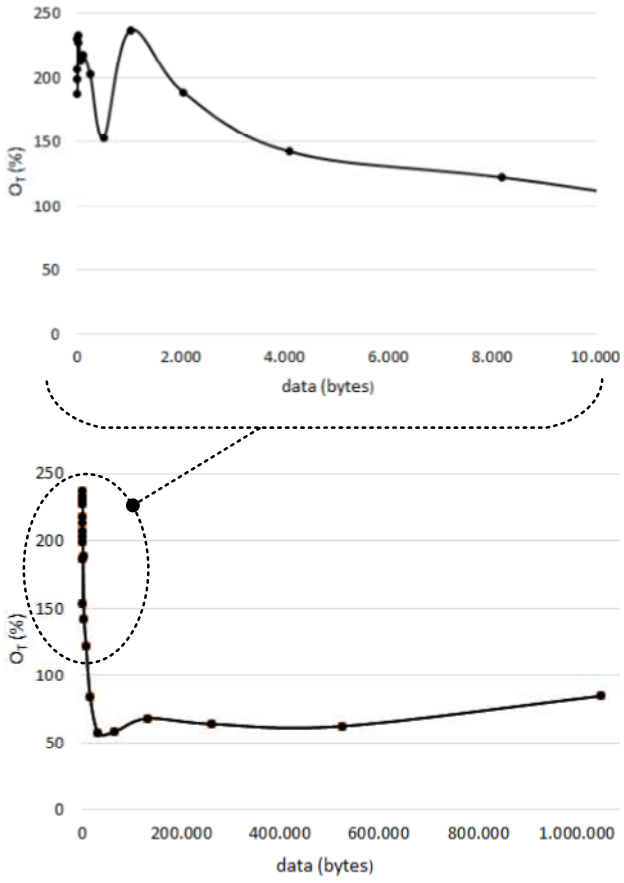


Figure 6. Data transfer time overhead of the *Websocket* protocol in relation to the TCP protocol

In our second experiment, we measure the data transfer time between client and server once the TCP connection and the *Websocket* session have been established. Client generates a given amount of payload data and sends them to the server. Upon receiving all the data, server echoes the same data back to the client. Client measures a round trip time. Results are given in Fig. 6. On vertical axis, we show a relative time overhead O_T a *Websocket*-based communication incurs over a plain TCP-based communication. This overhead is calculated by (4), where T_{TCP} is the time required to transfer the data using plain TCP protocol, while T_{WS} the time required to transfer the same amount of data using the *Websocket* protocol.

$$O_T = \frac{T_{WS} - T_{TCP}}{T_{TCP}} \cdot 100\% \quad (4)$$

Here we notice that, despite of the negligible difference in the overall amount of data being transferred (payload + *Websocket* frame headers), the *Websocket* protocol performs significantly slower than the TCP through the whole spectrum of the test cases. The performance drop is more significant for small messages and gets better as message size increases. The overhead of the *Websocket*-based communication fluctuates between 150 % and 250 % for very small messages. For bigger messages, as message size becomes more and more dominant factor in data processing and data transfer, the overhead gets stable at roughly 60-70 %.

There are several factors that impact the performance of the *Websocket* protocol. First reason is the relationship of the TCP and the *Websocket* protocols in the Internet protocol stack. The *Websocket* operates at a higher level of the stack and consumes more computational resources that deliver the data from the TCP layer to the upper layers. Second, opposed to TCP, the *Websocket* protocol uses an event-driven callback-based API to deliver data to Web applications, which requires additional application-level data handling. Third, every message sent from a client to a server requires additional processing to mask the data with a 32-bit masking key on a client, and to unmask them on a server. One more reason, yet unconfirmed in our experiments, might be the implementation of the protocol itself. While TCP protocol is intensively used for decades and highly optimized libraries are available for almost any programming language, the *Websocket* protocol emerged just a few years ago, with little production systems deployed so far, and with early implementations not yet verified by wide research and developer community.

IV. CONLUSION

In this paper, we have analytically and experimentally compared the *Websocket* and the TCP protocol against two performance-related dimensions: amount of generated

network traffic and data transfer time. As expected, TCP protocol slightly outperforms the *Websocket* protocol in both of these dimensions because the *Websocket* operates at the application layer and uses the underlying TCP as a transport layer protocol. While the difference in the amount of generated network traffic is minimal and negligible for any practical use, the degradation in data transfer time is still remarkable due to operation at a higher level of the Internet protocol stack, data masking, and event-driven callback-based interface to applications. However, the advantage of the *Websocket* protocol is its alignment with the existing Web infrastructure, where low-level TCP protocol is not directly applicable.

Since the *Websocket* protocol augments the Web with a TCP socket-like communication capabilities that were not available or were only inefficiently simulated in HTTP-based applications, slight drop in performance in relation to the plain TCP protocol still makes it a powerful mechanism for implementation of full-duplex asynchronous Web-based data streams. If used as intended, in long-running Web sessions, where data are continuously streaming between client and server, the cost of initial HTTP handshaking is split over the large number of optimized payload transfers and gets negligible even after few client-server interactions.

ACKNOWLEDGMENT

This work is sponsored by the Ministry of Science, Education, and Sports of the Republic of Croatia under the research grant agreement 036-0362980-1921.

REFERENCES

[1] D. Gourley et al., HTTP: The Definitive Guide. Sebastopol, CA: O'Reilly Media, Inc., 2002.

[2] R. Fielding et al. (1999, June). *Hypertext Transfer Protocol – HTTP/1.1* [Online]. Available: http://tools.ietf.org/ search/rfc2616

[3] S. Loreto, P. Saint-Andre, S. Salsano, and G. Wilkins. (2011, April). *Known Issues and Best Practices for the Use of Long Polling and Streaming in Bidirectional HTTP* [Online]. Available: http://tools.ietf.org/search/rfc6202

[4] P. McCarthy and D. Crane, Comet and Reverse Ajax: The Next-Generation Ajax 2.0. New York, NY: Apress, 2008.

[5] I. Paterson, D. Smith, P. Saint-Andre, and J. Moffitt. (2010, July). *XEP-0124: Bidirectional-streams Over Synchronous HTTP (BOSH)* [Online]. Available: http://www.xmpp.org/extensions/ xep-0124.html

[6] V. Wang, F. Salim, and P. Moskovits, The Definitive Guide to HTML5 Websocket. New York, NY: Apress, 2013.

[7] I. Fette and A. Melnikov. (2011, December). *The Websocket Protocol* [Online]. Available: http://tools.ietf.org/search /rfc6455

[8] I. Hickson. (2012, September). *The Websocket API* [Online]. Available: http://www.w3.org/TR/websockets/

[9] R. Berjon et al. (2014, February). *HTML5: A Vocabulary and Associated APIs for HTML and XHTML* [Online]. Available: http://www.w3.org/TR/html5

[10] *Python Websocket API* [Online]. Available: https://pypi.python.org/pypi /websocket/0.2.1

[11] J. Vos. (2013, April). *JSR 356 Java API for Websocket* [Online]. Available: http://www.oracle.com/technetwork/articles/ java/jsr356-1937161.html

[12] *C++ Websocket API* [Online]. Available: http://www.zaphoyd.com/websocketpp

[13] *.NET Websocket API* [Online]. Available: http://msdn.microsoft.com/en-us/library/system.net.websockets. websocket(v=vs.110).aspx

[14] *Jetty* [Online]. Available: http://www.eclipse.org/jetty/

[15] *Autobahn* [Online]. Available: http://autobahn.ws/

[16] P. Lubbers and F. Grecco. *HTML5 Web Sockets: A Quantum Leap in Scalability for the Web* [Online]. Available: http://www.websocket.org/quantum.html

[17] M. Laine and K. Saila, "Performance evaluation of XMPP on the Web," Finland, 2012.

[18] A. Almasi and Y. Kuma, "Evaluation of Websocket communication in enterprise architecture," Sweden, 2013.

[19] L. Huang, E. Y. Chen, A. Barthy, E. Rescorlaz, and C. Jackson, "Talking to yourself for fun and profit," in Web 2.0 Security and Privacy 2011, Oakland, CA, USA, 2011.

[20] A. Herzberg and H. Shulman, "Socket overloading for fun and cache-poisoning," in Annual Computer Security Applications Conference 2013, New Orleans, LA, USA, 2013.

[21] (1981, September). *Transmission Control Protocol* [Online]. Available: http://www.ietf.org/rfc/rfc793

Improving thread scheduling by thread grouping in heavily loaded many-core processor systems

Luka Milić* and Leonardo Jelenković**
* University of Zagreb, Faculty of Organization and Informatics, Varaždin, Croatia
**University of Zagreb, Faculty of Electrical Engineering and Computing, Zagreb, Croatia
luka.milic@foi.hr, leonardo.jelenkovic@fer.hr

Abstract – Basic scheduling techniques currently employed in operating systems usually don't account for hierarchy of processors cores and caches resulting in suboptimal system efficiency, especially on heavy loaded systems. In this paper we explore possible improvements in scheduling for such heavy loaded many-core systems. Basic idea is to group threads of the same process as close as possible, preferably even on the same processor. We expect that doing so may improve efficiency of processor's cache usage, resulting in better overall performance. The idea is tested by adapting the Linux's simplest scheduler (in code) – a round-robin scheduler. Achieved results are presented in this paper.

I. INTRODUCTION

Today's operating systems schedulers mostly do not differentiate multiprocessor system where every processor is simple (one core) from many core system where many cores are coupled on single processor sharing its caches and connections to system busses. Virtual cores (hyper-threading) are also usually treated the same (as regular processor). That is not really a problem until there are not many cores/processors in the system. In the future that is very likely going to change as most computers working non-trivial jobs will have many cores [1].

Architectures of multi and many core systems are various. Most personal computer systems (and some smart phones and tablets) have a single multi core processor. Workstations and servers that require more computing power may have several multi core processors. In the future multi core processors (up to 8 cores) will probably be replaced by many core processors (from 8 cores and up). Computing power of many core processor systems is not relaying only on raw processor's core capabilities and number of cores but also on data acquisition. Cores in many core processor are usually arranged in local groups that share local cache memory. Although any core can fetch data from anywhere (local cache, non-local cache, main memory), delay such fetch operation request will be lesser if data is closer.

Thread scheduling algorithm should take thread's properties and inter-thread interactions into account for achieving more efficient use of processor's cache and thus improve overall system performance [2][3]. Usually, threads of the same process operate on shared data and have high level of interactions. We will call such threads as connected threads (co-threads). Although threads from different processes can also be connected (i.e. use same shared resources), that is rarely, more exception than a rule.

Scheduling algorithm should be aware of connected threads as well as processor's core architecture where those threads are to be placed.

In systems in which the average number of ready threads is not much greater than the number of processor cores, scheduling procedures should keep connected threads as close as possible, on cores that share same caches (L2/L3).

In heavier loaded systems (at least twice as much threads than cores) different way of scheduling may be appropriate. Considering that it is more difficult to achieve a synchronous work of interconnected threads on multiple cores, which may be the optimal scheduling for those jobs (and maybe generally), the alternative could be to collect such threads on the same core. The advantage of this approach is in efficient use of the cache of such cores because it is expected that such interconnected threads will use the shared data probably already present in cache. In this way we will lose potential for parallel execution, but on heavy loaded system probability for such parallelism might be very small, and efficient use of cache could prove more beneficial.

Since schedulers don't differentiate simple single core processor from one core of multi/many processor, when describing existing schedulers we use term processor as synonym for both. In same context, term multiprocessor is used for describing systems that have more processor cores, standalone ones or in multi core package.

For testing such hypothesis a rather simple scheduler `SCHED_RR` from the Linux operating system is used and adapted. Primary reason for choosing `SCHED_RR` even if its real-time scheduler over `SCHED_OTHER` default scheduler for general thread is in source code simplicity. `SCHED_OTHER` that implements Completely Fair Scheduling algorithm is far more complex and hard to adapt. If proposed adaptation don't give positive results for `SCHED_RR` (using threads with same priority, which de facto turns off any real-time properties of scheduler) there is no reason for going into adapting `SCHED_OTHER`.

In section II we first describe scheduler `SCHED_RR`. Section III describes our modification of `SCHED_RR` scheduler, and section IV presents simulation environment and achieved results.

II. LINUX SCHED_RR SCHEDULER

Linux, like most operating systems use different scheduling policies for different types of threads, particularly differentiate normal from real time threads.

A. Linux scheduling policies

Currently, for normal threads Linux supports scheduling policies `SCHED_OTHER`, `SCHED_BATCH` and `SCHED_IDLE`. Those scheduling policies are based on a completely fair distribution of the virtual processor time. This virtual time is a rather complex function of the fair thread priority, real time and the processor consumption by fair threads [4].

Real time scheduling policies `SCHED_FIFO` and `SCHED_RR` are priority based schedulers. Thread with highest priority is always scheduled first. Difference between `SCHED_FIFO` and `SCHED_RR` policies are in scheduling of more than one highest priority threads.

For example if we have ready threads A and B with same highest priority, `SCHED_FIFO` will pick one that become ready first and execute it until it finished or block itself voluntarily. `SCHED_RR` will alternate threads A and B on processor, giving each a time interval before switching to next. On multiprocessor system (or multi/many core) both threads will be running in parallel, each thread on its processor.

On multiprocessors, real time thread scheduling is based on a system-wide strict priority scheduling. Since we use this mechanism to adapt scheduling, system-wide strict priority scheduling is detailed in next subsection.

B. Push-pull algorithm

Strict priority scheduling in a multiprocessor system with *N* processors must ensure that *N* highest priority threads are always chosen as active on those processors (one thread per processor) [5].

The priorities of real time threads are in range from 1 to 99, where higher number represent higher priority. For each processor there is separate data structure containing 99 thread queues, one thread queue per priority. Every ready real time thread is placed into one thread queue, one that matches thread's priority. Active thread for each processor is chosen from its ready queues (thread from highest non-empty queue), and then removed from that queue and marked as active.

System-wide strict real-time priority scheduling is implemented by using *push-pull algorithm* on overloaded thread queues. A real time thread queue is considered to be *overloaded* if it holds at least one another thread (besides currently active) which could be migrated to another processor according to that thread's processor affinity mask.

Push part of the algorithm is performed on processor which uses overloaded thread queues. Its purpose is to push overloaded threads to other processors that will immediately execute them. Such processors must be acceptable by such threads (they must be in thread affinity mask) and currently running lesser priority threads.

More precisely, push algorithm is activated on particular processors only after events:

1. active real time thread is changed (replaced with higher priority thread or its priority is lowered);
2. a new real time thread is created on current processor;
3. a real time thread is waked on current processor.

The push algorithm splits into three algorithms: *pushing*, *finding* and *searching*.

The *pushing* algorithm is activated when a queue on a given (current) processor is overloaded. Highest-priority thread from overloaded queue is then passed to *finding* algorithm which might find a processor where migrate such thread.

The *finding* algorithm use *search* algorithm to find a processor on which the thread will be migrated. Algorithm is repeat up to three times because of scheduling data structures locking semantics [6].

The *searching* algorithm looks processors that are in observed thread's processor affinity map. From that set of processors algorithm search for a lowest priority subset of processors, i.e. processors that currently run threads with currently lowest priority in system. If there are more processors in this subset, choose processor on which thread was run last time. If there is not such processor in subset, then choose one of the closest to processor on which thread was run last time (by scheduling domains).

Similarly to push algorithm, *pull* algorithm is activated when a thread on single processor is finished or blocked or its priority is lowered. Pull algorithm looks at ready queues of other processors (overloaded queues) and if it finds a real time thread with higher priority than the highest priority thread in his queues, it pulls such thread to this processor where such thread will immediately become active. Only threads that can be migrated to current processor are observed. If there are more such threads, only one with highest priority is chosen and pulled.

III. ADAPTING SCHED_RR

The goal of adaptations was to try to group threads of the same process (here called co-threads). The solution was implemented by four ideas, i.e. changes, for push-pull algorithm.

Push algorithm, i.e. its searching algorithm is modified as follows.

1. First try to find a processor on which are threads that belong to the same process (co-threads) but only if there are none on current processor. If such processor is found push thread to its thread queue.
2. If processor where thread was last executed is in thread's available processor subset push thread there.
3. Search for processor closest (by scheduling domains) to processor with co-threads (processor

that is not thread's available processor subset). If one is found push thread there.

4. Search for processor closest (by scheduling domains) to processor where thread was last executed.
5. Return any processor from thread's available processor subset.

In the pull algorithm, pull is additionally performed in situations when remote thread has the same priority as top priority local one if this remote thread has co-threads here and not on remote processor.

The scheduling of real-time threads is not actually changed with this adaptations – system-wide strict real-time priority scheduling is preserved.

Adaptations in Linux source code was done by adding a new scheduling policy `SCHED_RR2` based on `SCHED_RR`. Since changes to `SCHED_RR` are minimal, adaptations (additions) are preformed in same code.

Files included in adaptations (modified files) are:

- `include/linux/sched.h`
- `kernel/sched/sched.h`
- `kernel/sched/core.c`
- `kernel/sched/cpupri.c`
- `kernel/sched/rt.c`.

In kernel code, modifications regarding `SCHED_RR2` scheduling (in respect to vanilla `SCHED_RR`) are protected with a macro. To use `SCHED_RR2` macro must be set. Otherwise, when macro is not set vanilla `SCHED_RR` is used (all modifications for `SCHED_RR2` are not compiled). Experiments were performed on both kernel: one compiled with defined macro (using proposed modifications, i.e. `SCHED_RR2` scheduler) and one without macro (vanilla `SCHED_RR` scheduler).

IV. Experimental results

The processor on which experiments are performed has Intel's processor of *Sandy Bridge* architecture, namely the model *i7-2670QM*. This processor has four hyper-threaded cores (8 logical cores) from which every physical core has 32 kB L1 cache and 256 kB L2 cache. All cores share 6144 kB L3 cache. The processor itself runs on a frequency of 2.2 GHz. The kernel version used in experiments was Linux 3.3.0-rc7 [7].

Test program constructed for evaluation of adapted scheduler try to mimic general multithreaded program whose threads usually operate on shared data. Program first spawns given number of processes. Each spawned process further creates given number of threads producing large number of threads that simulate heavy loaded system. Since our test system has only four physical cores, size of shared data is kept small to amplify cache utilization in achieved improvements.

Test program was executed on unmodified version of kernel using `SCHED_RR` policy, and on modified kernel using adapted `SCHED_RR2` policy. Each run was timed to 10 seconds when progress was printed (number of predefined operations on shared data). Tests are performed many times and average results are presented in Table I.

TABLE I. Experimental results

Test No.	*Processes per program*	*Threads per process*	*Shared data size*	*Efficiency Improvement*
1	7	1	small	0.37%
2	11	5	small	16.02%
3	9	3	small	14.29%
4	9	3	medium	12.15%
5	9	3	larger	9.93%

Results are surprisingly good. Of course, this is synthetic test and maybe not a real measure for real application, but improvements of 10% and more are encouraging. As expected, when there are small number of threads (test no. 1) our adaptation didn't change performance (0.37% may be counted as statistical error). But when the number of threads is significantly greater than number of processors (all other tests) expected improvements did happen and surpass our expectations. Those improvements are greater with more threads and smaller shared data. As shared data grows, less cache utilization is achieved, lesser improvement is obtained.

Once again we must say that used scheduler is adaptation of real time scheduler and we compare obtained results by original real time scheduler. Real time scheduler is not concerned with fairness and maximal resource utilization. Therefore, we expect that if same ideas of grouping were implemented in scheduler for normal threads (policy `SCHED_OTHER`), improvements, if any, will be much lower since that scheduler uses more sophisticated load balancing algorithms. However, on heavily loaded many core systems we still expect that improvements in efficiency with proposed thread grouping should grow as number of cores grows.

V. Conlusion

This paper presents an idea to improve scheduler for many core systems when such systems are heavy loaded with lots threads. Idea is to group threads of same processes on same processor and improve efficiency of processor's local caches. In simulated environment adaptation to simple round robin scheduler (using `SCHED_RR` on Linux) show significant improvements up to 15%, which encourage further research in this direction. Further research should be oriented to adapting more complex schedulers, e.g. ones that are used for scheduling of normal threads. Furthermore, test program could also be improved to better mimic variety of real multithreaded applications.

REFERENCES

[1] T. Zangerl, “Operating system scheduling on multi-core architectures,” University of Innsbruck, Seminary thesis, June 2008.

[2] T. A. Anderson, M. Rajagoplan, B. T. Lewis, “Thread scheduling for multi-core platforms”, Technical report, Programming Systems Lab, Intel Corporation, 2007, http://static.usenix.org/event/hotos07/tech/full_papers/rajagopalan/rajagopalan.pdf.

[3] A.Fedorova, M. Seltzer and M. D. Smith, “Cache-fair thread scheduling for multicore processors”, Technical report, Harvard University, 2006, http://www.eecs.harvard.edu/~fedorova/papers/cache-fair.pdf.

[4] C. S. Pabla, “Completely Fair Scheduler”, Linux Journal, vol. 184, August 2009.

[5] A. Garg, “Real-time linux kernel scheduler,” Linux Journal, vol. 184, August 2009.

[6] R. Love, “Kernel locking techniques”, Linux Journal, 2002.

[7] Intel Corporation, “Intel Core i7-2670QM Processor”, Technical documentation, 2011.

Raspberry Pi as a Wireless Sensor Node: Performances and Constraints

Vladimir Vujović* and Mirjana Maksimović*
* Faculty of Electrical Engineering, East Sarajevo, Bosnia and Herzegovina
vladimir_vujovich@yahoo.com, mirjana@etf.unssa.rs.ba

Abstract - The main building block of Wireless Sensor Network (WSN) is sensor node consisted of four basic elements: the sensor unit, processing unit, communication and power units. In this paper we propose the Raspberry Pi, cheap, flexible, fully customizable and programmable small computer board and abilities of its usage as WSN node and SensorWeb node. The Raspberry Pi brings the advantages of a PC to the domain of sensor network, what makes it the perfect platform for interfacing with wide variety of external peripherals. Comparative analysis of its key elements and performances with some of current existing wireless sensor nodes have shown that despite few disadvantages, the Raspberry Pi remains an inexpensive computer with its very successfully usage in sensor network domain and diverse range of research applications.

I. INTRODUCTION

A wireless sensor network (WSN) is composed of spatially distributed nodes equipped with sensing devices to monitor and to measure characteristics of the physical environment at different locations. WSNs are designed and deployed for different purposes by various organizations. WSN based monitoring applications range from simple data gathering, to complex Internet-based information systems. In other words, the observations obtained from sensor networks may be helpful in many software applications like environmental, industrial and meteorological monitoring, building and home automation, medicine, urban sensor networks, intelligent transportation, security, military defense, etc [1].

Sensor nodes, as building blocks of WSN, are consisted of four basic elements shown in Fig. 1: the sensor unit, processing unit, communication and power units.

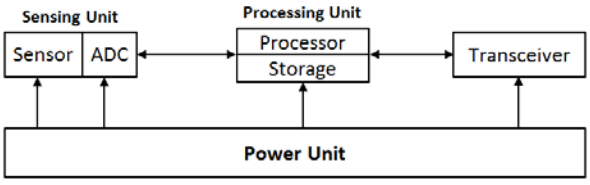


Figure 1. Typical sensor node architecture

Sensor nodes are the small, low power single board computers with a radio for wireless communication. Number and types of sensors depends on the applications. Sensor nodes collect and transfer data using four stages: collecting the data, processing the data, packaging the data and communicating the data [2].

Currently in the market there are many commercially available sensor node platforms. In this work emphasis will be on Raspberry Pi computer making a comparative study of its performances and constraints with current popular wireless sensor nodes presented in [2, 3]. The main goal of this research is to define and present advantages and disadvantages of Raspberry Pi and abilities of its usage as a sensor node.

The rest of this paper is organized as follows. The basic information about Raspberry Pi and literature review of its usage in sensor networks are presented in Section 2. Raspberry Pi's performances, constraints and abilities of its usage as a wireless sensor node and SensorWeb node are shown in Section 3. Finally, Section 4 provides conclusion remarks and outlines directions for future work.

II. WHAT IS RASPBERRY PI ?

Raspberry Pi is a small, powerful, cheap, hackable and education-oriented computer board introduced in 2012 (Fig. 2). This credit card-sized computer with many performances and affordable for 25-35$ is perfect platform for interfacing with many devices.



Figure 2. Raspberry Pi

The Raspberry Pi board contains a processor and graphics chip, program memory (RAM) and various interfaces and connectors for external devices (Fig. 2). Some of these devices are essential, others are optional but all Raspberry Pi models have the same CPU named BCM2835 which is cheap, powerful, and it does not consume a lot of power [4]. Raspberry Pi operates in the same way as a standard PC, requiring a keyboard for command entry, a display unit and a power supply. SD

Flash memory card normally used in digital cameras is configured in such a way to 'look like' a hard drive to Raspberry Pi's processor. The unit is powered via the micro USB connector. Internet connectivity may be via an Ethernet/LAN cable or via an USB dongle (WiFi connectivity) [5, 6].

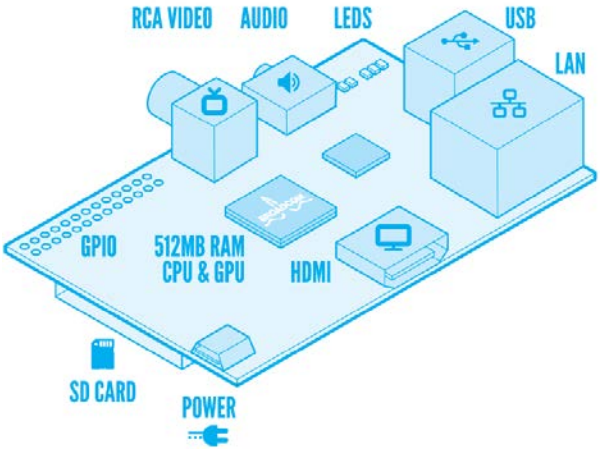


Figure 3. Raspberry Pi components [13]

Like any other computer, the Raspberry Pi also uses an operating system and the "stock" OS is a flavor of Linux called Raspbian. Linux, as a free and open source program, is a great match for Raspberry Pi. On one hand, it keeps the price of the platform low, and on the other, it makes it more hackable. There are also a few non-Linux OS options available [5]. The additional hardware and software requirements can be achieved by already existing hardware modules and open source software.

One of the great things about the Raspberry Pi is that it has a wide range of usage. Numerous recipes of Raspberry Pi usage with sensors, displays and motors is given in [7]. In the workshop [8] a Raspberry Pi is used to read sensors (inputs), store their vales in a database for historical trending and turn relays (outputs) on and off when a sensor value goes outside of a certain range. In other words, a system that turn on a light if the door is open or motion is detected in an area is built. In [9] is stated that wireless sensors and a Raspberry Pi open up a world of possibilities. Author of [10] also state that sensor networks are just one example of how these small, powerful, and inexpensive components can be used. On the example of XBee–ZigBee Mesh network author shows how to with a moderate skill build own sensor network. The possibility to use Raspberry Pi board as a ZigBee home automation server, which can be set up to support a range of tasks, including remote controlled air conditioning and lighting, and checking whether doors are open or closed is presented in [11]. Works [7, 12-16] propose combining the Raspberry Pi with the power of the open source Arduino platform, and present numerous home automation and sensor networks projects.

The performances and constraints of Raspberry Pi will be described in rest of the paper.

III. THE RASPBERRY PI'S PERFORMANCES AND CONSTRAINTS

The Raspberry Pi performances will be compared with following wireless sensor nodes [2, 3]:

- MicaZ – is based on the Atmel ATmega128L which is a low-power microcontroller and runs MoteWorks from its internal flash memory. The MICAz (MPR2400) IEEE 802.15.4 radio offers both high speed (250 kbps) and hardware security (AES-128).
- TelosB - bundles all the essentials for lab studies into a single platform including: USB programming capability, an IEEE 802.15.4 radio with integrated antenna, a low-power MCU with extended memory and an optional sensor suite. This platform delivers low power consumption allowing for long battery life as well as fast wakeup from sleep state.
- Iris – is used for enabling low power WSN. Iris provides users a wide variety of custom sensing applications providing up to three times improved radio range and twice the program memory over previous generations of MICA Motes.
- Cricket - is a location aware version of the popular MICA2 low-power Processor/Radio module. The Cricket Mote includes all of the standard MICA2 hardware and an ultrasound transmitter and receiver.
- Lotus - is based on the NXP LPC1758, 32-bit ARM Cortex-M3 based microcontroller. A single processor board can be configured to run sensor application/processing and the network/radio communications stack simultaneously. Lotus, as all previously mentioned sensor node platforms, has the data rate of 250 kb/s.

A. Size & Cost

The physical size and cost of each individual sensor node has a significant and direct impact on the ease and cost of deployment. Physical size impacts the ease of network deployment because smaller nodes can be placed in more locations and used in more scenarios. One of the main goals of every network is to collect data from as many locations as possible without exceeding fixed budget. A reduction in per-node cost will result in the ability to purchase more nodes, to deploy a collection network with higher density, and to collect more data [17]. The comparison of size, weight and cost of basic models of Raspberry Pi and above mentioned wireless sensor nodes is given in Table I (the smaller values are better).

TABLE I. THE COMPARISON OF SIZE, WEIGHT AND COST

Name	Size (mm)	Weight (g)	Cost per node US$
Raspberry Pi	85.6*53.98*17	45	**25-35**
MicaZ	**58*32*7**	**18**	99
TelosB	65*31*6	23	99
Iris	**58*32*7**	**18**	115
Cricket	**58*32*7**	**18**	225
Lotus	76*34*7	**18**	300

The values presented in Table I show that Raspberry Pi's advantage against other systems lies in its smallest per unit price.

B. Power and Memory

To meet the multiyear application requirements individual sensor nodes must be incredibly low-power. The average power consumption of WSN nodes should be measured in micro amps. Ultra-low-power operation can only be achieved by combining both low-power hardware components and low duty-cycle operation techniques. Also, algorithms and protocols must be developed to reduce radio activity whenever possible. This can be achieved by using localized computation to reduce the streams of data being generated by sensors and through application specific protocols. One of the solutions is to combine together events from multiple sensor nodes by a local group of nodes and then transmit a single result across the sensor network [17].

The CPU is the main component of the Raspberry Pi, responsible for carrying out the instructions of a computer program via mathematical and logical operations. The processor of Raspberry Pi is a 32 bit, 700 MHz System on a Chip (SoC), which is built on the ARM11 architecture and can be overclocked for more power. ARM chips come in a variety of architectures with different cores configured to provide different capabilities at different price points. This means that the vast majority of the system's components – its central and graphics processing units, audio and communications hardware along with 256 – 512 MB memory chip, are built onto single component. The ARM-based BCM2835 is the reason why the Raspberry Pi is able to operate on just the 5V 1A power supply provided by the onboard micro USB port. The Raspberry Pi for operating requires up to 700mA [18]. The unit of Raspberry Pi can be powered using a range of power sources (assuming they are able to provide enough current ~700mA) like [6]:

- Computer USB Port or powered USB hub (will depend on power output),
- Special wall warts with USB ports,
- Mobile Phone Backup Battery (will depend on power output),
- Alkaline batteries (six rechargeable AA batteries and a voltage regulator [7] or high capacity rechargeable batteries depending of Raspberry Pi power needs),
- Solar cells system (for outdoor usage).

The Raspberry Pi has four distinct power modes [19]:

- The run mode - the CPU and all functionality of the ARM11 core are available and powered up.
- The standby mode - the main core clocks are shut down (the parts of the CPU that process instructions are no longer running) although the power circuits on the core are still active. In this mode the core can be quickly woken up by a process generating a special call to the CPU called an interrupt. This interrupt will stop any current processing and do what the calling process has asked for. This mode is known as Wait for Interrupt, or WFI.
- The shutdown mode - there is no power.
- The dormant mode - the core is powered down and all caches are left powered on.

The two approaches of powering Raspberry Pi with batteries are discussed in [7]. The both approaches use the small capacity batteries which power Raspberry Pi model B and ensure Raspberry Pi working time of approximately 2 hours (depends of attached peripherals). Depending on power modes and active modules, power consumption can vary from 700 mA to 1 A. The Raspberry Pi's biggest limitation regarding the power supply is that no external device should draw more than 100 mA from any of its USB ports [4].

It is important to note there's no hard drive on the Raspberry Pi; everything is stored on a Secure Digital (SD) Card. Although large SD cards holding 32 GB, 64 GB or more are available, they are often prohibitively expensive, but the minimum required size of SD card is 2 GB depending on the distribution demands of the operating system. If possible, it is necessary to use the SD card class 10. This storage can be expanded by using devices that provide an additional hard drive upon using the USB ports. Known as USB Mass Storage (UMS) devices, these can be physical hard drives, solid-state drives (SSDs) or even portable pocket-sized flash drives [18] (Table II).

TABLE II. THE COMPARISON OF CPU AND MEMORY *

Name	Processor	RAM	External memory
Raspberry Pi	ARM BCM2835	**256-512 M**	**2-64 G**
MicaZ	ATMEGA128	4 K	128 K
TelosB	TI MSP430	10 K	48 K
Iris	ATMEGA1281	8 K	128 K
Cricket	ATMEL128L	4 K	512 K
Lotus	ARM NXP LPC1758	64 K	512 K

*(The higher value is better)

It is also important to note that the Raspberry Pi Model A has 256MB of RAM while the Model B has 512MB. Compared to other considered systems Raspberry Pi has the largest amount of memory (Table II) what generally leads to improved overall systems' performances.

C. Flexibility

For a wide range of usage scenarios WSN architecture must be flexible and adaptive. Each application scenario demands a slightly different mix of lifetime, sample rate, response time and in-network processing. Additionally, for cost reasons each device will have only the hardware and software it actually needs for considered application. The architecture must make it easy to assemble just the right set of software and hardware components. Thus, these devices require an unusual degree of hardware and software modularity while simultaneously maintaining efficiency [17].

One of the great things about the Raspberry Pi is that it is very flexible and there's no single way to use it. For

example, it can be used for: general purpose computing, learning to program or integrate it with electronics projects [5]. The following core components enable the wide range of its usage [20]:

- Two USB 2.0 ports allows connecting peripherals and storage devices while one micro USB serve for powering device.
- The 3.5mm analog audio jack allows connecting headphones and speakers to the Raspberry Pi what is especially useful for audio and media player based projects.
- Composite RCA port for attaching the yellow video cable from TV allows using TV as a monitor.
- The High Definition Multi-media Interface (HDMI) port allows the Raspberry Pi to be hooked up to high-definition televisions and monitors that support the technology. It is also used for streaming video and audio from the web to TV.
- Support for DSI (Display Serial Interface) - Raspberry Pi can be expanded with display.
- Support for CSI (Camera Serial Interface) - Raspberry Pi can be expanded using camera.
- The GPIO (general purpose input and output) pins on the Raspberry Pi are the main way of connecting with other electronic boards. It consists of 26 pins arranged in two rows containing 13 pins each. The left row contains the even-numbered pins, and the right row contains the odd-numbered pins (Fig. 4). The GPIO pins on the Raspberry Pi are the main way of connecting with other electronic boards. In other words, the GPIO pins can accept input and output commands and thus can be programmed on the Raspberry Pi. It is important to note that there is a certain difference between GPIO pins schedule among model A and model B of the Raspberry Pi. GPIO pins can be accessed for controlling hardware such as LEDs, motors, and relays, which are all examples of outputs. As for inputs, Raspberry Pi can read the status of buttons, switches, and dials, or it can read sensors like temperature, light, motion, or proximity sensors (among many others) [4]. Some of GPIO pins can be used as digital inputs/outputs and as interfaces for embedded protocols. Two most important protocols based on their widespread are:
  - I^2C – low-speed interface – Inter-Integrated Circuit (I^2C) is a serial bus interface which supports multiple devices and only requires two wires for communication. It's work on relatively low speeds [21].
  - SPI – Serial Peripheral Interface Bus (SPI) is a synchronous full-duplex (two way) serial connection [22].
- Extended GPIO – in addition to standard GPIO port, Raspberry Pi Model B Rev 2 has an expanded set of connectors. It is important to mention P5 header which is made up of 8 pins (+3.3 V, +5 V, two ground pins and four GPIO pins that can provide the second I^2C protocol) and P6 header with two pins – their short circuiting provides soft reset of BCM2835.

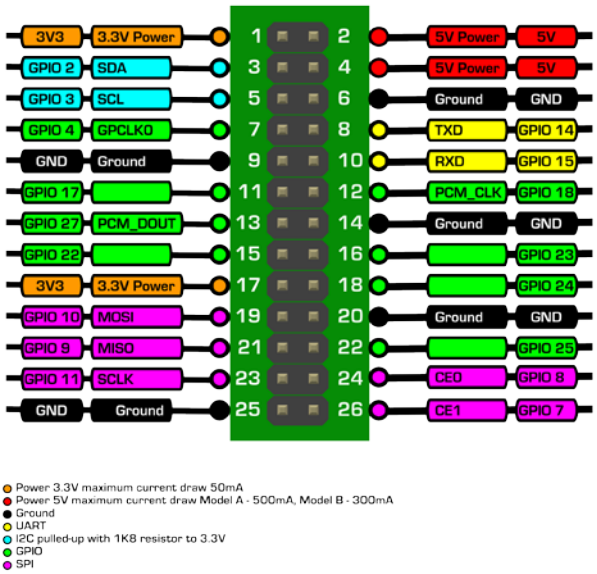


Figure 4. Raspberry Pi Model B Rev2 – GPIO connectors [20]

Analysis of the Raspberry Pi has shown that it, as above mentioned wireless sensor nodes, has expansion connector which supports Analog Inputs, Digital I/O, I^2C, SPI and UART what make easy their connection to a wide variety of external peripherals.

D. Communication

A key evaluation metric for any WSN is its communication rate, power consumption, and range. In order to create an interconnected network, nodes can not be placed too far apart. Most application scenarios according to desired sensing have natural node densities. If the radio communications range demands a higher node density, additional nodes must be added to the system in to increase node density to a tolerable level. The communication rate also has a significant impact on node performance. Higher communication rates mean less transmissions time and lower network power consumption. However, an increase in radio bit rate is often accompanied by an increase in radio power consumption. All things being equal, a higher transmission bit rate will result in higher system performance. But an increase in the communication bit rate has a significant impact on the power consumption and computational requirement of the node [17].

The Ethernet port is the Raspberry Pi's main gateway for communication with other devices and the Internet. The model B has a standard RJ45 Ethernet port while model A doesn't, but can be connected to a wired network by an USB Ethernet adapter. The Raspberry Pi's Ethernet port is auto-sensing which means that it may be connected to a router or directly to another computer (without the need for a crossover cable) [5, 6]. USB Ethernet adapter has two-speed mode, 10 Mb/s and 100 Mb/s. With a cable connected, the Raspberry Pi will automatically receive the details it needs to access the Internet when it loads its operating system through the

Dynamic Host Configuration Protocol (DHCP). This assigns the Raspberry Pi an Internet Protocol (IP) address on network, and tells it the gateway it needs to use to access the Internet (typically the IP address of router or modem). Internet connectivity of Raspberry Pi besides an Ethernet/LAN cable (standard RJ45 connector) may be via an USB WiFi adapter. Using such a device, the Raspberry Pi can be used for creating ad-hoc networks or to connect to a wide range of wireless networks, including those running on the latest 802.11n high speed standard [18].

Raspberry Pi can serve as static websites, but it can also generate dynamic content using databases and web applications. In addition, it can even provide access to its GPIO ports via web technologies. Also, Raspberry Pi can be used as SensorWeb node by connecting it to a network so it can be accessed from other computers. On the other side, the administration of Raspberry Pi can be done over Secure Shell (SSH), a network protocol for secure data communication [4].

E. Operating systems

Sensor nodes run embedded software that samples the physical environment, load data, aggregates and communicate with higher level (peers or gateways). Regardless of the hierarchical approach each sensor node still needs a program, and the most common approaches to programming each sensor node, is to either program it using some form of operating system or to choose a higher level of abstraction. The operating systems vary from traditional operating systems in terms of goals and technique and each system differs substantially in the approach to memory protection, dynamic reprogramming, thread model, real-time features, etc. [17].

Like every computer, the Raspberry Pi needs an operating system, and the preferred one for the Raspberry Pi is Linux distribution. That's partly because it's free, but mainly it's because it runs on the Raspberry Pi's ARM processor [4]. There are several distribution of Linux ported to the Raspberry Pi's BCM2835 chip, including Debian, Fedora Remix and Arch Linux [19]. The Raspberry Pi uses an operating system called Raspbian (Table III) based on Linux but there are also a few non-Linux OS options available.

TABLE III. THE COMPARASION OF USED OPERATING SYSTEMS

Name	Operating system
Raspberry Pi	RASPBIAN
MicaZ	TINY OS,MOTE RUNNER
TelosB	TINY OS,SOS,MANTISOS
Iris	TINY OS,MOTE RUNNER
Cricket	TINY OS
Lotus	RTOS,TINY OS

There are several reasons for deciding to go with the Raspbian operating system [19]:

- Raspbian has a desktop environment similar to Windows and Mac called Lightweight X11 Desktop Environment (LXDE), so it provides an easy transition for those not familiar with Linux command line.
- It comes pre-installed with software useful for writing codes.
- The operating system has been tailored to run on the Raspberry Pi. The code compilation is optimized for on-chip floating-point calculations (hard-float) rather than a slower software-based method.
- There is wide spread community support for the operating system.

F. Advantages

After analysis above performed, general conclusions about the Raspberry Pi advantages can be stated.

- First of all, it is important to mention that Raspberry Pi is a small independent computer that runs on the Linux operating system and can be programmed as needed.
- It has a very large working memory (many other sensor nodes do not have).
- It has expandable memory to store the data.
- It works on processor which supports a large set of instructions.
- It operates at speeds from 700 MHz to 1000 MHz.
- It has support for USB 2.0 which allows its expansion with a large number of peripherals.
- Depending of the needs it is possible to expand the Raspberry Pi with WiFi and Bluetooth adapters (power and range can be changed by changing the adapter).
- Expansion and communication with network devices over a LAN adapter are possible.
- It is possible to form an expandable system with various electronic components (sensors and electronic circuits) using digital inputs and outputs, I^2C or SPI protocols (most of today's devices use one of these methods of communication).

G. Disadvantages

Based on the above mentioned it can be noted that for small amount of money the Raspberry Pi comes with a lot of nice things, but it also lacks some useful features. The main disadvantages of Raspberry Pi are [4]:

- It does not have a real-time clock (RTC) with a backup battery. The fact of missing clock can be workarounded using a network time server, and most operating systems do this automatically.
- The Raspberry Pi always boots from an SD card. It means that even a perfectly valid installation of an operating system is available on a USB stick or an external hard drive, it can't be booted. In other words, external storage devices can be used but can't be used to boot the Raspberry Pi.
- It does not support Bluetooth or WiFi out of the box but these supports can be added by USB dongles.

- Unfortunately, most Linux distributions are still a bit picky about their hardware, so it should be first checked whether flavor of Linux supports particular device.
- It doesn't have built-in an Analog to Digital converter. External component must be used for AD conversion.
- Its power consumption varies depending on how busy Raspberry Pi is and what peripherals are connected. Typically, the model B uses between 700-1000 mA depending on what peripherals are connected, and the model A can use as little as 500 mA with no peripherals attached. Thus, it is necessary to provide a power supply that can provide enough current to power the device plus any connected peripherals.

IV. CONCLUSION

Based on the comparison of Raspberry Pi's key elements and performances with presented current existing wireless sensor nodes it is possible to summarize Raspberry Pi's pros and cons against to other systems:

- pros: processing power, memory, connectivity, multipurpose usage (USB), Linux,
- cons: power consumption.

The analysis above performed has shown that beside the power consumption issue, Raspberry Pi is ultra-cheap-yet-serviceable computer board. With support for a large number of input and output peripherals, and network communication it makes the perfect platform for interfacing with many different devices and using in wide range of applications. By coupling it with WiFi it can communicate remotely what the Raspberry Pi makes very suitable for the construction of wireless sensor nodes and SensorWeb nodes. Moreover, Raspberry Pi can be used as processing node in WSN networks, not just as sensor node but also as controller. In addition, data processing and decision making can be based on artificial intelligence.

Further, The Linux operating system usage provides additional advantages of using Raspberry Pi as a SensorWeb node. Programming in high-level languages such as C, C++, Python, or Java, solution implementation is quite simple and it is enabled to a large number of users, opposed to micro controller programming which usually depends of development kit.

By installing the Web Server on the unit and providing access to the Internet, Raspberry Pi becomes complete and ideal system (hardware and software) for building SensorWeb nodes. One of the possible Raspberry Pi usage scenarios, which is already implemented, is creation of hardware device that has implemented sensor units and communicate with Raspberry Pi via peripheral devices or via GPIO (I^2C) interface. The developed Raspberry Pi prototype SensorWeb node is based on RESTful services and created in order to build the infrastructure that supports fast critical event signaling and remote access to sensor data via the Internet (the detection of critical events is performed using fuzzy logic). The final evaluation of this prototype and Raspberry Pi usage as wireless sensor node and SensorWeb is planned to be performed by its implementing in home automation and Internet of things projects what are directions of our future work.

REFERENCES

[1] F. Ali, "A Middleware to Connect Software Applications with Sensor Web", *The International Journal of Technology, Knowledge and Society* Volume 6, Number 5, pp. 27-35, 2010

[2] M. Maurya, S. R. N. Shukla, "Current Wireless Sensor Nodes (Motes): Performance metrics and Constraints", *International Journal of Advanced Research in Electronics and Communication Engineering* (IJARECE), Volume 2, Issue 1, January 2013

[3] Memsic – Powerful Sensing Solutions, Available: http://www.memsic.com/wireless-sensor-networks, [15.12.2013]

[4] M. Schmidt, Raspberry Pi – A Quick Start Guide, The Pragmatic Bookshelf, 2012

[5] M. Richardson and S. Wallace, Getting started with Raspberry Pi, O'Reilly, USA, 2013

[6] Raspberry Pi Getting Started Guide, RS Components, Vsn 1.0, 2012

[7] S. Monk, Raspberry Pi Cookbox, O Reilly Media, CA, 2014

[8] Introduction to home automation with the Raspberry Pi, Available: https://vancouver.hackspace.ca/wp/2013/11/03/introduction-to-home-automation-with-the-raspberry-pi-sunday-november-24th-2013/, [21.12.2013]

[9] Home monitoring projects for Raspberry Pi, Available: http://www.projects.privateeyepi.com/home, [22.12.2013]

[10] C. A. Bell, Beginning Sensor Networks with Arduino and Raspberry Pi, Apress Media, 2013

[11] N. Heath, 10 coolest uses for the Raspberry Pi, European Technology, 2012, Available: http://www.techrepublic.com/blog/european-technology/10-coolest-uses-for-the-raspberry-pi/505/ [21.12.2013]

[12] A. K. Dennis, Raspberry Pi Home Automation with Arduino, Packt Publishing, USA, 2013

[13] C. A. Bell, Beginning Sensor Networks with Arduino and Raspberry Pi, Apress Media, 2013

[14] T. L. Warner, Hacking Raspberry Pi, Que Publishing, USA, November 2013

[15] S. Goodwin, Smart Home Automation with Linux and Raspberry Pi, 2nd Edition, Apress Media, 2013

[16] Electronic project - Simple home automation with the Raspberry Pi, Available: http://www.penguintutor.com/electronics/simple-homeautomation-raspberrypi, [22.12.2013]

[17] J. L. Hill, System Architecture for Wireless Sensor Networks, PhD Thesis, University of California, Berkley, 2003

[18] E. Upton, G. Halfacree, Raspberry Pi User Guide, Wiley, 2012

[19] B. Horan, Practical Raspberry Pi, Apres, USA, 2013

[20] Raspberry PI Rev2 – P1 Connector, Available: http://www.combinatorialdesign.com/boards/Raspberry_Pi/P1, [18.12.2013]

[21] UM10204 I2C-bus specification and user manual, Rev.5, October 2012

[22] Serial Peripheral Interface, MC68HC11A8 Technical Data, Motorola, Available: http://www-ee.eng.hawaii.edu/~tep/EE491E/Notes/HC11A8/HC11A8_SPI.pdf [5.12.2013]

Remote Firmware Update for Constrained Embedded Systems

Goran Jurković* and Vlado Sruk**
* Department of Intelligent Transportation Systems, Faculty of Transport and Traffic Sciences, University of Zagreb, Vukelićeva 4, HR-10000 Zagreb, Croatia
** Department of Electronics, Microelectronics, Computer and Intelligent Systems, Faculty of Electrical Engineering and Computing, University of Zagreb, Unska 3, HR-10000 Zagreb, Croatia
goran.jurkovic@fpz.hr vlado.sruk@fer.hr

Abstract - In this work, we present the architecture of firmware software update infrastructure, which uses a centralized server to distribute adequate firmware updates to a large number of embedded devices. These updates enable transparent proactive management of everyday bugs and other vulnerabilities inherent in software-based systems, as well as embedded system evolution. Presented firmware update mechanism utilizes a network connectivity nowadays already presented in a wide range of embedded systems.

I. INTRODUCTION

Ubiquity of embedded computer systems in every pore of human life gets the credits from flexibility and faster product development from an idea to the final product. Technological development of microcontrollers is focused on specialization to different type of embedded systems. One of the goals is fast and flexible replacement of embedded software also called firmware [1]. Due to this fact that firmware within the embedded system is largely defining product's role, capabilities and its functionality, it shortens time of a development from idea to the final product. In such environment the hardware can be designed as final product in early stages of this development bringing it faster to final stage, as software would in the end define this final product functionality. This creates another problem, as a very little time is left for detailed testing of a product and very often companies place mini batch of products on market which have not been thoroughly tested. In such cases end users become beta testers of the product. In such environment we need to provide an improved level of flexibility in automated and transparent firmware changes in the embedded system already installed at end users, to enable transparent and straightforward software update to prevent possible malfunctions.

Rising trend of wide spreading use of GSM and GPRS technology and its price drop opens possibility to affordably equip embedded computers with appropriate components to enable self-maintenance at any place. Presuming such environment, embedded system could achieve access to Internet and straight connection to the device manufacturer for maintenance purpose or dynamic replacement of a firmware [2].

In following section an algorithm for automatic software or firmware upgrade is described. Third section presents developed and implemented firmware upgrade algorithms. Forth section presents test environment and performance comparison.

II. RELATED WORK

The possibility of embedded system to upgrade running firmware over the Internet, also called FOTA [3] (Firmware Over The Air) is well known and widely used on very powerful embedded systems running ARM processors and operating system. However, this paper analyzes the use and the specifics of software agents in the process of upgrading software of the smaller embedded computer without operating system. Given the limitations of memory, it analyses the advantages and disadvantages of software agents with direct entry into the program memory and agents with the use of additional memory resources.

Implementation of the system software upgrade is based on software agents to enter directly into the program memory, and on-board computer was built in the system based on ATMEL processor [4] equipped with self-modifying program memory. Connectivity to the Internet was done with GPRS module that enables communication with the server. It describes the implementation of software algorithms, software agents, in detail. It evaluates the implementation process of dynamic upgrades of software described by a series of measurements that give insight into the properties and characteristics of deviation. Applications of dynamic software upgrade by software agents is also suitable for implementation in a limited region of embedded computer systems of limited resources.

III. FIRMWARE UPGRADE ALGORITHM

In first case we used embedded computer based on AVR family of micro-processors made by Atmel and GSM/GPRS module Wismo228 [5]. This case represents embedded systems where communication module doesn't contain memory for temporal storage, Figure 1. In this environment an algorithm for automatic software or firmware upgrade (called Agent) used an Intel HEX file format for transferring of the new software that was supposed to replace existing ones. In this case the binary

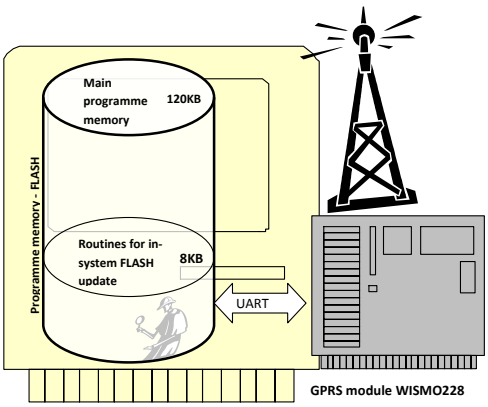


Figure 1. ATmega128 and communication module without memory buffer

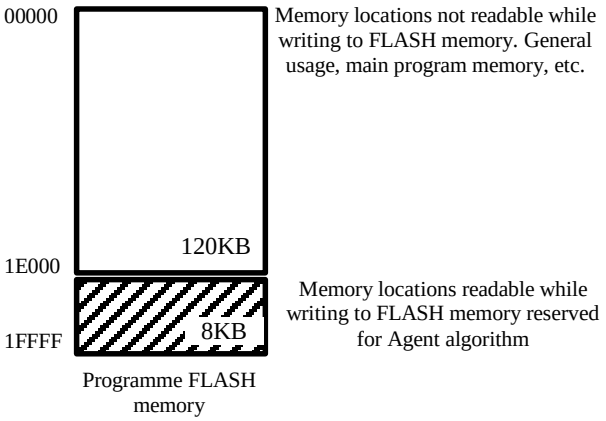


Figure 2. Flash update agent without memory buffer

firmware was packed into Intel Hex file format and downloaded from the server using Hyper Text Terminal Protocol. As used micro-processor ATmega128 is fairly small and inexpensive, but built on AVR core with strong processing capabilities and a lot of integrated features and interfaces, it is often used in industry for production of consumer electronics and other embedded systems. Since this was an extremely undemanding embedded computer without any additional memory, Agent downloaded and upgraded main software in segments sized 2KB, while main program and embedded system functions were stopped and the product in which computer is embedded in was put out of operation. Figure 2 presents memory organization for algorithm in case when there is not available memory buffer. Entire flash update agent without memory buffer is presented in Figure 3.

Replacing transferring format of new software into customized base64 encryption and storing it in segments of 2KB, where each has basic checksum, the file that is transferred via GPRS connection generally was reduced in size, almost halved. A binary segment of the 2048B (2KB) packed in Intel Hex format increased the binary to 5325B (5.2KB) while the same in customized Base64 format was reduced to 2735B (2.67KB).

The Agent itself occupies 8KB of program memory (FLASH memory) reserved for in system FLASH updating algorithm routines presented on Figure 6, and 120KB of program memory remains for the main software or firmware (from 128KB of FLASH which ATmega128 has). Although the processor has built in self changing FLASH memory with program routines inside the same FLASH, it has one limitation. 120KB of FLASH memory is not readable while calling FLASH writing routines. Only upper part of the FLASH program memory reserved for Agent routines is always readable and Agent program can run normally while writing FLASH.

Another advancement of an Agent was achieved by replacing GSM/GPRS module model WISMO228 with GSM/GPRS module model SIM908 [6] made by SimCom. This new GPRS module has built in http client engine and by using this integrated HTTP protocol with buffer memory of 160KB within a GPRS module; it became possible to download the entire new firmware, the main software at once without the segmentation, presented on Figure 8.

Also, it is now possible to initiate download within the main software while embedded computer and device performs main functions of the product in which it is built-in. Once new software was downloaded via GPRS and Internet network and stored in memory buffer of the SIM908 module, downloaded Base64 packed file would be checked segment by segment with checksum calculations for correct download of the new main software. Checking of a downloaded Base64 file takes

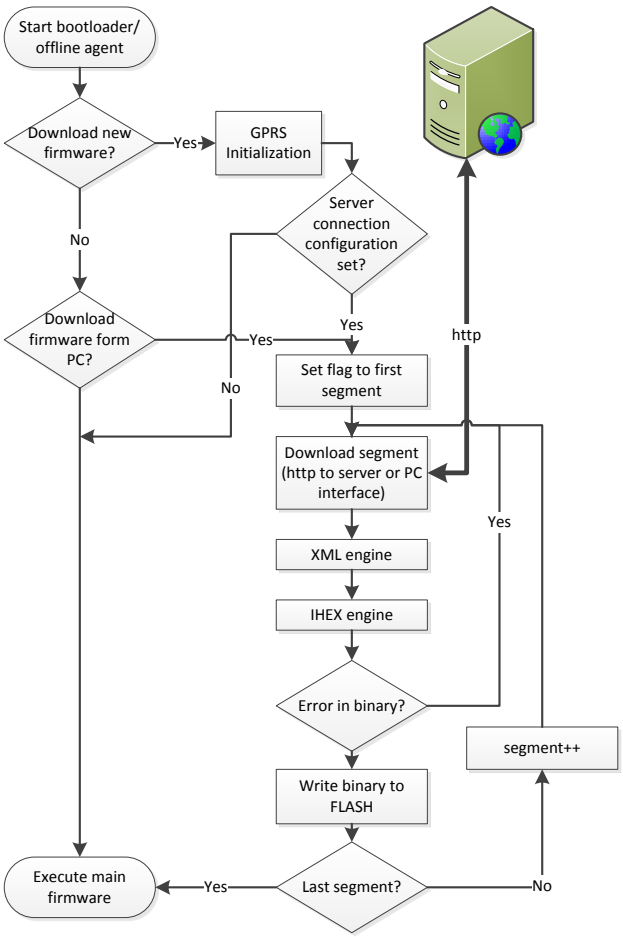


Figure 3. Flash update agent without memory buffer

about 23 seconds every time over local 64Kbps connection, and it doesn't depend on GPRS and network connection. This is just another precaution that takes a small amount of time and makes an Agent more robust and the software upgrade routines more transparent and reliable.

In a case when new downloaded software packed in Base64 is correct, the main program stops and calls stand-alone Agent routines to make program memory (FLASH) upgrade from a buffer which takes about 25 seconds over local 64Kbps connection that is a bit increased with a write FLASH time.

So it is known in advance exactly how much system would be out of operation (25 seconds for nearly the biggest program that fits FLASH) and it no longer depends on the effects of availability of GSM and GPRS networks and internet once the new software packed Base64 file is downloaded in SIM908 buffer.

Selected test embedded computer is normally built into the GNSS system for tracking, securing of vehicle from grand theft auto and remote engine start, called SMS Commander Genie made by Globis presented at Figure 4. This embedded system is quite flexible and used in a broad of applications. Embedded System is based on ATmega128 micro-processor by Atmel, GSM/GPRS/GPS Module model SIM908 by SimCom, switching power stabilizer for voltage regulation at 5V based MP1591 integrated circuit made by Monolithic Power Systems, 3V linear regulator for micro-processor based on TPS76930 by Texas Instruments, crystal oscillator of 7.37MHz frequency for micro-processors clock, MEMS motion sensor based on ADXL343 by Analog Devices, Lithium Polymer 3.7V backup battery, outside world outputs based on ULN2003D made by Texas Instruments, outside high power P-channel MOSFET outputs based on ZXMP6A18K made by ZETEX Semiconductors, CANBUS interface controller and logic converter based on MCP2515 and MCP2551 by Microchip, and other glue logic and interface to the outside world, which is the vehicle. As the embedded system operates 24/7 and it is supplied primarily from vehicle battery, maximum attention is dedicated to the smaller dimensions (108x54mm), and small energy consumption. To gain small energy consumption micro-processor is powered by 3V and MEMS motion sensor is used to disable GPS functions in GPS module SIM908 when the vehicle is not moving as the biggest power consumer is GPS function on the embedded system.

For testing and validating thesis of this paper only a micro-processor, oscillator, GPRS module and power supply circuit are needed and the electrical schematic for such embedded system is presented on the Figure 1.

IV. TESTING AND COMPARISON

Testing environment is presented at Figure 5, for live testing in a real world environment. Embedded system is equipped with an Agent. Tested embedded systems is additionally modified to enable a continuous 24 hour testing downloads. The real environment includes exploits commercial GPRS network and Internet, and as well on the other side server computer running Linux with Apache http service and PHP with MySQL database. Server is a real time production server used for embedded systems made by Globis, and it is only altered in a way to record starting and ending times of new firmware download. This way it recorded sample times of each download.

To test the new upgraded Agent versus previous work and results measurements from before, to compare upgraded Agent to the old one, the test was conducted using a new software binary sized 110KB just as the first test. But there is a big difference. The old Agent was working with binary files packed in Intel Hex format and the new Agent is working on binary files packed with customized Base64 format. So now the size of new file being transferred over the network is almost halved, and expected download time would be smaller. What goes in favor of the new upgraded Agent is a new way of downloading as it is no longer downloaded in segments, because of a memory buffer inside GPRS device. In this case download speeds could reach the theoretical maximum of GPRS connection class 10 which is 80Kbps for download and 20Kbps for upload.



Figure 4. GNSS capabale selected testing embedded system.

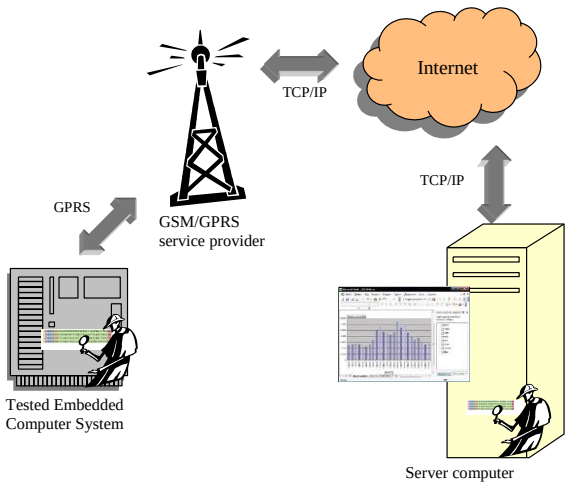


Figure 5. Recording data environment for testing embedded system Agent firmware download times.

TABLE I. FIRMWARE PACKAGE DOWNLOAD TIMES

Firmware package format	*Intel Hex*	*Base64*
Segment size:	2KB	N/A [a]
Firmware packed size (binary 110KB)[b]:	282KB	147KB
Theoretical download time (60/40kbps)[c]:	48.3s	25.1s
Theoretical download time (80/20kbps)[d]:	**25.5s**	**18.8s**
Best achieved download time:	**84.0s**	**19.0s**
Average achieved download time:	103.8s	36.9s
Worst download time:	1335s	219s
Number of tested samples:	624	313
Number of samples with retries:	86	N/A
Number of retries:	182	N/A

a. Because new embedded system uses GPRS module with internal buffer memory the software is transferred at once without segmentation

b. Binary size of a new software is 110KB but the packed size for Intel Hex and Base64 is different

c. Theoretical link speed of GPRS connection class 10 in normal mode

d. Theoretical link speed of GPRS connection class 10 in big files download mode

The standard GPRS connection class 10 uses 60Kbps for download and 40Kbps for upload, but it switches automatically to 80/20 mode in a case of transmission of bigger file. It is noticed that the old way of downloading files in segments, GPRS would never switch to 80/20 mode for faster download. So in comparison of times of older and upgraded Agent, not only the halved of firmware size would make the advantage. Also the switched 80/20 mode would make some difference in download times.

Testing samples were gathered by infinite loop calling an Agent to make download and starting and ending times were recorded in database. These downloads were conducted through 24 hours to check if a time of a day influence on the results because of the network bandwidth congestion. We can notice there are certain times of a day when the GPRS network is more utilized and therefore the spikes are recorded in tested samples. Figure 6 is showing segmented download times of firmware packed in Intel Hex format, while the figure 7 shows the same data, but downloaded at once in customized Base64 format.

In table I. the numbers of recorded samples are

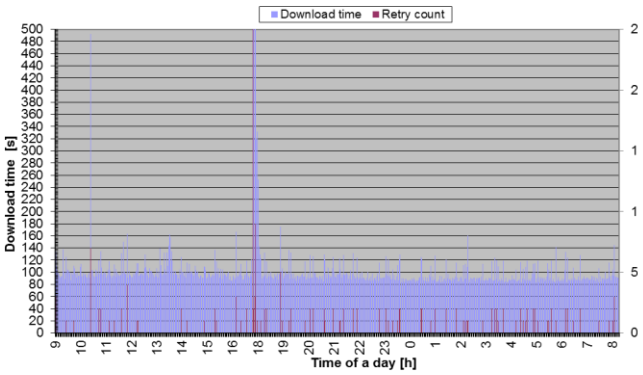


Figure 6. Recorded download times in a time of a day for Intel Hex packed binary firmware file originaly sized 110KB.

TABLE II. AGENT FUNCTIONS EXECUTION TIMES

Main firmware function:	*Agent time:*	*Intel Hex*	*Base64*
Online Agent Time [a]:	Best:		19.0s
	Average:		**36.9s**
	Worst:		219s
	Firmware check [c]:		23s
Offline Agent time (flash update)[b]:	Best:	84.0s	
	Average:	**103.8s**	
	Worst:	1335s	
	Flash write [d]:		**25s**

a. Main embedded system's firmware is running and Agent is working in parallel with the main firmware

b. Main embedded system's firmware and functions are down. Only the Agent is running.

c. After downloading the new firmware in a memory buffer built in SIM908, it is checked for errors before starting the Agent in an offline mode while the main process is stopped.

d. Flash is written from local buffer in new Agent with Base64 and SimCom, while on the old Agent, Flash is written while downloading segment by segment.

presented. There were 624 recorded samples for Intel Hex format and 313 recorded samples for Base64 format. There is a noticeable difference in download times because of the halved file in size in favor the new upgraded Agent, packed in customized Base64 format. But as it is presented in Table II, download time for not segmented file almost reached theoretically best possible download time for this type of a GPRS connection. While there is no possibility to achieve this kind of theoretical speeds in a segmented way of downloads. The best possible download time for Base64 packed file is 18.8 seconds, achieved 19.0 seconds. The best possible download time for old packed files are 25.5 seconds, but only achieved 84.0 seconds in the past. The difference is great. Also for average and worst download times, the new upgraded Agent is in noticeable advantage.

Also, there was no record of a retries in download for the upgraded Agent, because the GPRS module built in HTTP engine was taking care of this. So, if the retries did exist, they would manifest them self in a bit longer download time recorded.

The second advantage of the new agent over the old one is added local buffer memory inside the GPRS

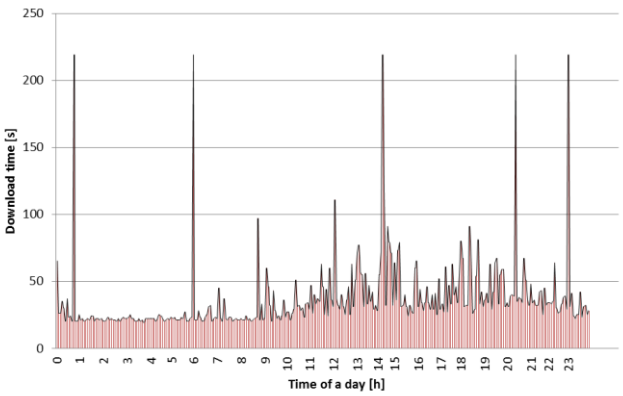


Figure 7. Download times in a time of a day for customized Base64 packed binary firmware file originaly sized 110KB.

module. This makes entire concept of the automated software upgrades more transparent, robust and reliable. As presented on the table I. the worst case scenario is very worrying. The main function of the embedded system, in this case a security against a grand theft auto was shut down for whole 1335 seconds or calculated in minutes, it is over 22 minutes. This is not acceptable in reliability sense. Therefore the new agent with local memory buffer wouldn't shut down the main process until it would made sure that the new firmware is correctly downloaded, and shut down time would always be exactly 25 seconds during the FLASH upgrade.

After the new agents receives a notification from GPRS SIM908 module that the HTTP download of a new firmware is completed, it would check it, that would take 23 seconds on 65Kbps connection, and shut down the main process and start up the offline Agent algorithm. The main process would be up and again running in 25 seconds with an upgraded firmware.

V. Conclusion and Future Work

The paper considered the possibility to upgrade software of an embedded system using the Internet to gain self-maintained software. Connectivity to the Internet was in our case is realized using GPRS module. Two software agents are presented: one with possibility of direct access into the program memory and agents with the use of additional memory resources. Considering the memory limitations, and processing power of the embedded system it is analyzed the advantages and disadvantages of.

Implementation of the system software upgrade is based on software agents to enter directly into the program memory, and on-board computer was built in the system based on ATMEL processor equipped with self-modifying program FLASH memory.

The paper compares implementation of the Agent without additional memory buffer for temporary saving the new firmware and the newer implemented Agent with additional memory buffer.

Although a compared embedded system mostly differs on built in GPRS module (with and without http engine and built in memory buffer) there was significant improvement in Agent algorithm implementation exploding different binary file packaging format. The simplest, easy to implement and most common data transferring format is Hyper Text Terminal Protocol, on the server side and client side. So, that is the reason for using HTTP protocol, as the client side was software implemented in Agents algorithms. But it has limitations and binary files transferred without controlling signals were not an option for Agent implementation. Instead, ASCII type file packaging was used.

On the older implementation of the Agent, Intel Hex file format was well known and already used for the micro-processor program memory upgrades. It has built in checksum In this case the segmented download and upgrade was only option due the lack of additional memory resources essential for temporarily saving downloaded firmware.

In second case advanced GPRS module with built in http engine and memory buffer used to store downloaded http page with a limited size of 160KB. In this case we exploited another way of packaging firmware files in a smaller and more compressed way. The decision favored base64 packaging format, as it used only ASCII signs and with addition of checksum and file segmentation with segments sized 2KB, it enabled both segmented and not segmented downloads on new software almost halved in size in comparison to Intel Hex file format.

Implemented Agent engine for parsing customized base64 file format is almost the same in size as implemented XML engine and Intel Hex parser.

The new upgraded embedded system with upgraded Agent algorithm is not only faster, but is more reliable and only fixed amount of time main process is down when upgrading program memory. Download is done as the background process in GPRS module and doesn't affect main process and embedded system in any way. This way updating FLASH time while the main program and routines are down is not affected in any way by GPRS and network bandwidth.

In conclusion, achieved solution is more advanced in hardware and algorithm of the software Agent. Achieved solution for dynamic upgrades of software described in this paper is faster, more robust and reliable which was presented by a series of measurements that give insight into the properties and characteristics of deviation.

Applications of dynamic software upgrade by software agents are also suitable for implementation in a limited region of embedded computer systems built on limited resources. The future work would be extending Agents with compression algorithm implementation. Since Agent has size limitation we will examine several possible candidates and its optimizations due to available program size and main memory required. Also, another there is possible improvement would be in improving program availability

References

[1] Goran Jurković, Mario Žagar, GSM Wireless Solutions in Distributed Embedded Systems Designed for Automatic Control, 46th International Symposium Electronics in Marine, ELMAR 2004., pp. 205-211, Zadar, Croatia, 16-18 June. 2004.

[2] Goran Jurković, Dynamic Software Update in Embedded Systems, Master Thesis, Faculty of Electrical Engineering and Computing, University of Zagreb, Zagreb, 2011.

[3] Over-the-air programming http://en.wikipedia.org/wiki/Over-the-air_programming/

[4] Atmel megaAVR Microcontrollers http://www.atmel.com/products/microcontrollers/avr/megaavr.aspx

[5] AirPrime Embedded Wireless Modules, Sierra Wireless WISMO228 GPRS module http://www.sierrawireless.com/

[6] SIMCom Wireless Solutions Co., Ltd SIM908 http://wm.sim.com/producten.aspx?id=1024

Performance-Occupation Trade-off Examination in Custom Processor Design

Danko Ivošević, Nikolina Frid
Faculty of Electrical Engineering and Computing,
Department of Electronics, Microeletronics, Computer and Intelligent Systems
Zagreb, Croatia
danko.ivosevic@fer.hr, nikolina.frid@fer.hr

Abstract - Trade-off between execution time and resource occupation arise in all kinds of digital system designs. Here we present such relation for FPGA-based custom processor design. Usually, the optimal tradeoff is directed by device sizing on all scales of the design, but for FPGA device, as predefined hardware platform, it is more focused on comparison of existing platforms organizations. The customization of processor architecture as a point of design performance improvement is usually focused on selection of parameter set that governs the most the design characteristics.

In this paper, the focus is on processor architecture datapath with predefined design template and relationship of its structure to final FPGA implementation it maps to. For purpose of evaluation of multiple design at the same time, the appropriate software flow is applied to construct the design space based on constraining of datapath functional units operation types. The data are collected throughout the whole design flow starting from input control and data flow characteristics of the application and ending with FPGA implementation data. The analysis of the design flow showed dependence of final implementation on datapath structure and its components complexities.

I. INTRODUCTION

As Moore's law is predicting for over a 50 years now, the miniaturization of electronic devices is constant, [1]. During this period great improvement in performance, i.e. speed, is achieved also, but lately the system designers community turned to finding other possibilities for future design advances. Multiprocessor System-on-Chip (MPSoC) became a paradigm in embedded system world for distributed cooperation of existing hardware cores since increase in single core speeds appeared to be more demanding task to accomplish in short period, [2]. During advances in submicron scales the phenomena as parasitic capacitances and inductances had always been an open issue to describe and neutralize during very large scale chip integration, [3, 4]. The tradeoff between system size and performance was always actual challenge to handle during system design, independent of design methodology and level of abstraction the designer had been looking. It arises in everyday embedded system design that is trying to satisfy design functionality in all kinds of environment usually with exact sizing and performance goals. It is valid not only for isolated hardware components the system is composed off, but also for their communication and interface to human what includes software and communication protocols embodied within the system. From the point of system-level design, that is motivated by controlling the whole design process from a high level perspective and targeting specific functionalities under strict design constraints, describing the expected tradeoff means even a harder task of predicting system behavior in earlier design phases.

Our previous work coped with processor based custom design. Especially our development corresponds to hardware synthesis or high level synthesis (HLS) as a one of system level design topics, [5]. High level synthesis usually assumes specification in programming language that is not of specialized type, but familiar to more potential users. C code is usual choice because of its acceptances in computer engineering community and easiness in presentation of application control and data flow, [6, 7].

Previous work in high level synthesis field, as in custom processor design we assumed, shows this task as very acceptable when trying to minimize overall design time. Also, the fact of coping with many questions that direct the final solution appeared to be very challenging, especially in early development phases. Every such design process was, besides the specification itself, governed with some kind of directives and constraints that describe target platform and execution environment features. The common goals are design minimization and performance maximization. As these goals are in contraposition for practically every digital design, the design space exploration takes place as natural assumption and consequence of research that governs the design process.

The foundation of custom processor assumed in this paper is the *No-Instruction-Set Computer* (NISC) concept, where datapath of processor architecture is fully customizable according to needs and wishes of the corresponding toolset user, [8, 9]. Since application specification assumed C coded algorithm our and some previous work effort was on automation of datapath design out of arbitrary C code written by user, [10-12]. During corresponding research the appropriate design

This work was supported by research grant No. 036-0362980-1929 from the Ministry of Science, Education and Sports of the Republic of Croatia.

flow was developed with several transformations of input C code until final FPGA implementation is achieved. The idea behind such development was in automation of the design that shortens design cycle time and eases the overall designer's work as are the light motives of Electronic Design Automation (EDA) approach. It assumes all kinds of software support for digital system design starting from high level specification to low level phenomena description. Development of software that supports embedded system design became dominant in designer's effort during last decade, [13].

One of the usual steps towards design size optimization that was recognized in HLS, and also in our approach, was in reducing the interconnect complexity by reduction of global interconnect seen through reduction of multiplexer inputs, [14].

FPGA device is the target hardware platform for many HLS tools and for our approach. It proved itself as appropriate platform for diverse design domains and development flows, [15]. The exploration of FPGA area-delay tradeoffs is however a specific task as many families of such devices are available at the market today. In fully custom hardware design the device sizing impact is the key characteristic that affects the signal delay because of electrical effects, while FPGA is predefined device to work with. Therefore, the comparison and choice of the appropriate device is the specific task when having different architecture organization, transistor and LUT sizing, [16].

In this paper the exploration of the design space does not have an approach of devices or components sizing variations, but variations from the point of processor datapath structure when mapped to the predefined hardware platform. On the side of processor customization, predicting the best design tradeoff primarily has a meaning of selecting the set of most meaningful design parameters [17], and here we vary the functional unit types as a base of datapath design.

In Section II the background of design space exploration is given, and in Section III the motivational example for examining occupation-performance tradeoff is presented. Section IV describes the principle of forming and analyzing the solution space and the results are illustrated in Section V. Section VI gives final conclusions.

II. BACKGROUND

The foundation for design space exploration is in processor datapath design algorithm, [12]. It is situated as the central point of the design flow where C code specification is conducted to FPGA implementation.

Basically, the datapath is designed with integration of application control flow blocks into unique solution. Before that, every block of the control flow is processes by its data dependencies to form appropriate datapath that enters the final integration into unique datapath. The Control and Data Flow Graph (CDFG) is used as the intermediate representation that exposes those two aspects of every application, [18]. The datapaths that belong to control flow blocks consist of functional units performing operations of every block and registers encapsulated within register files holding variables for those operations. Therefore, the integration of blocks contributions relies on their functional units. The functional units are added to the final datapath along with register files they interconnect. During this process, multiplexers are instanced at functional units and register files input ports if more than one connection enters any of those ports.

TABLE I. COMPARISON OF DESIGNS

Characteristic		Cnt1	CntMax	NoCnt
Datapath	# Adders	1	1	14
	# Subtractors	1	1	1
	# Shifters	1	1	5
	# ANDs	1	2	4
	# ORs	1	2	7
	# XORs	1	1	3
	# NOTs	1	1	1
	# Comparators	1	1	6
	# Assigns	1	1	10
	# RFs	26	27	47
	RFs by register count	10-6-6-2-1-1	11-6-7-1-1-1	44-0-0-1-1-1
	# Muxes	35	41	61
	Mux widths	6-10-13-5-1	8-18-10-3-2	23-23-14-1-0
	# Interconnect lines	235	248	354
Implementation	DM depth	202	198	198
	CW width	217	224	242
	# Exec. cycles	3850	3770	3770
	RAM36_EXPs	11	11	13
	Slices	753	856	1097
	Slice Registers	310	343	1149
	Slice LUTs	2797	3072	3910
	Min. period	15.2 ns	14.0 ns	10.5 ns
	Max. frequency	65.9 MHz	71.7 MHz	95.0 MHz

During the final datapath integration the designer can interfere with setting the top numbers of functional units per operation types allowed in the final datapath. When new functional unit is fetched from the processed block datapath it is firstly tested if its operation type count in the final datapath exceeds the limit set by the designer. If it does not exceed this limit a new functional unit is instanced into the final datapath, otherwise not. Thus the diverse datapaths can be produced for same application what naturally reflects in different end implementations.

III. MOTIVATIONAL EXAMPLES

Comparison of diverse resulting designs is presented in Table I as motivational design example. The application beneath is SHA-1 encryption algorithm, [19, 20]. The C code of this algorithm is not very long but has quite

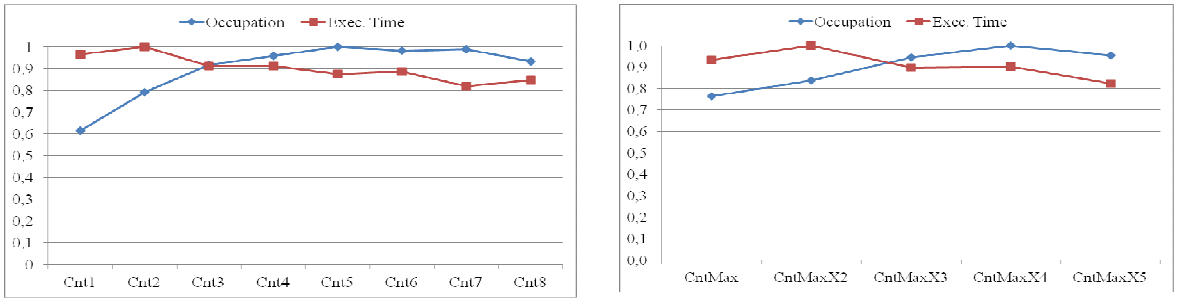
complex control flow consisting of 31 blocks. Moreover, it consists of nine different operation types: addition (ADD), subtraction (SUB), shifting (SHIFT), logical operations AND, OR, XOR and NOT, comparing (COMP) and assignment of values (ASSIGN). We examine its datapath design with some "milestones" defined by different operation type constraining styles:

- *Cnt1* as the design with limit of only one instance of every operation type.
- *CntMax* as the design with limits set to values calculated after analysis of all basic block schedules. Those values represent minimum numbers of all operation types needed to allow the full execution parallelism within blocks. For SHA-1 those values are two for logical operations AND and OR, and one for all others.
- *NoCnt* as the design without imposed limits on any operation type.

The designs in Table I are evaluated according to datapath and FPGA implementation features. The first nine rows of correspond to functional units instances of all operation types. For first two constraining cases they clearly correspond to set limits on allowed operation types, while for third they actually show the summed numbers of all blocks functional units. For last column there are no reuse of any functional unit or register, i.e. there is much redundancy in the datapath. There are much more functional units and register files and, consequently, multiplexers and interconnect lines. The first row below that describing register files count (*RFs*) shows distribution of registers across register files. Thus, there are 10 single registers (or "register file with single register"), six register files with two and three registers (each), two register files with four registers and one register file with five and six registers (each) for *Cnt1*, and even 44 single registers for *NoCnt* case. The row *Mux widths* shows the numbers of multiplexers per widths in format *Mux2-Mux4-Mux8-Mux16-Mux32*. As expected, the datapaths with less components use multiplexers with more inputs (in average) than those with more component redundancy. On the side of implementation the slices occupations of the first two are 20-30% smaller than the last one, but they are 30-45% slower than this last one.

Elaboration that is started by *Table I* data is further extended by establishing a set of constraints with clear tendency of increasing datapath sizes. Fig. 1 shows outcomes of two such constraining approaches. First, in Fig 1a is extension of *Cnt1* constraining style where *Cnt2* to *Cnt8* denote limiting the operation types to two to eight instances of every type. The second approach in Fig. 2 extends the *CntMax* constraining style where *CntMaxX2* to *CntMaxX5* denote limiting the operation types to double, triple, quadruple and quintuple numbers of every type compared to *CntMax* numbers. Graphs in Fig. 1 display existence of tendencies of rising occupations and performance looking from left to right.

Therefore, to investigate if such tendencies will be, and in what extent, recognized as common relation for our processor based design we implemented the software support that will in short time produce a range of diverse



(a) (b)

Figure 1. Correlation between occupation area and execution time under simple constraining variations

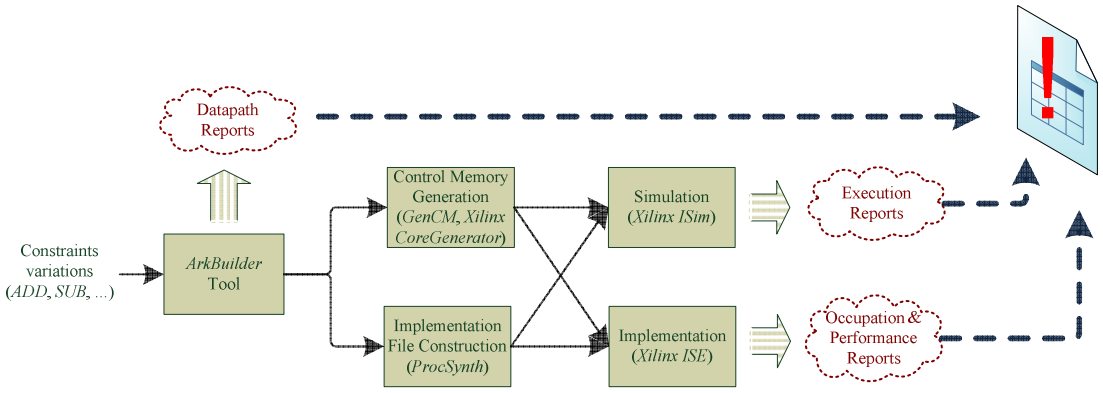


Figure 2. Design Space Exploration Flow

solutions for analysis.

IV. Design Space Exploration Principle

The design flow presented in [12] is supported with several scripts that automate generation of diverse final implementations, Fig. 1. Variations of allowed operation types numbers are generated for running datapath generation tool *ArkBuilder*. This is followed by control memory generation with *GenCM* and *Xilinx* CoreGenerator tools and Verilog implementation file construction with *ProcSynth* tool. Generated control memory cores and implementation file are then used to synthesize and simulate the final FPGA implementation. This flow is covered with several report types that serve as sources for deeper design analysis. Those are reports on:

- Datapath.
- Execution
- Occupation & Performance

Datapath reports include data on components and interconnect that form the datapath. Components data include data on component types: functional units, registers, register files, multiplexers, data memory, control unit. Functional units can be of 11 different types what is determined by operation type they perform. Register files can consist of any number of registers and registers can store any number of variables. Multiplexers can be of different widths (2, 4, 8, 16, etc.), while interconnect lines connect different components.

Execution report is generated during design simulation and has information on execution cycle count.

Occupation and performance reports assume files generated from *Xilinx ISE* implementation. Those include place and route (*.par* file), timing (*.twr* file) and power (*.pwr* file) data.

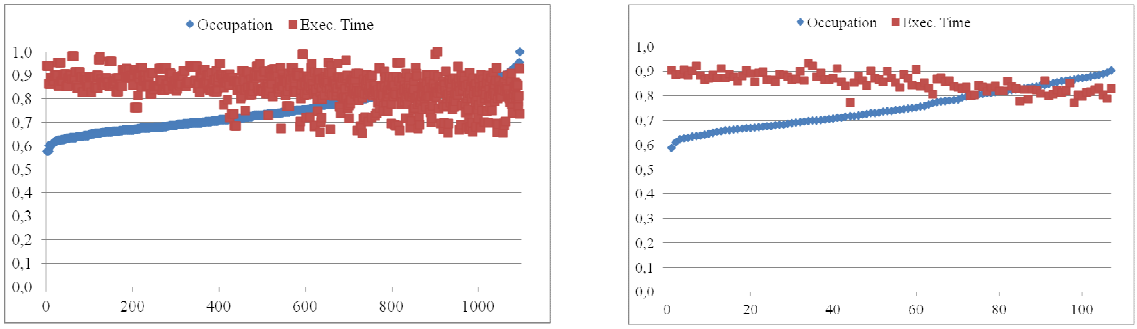
Therefore, all further analysis is based on those three report types.

V. Results

We generate three groups of cases according to different constraining levels:

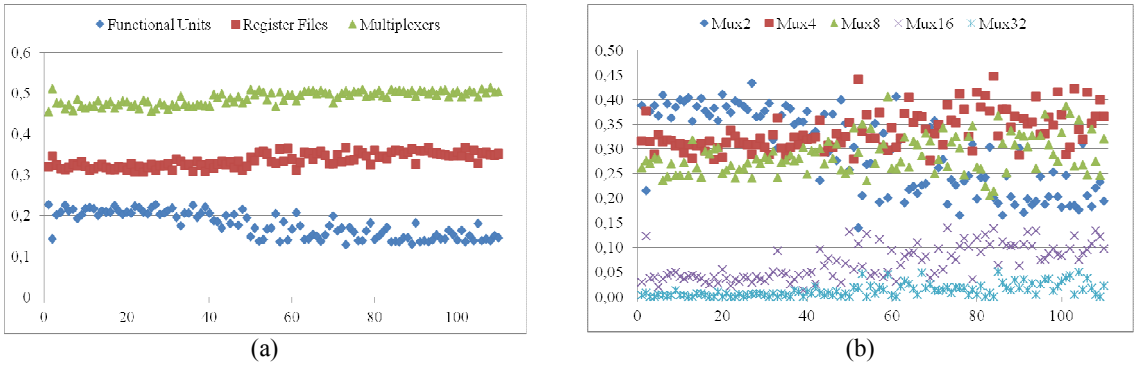
- *ordered12* resulting from variations of every operation type constraint between one and two. As there are nine different operation types, this set consists of 512 cases.
- *random14* as randomly generated 64 cases of constraints with minimum of one and maximum of four allowed instances of all operation types.
- *random17* as randomly generated 520 cases of constraints with minimum of one and maximum of seven allowed instances of all operation types.

Therefore, the complete solution set consists of 1096 datapaths and implementations, and Fig. 3a shows the relationship between occupation (in FPGA slices) and execution time for the complete set. The data are sorted by increasing occupation and show tendency of performance improvement with greater occupations, however with significant variations. This tendency is more emphasized



(a) (b)

Figure 3. Tradeoff between occupation area and execution time: (a) without grouping, and (b) when grouped by 10 constrained cases



(a) (b)

Figure 4. Dependence of execution time on: (a) component types shares (b) multiplexer widths shares

when solutions are grouped by 10 as in Fig. 3b. The same relationship between occupation and execution time when looked for constituents of the data set, i.e. separated *ordered12*, *random14* and *random17* groups, appears to differs only in steep of rising occupation points. It is steeper for cases with more relaxed constraints, i.e. *random14* and *random17* (not presented because of limited space).

Fig. 4 illustrates the solution space data grouped by 10 solutions where solutions are sorted by rising execution time. Starting points assume lower and ending points assume greater execution times, i.e. left sides of the graphs represent solutions with better performance (not explicitly drawn because of the illustrations clarity). Fig. 4a shows dependence of the execution time on component types, functional units, register files and multiplexers, shares in total component numbers. It can be noted that more functional units allow better performance as it is the main datapath design constituent allowing higher level of parallelism. However, it is not the only point of influence on execution time that is noticed. Greater number of functional units, even when having some redundancy, is better for performance when analyzing timing relationships between designs. The designs with more functional unit instances have more other component types also, i.e. register files and multiplexers. Their total count is greater, but their structures are simpler, i.e. register files have (in average) less registers and multiplexers use fewer inputs. Fig. 4b illustrates the execution times dependency on number of multiplexers per width showing their shares in total multiplexers count. The most significant in this graph is the drop of *Mux2* share that is happening along with performance drop. Reciprocally the other multiplexer widths types shares are greater when moving to the right of the graph what confirms the conclusion that components simplicity significantly influence the performance.

During system design it is usual to employ the concept of resource sharing not only by reuse of registers, but also the by use of functional units that can accomplish different operation types. Using such combined functional units reduces their overall number and numbers of other components that serve them by providing and storing data. Consequently it reflects in area occupation reduction. Therefore, the parallel solution space for *ordered12* solution set is formed by allowing combinations of different operation types within the same functional unit. Fig. 5. illustrates the relationships of area occupations (a) and execution times (b) between the cases with different functional unit sharing policies. With respect of area occupation it is noticed that in 92,5% of cases occupation is smaller when functional units sharing is allowed, but in 100% of such cases the execution speed is worse. This also leads to conclusion that complexity of datapath components is an important parameter for optimality of solution space as more complex units, i.e. combined functional units, have negative impact on timing and, consequently, design execution time.

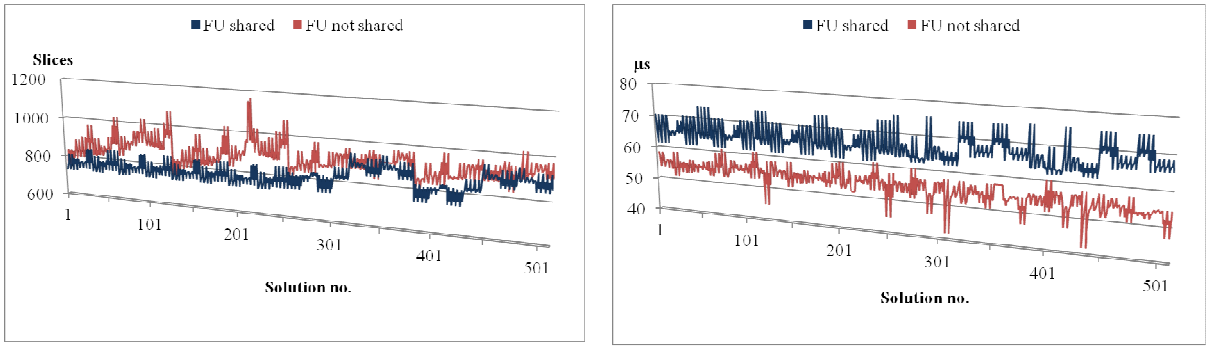
The simulation and implementation were accomplished in Xilinx ISE 14.3 Design Suite for Virtex XC5VSX50T device, [21, 22].

VI. CONCLUSIONS

The tradeoff between digital system occupied area and performance is open issue in every design process. Here we examined the design space evaluation for custom processor design that is based on automated C code algorithmic specification. This task is highly dependable on the quality of target hardware platform description and its correspondence with early design phases. The knowledge of implementation optimization principles is the to final solution occupation and performance estimation. Quality design space exploration and precise prediction of area-performance behavior depends on many parameters. Here we used control over functional units types to form the design space.

As the part of the overall custom processor design flow, we used proprietary IP tools for achieving FPGA device implementation so we could direct the design focus between area and performance only partially. However, we established the structure for design space exploration of custom processor design based development by constraining the functional unit component instantiation as a base datapath design component. Design space data collecting and reporting time was primarily dependable on particular tools run times and did not introduce noticeable time overhead.

We noticed that examined tradeoff is only partially impacted by execution cycles counts, but more on complexity of components.



(a) (b)

Figure 5. Comparisons of: (a) occupation area, and (b) execution time data with dependence on functional units sharing for *ordered12* set

Future work will be on exploiting established solution space to find best performance solution within strict occupation area constraint. Also, another point of future research interest is datapath interconnect reduction based on reduction of multiplexers widths to move the occupation-performance tradeoff focus closer to optimality within the solution space.

REFERENCES

[1] C. A. Mack, "Fifty Years of Moore's Law," IEEE Transactions on Semiconductor Manufacturing, vol.24, no.2, pp.202,207, May 2011

[2] W. Wolf, A. A. Jerraya, G. Martin, "Multiprocessor System-on-Chip (MPSoC) Technology," IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems , vol.27, no.10, pp.1701-1713, Oct. 2008

[3] N. D. Arora,, K. V. Raol,, R. Schumann, L. M. Richardson, , "Modeling and extraction of interconnect capacitances for multilayer VLSI circuits," IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, , vol.15, no.1, pp.58-67, Jan 1996

[4] K. L. Shepard, Zhong Tian, "Return-limited inductances: a practical approach to on-chip inductance extraction," IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol.19, no.4, pp.425-436, Apr 2000

[5] P. Coussy, D. D. Gajski, M. Meredith, A. Takach, "An Introduction to High-Level Synthesis," in *IEEE Design & Test of Computers*, vol. 26, no. 4, pp. 8-17, July-August 2009

[6] TIOBE Software, "TIOBE Programming Community Index", http://www.tiobe.com/index.php/content/paperinfo/tpci/index.html

[7] SourceForge, "The Transparent Language Popularity Index", http://lang-index.sourceforge.net

[8] M. Reshadi, B. Gorjiara, D. D. Gajski, "NISC Technology and Preliminary Results," Technical Report, University of California, Center for Embedded Computer Systems, Irvine, August 2005

[9] B. Gorjiara, D. D. Gajski, "Custom Processor Design Using NISC: A Case-Study on DCT algorithm," in *Workshop on Embedded Systems for Real-Time Multimedia*, pp. 55-60, 2005

[10] J. Trajković, S. Abdi, G. Nicolescu, D. D. Gajski, "Automated Generation of Custom Processor Core from C Code", Journal of Electrical and Computer Engineering, Vol. 2012, 26 p., March 2012.

[11] D. Ivošević, V. Sruk, "Automated Modeling of Custom Processors for DCT Algorithm", Proceedings of 34th International Convention MIPRO, Opatija, Croatia, pp. 762-767., 2011

[12] D. Ivošević, V. Sruk, "Unified Flow of Custom Processor Design and FPGA Implementation", Proceedings of International Conference on Computer as a Tool EUROCON, Zagreb, Croatia, pp 1721-1727, 2013

[13] C. Ebert, C. Jones, "Embedded Software: Facts, Figures, and Future", IEEE Computer, Vol. 42, No. 4, pp. 42-52., April 2009

[14] Taemin Kim; Xun Liu, "A Functional Unit and Register Binding Algorithm for Interconnect Reduction," Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on , vol.29, no.4, pp.641-646, April 2010

[15] J. J. Rodríguez-Andina, M. J. Mouré, M. D., Valdés, "Features, Design Tools, and Application Domains of FPGAs", IEEE Transactions on Industrial Electronics, Vol. 54, No. 4., pp. 1810-1823, July-August 2007

[16] I. Kuon, J. Rose, "Exploring Area and Delay Tradeoffs in FPGAs With Architecture and Automated Transistor Design," IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol.19, no.1, pp.71-84, Jan. 2011

[17] M. Zuluaga, E. Bonilla, N. Topham, "Predicting best design trade-offs: A case study in processor customization," Design, Automation & Test in Europe Conference & Exhibition (DATE), pp.1030-1035, March 2012

[18] A. Orailoglu, D. D. Gajski, "Flow graph representation," in Proceedings of the 23rd ACM/IEEE Design Automation Conference, pp. 503 - 509, 1986

[19] Secure Hash Standards (SHS). [Online]. Available: http://csrc.nist.gov/publications/fips/fips180-4/fips-180-4.pdf

[20] C to Verilog. [Online]. Available: http://c-to-verilog.com/howtos.html

[21] ISE Design Suite - Xilinx, http://www.xilinx.com/products/design-tools/ise-design-suite

[22] Virtex-5 FPGA Data Sheet: DC and Switching Characteristics, http://www.xilinx.com/support/documentation/data_sheets/ds202.pdf

Critical Path Method Based Heuristics for Mapping Application Software onto Heterogeneous MPSoCs

Nikolina Frid and Vlado Sruk
University of Zagreb, Faculty of Electrical Engineering and Computing, Zagreb, Croatia
nikolina.frid@fer.hr, vlado.sruk@fer.hr

Abstract – In this paper the authors propose new heuristics for automation of software partitioning and mapping onto heterogeneous multiprocessor System-on-Chip (MPSoC) platform – Longest Parallel Path mapping algorithm (LPP). In contrast with traditional approach to solving this NP-complete problem – the Integer Linear Programming (ILP), our method uses a modified version of Critical Path Method with additional heuristics that rely on greedy approach. The algorithm performs one-to-many mapping of application to platform with minimizing the overall execution time of the application as the main objective. Our experiments with generic application model and several different platform layouts show that the proposed algorithm provides an efficient mapping scheme enabling significant execution speedup. In addition, the comparison with another greedy mapping algorithm shows that LPP algorithm exploits available task level parallelism better.

Keywords: Design Space Exploration, Software partitioning and mapping, Task scheduling, Critical Path Method

I. INTRODUCTION

Modern embedded systems are above all ubiquitous: they are found everywhere - from simple electronic products, communication and entertainment devices to expensive and complex systems (car, plane). They are getting more and more complex, and have to be able to perform advanced computation and meet special requirements: real-time response, maximum performance, minimum size and weight, low power consumption. This leads to development of new generation of embedded systems – MPSoCs which must often be very heterogeneous to meet all requirements with minimal cost.

Designing this type of systems is a challenging and time consuming task. Automation of design process would significantly lower cost and time-to-market. However, because of MPSoC versatility and heterogeneity, this is easier said than done. Many different approaches exist and many algorithms have been developed but this issue is far from being solved yet.

Scheduling and mapping software to hardware on heterogeneous platforms is NP – complete problem, and finding near optimal solution in reasonable amount of time is challenging [8].

Traditionally this problem is solved using Linear Programming model [4], but high complexity of this approach and extremely high computational requirements significantly increase design time. There have been several attempts to solve this problem, including greedy deterministic [7][6] and evolutionary non-deterministic approaches [11]. Greedy approach is the most common approach and delivers the final mapping solution much faster than evolutionary approach but always gives a suboptimal solution.

The Longest Parallel Path Mapping Algorithm (LPP), proposed in this paper, aims to present another possible approach to solving task mapping problem on heterogeneous MPSoC architectures.

II. DESIGN SPACE EXPLORATION

Design Space Exploration (DSE) is a process to find out near-optimal system architectures for a given application, considering the design constraints and objectives. It involves task partitioning and mapping, architecture and processing element selection, and performance estimation before a hardware prototype is built [2]. The process starts with application specified using a model of computation, description of architecture instance and a set of constraints for the overall system. Those inputs are fed to mapping engine which performs partitioning and mapping of software to hardware. Finally, a performance analysis is conducted to determine whether the output design meets given constraints. If the result is positive the production phase begins, if not modifications to application and/or architecture are made

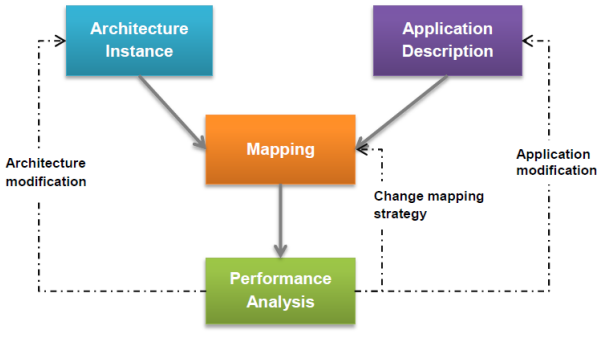


Figure 1. Design Space Exploration Flow

and the whole process is repeated. It is also possible to alter the mapping strategy and check if it produces better outcome. Figure 1 illustrates the described DSE flow.

A. *Related Work*

There is no commercial tool available to automate the whole DSE process. However, a tool is regarded as a DSE tool if it enables the system designer to alter the target architecture and task mapping easily, and it provides the estimated performance. Several tools exist that provide a general modelling and mapping framework to allow DSE [2]. Most prominent is a commercial solution - CoFluent Studio from CoFluent Design, a SystemC-based system-level design tool that supports DSE through Y-chart modelling. The designer specifies separate models for the application and the platform and combines them at the mapping stage. Then, the resultant design is evaluated by simulating the transactional SystemC code that is automatically generated. VisualSim from Mirabilis Design is a system-level design tool that supports diverse simulation capabilities in a single framework. Both algorithm simulation and architecture simulation can be performed at various abstraction levels. By changing the mapping information and architecture parameters, the user can explore the design space efficiently. However, in each of these tools task mapping must be performed manually.

Most prominent non commercial DSE tools include The Modular Performance Analysis (MPA) toolbox from ETH Zurich, which is a performance analysis tool based on real-time calculus, and Daedalus framework from University of Amsterdam, which has the capability of automatic mapping but the process takes several hours to finish [13]. However, much effort is put into further research and development of these tools and it is reasonable to expect that future improvement will resolve current issues.

B. *Task partitioning and mapping*

The key step in design space exploration process is partitioning application software into separate tasks and deciding where to allocate each task on the target platform. Due to heterogeneity of both application and platform this problem requires an NP-complete class algorithm to find the optimal solution. This means that the solution cannot be reached in polynomial amount of time. A faster alternative is using some form of specialized heuristics based on greedy strategy. This way a suboptimal solution is obtained much faster, but heuristics are usually rather specialized and are applicable to a smaller set of problems.

III. PROBLEM DEFINTION

The main problem can be summarized as: finding the near optimal solution to mapping of a generic type of embedded application to a heterogeneous multiprocessor platform in a reasonable amount of time.

The rest of this chapter gives a more detailed problem specification.

A. *Application Model*

Due to versatility of embedded systems, embedded applications come in a wide variety of forms. For the purpose of evaluation of the proposed algorithm a generic application model is constructed with certain assumptions and constraints in mind.

First, it is assumed that application has a flow from a single point of entry to a single point of exit and inside the application flow there are no cycles at all. However, the application as a whole can be cyclic, meaning that after exit point it starts all over again from start. Second, some level of functional parallelism in application is implied because otherwise the discussion about mapping parts of application to different processors is pointless, unless a certain task requires a certain type of special HW accelerator.

The application is modelled using a parallel model of computation and represented as a Directed Acyclic Graph (DAG), $G = (T, E)$. For each vertex $t \in T$ which represents one task, we define set of edges $E_t \subseteq E$ connected to t that represent communication between tasks. Figure 2 depicts an example application graph.

Vertex weight, given by a weight function:

$$w_t : T \to \mathbb{R} \tag{1}$$

stands for the number of operations executed in the process (in Mops) and the edge weight, given by a weight function:

$$w_e : E \to \mathbb{R} \tag{2}$$

is the amount of data transacted (in KB), both as reported by a profiler.

B. *Architecture Model*

The MPSoC hardware platform is also modelled as a directed graph $H = (P, F)$, each vertex $p \in P$ represents a processing element. Vertex weight is the computation

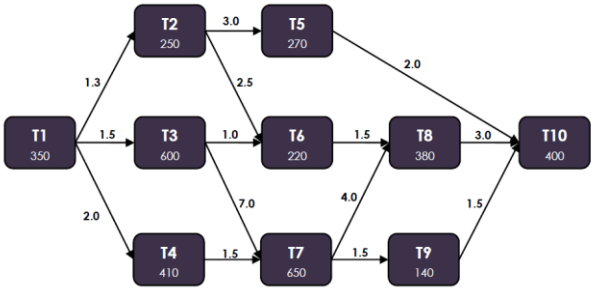


Figure 2. Application representation

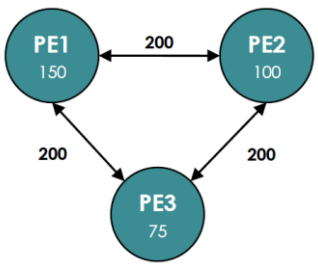


Figure 3. MPSoC platform representation

speed in Mops/s and is given by a weight function:

$$u_p : P \rightarrow \mathbb{R} \quad (3)$$

Each edge represents connection between PEs and weight is the throughput in kBps given by a function:

$$u_f : F \rightarrow \mathbb{R} \quad (4)$$

Figure 3 depicts an example platform graph with three different processing units.

In this model, processing elements differ from each other strictly by raw computational power, i.e. it is assumed that each PE is capable of performing all types of operations equally efficient.

It is also assumed that each processing element has its own private memory and communicate via interconnect network.

C. Mapping Problem

Given the application and architecture models, the mapping problem can be defined as assigning application components - tasks to architecture components in a way that the overall execution time required to execute the application is minimized. The final solution has to be found under the constraints that all the tasks are assigned a PE and one task is assigned only to one PE.

For n tasks (n=|T|) and m processing elements (m=|P|), where $n \geq m$, the cost of mapping of task $i := t_i$ to processing element $k := p_k$ can be defined as the time required to complete all of required computation and communication.

$$T = T_{comp} + T_{comm} \quad (5)$$

Computation time is calculated as:

$$T_{comp}(i,k) = \frac{w(t_i)}{u(p_k)} \quad (6)$$

Communication time is the sum of duration of communication between task i and all predecessor tasks mapped to a processing element different that k (communication cost between two tasks on the same PE is zero). The cost of communication between two tasks i and j mapped to processors k and l respectively is expressed in (7).

$$T_{comm}(i,j) = \frac{w(t_i, t_j)}{u(p_k, p_l)} \quad (7)$$

In order for a task to begin execution, all its predecessors must finish execution. After the predecessor j finishes execution, it would take $T_{comm}(i,j)$ time before the task i can use the results. This has to be calculated for all predecessors of task i and the maximum such value is the time when task i can start execution. The time the task i takes to execute $T_{comp}(i)$ depends on the PE it is assigned to.

IV. Longest Parallel Path Mapping Algorithm

In order to solve this resource allocation problem, a static scheduling algorithm [14] has been developed based on Critical Path Method [5] with additional modifications to develop fast and efficient heuristics. This algorithm's objective is to minimize the overall execution time of the application.

For each task the earliest and the latest possible start time governed by data dependencies can be defined. If both times for a certain process are equal that process is a *critical process.* A set of critical nodes (processes) and edges between them from source to finish form a *critical path.*

The mapping process, as shown on Figure 4, consists of four stages. In the first stage, first of all processing elements are sorted descending by their computation speed. Afterwards, using a modified version of Dijkstra algorithm [12], the longest path is identified and mapped to the fastest processor. As long as there are free PEs, the algorithm identifies the second critical path, the third etc. and maps all tasks in the path to the fastest free PE. Thus, communication costs between tasks on the same path are minimized.

Next, a preliminary schedule is made for mapped tasks, which enables the search for free gaps in PE task list where unmapped tasks will be mapped in stage 3.

Stage 3 comes in two versions. In the first one, mapping of remaining tasks starts with PEs *sorted ascending* by computation speed. For each unmapped task a time frame is calculated - it begins when all of task's mapped predecessors have finished and ends before the first mapped successor has to start. Starting with the slowest PE, mapper checks if it is free during time frame and fast enough to compute the task in time; if

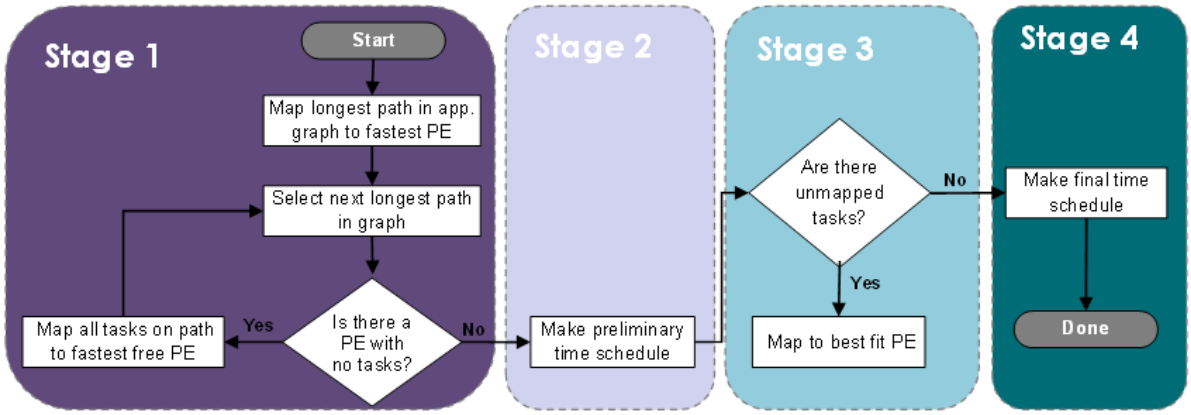


Figure 4. Longest Parallel Path algortihm flow

it is not, the overflow for that task and that PE (expressed as time) is stored for later evaluation.

In the second version, PEs are *sorted descending* by computation speed thus search for candidate PE starts with fastest processor.

When the fit PE is found the task is mapped to that PE – a perfect fit, and in case when there is no perfect fit the task is mapped to PE with least overflow.

The rationale behind the first version of Stage 3 mapping is that it is hardest to find a task that can be mapped to slowest PEs without disrupting entire schedule. On the other hand, mapping one task to slow PE usually means that it will consume the most of the available time frame and leave little chunks of free processor time that are not big enough for any other task to fit in. On the contrary, fast processors require much less time to finish the computation and it is probable that when a task is assigned in a free time frame, enough room will be left for at least one more task to be fitted in the same time frame.

The final schedule is calculated in the last stage of the mapping process.

Based on the algorithm discussed above, Automatic Mapping Tool has been developed in Java programming language and preliminary results of conducted simulations are discussed in the next chapter.

V. Experimental Results & Future Work

In this chapter we present the results of performance testing of LPP algorithm, make comparison with results obtained by using another greedy algorithm – LPT [7] and propose directions for future development.

A. Case Study

The performance of proposed mapping algorithm, in terms of total execution time estimation, was examined on nine different platform configurations for two generic applications. The configuration for each test case is described in Table 1. The numbers represent PE speed in Mops, and in each case communication costs between two processes on different PEs are the same.

1) The LPP algorithm was tested for both versions of Stage 3 (ascending and descending order). The execution times estimation of mapping generic application, represented by a DAG on Figure 2, are shown on chart on Figure 5. In the same figure we compare to the results obtained using LPT algorithm. LPP – ascending version results

TABLE I. Platform configurations for test cases

	1PE [Mops]	2PE [Mops]	3PE [Mops]
Configuration 1	150	150, 150	150, 150, 150
Configuration 2	100	150, 100	150, 100, 100
Configuration 3	75	150, 75	150, 100, 75

In all nine test cases LPP outperforms LPT by up to 130 percent. It is also important to note that LPP algorithm, in case of platform with two fast PEs (both 150 Mops), reaches optimal solution: total duration is equal to critical path duration, and LPT can never reach that situation. In that case LPP algorithm provides speedup of 1.56 times compared to a sequential solution. Even more interesting is that in Configuration 3, when two fast PEs (both 150Mops) are replaced by a set of PEs by a third and a half slower ('3PE' in Table 1) the speedup is still high - 1.5 times compared to sequential solution.

It is also important to note that time complexity of LPP algorithm is $T(n) = O(n \log n)$ which is a significant improvement compared to complexity of LPT algorithm which is $T(n) = O(n^3)$ [7].

2) LPP – descending version results

The results for the second version of LPP (Stage 3 PE search in descending order) are exactly the same as in the previous case. A possible explanation for this is that other features of the algorithm govern task scheduling in later stages. For example, the first stage of the algorithm works in a way that it puts most load to fastest processors which later have little room to host additional tasks. On the other hand slower processors are left with much more idle time after the preliminary scheduling and have a higher

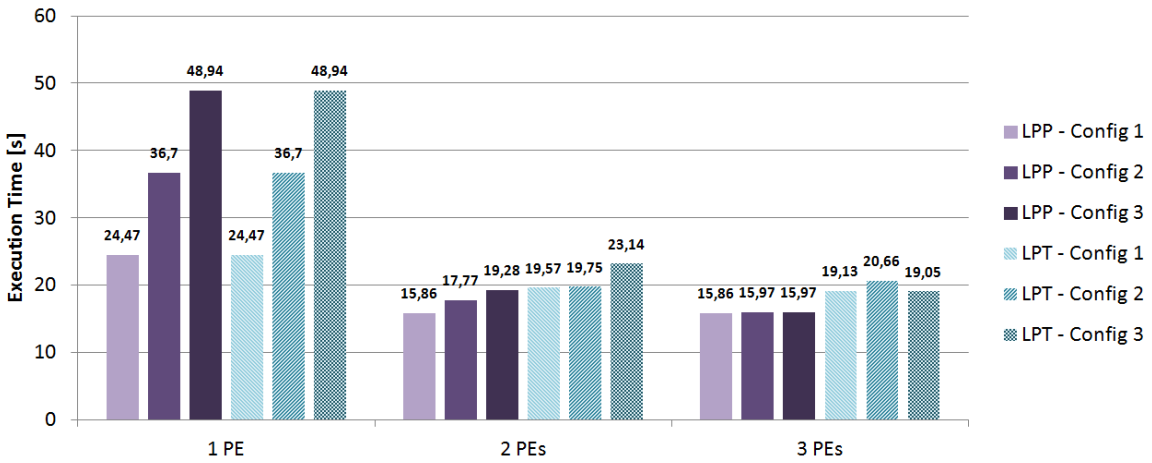


Figure 5. Total execution time comparison

possibility to contain a gap in scheduling that is big enough to fit a new task without overflow. Another possible reason for this, in case when all processors are left "crowded" after the preliminary scheduling, and overflow is almost iminent, is that faster processors will have less overflow (can compute the task faster) and will be filled with new incoming tasks first, regardless of the order of processor list traversal.

B. *Future development*

In future research, the LPP algorithm should be improved so it could tackle the following two issues: support for application with loops and recursions, and support for MPSoC platforms that have a more specialised type of components which are more suitable for a certain type of operation (e.g. floating point operations) – *mapping affinity problem.*

Another direction of development could address DSE from another point – giving recommendation for the platform architecture by selecting the most suitable configuration from a set of available components.

VI. CONCLUSION

In this paper we have given a short overview of Design Space Exploration, as a vital concept for modern heterogeneous multiprocessor embedded system, and related work. Further, we have addressed the issue of lack of quality mapping algorithms – crucial for automation of the entire DSE process by presenting new greedy strategy based on critical path method.

The presented LPP algorithm provides efficient automatic mapping strategy which exploits available task level parallelism in the given application. Conducted case study shows that the algorithm is able to schedule tasks on heterogeneous platforms with variable processor performance and achieve speedup very close to optimal solution in polynomial amount of time. In comparison to another acknowledged greedy strategy, LPP algorithm shows significant performance improvement – up to 130 percent. However, some issues concerning exploiting data level parallelism available in application and higher level of "platform awareness" remain and will be dealt in future research.

REFERENCES

[1] D. D. Gajski, S. Abdi, A. Gerstlauer and G. Schirner, *Embedded System Design. Modeling, Synthesis and Verification*. Ed. New York: Springer, 2009.

[2] H. Park, H. Oh and S. Ha, "Multiprocessor SoC Design Methods and Tools," *IEEE Signal Processing Magazine*, vol. 26, pp. 72 – 79, November 2009.

[3] W. Wolf, "Hardware and Software Co-design," in *High-Performance Embedded Computing. Architecures, Applications and Methodologies.* Ed. San Francisco, CA: Elsevier, 2007.

[4] J. Lin, A. Srivatsay, A. Gerstlauer, and B. L. Evans, "Heterogeneous multiprocessor mapping for real-time streaming systems," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, 2011, pp. 1605 – 1608.

[5] W.H. Kohler, "A Preliminary Evaluation of the Critical Path Method for Scheduling Tasks on Multiprocessor Systems," in *IEEE Transactions on Computers*, Vol C-24, pp. 1235 - 1238, December 1975.

[6] K. Vivekanandarajah, and S. K. Pilakkat, "Task Mapping in Heterogeneous MPSoCs for System Level Design," in *Proceedings of 13th IEEE International Conference on Engineering of Complex Computer Systems*, Belfast, UK, 2008, pp. 56 – 65.

[7] *V.* Zadrija and V. Sruk, "Mapping algorithms for MPSoC Synthesis" *in Proceedings of MIPRO 2010, 33nd International Convention*, Opatija, Croatia, 2010, pp. 624 - 629.

[8] D. Pierre, *Optimization Theory With Applications*, ser. Dover Books on Mathematics Series. Dover Publications, 1986.

[9] R. Garey and D. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*, ser. Series of Books in the Mathematical Sciences. W. H. Freeman, 1979.

[10] J. Hopcroft, R. Motwani, and J. Ullman, *Introduction to automata theory, languages, and computation*, ser. Addison-Wesley series in computer science. Addison-Wesley, 2001.

[11] M. Palesi, T. Girvagis,. "Multi-Objective Design Space Exploration Using Genetic Algorithms", *Proceedings of the Tenth International Symposium on Hardware/Software Codesign (CODES)*, 2002, pp. 67 – 72.

[12] R. Sedgewick, K. Wayne, Algorithms, 4th Edition, http://algs4.cs.princeton.edu/home/, December 2013.

[13] M. Thompson, H. Nikolov, T. Stefanov et al., "A Framework for Rapid System-level Exploration, Synthesis, and Programming of Multimedia MPSoCs", *Proceedings of the 5th International Symposium on Hardware/Software Codesign CODES+ISSS'07, Salzburg, Austria, 2007*, pp. 9-14.

[14] N. Frid, V. Sruk. "Longest Parallel Path Mapping Algorithm for Heterogeneous MPSoCs", ACM womENcourage Europe Conference, March 2013, accepted for publication.

Event Calendar for Internet Data Sources

P. Rantanen* and P. Sillberg*
* Tampere University of Technology/Pori Unit, Pori, Finland
{petri.rantanen,pekka.sillberg}@tut.fi

Abstract – Several nations in Europe have released formerly closed data sources collected by government agencies and municipalities. These open data initiatives have enabled new and innovative services. The research presented in this paper is funded by the European Union Regional Fund as part of the Satakunta Regional Centre for Economic Development, Traffic and Environment funding program, and by the City of Pori. One of the goals of the research is to study how open data can be utilized technically. This paper presents technologies found useful for the implementation of a service utilizing open data. These technologies are presented in this paper as part of an event calendar service that uses open data sources available in the region of Satakunta, Finland. The overall architecture of the service is described briefly, and the paper focuses on the open source components used to implement the user interface, front-end service, and database. Despite the use of region-specific data sources, the utilized solutions i.e. hybrid databases, search platforms, and HTML5 technologies can be exploited in the design of other services.

I. INTRODUCTION

The research presented in this paper is part of the AVARAS project. AVARAS is an Open Data Project in the Satakunta Region in Western Finland conducted by Tampere University of Technology, Pori Unit. The idea of it is to study and implement business opportunities related to Open Data. It is a regional activity with close connections to more than ten SME-sized software companies that are used to guide and direct the project activities. The project is funded by the European Union Regional Fund as part of the Satakunta Regional Centre for Economic Development, Traffic and Environment funding program. [1]

One of the main goals of the project is to study how open data sources can be utilized technically. The research is conducted using pilot applications, which will be openly available for public use. Their purpose is especially to publish interfacing technologies and interface specifications, which can be re-used in new applications. Similar activities are taking place in other regions in Finland [2], [3]. This paper presents one of the pilot applications designed in the AVARAS project.

The application presented in this paper is an event calendar service, which utilizes event data available on the public Internet. The service does not require any input from the event publishers, and can be set to crawl any number of sources automatically. There are many event calendars available on the Internet and the idea of an event calendar is in no way a new one, even though one could argue that there are not that many calendars that can list external events without some kind of input from the event publisher or event provider. Nevertheless, the main reason for choosing an event calendar application in particular was the type of container it provides. Basically any data can be thought of as "an event": the daily weather forecast, road accident reports, sports events, or movie showtimes – which are used as an example case in Section V of this paper. In a way, the idea of an event calendar can be generalized to contain almost any kind data. This paper does not go into too much detail when describing the basic calendar functionalities, but attempts to illustrate how the chosen technologies could be used both in this use case and perhaps in a more general way outside the event calendar context by describing the advantages and disadvantages of the chosen technological solutions.

As mentioned above, the example shown in Section V uses details of movies from the local cinema as a data source, though any other data source could have been chosen as well. The other sections of this paper are as follows: Section II gives a short overall presentation of the chosen system architecture; Section III describes the chosen database solution; Section IV explains the designed REST interface; Section VI discusses some of the issues related to the application; and finally, Section VII summarizes the paper.

II. SYSTEM OVERVIEW

The system introduced in this paper is an effort to utilize publicly available data in the region of Satakunta, Finland, and also test the usability of open source software components and libraries. The selection was made partly by suggestions from the steering committee of the project, and partly guided by academic ambitions. Fig. 1 illustrates the basic organization of the system. There are three distinct components: data storage, front-end service, and client, which are described in more detail in Sections III, IV and V, respectively.

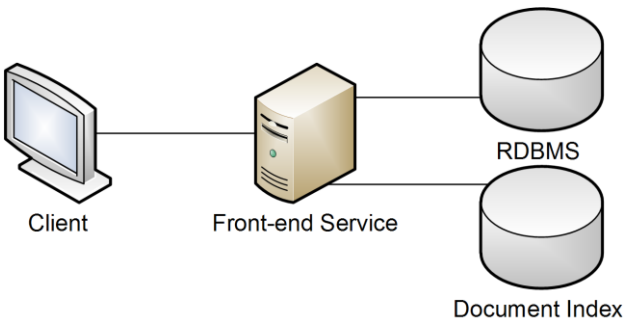


Figure 1. System overview

The *Client* is in principle an HTML web page, which accesses the *Front-end Service's* RESTful (e.g. [4], [5]) interface to retrieve data, and then visualizes this data in a calendar view. The client allows the user to narrow down the dataset by entering filtering terms such as date, time, keywords and also free text searches.

The *Front-end Service* is an access point to the features of the system. It provides user authentication, RESTful API for data in XML format, and adapts the data sources to make them suitable for indexing and storing. The service is written in Java and the main open source components it utilizes are Apache Tomcat [6] and Spring Framework [7].

Our database solution includes features from both the traditional RDBMS and NoSQL [8] approaches. The *Document Index* uses the Apache Solr [9] search platform as it is specialized in indexing and searching, while the *RDBMS* is based on MySQL [10] and is used to store the document relationships and basic structural information such as user credentials.

III. DATABASE

The searching and indexing of data provided by open data sources is the key point of the system. The database solution must have good search capabilities, but it also must be able to handle loosely structured data. Different data sources are often provided in various formats, thus it may be difficult to parse the same data from them all. For example, one source could simply list the names of the performers while another source may contain performer names, email addresses, and groups they belong to.

The SQL approach with its well-defined table-column structure was considered too rigid for the data we might encounter. Of course, we could define all possible columns and leave them null if not available, but that would not be very practical. From the NoSQL approach we found a solution, Apache Solr, which had the capability to index and store the data, as well as sophisticated text search features.

The most interesting feature of Solr is its ability to index any kind of text in various ways. With relatively small changes and additions to the Solr core's basic setup, it is possible to enable request handlers such as full text search, "more like this," and auto completion [11]. As a security measure, the database section in our system is behind the front-end service. This means that the client cannot access features of the database directly. A simple call through API was implemented on the front-end service to handle requests.

There is often a need for storing relationships between documents, and implementing this feature was cumbersome with Solr alone. For storing simple data types, such as strings, texts, dates, or a list of strings, Solr can manage with no problems. Issues arise if there is a need to store complex and denormalized data within the same document. For example, take the situation where there is an event involving many performers in it, and for every performer we need to store a name, email address and URL for the website. There are several possible approaches on how to solve the issue:

1. Store performer data in synchronized lists inside the same event document. This requires some extra effort in marshalling objects, but Solr can still index the document effectively. Maintenance of the document or parts of it suffers from the denormalized data format and from the way Solr handles updates (complete overwrite, i.e. no partial updates).
2. Store performer data in another Solr core as separate documents, and store the relationship (i.e. performer ID) in the original event document. This approach allows easier maintenance and is closer to the traditional relational approach. Solr cannot handle joins as fluently as SQL does, so extra lookups and queries are required. Extra work is required for fully utilizing search and indexing in the Solr multi-core setup.
3. Store performer data in another Solr core as separate documents, but store the relationship (i.e. performer ID and event ID) in another database (e.g. MySQL). If the only task of this other database is mapping of relationships, a graph database (e.g. [12], [13]) might be an alternative choice for a SQL database. This hybrid – or polyglot [8] – approach is much like the previous case, but the handling of relationships might be easier. This increases complexity due to the use of another database technology.
4. Store both the event data and the performer data in the same Solr core, but store the relationship (i.e. performer ID and event ID) in a separate database. Use of a single Solr core simplifies indexing and searching. Storing of multiple types of documents in a Solr core can be achieved as Solr allows dynamic field assignments. The type of document must be known when the data is marshaled. Compared to the first case, maintenance of these smaller documents may be easier.

A combination of multiple databases is called the polyglot persistence model by reference [8]. Together they may create an ecosystem that is more powerful than the sum of its parts [8]. In theory, one could choose any combination of database techniques if using them together generates more benefits than drawbacks (e.g. code complexity, time to develop).

We chose the hybrid approach (3) for the event calendar implementation. In our use case the main event details and minor event details are stored in separate Solr cores, and the relationships are stored in a MySQL database. The minor event details can contain information such as event location and event organizer, or sub events which belong to the main event. In principle the minor event details could be stored directly to relational database, but this would prevent using Solr's text indexing and search functionality. On the other hand, the details could be merged with the main event details, and stored in a single core as described in the approach 1 above. In practice this would require flattening of very complex objects to key-value pair lists, which is more difficult than separating the less important data into individual cores.

Naturally, dividing the data between multiple cores will have an effect on the complexity of search queries and the search performance if the queries need to be targeted to multiple cores. In addition to a performance gain, in our use case there is also an additional advantage achieved by flattening certain important data with the main event details, which is perhaps not immediately evident. Event details often contain various performer details, such as description of a band, movie studio or names of the artists. These descriptions generally vary depending on the data source and analyzing the details to form a single combined entity is difficult. A simple solution to this problems is to ignore the performer-event relations altogether and flatten the data into a single core. This will cause all details to be duplicated and appear slightly different for each event, but it will also enable the use of Solr's text search features when querying for the event details. From the text indexer's point of view these minor differences in the descriptions are irrelevant.

IV. FRONT-END SERVICE

The front-end service implements interfaces for three kinds of functionality: user management, suggestions, and search. In our demo implementation the user management is limited to basic functionality, i.e. creating a new user and modifying user details (such as password and email address). User authentication is provided using the Spring Security [14] framework and in principle supports all features supported by the framework. The reason for implementing only a minimal set of user authentication and authorization features is the nature of the processed data, which consists of publicly available content. There is no need to check user authorization when accessing the content. Of course, the service could be extended with the possibility to add user-created content, but this would complicate the implementation and is not the focus of our current demo implementation. The user management could also be utilized for personalized calendars, but this functionality is not currently implemented.

The suggestion API provides the user with search suggestions. It is implemented using the text indexing features found on Solr and works similarly to the functionality found on many search engines. For example, if the user types "Hobb", based on the indexed content the system may suggest search terms "Hobby" or "Hobbit".

The search API provides the functionality required for content queries and offers features such as free-text search, content filtering, sorting, and paging. Many of the currently available open source databases directly offer interfaces for the aforesaid features [15], [9], which can be easily used over HTTP by using GET or POST. In this case content filtering refers to filtering by values (e.g. only return events taking place between January 1st and January 10th), though similar functionality could also be used to implement filtering for improper content (e.g. do not show adult content to anonymous users or to children).

Depending on the use case, these interfaces might be all that is necessary. The limitations of the APIs are not so much on the basic functionality, but more on the user management. The NoSQL database APIs often do not enforce – or even provide only a very limited – user authentication. In a way this is a shame, as in many cases the REST interface provided by the NoSQL databases could be all that is needed for the query operations. Then again, perhaps the idea is, that it is not the task of the NoSQL database to perform user authentication. Of course, in many cases this does not matter so much as it is often the case with web applications, that the user authentication is performed by the front-end service or by some kind of mediator interface, and the user does not directly interact with the database. The similar approach is possible with Solr and with any other NoSQL database. If a complex authentication scheme is required it might be better to implement a front-end service with interfaces combined with user authentication and authorization. The demo implementation presented in this paper uses the aforementioned design pattern by providing a separate REST interface for query operations, which in turn will delegate the requests to the database when appropriate.

Another advantage received from implementing a separate front-end service is the additional control gained over the requests. This enables simplifying the search queries or limiting the set of available search filters or even resolving the XML element (or JSON object) names given in a query to database field names to be used for sorting and filtering. In particular, mapping element names to database fields is often required when the actual field names defined in the database are hidden to end users or developers using the API. As document databases generally offer output in the form of XML or JSON, it is possible to use the field names defined in the document directly in the response returned to the user, but this is generally not a good practice – it is much easier to change the internal implementation when the fields defined in the database layer are not directly being used by the users of the service API. Also, defining separate field names and XML element names is often a minor task, for example by using annotations in Java. Both the Solr Java implementation (SolrJ) [16] and Spring Data framework [17] provide annotation support for defining Solr field names for Java objects, and the annotations can be combined with Java Architecture for XML Binding (JAXB) [18] or Google GSON [19] to provide XML and JSON responses. This way the response Java objects that are deserialized from Solr query responses can be directly serialized to XML or JSON for external service API responses.

On Solr, database queries cannot modify the database content and thus the risk of code injection can be mitigated by rejecting all other user requests except queries. In our use case, there is no risk of users retrieving content they are not authorized to view by running a malicious query, as all of the content is publicly available. Also, it should be noted that implementing user preferences does not necessarily require user authentication. For example, if the user always wants to see the latest events or only events of a certain type (such as sports), the preferences can be stored on the client software or in local storage [20] on a web browser and from these preferences the required filter parameters can be generated and appended to the queries.

V. CLIENT

Fig. 2 illustrates a calendar with events retrieved from XML API [21] provided by a local cinema in the City of Pori. The calendar view could include any events indexed by the front-end service, but for simplicity only movie events are shown in this example. The view in the figure is a screenshot from a web browser. The main page layout is created using the w2ui JavaScript UI Library [22], which consists of plugin extensions for the popular jQuery library [23] and the calendar view itself is an embedded component created with the FullCalendar jQuery plugin [24]. There are many open source alternatives that can be used to implement web user interfaces (e.g. [25], [26]) and calendars (e.g. [27], [28]), and it is possible to implement a new user interface from scratch using basic JavaScript or jQuery, but in many cases the pre-made plugins and libraries provide a better starting point. The choice between the various alternatives is partly use case-specific as not all libraries provide similar functionality, and partly based on the developer's preference or pre-knowledge of a particular library. In this use case we use movie details from the local cinema.

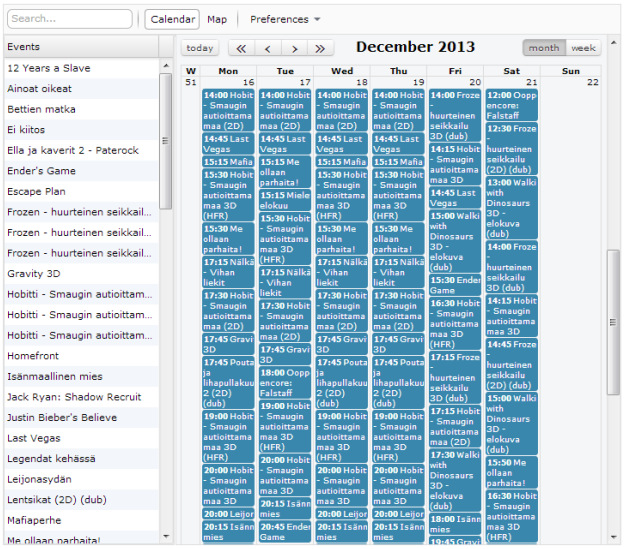


Figure 2. Calendar view

The calendar view in Fig. 2 shows all events currently listed in the events panel on the left. In the example case, no search terms have been given and thus all movie events known by the system are shown. The search field at the top left accepts both free text search and the basic query syntax of the Solr query parser. The Solr query syntax is based on the Lucene Query Parser syntax, which can be somewhat cryptic for end users. There are alternatives for the default parser such as the Extended DisMax Parser [21], which can provide better usability. However, in general, users will most likely prefer free text search over syntax-based queries, and it is better to create filter parameters and additional search terms programmatically and append them to the search query given by the user. The additional parameters can originate from, for example, checkbox selections or user preference options.

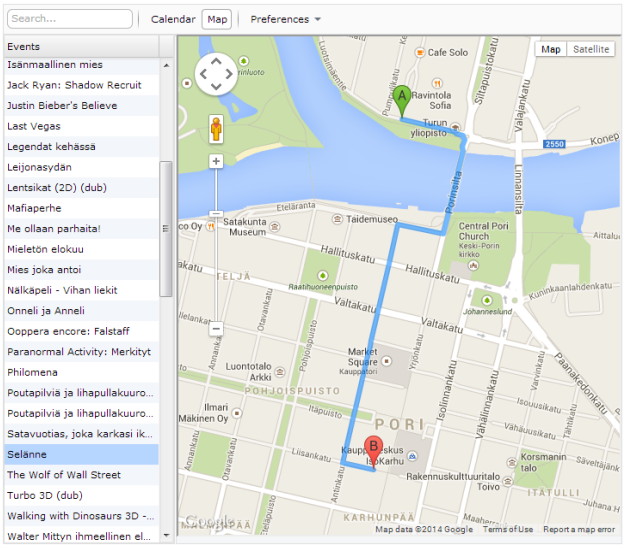


Figure 3. Map view

Fig. 3 shows the location of a selected event on a map using the Google Maps API [30] embedded in a w2ui panel. Showing a map and directions is often quite simple, but they can provide a lot of additional value for the end user. Most of the online map services provide JavaScript APIs, which can be used to show a location on a map and provide guidance to the location from the user's current location [31] on foot, by car or by public transportation. Unfortunately, in most cases the directions are not turn-by-turn capable and cannot as such be used to navigate directly to the target location except by continuously re-calling the mapping service, which can be very resource-intensive. To enable turn-by-turn navigation, the use of a native map application found on a mobile device is often required. If more complex map operations are required, the use of a native map application may be a better choice than a web-based client. In addition to extended functionality, native applications generally perform much better on mobile devices when compared to JavaScript web applications with similar functionality.

VI. DISCUSSION

This paper presented an event calendar service, which can be used to show data retrieved from various data sources, and provided some technological insight on how to implement the aforesaid service. Nevertheless, there remain several problems with the current implementation, and these problems are not strictly related only to the application in question. As mentioned above, the newer database technologies have not yet been fully utilized, and because of this it is often hard to find ready-made solutions for the problems. This is especially true when compared to traditional relational databases – several good books have been written and best practices developed to solve various implementation problems related to relational databases. The lack of best practices naturally creates problems for adapting new technologies, but on the other hand it makes the subject an excellent research topic. When adapting to a new technology, there is also a great temptation to abandon the old models and use new technology for every use case – even when it is not appropriate to do so. For example, it is possible to use a

key-value database or document database to map relationships, even though that is not what they are designed for, while a relational database would be a more appropriate choice.

Another problem is finding the event sources from the Internet. It is relatively easy to find listings (e.g. [32]) of usable event APIs by using a simple web search, but utilizing the sources requires implementing parsers for each of the various APIs. Implementing a parser is generally a straight forward task, but it does require one to study the API documentation and perform some programming work. Another option is to search events from popular web pages such as Facebook or Google Calendar – or use the APIs provided. The availability of events is dependent on the services used, and extracting the events may require special permissions from the publishers of the events whether they are the users of the service or the service operators. In practice it may be better to use the service APIs directly instead of crawling the web in hope of finding up-to-date event details in sensible format is difficult. Also, in many cases the online calendar applications are dynamically populated with, for example, JavaScript, and can be very challenging to parse. Extracting content from dynamically created web pages has been researched by e.g. [33] and [34].

In addition to purely technical problems, there are also other issues related to implementation and usage. In many cases it is often difficult to figure out whether the data in question is actually *open* data or merely *openly* available [35], [36]. This can be illustrated by using the example chosen in this paper. If the cinema does not show copyright notices or present license agreements on its web page nor are there any present in the provided data, does that mean that the data is truly open and free to use? Should permission be asked from the cinema owner, or maybe from the company who made the movies, or even from the company who made the local translations of the movie descriptions? In our use case one might assume that showing the movie details and showtimes on a public calendar is only good publicity for the cinema and a way to advertise in a way the cinema did not think of, but legally this may not be the case.

Also, many open data sources do offer data freely, but with certain limitations. The user may have to register first, or maybe the data cannot be shown openly and freely in all possible use cases. Finding these license agreements is not always a simple task. The service presented in this paper mostly ignores these issues by being a technology demo, but when designing an actual implementation, which will be a publicly available service, one should be careful to check that the apparently free and open data is truly free and open.

VII. SUMMARY

This paper presented an event calendar service, which can be used to visualize open data. The overview of the service was presented, and the chosen technical solutions were explained. The issues encountered while implementing the service were also dealt with briefly. The pilot application proved that the chosen technologies were promising candidates for future studies. In the future, our research will continue to focus on the new database technologies. We will also continue to improve the calendar implementation, with a possibility of running a user trial on a local summer festival. It was also found that there is a need for a separate study on the issues concerning the licensing of open data.

REFERENCES

[1] Tampere University of Technology, AVARAS Regional Initiative to Reveal the Opportunities of Open Data, https://www.tut.fi/avaras/?page_id=424. Retrieved February 24th, 2014.

[2] Open Data TRE, Open Data TRE, http://www.hermiagroup.fi/opendatatre. Retrieved February 24th, 2014.

[3] Helsinki Region Infoshare, Helsinki Region Infoshare | Open regional data, http://www.hri.fi/en/about. Retrieved February 24th, 2014.

[4] R. T. Fielding, Architectural Styles and the Design of Network-based Software Architectures. Doctoral dissertation, University of California, Irvine, 2000.

[5] Wikipedia.org, Representational state transfer, http://en.wikipedia.org/wiki/Representational_state_transfer. Retrieved February 24th, 2014.

[6] The Apache Software Foundation, Apache Tomcat, http://tomcat.apache.org. Retrieved February 24th, 2014.

[7] GoPivotal Inc., Spring Framework, http://projects.spring.io/spring-framework. Retrieved February 24th, 2014.

[8] E. Redmond and J. R. Wilson, "Seven databases in seven weeks," O'Reilly Media, ISBN 978-1-934356-92-0, 2012.

[9] The Apache Software Foundation, Apache Lucene - Apache Solr, http://lucene.apache.org/solr. Retrieved February 24th, 2014.

[10] Oracle Corporation, MySQL, http://www.mysql.com. Retrieved February 24th, 2014.

[11] The Apache Software Foundation, Suggester - a flexible 'autocomplete' component, http://wiki.apache.org/solr/Suggester. Retrieved February 24th, 2014.

[12] Neo Technology, Neo4j, http://www.neo4j.org. Retrieved February 24th, 2014.

[13] I. Robinson, J. Webber, and E. Eifrem, "Graph databases," O'Reilly Media, ISBN 978-1-44935-626-2, 2013.

[14] GoPivotal Inc., Spring Security," http://projects.spring.io/spring-security. Retrieved February 24th, 2014.

[15] MongoDB Inc., mongoDB – HTTP Interface, http://docs.mongodb.org/ecosystem/tools/http-interfaces. Retrieved February 24th, 2014.

[16] The Apache Software Foundation, SolrJ, https://wiki.apache.org/solr/Solrj. Retrieved February 24th, 2014.

[17] GoPivotal Inc., Spring Data, http://projects.spring.io/spring-data. Retrieved February 24th, 2014.

[18] E. Ort and B. Mehta, "Java Architecture for XML Binding (JAXB)," http://www.oracle.com/technetwork/articles/javase/index-140168.html, 2003. Retrieved February 24th, 2014.

[19] Google Inc., google-gson, http://code.google.com/p/google-gson. Retrieved February 24th, 2014.

[20] World Wide Web Consortium, "Web Storage," http://www.w3.org/TR/webstorage, 2013. Retrieved February 24th, 2014.

[21] Finnkino Oy, XML Services, http://www.finnkino.fi/xml. Retrieved February 24th, 2014.

[22] w2ui, JavaScript UI – w2ui, http://w2ui.com/web. Retrieved February 24th, 2014.

[23] The jQuery Foundation, jQuery, http://jquery.com. Retrieved February 24th, 2014.

[24] A. Shaw, FullCalendar - Full-sized Calendar jQuery Plugin , http://arshaw.com/fullcalendar. Retrieved February 24th, 2014.

[25] Telerik, jQuery-powered framework for desktop-style apps and mobile sites | Kendo UI Web, http://www.telerik.com/kendo-ui-web. Retrieved February 24th, 2014.

[26] PrimeTek, PrimeFaces, http://primefaces.org. Retrieved February 24th, 2014.

[27] K. Stetz, CLNDR, http://kylestetz.github.io/CLNDR. Retrieved February 24th, 2014.

[28] Bic.cat, BIC Calendar, http://bichotll.github.io/bic_calendar. Retrieved February 24th, 2014.

[29] The Apache Software Foundation, ExtendedDisMax, http://wiki.apache.org/solr/ExtendedDisMax. Retrieved February 24th, 2014.

[30] Google Inc., Google Maps, https://maps.google.com. Retrieved February 24th, 2014.

[31] World Wide Web Consortium, "Geolocation API Specification," http://www.w3.org/TR/geolocation-API, 2013. Retrieved February 24th, 2014.

[32] W. Santos, "93 Events APIs: Eventful, Upcoming, Google Calendar," http://blog.programmableweb.com/2012/05/01/93-events-apis-eventful-upcoming-google-calendar. Retrieved March 28th, 2014.

[33] C. Duda, G. Frey, D. Kossmann, and C. Zhou, "AJAXSearch: crawling, indexing and searching web 2.0 applications," in Proceedings of the VLDB Endowment, New Zealand, vol. 1, issue 2, pp. 1440-1443, August 2008.

[34] S. Choudhary et al., "Crawling rich internet applications: the state of the art," in Proceedings of the 2012 Conference of the Advanced Studies on Collaborative Research, Canada, pp. 146-160, November 2012.

[35] Open Knowledge Foundation, Open Definition, http://opendefinition.org/od. Retrieved February 24th, 2014.

[36] Open Knowledge Foundation, Open Licenses Service, http://licenses.opendefinition.org. Retrieved February 24th, 2014.

Simulation Library for Resource Constrained Project Scheduling with Uncertain Activity Durations

Mario Brčić* and Nikica Hlupić *
* Faculty of Electrical Engineering and Computing, Zagreb, Croatia
mario.brcic@fer.hr, nikica.hlupic@fer.hr

Abstract – Stochastic Resource Constrained Project Scheduling (SRCPS) is among the hardest combinatorial problems. Exact calculations of interesting measures, such as expected project duration and the probability of satisfying the deadline, using known probabilities are in #P even for relaxed instances of the problem where resource constraints are ignored. The most common approach is to use substantial simulation to evaluate candidate solutions. All of the work so far uses ad-hoc developed simulation environments with prevalent use of a priori generated activity duration scenarios. This paper describes the discrete-time simulation library aimed to support the creation of simulation-based algorithms for solving SRCPS problems with known probability distributions of activity durations. The library is designed, in the first instance, with shared memory parallelization of simulation, using OpenMP. Runtime parallelized generation of random activity duration scenarios is supported and we deal with "inconveniences" that are otherwise elegantly avoided using a priori generated activity duration scenarios. However, in some approaches that is not feasible and runtime scenario generation has to be used. We propose modular organization of simulators that enables better reuse of basic intrinsic project scheduling functionality.

I. INTRODUCTION

Stochastic Resource Constrained Project Scheduling Problem (SRCPSP) is a relevant generalization of previously well-researched deterministic problem [1]. The model and its variants are general enough to model real-life projects. It uses random variables with known probability distributions to model uncertain elements.

The solution for deterministic variant of the problem is a schedule calculated offline, before the project start, that is executed exactly as planned using the trivial policy that starts activities according to the schedule. This is not the case for stochastic variant. Instead, the solution is a strategy, i.e. policy, that guides creation of schedules using the information that becomes available during the project execution – a form of contingency planning is necessary.

Unfortunately, these problems are computationally very hard and evaluation of candidate solutions is much harder than evaluation of candidate solutions for the deterministic variant where it is in low polynomial complexity. It has been shown in [2] that, even in the case of project with no resource constraints and discrete activity duration distributions with just two values in the support, the exact calculation of expected project makespan is #P-complete. Simulation is used as a way of obtaining evaluations of measures of interest such as: objective function, execution statistics, etc. Of course, depending on the problem, each simulation-obtained estimate is noisy due to the simulation error. When using simulation for evaluation and choosing among several options through comparisons, we are trying to achieve as precise estimates as possible for each pairwise difference between the estimates in order to make a selection. Due to the mathematical properties of variance of the difference of random variables and based on assumptions that the estimate errors of are positively correlated [3], using the identical set of scenarios for simulation of each option reduces the variance of differences between the estimates for options. This is the basis for using the a priori generated scenarios for all the simulations where the procedure permits. Also, these scenarios are usually generated in serial mode because the cost of generating them is much smaller than the simulation cost which has to be parallelized. In serial mode, all the benefits of existing pseudo-random number generators and their mathematical properties are reaped.

There has been substantial work in the area of SRCPSP with uncertain activity durations. Three basic types of solving procedures have been applied: predictive, proactive, and reactive scheduling. Predictive scheduling ignores uncertainty and uses point estimates in calculation of the schedule. There are two types of reactive methods: completely online and schedule repair. Online methods create schedule in steps without using baseline schedule. Schedule repair methods use the baseline schedule and perform the rescheduling when the baseline is made infeasible due to random events. Both reactive procedures search for policies that govern the runtime decisions depending on the current project situation. Proactive methods try to create schedules that are robust with respect to some measure of robustness. Most commonly used measures of robustness are quality and stability robustness. Although protected against some uncertainty, proactive schedules, just as predictive ones, can be yielded infeasible due to unfavorable activity duration scenarios. In all cases, a scheduling policy is a necessary component of the solution that makes project execution possible and

ensures project termination. Theory of solving these problems is based on stochastic dynamic programming [4] and, due to the "curses of dimensionality", the solving efforts usually steer to suboptimal methods [5]. One approach is to define policy families, their parameterization, and use some search method to find good enough solution inside that family. Other approach is to work with the value function approximations.

Of existing works in SRCPSP that used simulation, Stork in his PhD thesis [6] used simulation in branch and bound procedure to evaluate objective value of nodes in the search tree. Van de Vonder et al. in [7] used the simulation in improvement procedure over initial buffered schedule obtained by Starting Time Criticality (STC). Deblaere et al. in [8] used simulation based descent to search the family of proactive resource-based policies with release times. Ashtiani et al. in [9] defined PP-policies and searched the policy space using the two-phase genetic algorithm and simulation. Authors in [10] used simulation with electromagnetic heuristic. In the related field of machine scheduling, authors in [11] used genetic programming and evolutionary algorithms in order to search complex policy space. Additionally, authors in [12] used support vector regression in value function approach for finding good rescheduling policies with pre-generated scenarios. But, all these works used simulators made specifically for the problem at hand where certain number of scenarios is generated a priori, before the optimization, and later used for all evaluations. Csaji and Monostori in [13] used fitted Q-learning to find good dispatching policies for machine scheduling. They also used distributed sampling, but didn't specify the used random sampling mechanism in their ad-hoc approach.

In order to make the future developments on dynamic optimization of SRCPSP with activity duration uncertainties, the area of SRCPSP research that withstood the greatest developments, we propose a simulation library that can be used as a basis of future research as well as the applications of the existing work in practice.

Contributions of this paper are: general-use simulation library for dynamic optimization of SRCPSP with uncertain activity durations, runtime parallelized scenario generation that enables sampling from the conditional discrete distributions, model separation of simulator into the simulator/sub-simulator components, and the MDP model builder in order to measure the optimality gap for small toy problems.

This paper is organized as follows. General structure of the library is given in the section 2, with detailed description of components in the subsections. Section 3 gives basic performance measurements. Finally, conclusion is made in the section 4.

II. LIBRARY STRUCTURE

The library consists of several important components:

- *Project scheduling problems* – contains the basic project scheduling problems
- *Pseudo-random number generators (PRNG) and probability distributions*
- *Simulators* – simulate the project execution with defined simulator-dependent state space and action space
- *Controllers* – implement the policies, i.e. make decisions based on the current project state
- *Experiments* – contains facilities for the most commonly used parallelized simulations
- *Optimization algorithms* – basic collection of algorithms that can be used as basis for new algorithms or for comparisons with them
- *Model builder* – builds Markov Decision Process (MDP) model from certain types of simulators that have factored project state representations

As in many cases simulations are easily parallelizable, the library is designed to support shared memory parallelization through OpenMP. The library takes care to support parallel random number generation based on the existing work in literature. In one typical application scenario, there can be more than one levels of parallelization, but for performance purposes nested parallelization is by default disabled. The first prototype of the proposed library was implemented in Python with the basic multithreading supported by the language. The prototype is implemented in C++11 for better performance and all of the components except optimization algorithm SBD and model builder are re-implemented with the letter components in plan for the near future work. Due to constrained space, UML diagrams contain only the parts relevant for the exposition.

A. Project Scheduling Problems

SRCPSP is a tuple *(V, E, p, R, B, D, c)*, where *V* is the set of activities, *E* defines the precedence relation, *p* is the joint probability distribution of activity durations, *R* is the set of resources, *B* is the vector of resource availabilities, *D* the matrix of activity-resource demands, and *c* is the objective function, usually the expected makespan. For more details on the formal definition, check [1].

The library holds definitions for the deterministic RCPSP, SRCPSP, and proactive variant where the objective function depends on deviations of realized and predicted baseline activity start times. All problems are realized in accordance to the formal model with a class hierarchy as shown in Fig. 1, where only parts of the classes pertinent to our exposition are displayed. Input data for test problems is available in benchmarks. Standard benchmark data is available for deterministic

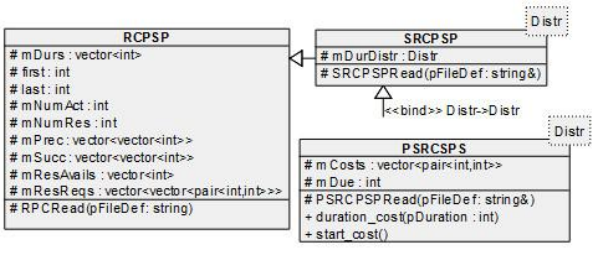


Figure 1. Project problem classes

variant, mostly in the Patterson format. The most commonly used benchmarks are PSPLIB [14] and the newer dataset RG300 [15], both in Patterson format. The simulator library enables reading and parsing the input files. Also, we define an extension on Patterson format to include the data about probability distributions and various costs for more complex SRCPSP variants, such as proactive ones that include costs for activity start deviations with respect to the baseline schedule. In order to obtain test instances for specific problems, the input data for deterministic instances can be "enriched" with the rest of the necessary data for more general instances. The library provides the option of instance enrichment using random data. Enriched problem instances can be saved in the extended form for later usage.

The described project scheduling problem classes are complex and have a big memory footprint. The most of the data is constant and immutable during the project execution. The only data that can change during the project execution is the realized activity duration vector, due to revealing actual activity durations, and in some applications the activity duration joint probability distribution, due to the conditioning on the available data. For that reason, in order to reduce the copying overhead of project scheduling problem definitions when creating new instances of simulators, one project description is shared among the all using objects. All evolvable/mutable structures are isolated into special classes. *EvolvableDuration* class holds and tracks the data on evolution of the activity duration vector. Objects of this class have small memory footprint relative to the project definition objects and the gains in speed are considerable when simulators use them, as is demonstrated in the measurements section.

B. PRNGs and probability distributions

As mentioned earlier, most of the existing work in the area of SRCPSP uses pre-generated random scenarios during the whole simulation and optimization process. In that case we can generate all the scenarios in the main thread and use the generated data in parallel simulations. However, when we have to generate random numbers during parallel simulations, as several works warn [16, 17], we have to be careful which PRNGs we use and how we initialize them. In parallel simulations, in order to avoid bottlenecks if we were using the singleton thread-safe random number generator, we need to generate random variates from several streams. The problem with the most of existing, conventional, PRNGs is their serial nature, as they operate in a sequential fashion through the states in the state space where the next state depends on the process history that contains visited states. There are several options to achieve that with conventional PRNGs in a way that each random stream would be as independent as possible: random seeding, parameterization, block splitting, and leapfrogging [18]. The former two are multi-stream, while the latter two are sub-stream approaches. Random seeding of PRNG seeds, although the easiest, is a procedure without any theoretical guarantees on independence of different streams and is feasible only for PRNGs with very long periods. Parameterization approach uses as many PRNGs of the same type as needed and all of them have carefully chosen parameters, but theoretical guarantees for this approach are not strong. Block splitting splits one stream of random numbers into consecutive blocks of a priori known and set length. Leapfrogging also uses one stream of random numbers which is split into sub-streams in the way that k-th sub-stream gets each k-th number from the main stream.

In the library, we have implemented support for the two types of parallel PRNGs (Fig. 2) using publicly available implementations: Mersenne twister with carefully constructed parameterizations of all streams as presented in [19] and yarn PRNG that uses leapfrogging as presented in [18]. In future we intend to implement several other parallel RNGs. L'Ecuyer et al. in [20] presented a solution to create multiple long streams and substreams using combined multiple recursive generator that can be used in parallel. Salmon et al. in [21] present a very different and promising approach to conventional PRNGs, introducing counter-based PRNGs which can be easily parallelized in both multi-stream and sub-stream modes. This library follows the design choices present in libraries GNU Scientific Library (GSL) and Boost where the pseudo-random number generator is decoupled from the transformation algorithms necessary for getting variates from variety of probability distributions. This enables the easy validation of results using different PRNGs.

In order to enable modeling of heterogeneous joint probability distributions, we have modeled probability distributions inside the class hierarchy with the top class Distribution as is visible in Fig. 3. The univariate probability distributions in the library are designed using adapter pattern around probability distributions in other libraries and expose functionalities both for pseudo-number generation transformations and statistical tests. This enables uniform behavior even if using different support libraries in the background. The library, at this moment, has implementations of the following univariate distributions: beta, general discrete, triangular, uniform continuous, and uniform discrete. Also, the library has several classes of independent joint distributions which are important for modeling activity durations. *JointIndepDistribution* is the class of independent joint distributions in which independent component distributions do not necessarily belong to the same distribution family. *JointIndepIdentDistribution* is, on the other hand, the template class of independent joint distribution with all independent component distributions belonging to the same distribution family. Among the

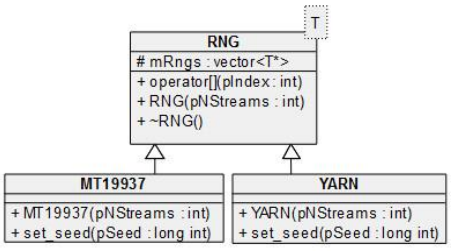


Figure 2. Supported PRNGs

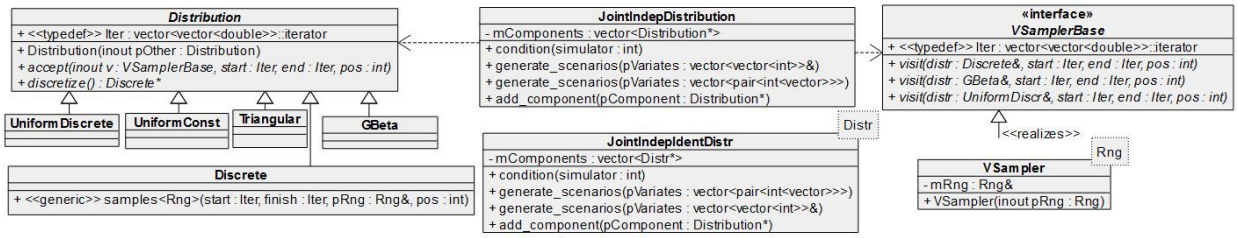


Figure 3. Probability distributions and related classes in the library

latter ones, the joint independent distributions of general discrete distributions as components is the most important for our purposes as, by the definition of SRCPSP, the activity duration distributions are discrete with the support in N_0.

All the existing probability distributions are discretizable so that they can be transformed to the general discrete distribution and used in homogeneous joint distribution for the simulation. Joint distributions can be conditioned on the simulator state. In the case of joint independent distributions, that means that each component for currently running activities can be conditioned on the time elapsed since the activity start, i.e. can get truncated. Truncation at the moment is supported only for general discrete probability distribution, as that was the priority for SRCPSP. For other distributions either rejection sampling or Markov chain Monte Carlo (MCMC) methods can be implemented in the future. In order to enable random sample generation from heterogeneous joint distributions using different PRNGs, a variant of Visitor pattern is used. A template member function *samples* is added to all classes in the hierarchy of basic probability distribution classes except the *Distribution*. Class *VSamplerBase* and class template *VSampler* are part of the pattern. *VSampler*, together with its purpose to implement double dispatch, also decides on which stream of the PRNG object to use in sample generation. The thread ID is used as an index of the stream.

C. Simulators

The specific feature of this simulation library is separation of the simulation engine into sub-components. The library contains basic simulators concerned only with the basic scheduling and schedule book-keeping – very specific and common project scheduling functionality dealing with resource and precedence constraints. These basic simulators can be used as sub-simulators in more complex simulators that have specific features needed for supporting complex project scheduling decision making problems. In the view of the project as a multistage decision making process [4], decisions are made in successive timepoints. This inevitably leads to the fact that complex simulators will use as sub-simulators various variants of parallel schedule generation scheme (SGS). Examples of such simulators are given in Fig. 4.

SerialSimulator and *TPSimulator* are the most basic simulators implemented in the library. The former is the simulator that supports serial SGS and inserts activities in the first feasible timepoint in the existing partial schedule. The latter is transactional parallel simulator and supports

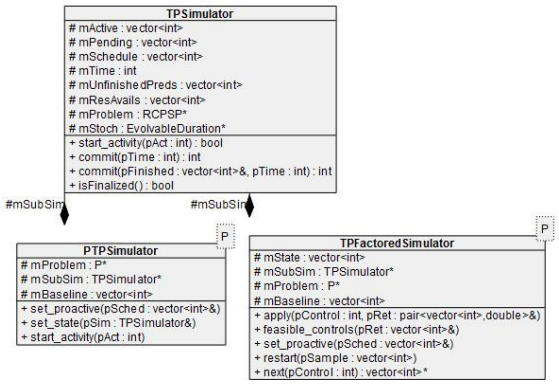


Figure 4. Parallel simulators

making decisions on starting activities in series of increasing timepoints. The decisions in each timepoint have to be committed in order to jump to the next decision timepoint. Both basic simulators have set-based representation of their states.

The complex simulator *TPFactoredSimulator* is the built on the top of *TPSimulator* and has a factored state representation, i.e. the state is represented by a tuple. That simulator also calculates costs for transitions between the states and it is suitable for value function approaches as this state representation is tailored for the value function approximation. *PTPSimulator* is the proactive extension of *TPSimulator* with the calculation of tardiness-earliness costs after each applied control.

Other simulators implemented in future can include more complex control space that is the superset of its sub-simulator's control space, inclusion of the baseline schedule in predictive/proactive-reactive approaches, use of different objective measures, etc.

D. Controllers

Controllers embody policies and in the given state they provide the next control for the decision making process, as well as apply the control to the simulator. The library, at the moment, contains three basic controllers: parallel schedule generation scheme (SGS) using activity priority list (in Fig. 5), serial SGS using activity priority lists, and parallel SGS using resource-based policies with release times. Each controller at each decision timepoint gets, as the input, the whole simulator that is the holder of the project state. It is up to the controller to choose if it will be non-anticipative, as it could also see some future data available in the simulator and that can be useful if modeling problems with objective functions in the form of regret. This unconstrained access to the simulator leaves to the policy to define which conditions and

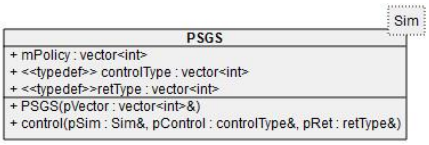


Figure 5. Parallel SGS controller

constraints it will satisfy. The parallel SGS variants are the basis of stochastic scheduling where policies satisfy non-anticipativity and non-retroactiveness, which is in accordance with scheduling decision making in real-world projects. Basic serial SGS does not satisfy non-retroactiveness nor non-anticipativity. The latter is mainly used in deterministic scheduling as it has been shown in [22] that in deterministic RCPSP, unlike serial SGS, greedy parallel SGS is sometimes unable to reach an optimal solution.

E. Experiments

Experiments embed simulations undertaken to estimate some measure of interest. This is the part that is highly parallelizable, and the rest of the library is realized as a support for the parallelization purposes here. Each experiment uses pre-generated set of scenarios as an input. This enables us to choose if and when to run the experiments with the same scenarios. By embedding experiments within each-other we can implement runtime random number generation. Although experiments can be written on case-by-case basis, we have modeled integrated approach to experiments. The experiments are modeled through a static template member method *experiment* in the utility class *Experiments*. The body of that template method has a basic skeleton which uses blocking parallelization. It runs the project evaluations using the controller leaving to the experiment policies to define the necessary data recording. Two experiment policies have been implemented: *ExperimentStartsPol* and *ExperimentAvgPol*. The former records start times of all activities, while the latter evaluates the average project duration under policy defined by the controller. The pseudocode of the experiment body is given below:

```
Experiment(simul,controller,ExpPolicy,scenarios,num_repeats,
num_threads)
    ExpPolicy.init(result)
    Parallel block
        Calculate the task block;
        ExpPolicy pol;
        For each task in block
            Create simulator sim with unused scenario;
            While sim not finished
               Control,cost=Controller.control(sim);
               pol.update1(sim,control,cost,result, taskID);
            pol.update2(sim, result, taskID);
        critical section
           pol.update3(result,task_block data);
    ExpPolicy.post_process(result);
    return
result;
```

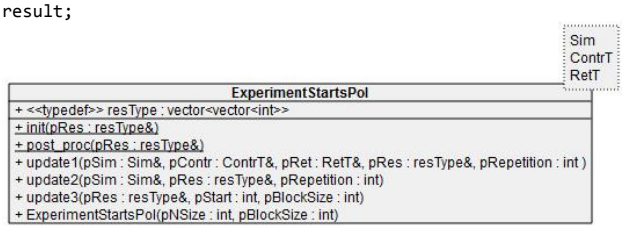


Figure 6. Experiment policy class example

F. Optimization algorithms

The input to the algorithm is a project definition and the output is a controller with good performance. The library, at the moment, contains *GADet* and *SBD*. *GADet* is class that implements the VanHoucke-Debels bi-population genetic algorithm for solving deterministic RCPSP as described in [15]. Many algorithms for solving SRCPSP need the deterministic schedule as an input [8, 23] and this algorithm finds high quality solutions to big problem instances. *SBD* was mentioned in the introduction. It searches for the good enough policy in the family of resource-based policies with release times in order to minimize expected stability and project earliness-tardiness costs in proactive-reactive scheduling problems.

G. Model Builder

For very small problems exact model can be built using *ModelBuilder* as a Markov Decision Process (MDP) represented by the class *MDPModel*. This component uses the explicit state-based simulator (for example *TPFactoredSimulator*), the data about activity duration joint probability distribution, Bayes' theorem, and the depth first search in order to build state indexes, the transition table, and the cost vector. The created MDP model can be solved exactly in order to compare the performance of simulation based approaches relative to the optimum. Possible solving methods are: generalized policy iteration, value iteration, and linear programming (LP) [24]. Former two methods work directly with the MDP model. For the linear programming, there is *LPModeler* class that takes MDP model as the input and creates the linear program that can be solved using some of available LP solvers. This component is implemented in the python prototype library and is planned to be implemented in C++ in the future work.

III. COMPUTATIONAL RESULTS

There are two relevant types of errors in simulation based optimization; model error, which depends on discrepancies between reality and the used model and data, and simulation error. The former is under control of project modeler, and this library supplies several general models with the possibility of extension. The latter can be reduced by increasing the number of experiment replications and parallelization helps in that regard. For the test setting we used all 600 PSPLIB instances with 120 activities enriched with the stochastic data in identical way as in [8]. The program code was written in C++ and all time-measurement results were obtained on a laptop equipped with Intel i7-2670QM quad-core processor. In order to examine the performance and the benefits of parallelization, we have made measurements on simulator that uses *EvolvableDuration* and shared project descriptions (labeled 'Evolvable' in Fig. 7), and the simulator variant that works directly with the project definitions (labeled 'Project' in Fig. 7). The trivial priority vector was constructed and the resulting parallel SGS policy was evaluated using *ExperimentStartsPol* with 1000 scenarios on each project. The graph in Fig. 7 shows experiment runtimes, without experiment setup, averaged over 600 projects with different experiment settings. It is evident that simulators that isolate mutable

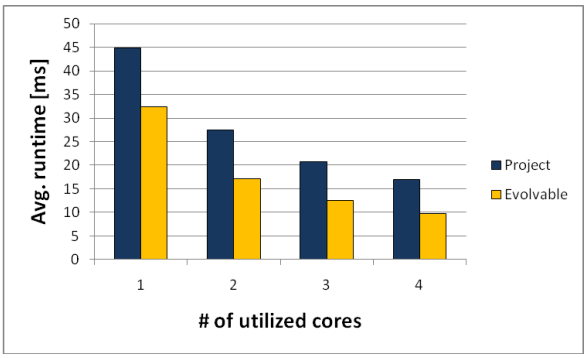


Figure 7. Average runtimes versus the number of utilized cores

structures have better performance. The speedup is close to linear with respect to the number of utilized cores.

IV. CONCLUSION

The general simulator library for SRCPSP has been presented. As the most of the work so far focused only on ad-hoc approaches to simulation, this brings an approach that makes possible and easy to work on future simulation based optimization algorithms without having to "reinvent the wheel". Parallelization is very important issue in simulation, where measures of interest are calculated through averaging many samples. The most common approach is to generate scenarios a priori, before the simulation and, due to favorable mathematical properties of variance reduction of comparisons, compare candidate solutions only with samples based on these scenarios. However, in some approaches samples have to be generated during the runtime, and if the scenario generation is done in parallel there are potential problems with correlated random number streams. We proposed inclusion of several different solutions into our library to circumvent the potential issues. We presented a modular organization of simulators where more complex simulators build on simpler ones, and that enables better reuse of common project scheduling functionalities.

Future work includes empirical data on performance of the library using newly developed as well as existing algorithms. Other models of parallelization in heterogeneous ways using Message Passing Interface (MPI), General purpose computing on graphical processing units (GPGPU), and combinations of approaches could be tried out and used. It would be interesting to see if it is opportune to use GPGPU for calculations of this type because relatively big data structures could hinder the opportunities for the speed-up. Research into the effects of using different PRNGs in parallel simulations might provide interesting insights and recommendations for the practical usage.

REFERENCES

[1] M. Brcic, D. Kalpic, and K. Fertalj, "Resource Constrained Project Scheduling under Uncertainty: A Survey," in *Proceedings of the 23rd Central European Conference on Information and Intelligent Systems,*, Varazdin, Croatia, 2012, pp. 401–409.

[2] J. N. Hagstrom, "Computational complexity of PERT problems," *Networks*, vol. 18, no. 2, pp. 139–147, 1988.

[3] D. P. Bertsekas, *Dynamic programming and optimal control, vol.1*, 3rd ed., vol. 1, 2 vols. Belmont, Mass.: Athena Scientific, 2005.

[4] R. H. Möhring, F. J. Radermacher, and G. Weiss, "Stochastic scheduling problems I — General strategies," *Z. Für Oper. Res.*, vol. 28, no. 7, pp. 193–260, Nov. 1984.

[5] W. B. Powell, *Approximate Dynamic Programming: Solving the Curses of Dimensionality, 2nd Edition*, 2nd ed. Wiley, 2011.

[6] F. Stork, "Stochastic resource-constrained project scheduling," PhD Thesis, Technical University at Berlin, Berlin, Germany, 2001.

[7] S. Van de Vonder, E. Demeulemeester, and W. S. Herroelen, "Proactive heuristic procedures for robust project scheduling: An experimental analysis," *Eur. J. Oper. Res.*, vol. 189, no. 3, pp. 723–733, 2008.

[8] F. Deblaere, E. Demeulemeester, and W. Herroelen, "Proactive policies for the stochastic resource-constrained project scheduling problem," *Eur. J. Oper. Res.*, vol. 214, no. 2, pp. 308–316, Oct. 2011.

[9] B. Ashtiani, R. Leus, and M.-B. Aryanezhad, "New competitive results for the stochastic resource-constrained project scheduling problem: exploring the benefits of pre-processing," *J Sched.*, vol. 14, no. 2, pp. 157–171, Apr. 2011.

[10] P. Godinho and F. G. Branco, "Adaptive policies for multi-mode project scheduling under uncertainty," *Eur. J. Oper. Res.*, vol. 216, no. 3, pp. 553–562, Feb. 2012.

[11] C. W. Pickardt, T. Hildebrandt, J. Branke, J. Heger, and B. Scholz-Reiter, "Evolutionary generation of dispatching rule sets for complex dynamic scheduling problems," *Int. J. Prod. Econ.*, vol. 145, no. 1, pp. 67–77, Sep. 2013.

[12] K. Gersmann and B. Hammer, "Improving iterative repair strategies for scheduling with the SVM," *Neurocomput.*, vol. 63, pp. 271–292, Jan. 2005.

[13] B. C. Csáji and L. Monostori, "Adaptive stochastic resource control: a machine learning approach," *J. Artif. Intell. Res.*, vol. 32, no. 1, pp. 453–486, 2008.

[14] R. Kolisch and A. Sprecher, "PSPLIB - A project scheduling problem library: OR Software - ORSEP Operations Research Software Exchange Program," *Eur. J. Oper. Res.*, vol. 96, no. 1, pp. 205–216, Jan. 1997.

[15] D. Debels and M. Vanhoucke, "A decomposition-based genetic algorithm for the resource-constrained project-scheduling problem," *Oper. Res.*, vol. 55, no. 3, pp. 457–469, 2007.

[16] P. Hellekalek, "Don'T Trust Parallel Monte Carlo!," in *Proceedings of the Twelfth Workshop on Parallel and Distributed Simulation*, Washington, DC, USA, 1998, pp. 82–89.

[17] M. Matsumoto, I. Wada, A. Kuramoto, and H. Ashihara, "Common Defects in Initialization of Pseudorandom Number Generators," *ACM Trans Model Comput Simul*, vol. 17, no. 4, Sep. 2007.

[18] H. Bauke and S. Mertens, "Random numbers for large-scale distributed Monte Carlo simulations," *Phys. Rev. E*, vol. 75, no. 6, p. 066701, Jun. 2007.

[19] M. Matsumoto and T. Nishimura, "Dynamic Creation of Pseudorandom Number Generators," presented at the Proceedings of the Third International Conference on Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing, 1998, pp. 56–69.

[20] P. L'Ecuyer, R. Simard, E. J. Chen, and W. D. Kelton, "An Object-Oriented Random-Number Package With Many Long Streams And Substreams," *Oper. Res.*, vol. 50, pp. 1073–1075, 2001.

[21] J. K. Salmon, M. A. Moraes, R. O. Dror, and D. E. Shaw, "Parallel random numbers: As easy as 1, 2, 3," in *High Performance Computing, Networking, Storage and Analysis (SC), 2011 International Conference for*, 2011, pp. 1–12.

[22] R. Kolisch, "Serial and parallel resource-constrained project scheduling methods revisited: Theory and computation," *Eur. J. Oper. Res.*, vol. 90, no. 2, pp. 320–333, Apr. 1996.

[23] S. Van de Vonder, E. Demeulemeester, R. Leus, and W. Herroelen, "Proactive-Reactive Project Scheduling Trade-Offs and Procedures," in *Perspectives in Modern Project Scheduling*, vol. 92, J. Józefowska and J. Weglarz, Eds. Springer US, 2006, pp. 25–51.

[24] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, 1st ed. Wiley-Interscience, 1994.

Methods and software for estimation of information system dependability

Nikola Šamec *, Alen Jakupović **
* National Protection and Rescue Directorate / National Protection and Rescue Information and Communication System, Zagreb, Croatia
** Polytechnic of Rijeka, Business Department, Rijeka, Croatia
nikola.samec@duzs.hr
alen.jakupovic@veleri.hr

Abstract - Information system dependability reflects the user's degree of confidence in the system. Today's ICT allows connections between information systems in different applications and in different geographic locations. These information systems are called dependable because connections between them cause certain events in one information system to be reflected in the other. Therefore, it is important to have a method for the estimation of information system dependability that will be used in the information system development phase. This estimation will provide arguments for a possible modification of information systems before their final commissioning. The purpose of this paper is to provide an overview of the methods and software for the estimation of information system dependability. Five qualitative and two quantitative methods for estimation of information system dependability are presented. A list of ten estimation software it has been made and five of them are presented in detail.

I. INTRODUCTION

The protection and survival of complex information systems that provide services for the infrastructure of advanced society has become a national and world-wide concern of the highest priority. [30]

Real-time systems are used in the critical applications such as space flight, nuclear technology, aviation, etc. Failure of such systems can lead to major damage, loss of life and environmental impact. The common denominator of these types of failure in organizations is the loss of value in the capital market. Research in [3] states that a failure of information systems can cause a 2% decrease share values for a period of two days.

On the basis of functions and scope of the failure of the system in real time, there are three types of systems:

The first is *Safety-critical* systems required to ensure the safety of EUC (equipment under control), people and environment. Examples of such systems are: shutdown of the nuclear reactor, digital aircraft flight control, etc. [1]

A second type is *Mission-critical* systems whose failure results in failure of the mission. For example, control and coding unit (CCU) of an avionic navigation system or spacecraft, etc. [1]

The third type is *Economically-critical* systems whose failure results in the unavailability of EUC (equipment under control) or a failure such as a server or communications equipment in a business environment (system control reactor nuclear power plant, unavailability of servers in the banking sector due to a fault on the router etc.). [1]

The purpose of this paper is to present methods and software for estimation of information system dependability and to present short review of several important concepts relevant to the understanding of term dependability. Such concepts to be reviewed are the concept of dependability and terminological structure that is published by Jean-Claude Laprie in his work "Dependability: Basic Concepts and Terminology". [15] Other concepts are methods for estimation dependability with emphasis on Markov chains, methods for estimation reliability, availability, maintainability and safety, and software to support these estimation methods.

II. INFORMATION SYSTEM DEPENDABILITY

Dependability of an information system is the ability to deliver service that can justifiably be trusted. [5] There is also an alternative definition by which dependability is defined as the ability of the system to avoid frequent and severe inaccessibility of services than what is acceptable.

Dependability studies two or more systems which interact in a dependent relationship, or how an event in one system affects the operation of other systems and the overall system. Evolution of the concept of dependability through four decades is resulted on terminology that is presented by Jean-Claude Laprie.[15]

The three basic elements of observation are: attributes, threats and methods for achieving dependability.

A. Attributes

Information system dependability is a complex information system characteristic which consists of the following attributes:

- *Availability* is the system's ability to deliver the correct service. [5], [4], [2], [14]
- *Reliability* is the system's ability to continually deliver the correct service. [5], [4], [2], [14]

- *Safety* is the system's ability to operate without catastrophic consequences on the user and environment. [5], [12], [7], [2], [14]
- *Confidentiality* is the system's ability to not permit unauthorized disclosure of information. [5], [4], [2], [14]
- *Integrity* is the system's ability to prevent improper modification of the system. [5], [4], [14]
- *Maintainability* is the system's ability to be easily repaired and maintained. [5], [4], [2], [14]
- *Security* is the system's ability to protect itself from accidental or deliberate attack (is formed by combining the attributes of availability, confidentiality and integrity). [19], [2], [14]

B. Threats

Threats of the information system dependability are:

- *Fault* is improper deviation of at least one characteristics of the system. [21], [14]
- *Error* is the state of the system that can cause *failure*. Errors can be latent or detected [5], [4], [14]
- *Failure* is the permanent cessation of the system's ability to perform a required function. [21], [14]

C. Methods for achieving dependability

There are four basic methods for achieving dependability:

- *Fault prevention* is to prevent the occurrence of a fault, and achieves the quality control during the design and production of hardware and software. [5], [4], [14]
- *Fault tolerance* refers to information systems that deliver the right service due to the existence of an active fault and have well-developed immunity to malfunction. This technique is commonly used for error detection and subsequent system recovery. [5], [4], [2], [14]
- *Fault forecasting* is performed by estimating the behaviour of the system with respect to the occurrence of a failure. It is estimated the current situation, future failures and the consequences of failure [5], [4], [14]
- *Fault removal* is applied in the development and working phase of information systems. The development phase includes testing, diagnosis and repair, while the working phase includes corrective and preventive maintenance. [5], [4], [14]

III. METHODS FOR ESTIMATION OF INFORMATION SYSTEM DEPENDABILITY

These methods perform estimations of the current state of the probabilities of failure and estimations of the probability of the consequences of failure. Using these methods it is possible to estimate the degree of confidence on the ability of the system to meet specific goals. These methods are used to [19]:

- compare possible solutions
- estimate level of system resistance in the work
- estimate system confidence, required resources and costs

Estimations of system dependability can be observed through two approaches:

- *Qualitative* or *ordinal approach*, wherein the identification and classification of the system failure or combination of events which cause system failure is performed [4]
- *Quantitative* or *probabilistic approach* involves estimation of the probability to what extent measured attributes meet requirements. An important component of this approach is the measurement of attribute. [4]

A. Qualitative (ordinal) estimation methods

The following methods are shown: FMECA method, Reliability block diagram (RBD), Fault tree, Attacks tree.

a) FMECA method

FMECA or Failure Modes, Effects, and Criticality Analysis is method that was intended to be used for modelling hardware, but subsequently began to be applied on software (SEEA: Software Error Effect Analysis). Originally FMECA was developed in the '40s for the U.S. military under the standard name MIL-P-1629th. [26]

In the early 60's this method was adopted by NASA to develop space programs such as Viking, Voyager, Magellan and Galileo. [20]

Soon after, the method was extended to civil aviation and auto industries.

The method is used to identify each component or function, discovering ways in which failure occurs and what are the consequences caused by the system. It is possible to make estimations of any possible case of failure and based on the weight of the consequences caused by failure to recognize and execute a prioritization. FMECA process recognizes the critical failure modes and thus enhances the formal recognition of risk for the project and provides an incentive to change the design of the system. [19]

b) Reliability block diagram (RBD)

Reliability Block Diagram executes analysis of reliability and availability of large and complex systems using block diagrams to show network connection. The method was developed by the U.S. military under the standard name MIL STD-756B.

Graphical method consists of two types of components: blocks (representing system components) and dummy nodes (for connections between the components). Graphic topology describes how the reliability of individual components affects the reliability of the system. Dependability of blocks and nodes models is dependent on the condition of its components. The system is considered operational if the ultimate dummy nodes are connected at any point of time. If it does not, the system is considered to be faulty. Serial systems are considered not redundant while parallel systems are considered redundant systems. [9]

c) Fault tree

This method was developed by H.A. Watson from Bell Laboratories in cooperation with the U.S. Air Force, and

later, this method was extended to the field of civil aviation. [8] A model presents a graphical representation of a combination of events that cause the occurrence of the event. Tree can be used for modelling of hardware and software failures, human errors, errors in system maintenance and environmental impact on the system. The model identifies the relationships between undesired events in the system and failure of subsystems that contributes to failure of the entire system. A model for estimation of the reliability is developed by "top-down" principle, and the method can be applied at each stage of system design. The method provides qualitative and quantitative estimations of the reliability of the system. Fault tree represents an acyclic graph with internal nodes (having logic gates such as AND or OR), external nodes (leaves or basic events, which are the components of the system) and edges that represent the flow of information failure in terms of Boolean entities (TRUE or FALSE, 0 or 1). The links that are connected to the edges define the operational dependence on the system components. At any point of time, the logic value of the root node determines whether the system will be operational. [9]

d) Attacks tree

Attacks tree is based on modelling with the help of graphical and mathematical methods. This method was in its original form developed for intelligence agency and for the first time the idea of logic trees threats appeared in the literature in the late 80s. Later, B. Schneier published the concept of attack tree [6].

The method of attack tree is closely related to the concept of the idea of fault tree method because it describes a set of events that can lead to system failure. This tree can be used for modelling of all possible attacks on the system, which provide a formal, methodical way to describe the security system that is based on the different types of attacks. Attacks on the system are displayed in a tree structure, where the roots are nodes of attack targets, and methods of attacking are the leaf nodes. Safety of large systems can be modelled with a set of attack trees, where the root of each tree is an attack that can significantly damage the system. In the outline view, a tree is shown with two types of nodes (logical AND or OR). Attacks tree provides a systematic description of security vulnerabilities, thus making it possible to assess the risks and make safety decisions. [9] [25]

B. Quantitative (probabilistic) estimation methods

The big disadvantage of the above-described qualitative estimation methods is assumption of stochastic independence between components in a system which has resulted in the inability of its applications in complex systems. [17]

For this purpose the quantitative (probabilistic) estimations methods are used. These methods are based on the state space e.g. Continuous Time Markov chain, Markov reward models, Petri nets, etc.

Markov chain is a discrete stochastic process that is commonly referred to as the state transition diagram. Markov "law" says that the next step depends only on the current situation, which can be represented by the following equation (1):

$$P(X_{n+1}|X_1, X_2, ..., X_n) = P(X_{n+1}|X_n) \quad (1)$$

Markov chains know the following space states: initial, transitional and final.

Continuous time Markov Chain (CTMC) is a mathematical model that allows the state change in any moment in time. Transition matrix (describe transitions between the state) expresses transition rates instead of probabilities. Description of the condition of continuous time Markov chain can be explicitly used to monitor the status of components and subsystems for observed system. Each state represents particular error state, while transition represents component failure rate. States show the number of failures of components over time. CTMC is time homogeneous process in the way that events such as failures and repairs are independent of each other. [10]

Discrete time Markov chain (DTMC) is a mathematical model that executes a change of state of the system after a specified time. The transition to the next state depends on the transition probabilities in the transition matrix. Each row of transition matrix represents an output from the state, and each column represents the input to the state. This model is used to predict the probabilities of hardware failure. [10]

Petri Nets is a graphical form of the formal logical description of the interaction between the components or the flow of activities in a complex system. The original Petri Nets have no time dimension. For the study of dependability is necessary to introduce the duration of events associated with transient conditions. Petri Nets can be extended with the transition duration which results in depicting temporal dimension. A special case of time Petri Nets are Stochastic Petri Nets (SPN) where the trigger time is considered random variables. SPN can be automatically converted to the basic Markov model. SPN in graphic terms consists of two types of nodes - places and transitions. Places typically represent conditions in the system while transitions represent events that cause changes in system conditions. Tokens are points associated with the place and represent the status of the place. Arcs connect places and transitions. The places from which an arc runs to a transition are called the input places of the transition. The places to which arcs run from a transition are called the output places of the transition. [16]

a) Markov chains in estimation of information system dependability

Markov chains are used to estimate the availability and reliability of complex hardware systems. This probabilistic method is applied also for modelling software reliability. Approach is suitable for predicting the reliability already at the design phase of the information system, even before a "black boxes" models and real components or software subsystems become available.

Markov models can serve as a basis for Markov Reward Models (MRM), which is used to estimate the reliability of the system. It allows "rewarding" the system for time spent in states that represents readiness of the system. [13]

b) Stochastic Petri Nets in estimation of information system dependability

Stochastic Petri Nets (SPM) provides a valuable tool for describing and analyzing the system. Since the beginning of its application, this method is used to solve problems in the estimations of reliability, availability, performance and analysis of software and hardware systems. [17]

In [28] is described Stochastic Activity Networks (SAN) that is based on Stochastic Petri Nets. This method is used to estimate reliability and availability. The main concept is activities (transition in Petri nets) that have impact to the system's ability to perform. Estimation is based on the analysis of which activities are enabled in a certain state of the system, and what is the next state of the system upon completion of activities.

IV. SOFTWARE FOR ESTIMATION OF INFORMATION SYSTEM DEPENDABILITY

There are several software packages for modelling dependability. Some of them are: SURF-2, GREAT-SPN, Ultra SAN, Möbius, Sharp, DrawNet ++, SPNP, Deem, TimeNeto, DSPNexpress, adviser, ARIES, CARE III, METFAC, SAVE, SURE, ASSIST, HARP, etc. Table 1 shows some software for modelling and estimation of dependability.

A. SURF-2

Surf 2 is a tool for estimation of hardware and software systems dependability. It is based on a strict construction, validation and numerical solving of Markov models. The software was developed in the laboratory LAAS, France 1996. The system behaviour is modelled with Markov chains and Stochastic Petri Nets. The main idea of the software is to have a simple method that compares the reliability of various system architectures. [24]

B. Great-SPN

GreatSPN is a tool that supports the design and the qualitative and quantitative analysis of Stochastic Petri Nets and Stochastic Well-Formed Networks (SWN). GreatSPN appeared in the late 80s of the last century. Since then, several different versions of this tool have been developed.

GreatSPN2.0 package consists of different software that collaborates in the construction and analysis of PN models. Using the capabilities of communication networks, it is possible to perform various analyzes on different machines in a distributed computing environment. GreatSPN2.0 modular structure enables the addition of new modules for the analysis, as well as new research results. All modules are written in the C programming language, which ensures portability and efficiency on various UNIX machines. [11], [22]

TABLE I. SOME SOFTWARE FOR ESTIMATION OF INFORMATION SYSTEM DEPENDABILITY [19]

Software name	Used methods	Designer
SURF-2	GSPN, Markov chain	LAAS, France
Great-SPN	GSPN, SWN	Torino, Italy
UltraSAN	SAN	UIUC, USA
Möbius	SAN, Markov chain …	UIUC, USA
SHARPE	Fault Tree, Markov chain, SAN…	DUKE,SAD
SPNP	SPN, SRPN, non Markov chain	DUKE,SAD
DRAWNET++	Fault Tree, SWN	U. del Piemonte orientale, U. Torino, U. Napoli, Italy
DEEM	SPN,	UNIFI-PISA, Italy
Time NET	non SPN	Hamburg, Germany
DSPNexpress	Deterministic and stochastic Petri nets	Dortmund, Germany

C. UltraSAN

UltraSAN is a software tool used for evaluation of the system based on the Stochastic Active Networks (SAN). SAN has the characteristics of Stochastic Petri Nets and queuing models. Using a variety of analytical and simulation modules, it is possible to determine following: efficiency, maintainability and feasibility. UltaSAN enables the graphical presentation of the results. To determine the valid model, it is necessary to specify a set of subnets using SAN editor and classify them according to the hierarchy. UltraSAN offers six analytical techniques for solving transient and steady state. [29]

Three techniques that allow solving the steady state are: direct steady-state, iterative steady state, and iterative deterministic steady state. To resolve transients state, transient instant of time method, PDF interval-of-time and expected interval-of-time techniques is used. [27]

D. Möbius

Möbius (Model-Based Environment for Validation of System Reliability, Availability, Performance and Security) is a software tool for modelling behaviour of complex systems. The tool was designed at the University of Illinois, USA, and was originally designed to study the reliability, availability and efficiency of computer and network systems. Its flexible approach allows engineers and scientists to present their systems in a language model appropriate to their problem area, and then accurately and efficiently solve systems using solving techniques that best suit the size and complexity of the problem. The tool supports Stochastic Petri Nets, Markov chains and Stochastic Process algebra. Models are presented numerically and graphically, and are made with the right level of detail, and have the possibility to adapt to the behaviour of the system of interest. This tool can build detailed mathematical expressions that measure the correct information about the system (e.g., reliability, availability, performance and security). Measurement can be carried out at certain points of time (some time period or time when the system reaches a steady state). [18]

E. SHARPE

SHARPE (Symbolic Hierarchical Automated Reliability and Performance Evaluator) is a tool that provides a specification and solution methods for the most commonly used type of estimation model for performance, reliability and feasibility. Models that are present in the SHARPE are Fault Tree, Model Queuing and State Model (Markov chains, semi Markov chain with reward and Stochastic Petri Nets). For each model SHARPE has preferred analytical algorithms. This tool allows measurement of models that can be used as a parameter to another model. Therefore it is considered to be a hierarchy oriented tool. The user interface supports command-line and graphical user interface that is made in Java. SHARPE is developed at Duke University, USA. [23]

V. CONCLUSION

The term dependability is relatively new and is based on the real needs of reliable and available information system that is easily maintained, has protected confidential information, is secure from external and internal threats and is safe for environment.

First estimation methods for information system dependability are developed for the needs of mission and safety critical systems in the space programs of NASA, ESA and in the military industry. These methods are now customized for use in economically-critical systems.

The fact that the failure of information system in economically-critical systems may result in significant costs, has led to a need to estimate dependability before information system is implemented. For this purpose, there is a space for development of a new methods and appropriate software for estimation of information system dependability in economically-critical systems.

REFERENCES

[1] A. K. Verma, S. Ajit, M. Kumar, Dependability of Networked Computer-based Systems. Springer, 2011,1; pp 3-5

[2] A. Avizienis, J.C. Laprie, B. Randell, Dependability and its Threats: A Taxonomy, 18th IFIP World Computer Congress, Toulouse 2004

[3] A. Bharadwaj, M. Keil, M. Mähring, Effects of information technology failures on the market value of firms. Journal of Strategic Information Systems. 18:66-79. 2009

[4] A. Avizienis, J.C. Laprie, B. Randell, C. Landwehr. (2004) Basic concepts and taxonomy of dependable and secure computing. IEEE Transaction Dependable and Secure Computing, January-March 2004 (vol. 1 no1.), pp. 11-33

[5] A. Avizienis, J.C. Laprie, B. Randell. Fundamental concepts of dependability. In: Proc. of 3rd Information Survivability Workshop, pp 7-11, October 2000

[6] B. Schneier. Attack trees: Modelling security threats. Dr. Dobbs journal. pp. 21-29. December 1999.

[7] C.B. Jones, B. Randell, The role of structure: a dependability perspective. In Structure for Dependability: Computer –Based Systems from an Interdisciplinary Perspective. pp. 3-15.Springer 2005

[8] C. A. Ericson II. Fault Tree Analysis-A History from the Proceedings of The 17th International System Safety Conference, 1999

[9] D. M. Nicol, W. H. Sanders, and S. K. Trivedi. Model-based evaluation: From dependability to security. IEEE Trans. on Dependability and Security, 1(1):48-65 2004.

[10] Dr. P. Tröger, Dependable Systems, State-Based Dependability Modeling, Dependable System Course 2013, pp 8-29

[11] GreatSPN, http://www.di.unito.it/~greatspn/index.html, 30.01.2014.

[12] IEC 61508: Functional safety of electric/electronic/programmable electronic safety-related systems, Parts 0-7; Oct. 1998-May (2000)

[13] I. Eusgeld, B. Fechner, F. Salfner, M. Walter, P. Limbourg and L. Zhang. Hardware Reliability. In Dependability Metrics. 9; pp. 59-103. Springer 2008

[14] A. Jakupović, Utjecaj oslonjivosti informacijskog sustava na poslovne organizacije, Zbornik Veleučilišta u Rijeci. 1 (2013) , 1; pp. 165-178

[15] J.C. Laprie, A.Avizienis, H.Kopetz. Dependability: Basic Concepts and Terminology, Springer-Verlag, 1995

[16] J. Happe, Analytical Performance Metrics. In Dependability Metrics, Springer 2008, pp214-218

[17] J. K. Muppala, R. M. Frics and K. S.Trivedi, Techniques for system dependability evaluation. In Computational Probability. pp 9-10. Springer, 2000

[18] Möbius, https://www.mobius.illinois.edu/, 30.01.2014.

[19] M. Kaâniche, K. Kanoun, J.C. Laprie, Dependability and Security evaluation of computer-based systems 2009

[20] NASA, Failure Modes, Effects and Criticality Analysis (FMECA), Practice no. PD-AP-1307. 1999

[21] R. Isermann, Fault-Diagnosis Systems: An Introduction from Fault Detection to Fault Tolerance, Springer 2006

[22] S. Baarir, M. Beccuti, D. Cerotti, M. De Pierro, S. Donatelli and G. Franceschinis,The GreatSPN Tool: Recent Enhancements, ACM Performance Evaluation Review Special Issue on Tools for Performance Evalaluation, Volume 36, Issue 4, September 2009

[23] SHARPE, http://sharpe.pratt.duke.edu/, 30.01.2014.

[24] SURF-2, http://homepages.laas.fr/surf4tst/what-uk.html#modele, 30.01.2014

[25] T. R. Ingoldsby. Attack Tree-based Threat Risk Analysis. Amenaza Technologies Limited. 2010, pp. 3-9

[26] U.S. Department of Defense. Procedures for Performing a Failure Mode, Effects and Criticality Analysis. MIL–P–1629, 1949

[27] W. D. Obal II, M. A. Qureshi, D. D. Deavours, W. H. Sanders, Overview of UltraSAN, in Computer Performance and Dependability Symposium, 1996., Proceedings of IEEE International, 4-6 Sep 1996

[28] W. H. Sanders and J. F. Meyer, „A Unified Approach for Specifying Measures of Performance, Dependability, and Performability", in Dependable Computing for Critical Applications, Vol 4: of Dependable Computing and Fault-Tolerant Systems (ed., A. Avizienis and J. Laprie), Springer-Verlag, 1991

[29] W.H. Sanders, W.D.Obal II, M.A.Qureshi and F.K. Widjanarko, UltraSAN Ver3: Architecture, Features, and Implementation, roceedings of the AIAA Computing in Aerospace 10 Conference, San Antonio, TX, March 28-30, 1995, pp. 327-338

[30] Wilson, STRATUS Computer System", in Resilient Computing Systems, 1985, pages 208-231

Internet of Things Cloud Mediator Platform

Kresimir Misura, Mario Zagar
University of Zagreb
Faculty of Electrical Engineering and Computing
Unska 3, HR-10000, Zagreb, Croatia
{kresimir.misura, mario.zagar}@fer.hr

***Abstract*—Internet of things (IoT) will connect a huge amount of devices to the Internet. There will also be many applications that will use the generated data to provide a service to end-user. This presents a problem in situations when the application owner is not the device owner. Application owner needs to find and contact the owners of the devices that provide the data he needs. When done by a human, this process lasts a long time and costs a lot of money. In this paper, we propose a middleware architecture that connects the appropriate devices and applications. It is based on software agents representing devices and applications negotiating between each other on the terms by which the data can be used.**

I. Introduction

Internet of things market can be divided in two basic areas. The device owners and the application owners. Although this distinction is still in its nascent stages, it will grow only more important in future. The device owners are people who go to the store and buy an Internet connected thermometer, pollution sensor, heart rate tracker or any other kind of Internet connected device. Application owners are the people who want to use data from these devices to create new applications. For example, a pharmaceutical company could want to build an application that tracks the average heart rate of the people in a city and publish it in real time on the web. This could be a part of their marketing efforts, they could advertise it as "The heart rate of Los Angeles".

There are a couple of reasons for creating an Internet of things cloud platform. One of them is to overcome the problem of each device speaking it's own language. That way, the cloud can speak all languages, and allow applications to access all of the devices using a unique interface.

Also, the cloud platform can aggregate the data, store it for later retrieval, perform complex operations on that data and enable device and data discovery. These problems are already being solved by the emerging internet of things cloud solutions and numerous startup companies, examples of which are given in section II.

However, there is another service the cloud platform could provide, and that is the mediation between application and device owners. In this context, mediation is used for finding the conditions of data provision that both device and application owners find acceptable. For example, a person could create an IoT application similar to the city heart rate monitor described above. But, for that to work, the application needs a lot of devices to provide it data. Deploying new devices for this specific purpose is not financially viable, and personally contacting many device owners to find the ones interested in exposing the data and then negotiating with them on the price, schedule and other conditions is tiresome and unscalable. Creating a mediation framework that connects the compatible devices and applications and performs the negotiation automatically would ease this process significantly, and could create many new economically viable use cases for the IoT. This paper presents one approach for implementing this kind of service.

The remainder of this paper is organized as follows. Section II examines the previous research done on this and similar subjects. A platform that is used to connect IoT applications with devices is presented in section III.

More details on the actual negotiation process between applications and devices are given in section IV. Section V proposes directions for future work, and concluding remarks are given in section VI.

II. Previous Research

This paper deals with a number of topics from various research areas - including IoT middleware platforms, software agents, agent negotiation protocols, semantics of IoT devices, etc. This section will give a brief overview of existing IoT platforms while the specifics of agent negotiation will be presented in a separate section.

There are a lot of startup companies creating IoT platforms enabling the cloud management of IoT data. Some of these platforms are: Axeda[1], Carriots[2], Evrythng[3], GroveStreams[4], Device Cloud by Etherios[5], Nimbits[6], One Platform by Exosite[7], Xively (Pachube)[8], SensorCloud[9], etc.

These platforms usually enable remote control, management and monitoring of IoT devices. They allow device owners to connect their devices to cloud and then do various operations with incoming data. Usually, they can make the data public so IoT application owners can use that data in their applications. However, they don't allow automatic negotiation on various aspects of the data the device provides to applications.

In [1] authors propose an agent based negotiation strategy for service provision in IoT domain. However, the purpose and

[1] http://www.axeda.com
[2] https://www.carriots.com/
[3] http://www.evrythng.com/
[4] https://grovestreams.com/
[5] http://www.etherios.com/products/devicecloud/
[6] http://www.nimbits.com/
[7] http://exosite.com/products/onep
[8] https://xively.com/
[9] http://www.sensorcloud.com/

domain area are a bit different than in our case. The authors are mostly concerned with provisioning optimized services to mobile devices users. Agents and negotiation are used to overcome the possible semantic communication differences between various devices. Also, there are two other differences. The service provider and service consumer agent in [1] share a common goal, while in our problem, goals are often conflicting. The other difference is that in [1] the negotiation is of a bilateral type, where service consumer negotiates directly with the service provider in order to achieve the optimal service. In the case of IoT mediator, the application agent is negotiating with many device agents at the same time, and the outcome of each one of those negotiations can impact the other.

Global Sensor Network (GSN) [2] is a platform which provides a scalable infrastructure for integrating heterogeneous sensor network technologies using a small set of powerful abstractions. It supports the integration and discovery of sensor networks and sensor data, provides distributed querying, filtering, and combination of sensor data, and supports the dynamic adaption of the system configuration during operation through a declarative XML-based language. The sensor services are published through a peer-to-peer directory and searched based on their properties. Applications can discover sensor networks using the registry and subsequently access sensor networks by using a standard data access interface. However, in GSN it is assumed that sensor networks that produce data share it freely with all interested members of the network. This is different than IoT mediator, where the applications need to negotiate with sensor services on the conditions under which the sensors will provide data.

Hourglass [3] is an Internet-based infrastructure for connecting a wide range of sensors, services, and applications in a robust fashion. The infrastructure consists of an overlay network of well-connected dedicated machines that provide service registration, discovery, and routing of data streams from sensors to client applications. In addition, Hourglass supports a set of in-network services such as filtering, aggregation, compression, and buffering stream data between source and destination. When a service (sensor) announces itself to the Hourglass system, it affiliates with one or more topics and further describes itself using a set of predicates. These predicates are used to match services against application queries. More than one service may match the set of predicates and topics. In this case, it is up to the Hourglass system to decide which service to use. It makes this decision based on a pre-scripted behavior (e.g. trying to accomplish load balancing, pick a geographically close server, etc.), and unlike IoT mediator, there is no negotiation going on between the applications and devices (services) on which device is willing to offer the best usage terms.

III. IoT Mediator Platform

A prototype platform was created to solve the problem presented in the introductory section. Applications and devices can communicate with the platform through an HTTP interface. The basic use case is when the application asks for the appropriate devices, the platform returns a list of contracts with the specific devices from which the application can get the data it needs.

Platform architecture is presented in figure 1. The platform exposes an HTTP interface that applications and devices can use for registration, device discovery, contract making and data retrieval.

When a device is registered, all of the important details such as the type of data it provides, or the negotiation strategy used by the device agents are specified in an xml file. An example of this xml is given in figure 2. Negotiation strategy selection is particularly important. Owners can select from a list of template strategies, and change their configuration to best suit their needs. In the example in figure 2, a minprice strategy is used. This means that the device agent will provide the data only if the price is higher than the specified limit.

Once a device is registered, the platform saves the details in the database for later use. When an application requests the discovery of devices that match it's criteria, a negotiation module is utilized. This module consists of two submodules: Preselection and JADE Container. Preselection module first selects all of the devices that could potentially provide the data that the application needs. The purpose of this step is to speed up the actual negotiation process. Preselection is done to eliminate all of the devices that are outside the desired location, measure different properties than requested, or can be excluded from the negotiation process for any other reason. After the preselection step, there is a small number of potential devices remaining, and it is now only needed to select the ones that can give the best terms to the requesting application. This is done using autonomous agents. A JADE container is created and populated with a single agent for every device and application. This agents then negotiate to determine which devices will provide the data and at what conditions.

Typically, a separate container is created for each negotiation. Thus, there will usually be only one application agent and multiple device agents per container. Agents use negotiation algorithms and other constraints that were specified at the time that applications/devices were registered. Application agents have goals such as "find as much consistent data as possible given a limited budget" or "find all of the needed data using an unlimited budget". Device agents have goals such as "conserve as much device battery as possible", "achieve the highest possible price for the data", etc.

Applications search for the devices that are able to provide the appropriate Schedule, Location and Data descriptions.

Schedule is a single point in time, or a recurring event. The application agent can leave a schedule field blank, specify a time range, a list of possible schedules, or require an exact schedule.

Location is a geographical area containing all acceptable places for deployed devices. This can be irrelevant for some applications, in which case the agent would simply leave this field blank.

Data description is a description of data that the device should provide. This is usually not open to negotiation since most devices have a predefined type of data they provide. Although this field is named "data description", it contains everything related to sensors, their parameters and the properties they measure.

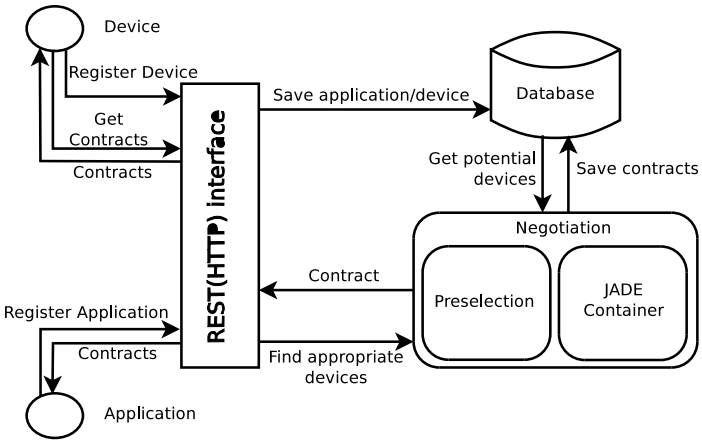


Fig. 1. Architecture overview

```
<device>
    <name>kitchen_thermometer</name>
    <type>HUMIDITY</type>
    <location>
        <latitude>45.8167</latitude>
        <longitude>15.9833</longitude>
        <elevation>1.2</elevation>
    </location>
    <algorithm type="minPrice">
        <minPrice>22.85</minPrice>
    </algorithm>
</device>
```

Fig. 2. Device registration xml example

Agents negotiate on several properties: data pricing, reading schedule, and in some cases, the reading locations (for devices that are mobile). Once the negotiation is finished, if both application and device agent are satisfied, the platform creates a contract between each device and application. The contract defines at what times will the device provide data, what is the price, and all the other relevant parameters. A copy of the contract is sent to both the device and the application. Once the contract is made, the device is obliged to provide the agreed data to the application respecting the contract terms. Application (i.e. application owner) is obliged to pay the agreed compensation for the provided data.

An example of a contract is given in figure 3.

Once contract is made, devices (i.e. device owners) are obliged to send the data to the platform at the times that were specified in the contract. Applications can use the HTTP interface to retrieve the generated data. Although it would be possible for devices to send data directly to applications, platform proxy is necessary so that the mediator platform would be able to determine if some applications or devices are not holding to their end of the contract.

IV. Agent Negotiation

Autonomous agents are a well researched area and are well suited for the problem of negotiation [4]–[6]. An architecture overview, describing how autonomous agents are used to reach

```
<contract>
   <application-id>126</application-id>
   <device-id>1121</device-id>
   <location type="exact-location">
      <coordinates>
         <lat>45.8167</lat>
         <lon>15.9833</lon>
      </coordinates>
   </location>
   <schedule>
      <repeating>
         <start>1376863200000</start>
         <period>3600000</period>
         <repeat-count>20</repeat-count>
      </repeating>
   </schedule>
   <data-type>temperature</data-type>
   <price type="per-measurement">
      <amount>0.5</amount>
      <currency>USD</currency>
   </price>
</contract>
```

Fig. 3. Contract example

agreements between applications and devices in IoT mediator platform, is given in figure 4.

By considering the complexity of negotiation environments, agent negotiation can be classified into three levels, which are the Bilateral Negotiation Level, the Multilateral Negotiation Level, and the Multiple Related Negotiation Level [7]. Application-device negotiation is an instance of Multiple Related Negotiation Level so it will be described further. This level represents a situation in which an agent needs to perform several negotiations in order to reach a global goal, and relationships between these negotiations are somehow related [7].

In device-application negotiation, the agents need to perform more than one negotiation with different opponents for different goals at the same time. These goals are often not independent, and these multiple negotiations are somehow related. For instance, a device agent could be involved in negotiations with multiple application agents, but the actual device might not have enough battery power to satisfy all of the applications. Or, a device might be interested in serving all of the requests at the same point in time, because in that case it needs to activate less often and thus can conserve battery power.

Most existing approaches to agent negotiation [8]–[10] separate the related negotiations and treat each of them individually. IoT mediator platform also uses this technique to simplify the overall negotiation process. The exact algorithm is described in the rest of this section, and a diagram can be found in figure 5.

The application has a list of measurement requirements. Each of these represents a separate requirement, for example, in an application that correlates current weather data to citizens heart rates, one measurement requirement is to get the weather data, and the other is to get the current average citizens heart

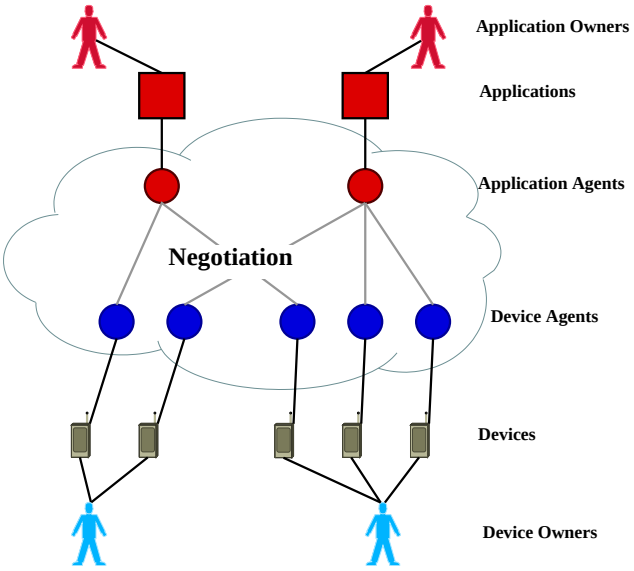


Fig. 4. Negotiation agents in M2M communication

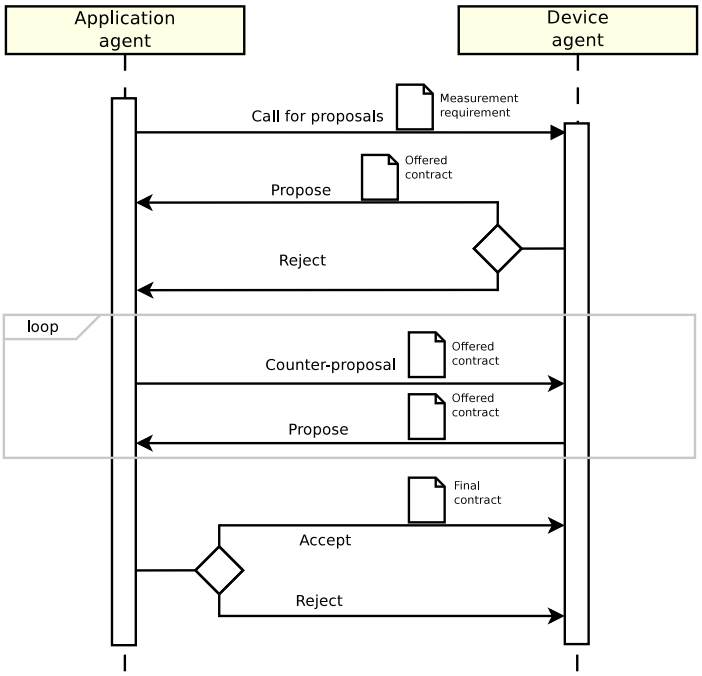


Fig. 5. Negotiation protocol

rate.

Negotiation for each of these measurement requirements is conducted separately, and if successful, temporary agreements are made. Once all of the measurement requirements are satisfied, temporary agreements are turned into obligatory contracts.

First, the application agent sends a call for proposal to all agents providing the appropriate data. This message contains the range of values that the application agent finds acceptable. For example, it contains the geographical area, a list of possible timestamps when the device should provide the data, acceptable price ranges, etc.

The devices respond with offered contracts. Each response contains the exact timings, locations, and prices at which the device is prepared to provide the data. At this point, the application can accept the proposal, reject it, or make a counter-proposal. This process can repeat until the device and application agents agree, or find out that the deal isn't possible.

Once the application has made all of the temporary agreements it needs to satisfy all of it's measurement requirements, the temporary agreements are turned into obligatory contracts. If the device has made temporary agreements for some measurement requirements, but isn't able to fulfill other requirements, it can cancel the temporary agreements. This situation could happen, for example, in an application that correlates average heart rate of citizens with current weather data. If this application creates temporary agreements for weather data, but is unable to find the appropriate providers of heart rate data, it can cancel the agreement on weather data because this data alone holds no value to the application.

The actual agent negotiation module was implemented as a multi-agent system using JADE (*Java Agent Development Framework*) agent platform [11]. Agents communicate by exchanging ACL (*Agent Communication Language*) messages [12].

V. Future Research

The data description is currently given as a simple enumeration of all allowed data types. So, a device owner selects if his device measures temperature, humidity, heart rate, etc. This could be refined so that data could be described using a standardized ontology, most convenient of which is the one created by the W3C Semantic Sensor Network Group [13].

Recognizing devices and applications that are not holding to their contract obligations would be very important in a real-world scenario. There are some existing approaches based on Web Service Level Agreements (WSLA) [14], but a more lightweight approach based on rating the credibility of the device/application owners might be more suitable. A similar mechanism is successfully implemented in online marketplaces such as ebay. Thus, application agents will favor devices with high credibility and vice versa.

Currently, application and device owners define all of the details about their devices/applications in an xml file. This xml is then sent to the platform when the device/application is registered. This is not suitable for technically inexperienced users. Thus, it would be useful to create a web based application that would give the users an intuitive graphical interface for registering devices and applications.

Platform distribution across multiple servers is also an interesting problem, and would become even more significant if the number of connected devices increases. How would the described mechanism of agent negotiation function on a distributed platform is an open question.

Currently, the mediator platform focuses mostly on sensors, but in the future it could be used just as well to accommodate actuators. Addition of actuator support would open up a whole range of new possibilities. One example of an interesting

actuator device is the electronic billboard that could connect and negotiate with different companies about showing their ads. They could negotiate on the times when the ad will be shown, pricing, etc. Interesting use cases could emerge, for example, companies will be willing to pay more if there are a lot of people standing in front of the ad, ice-cream companies would be prepared to pay more to show their ad on a hot day, etc.

VI. Conclusion

Creating a framework for automatically connecting applications with appropriate devices creates new use cases for the IoT. It provides a financial incentive for the end users to buy and connect more devices with the final goal of making a profit by the data these devices provide. It also creates possibilities for the application owners to get the necessary data without the need to purchase new devices, but instead to reuse the already available data. This could hopefully help create a feedback loop where more devices make the applications more interesting which in turn gives the financial reason for the device owners to buy even more devices.

The prototype implementation described in this article provides one way of implementing this service. It uses autonomous agents to implement the actual negotiation between devices and applications, which allows us to use a wealth of research done on agent negotiation. It also uses standard protocols and simple data descriptions to make the usage of platform as easy as possible.

Acknowledgment

This work is supported by the Croatian Ministry of Science, Education and Sport, under the research project ZP0361965 Software Engineering in Ubiquitous Computing.

References

[1] P. Jeon, J. Kim, S. Lee, C. Lee, and D.-K. Baik, "Semantic negotiation-based service framework in an m2m environment," in *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2011 IEEE/WIC/ACM International Conference on*, vol. 2, 2011, pp. 337–340.

[2] K. Aberer, M. Hauswirth, and A. Salehi, "The global sensor networks middleware for efficient and flexible deployment and interconnection of sensor networks," *Ecole Polytechnique Fdrale de Lausanne (EPFL), Tech. Rep. LSIR-REPORT-2006-006*, 2006.

[3] J. Shneidman, P. Pietzuch, J. Ledlie, M. Roussopoulos, M. Seltzer, and M. Welsh, "Hourglass: An infrastructure for connecting sensor networks and applications," Harvard Technical Report TR-21-04, Tech. Rep., 2004.

[4] P. Huang and K. Sycara, "A computational model for online agent negotiation," in *System Sciences, 2002. HICSS. Proceedings of the 35th Annual Hawaii International Conference on*, 2002, pp. 438–444.

[5] S. Kraus, *Strategic Negotiation in Multiagent Enviroments*. The MIT press, 2001.

[6] S. Parsons, C. Sierra, and N. Jennings, "Agents that reason and negotiate by arguing," *Journal of Logic and computation*, vol. 8, no. 3, pp. 261–292, 1998.

[7] F. Ren, "Autonomous agent negotiation strategies in complex environments," Ph.D. dissertation, School of Computer Science and Software Engineering, Faculty of Engineering, University of Wollongong, 2010.

[8] M. He, A. Rogers, X. Luo, and N. R. Jennings, "Designing a successful trading agent for supply chain management," in *Proceedings of the fifth international joint conference on Autonomous agents and multiagent systems*. ACM, 2006, pp. 1159–1166.

[9] S. Proper and P. Tadepalli, "Solving multiagent assignment markov decision processes," in *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*. International Foundation for Autonomous Agents and Multiagent Systems, 2009, pp. 681–688.

[10] X. Zhang and V. Lesser, "Meta-level coordination for solving negotiation chains in semi-cooperative multi-agent systems," in *Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems*. ACM, 2007, p. 15.

[11] F. B. G. Caire, A. Poggi, and G. Rimassa, "Jade. a white paper," 2003.

[12] A. Fipa, "Fipa acl message structure specification," *Foundation for Intelligent Physical Agents, http://www.fipa.org/specs/fipa00061/SC00061G.html (30.6. 2004)*, 2002.

[13] L. Lefort, C. Henson, K. Taylor, P. Barnaghi, M. Compton, O. Corcho, R. Garcia-Castro, J. Graybeal, A. Herzog, K. Janowicz *et al.*, "Semantic sensor network xg final report," *W3C Incubator Group Report*, vol. 28, 2011.

[14] A. Keller and H. Ludwig, "The wsla framework: Specifying and monitoring service level agreements for web services," *Journal of Network and Systems Management*, vol. 11, no. 1, pp. 57–81, 2003.

Second Screen User Experience: A New Digital Television Frontier

Nenad Šoškić*, Nikola Kuzmanović **, Milan Vidaković* and Goran Miljković*
* RT-RK Computer Based Systems, Novi Sad, Serbia
** Faculty of Technical Sciences, University of Novi Sad, Novi Sad, Serbia
nenad.soskic@rt-rk.com

Abstract - Availability and increased hardware performance of handheld devices with high resolution screens encourage users to exploit them as the second screen while watching television. Second screen is a term that refers to an additional screen of a handheld device such as a tablet or a smart phone, with an operating system capable of running desired installed applications. Desire to make additional DTV information available to the user which can be presented on the second screen tends to push the limit of digital television in that direction. This paper presents a framework to distribute and present both premium and free-to-air television content. The framework is based on server/client architecture and includes: TV remote control, streaming of television services, additional content (Electronic Program Guide, reminders...), data protection of streamed services to client devices. The system architecture consists of a media server and its clients. Media server is a Set-Top box with Android OS. Main duty of media server is to receive and process DTV content received from broadcasters and distribute the content to clients (second screen devices). Client application is developed on various OS/Platforms: Android and iOS. Multimedia content can be streamed using supported protocols: HTTP, RTP/RTSP, and adaptive streaming protocols (HLS and MPEG-DASH). Streamed content is protected using DTCP-IP. The main objective of the paper is to describe the utilization of before mentioned standards in a framework to use the DTV data and functions beyond the scope of the main TV screen. The client application also implements the remote controller GUI in order to use one device to control the TV.

I. INTRODUCTION

During people's leisure time, television (TV) is primary media source. With the expansion of new technologies (social services, handheld devices) and their integration with the TV, watching TV is drastically changed. Connection through the internet and using social networks allows people to be more connected than ever. Participation and involvement with media is possible through various platforms [1]. TV is easier to access over different devices, and usage of additional devices during the TV watching has become an usual habit [2]. Even media consumption which is relevant to the watched TV service has changed [3]. New technologies allow for personalized and more flexible media consumption [4]. Forms of media and associated devices are widespread in domestic settings. They alter the types of entertainment and information-centric activities that people engage within the domestic media environment, with multiple media appliances in their homes (laptops, smartphones, tablets) [5]. These media devices can be used as a second screen in the second screen TV environment.

Second screen allows television audience to interact with the content they are consuming such as TV shows, music or video games. While watching the TV, data which would have not been shown otherwise (teletext, electronic program guide (EPG), etc.) can be shown on the second screen. Without any interference with other users which are viewing the TV on the main screen, applications for second screen can integrate interactive television (iTV) services and TV-centric applications [6] on the second screen. In the research presented in [7] it has been concluded that TV watchers frequently use their laptop to surf the web, check their e-mail or shop online while watching the TV. In that same research it has been said that there is a priority to have access to different services on multiple screens, rather than putting all the information on one shared screen.

Remote controllers which are more flexible and have more capabilities will help viewers to become more active in the iTV environment, changing the traditional way of watching the TV, as it is highlighted in the studies [6,7,8]. In [8] it has been shown that remote controllers are often considered less useful. Furthermore, results indicate that flexibility of portable devices can be used as remote controllers and to offer more advanced features. From [6] it has been concluded that more sophisticated form of control is needed to be presented for iTV to gain its full potential.

The idea for using a second screen in the television environment is not new. One of the first systems where handheld devices were used for interacting with the television is presented in [9]. Research study [10] shows that staggering 77% of TV viewers is using an additional screen while watching TV. Such results undoubtedly imply that future development of DTV will promote second screen environment even further. In order to improve second screens' functionality majority of features are transferred from set-top-box (STB) to second screen device [11]. A widely accepted way of connecting different devices in home network is through the use of Digital Living Network Alliance (DLNA) protocol.

The most important feature of second screen concept is the security of the audio and video content. Since the signal is Digital Video Broadcasting (DVB) signal,

regulated by DVB standards, therefore security is provided by Conditional Access Module (CAM) which uses Conditional Access techniques presented in [12,13,14]. As DVB signal can be broadcasted through various mediums, different DVB standards are applied. We can distinguish two different phases in content protection: acquisition and post-acquisition as it was stated in [15]. While the content security during the acquisition phase is regulated by DVB standards through CAM, it is up to developer to provide content protection in post-acquisition phase in order to stream premium content to second screen [16].

This paper presents a framework to distribute and present both premium and free-to-air television content. The framework is based on server/client architecture and includes: TV remote control, streaming of television services, additional content (Electronic Program Guide, reminders...), data protection of streamed services to client devices, and various second screen concepts.

II. System Architecture

A second screen environment implies that there are more than one visualization devices that are served with content by a primary (source) device. Therefore, server-client architecture is necessary. It is presented in fig. 1.



Figure 1. Server/Client architecture

Because of the recent development, tablets represent a convenient solution for target device on client side. Server receives broadcasted DVB content and passes it through to clients when all security conditions have been met. Server is a set-top box device based on Android OS [17] that has a DTV service server and provides all of the content for its clients. Clients are second screen devices in form of a tablet or smartphone.

In our system CAM is an integral part of host, as it is in fact a piece of hardware which is usually integrated in STB. In order to control the CAM, Common Interface Plus (CI+) stack is used. Streamed content between server and client is protected using Digital Transmission Content Protection (DTCP-IP) when it is needed. DTCP-IP was chosen because it is widely accepted and it is part of DLNA protocol that is supported by majority of second screen devices. Moreover, CI standard encourages the use of DTCP-IP for protection of streamed content.

III. Second Screen TV and Mulitmedia Playback

Live TV can be watched on the second screen device independently from the main screen. Channels watched on the second and the main can be different or duplicated (same channel on both screens). On the second screen, while watching live TV, user can switch channels, control volume, view services provided by the watched channel and all of that without interfering with the main screen. Some controls like channel zapping and volume control are possible with finger gestures. For example, swiping left or right zaps to the next or the previous channel.

In order to watch the live TV on the second screen, player had to be implemented. That player supports both premium and free-to-air TV content. TV channels are streamed from the server for the clients. The solution for problem of content protection for premium content is presented in fig. 2.

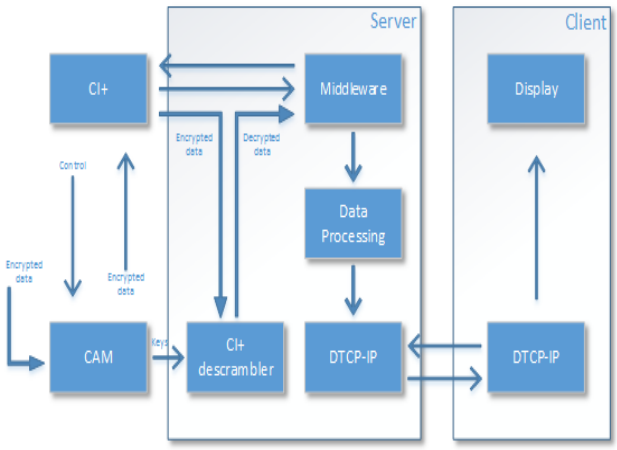


Figure 2. Content Protection

This solution provides the user with a significant amount of freedom regarding the format of streamed content, due to the fact that all the processing and descrambling is done on the server side. After the data is descrambled, all wanted alterations can be done regarding the format of the content. The codec of audio and video data could be changed in order to make it suitable for particular second screen device, for example. That allows for greater system flexibility. This solution is compliant with practically all second screen devices, as the data format can always be changed on the host side. However, this solution requires more processing on host side that uses host's bandwidth. As a consequence a powerful hardware is needed.

Multimedia content is streamed over HTTP, RTP/RTSP and adaptive streaming protocols such as HLS or MPEG-DASH. All of the videos or music on the server could be streamed to the second screen device. User can browse through the set-top-box's music and video library and play the content he or she desires.

IV. Additional Content

Multiple examples of additional content are presented in this section. Some of the use cases can be seen in figure 3.

A. Electronic Program Guide

For any channel available, the user can view EPG on the second screen. User can select and preview info about any channel's current, past or future event while browsing EPG. From the EPG presentation channels can be changed or added to the favourites list. Details on any EPG event can be viewed by selecting desired event. Selected EPG event and EPG in the background can be seen in fig. 4.



Figure 3. Second Screen use cases

Because it does not have to be the channel currently viewed on the main screen, it enables personal EPG browsing without interfering with other TV watchers or with the main screen of the DTV. EPG can be activated from the Remote Controller for the main screen, or from the mosaic and live TV on the second screen. From the EPG user can also set the reminders on both main and the second screen for pop-up reminder to remind user to watch some channel later. The reminder will pop on both main screen and the second screen when the show is about to start.



Figure 4. EPG with deatiled view of an EPG event

B. *DLNA Multimedia Browser*

From their second screen device, users can browse DLNA servers using the DLNA browser implemented in the second screen application. If user chooses, browsed media content (pictures, audio and video) can be reproduced on the second screen. Also, multimedia content on the TV set can be browsed and reproduced including videos recorded with PVR (Personal Video Control). The user can choose whether to reproduce multimedia content on the second screen or main screen.

C. *Personal Video Control*

It is possible to browse and watch previously recorded content on the second screen or previously recorded content on media server and control it. If the recording is activated on the second screen device, the user can designate amount of memory which can be used for storing the data. Controlling the content includes play, pause, stop, rewind and fast forward functions. These controls can be used to time shift live TV channels. Personal Video Control (PVR) is done via remote controller module of the second screen application.



Figure 5. Remote Controller modules

D. *Remote Controller*

Second screen application offers remote controller UI and can act as a secondary remote controller for the set-top box. The idea is that remote controller on the second screen becomes a lot more flexible as it is a software application and can be used to control more than the TV set. Remote controller is used simultaneously with classic remote controller which comes with the TV, but considering that all of the other options are present on the second screen, software remote controller is accepted with ease. By using the Controller unit, it is possible to use the application on Android powered set-top boxes. Remote Controller consists of three modules: Keyboard, Touchpad and Remote Controller which can be seen in figure 5. Touchpad and keyboard modules are used for browsing content on the Android part of the set-top box. Remote Controller is used in a classic way of interacting with the TV and it consists of three modules: Navigation, Numbers and Classic Remote Controller.

V. CONLUSION

In this paper second screen framework for live TV is presented. The framework includes additional data such as EPG, reminders, PVR control. Second screen live TV which supports both premium and free-to-air TV is the key part of the solution for the problem of content protection for second screens. Future work is planned on

expanding additional second screen features and fully implementing SAT>IP and ChromeCast in the second screen environment. Also, connecting second screen application with social networks in order to connect second screen users is planned thus making this second screen environment even more user friendly.

ACKNOWLEDGMENT

This work was partially supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia under Grant 32031.

REFERENCES

[1] S. Livingstone, "The challenge of changing audience: or, what is the audience researcher to do in the age if the internet?", European Journal of communication, 2004, 19(75), pp.75-86.

[2] M. Flávio, P. Filipa, M. João, "SentiTVchat: sensing the mood of social-tv vievers", Procedings of the 10th European conference on Interactive tv and video, 2012, pp.161-164.

[3] J. Hess, B. Ley, C. Ogonowski, L. Wan, V. Wulf, "Understanding and supporting cross platform usage in the living room", ACM 2012.

[4] E. Tsekleves, R. Whitman, K. Kondo, A. Hill, "Bring the television experience to other media in the home: an ethnographic study", ACM 2009.

[5] E. D'heer, C. Courtois, S. Paulussen, "Everyday life in (front of) the screen the consumption of multiple screen technologies in the living room" ACM 2013.

[6] L. Cruickshank, E. Tsekleves, R. Whitham, A. Hill, K. Kondo, "Making Interactive TV Easier to Use: Interface Design for a Second Screen Approach", The Design Journal, vol. 10, Number 3, November 2007, pp.41-53.

[7] W. Seager, H. Knoche, M. Sasse, "TV-centricity—requirements gathering for triple play services", Adjunct Proceedings of the European Conference on Interactive Television, pp. 274–278, 2007.

[8] R. Bernhaupt, M. Obrist, A. Weiss, E. Beck, M. Tscheligi, "Trends in the living room and beyond", Proceedings of the European Conference on Interactive Television, pp.146–155, 2007.

[9] Robertson, S., Wharton, C., Ashworth, C., Franzke, M., "Dual device user interface design: PDAs and interactive television", CHI '96: Proceedings of the SIGCHI conference on Human factors in computing systems, pp.79-86, 1996.

[10] Google, "The New Multi-screen World: Understanding Cross-platform Consumer Behaviour", online: http://services.google.com/fh/files/misc/multiscreenworld_final.pdf, August 2012.

[11] N. Soskic, D. Popov Tapavicki, M. Subotic, N. Kuzmanovic, M. Savic, "A proposal for second screen DTV and remote control application", ICCE Berlin, pp.391-393, 2013.

[12] ETR 289: "Digital Video Broadcasting (DVB); Support for use of scrambling and Conditional Access (CA) within digital broadcasting systems", European Telecommunications Standards Institute, October 1996.

[13] EN 50221: "Common Interface Specification for Conditional Access and other Digital Video Broadcasting Decoder Applications", CENELEC, CENELEC TC 206, February 1997.

[14] TS 101 699 V1.1.1: "Digital Video Broadcasting (DVB); Extensions to the Common Interface Specification". European Telecommunications Standards Institute, November 1999.

[15] Diaz-Sanchez, D.; Marin, A.; Almenarez, F.; Cortes, A., "Sharing Conditional Access Modules through the Home Network for Pay TV Access", Consumer Electronics, IEEE Transactions on (Volume:55 , Issue: 1), February 2009, pp 88-96.

[16] Sakamoto, N.; Muguruma, K. ; Koshino, N. ; Chiba, S. ; Sakurai, M., "A digital HDTV receiver with home networking function and digital content storage", Consumer Electronics, IEEE Transactions on (Volume:51 , Issue: 3), September 2005, pp 831-835.

[17] M. Vidakovic, N. Teslic, T. Maruna, and V. Mihic, "Android4TV: a proposal for integration of DTV in Android devices", *30th IEEE* International Conference on Consumer Electronics (ICCE), January 2012, pp. 441-442.

Device Cloud Platform with Script Based Agents for "Anywhere Access" Applications Development

M. Z. Bjelica*, N. Ignjatov**, I. Papp** and N. Teslic*
* University of Novi Sad, Faculty of Technical Sciences, Novi Sad, Serbia
** RT-RK Institute for Computer Based Systems, Novi Sad, Serbia
milan.bjelica@rt-rk.com

Abstract – Providing access to devices via Internet is a growing trend allowing consumers to monitor and control their home appliances, heating, air conditioning and multimedia devices remotely from their tablets or laptops. Increasing number of "anywhere access" applications emerged to provide the users with cutting edge monitoring and control possibilities, including scheduling, event linking or even in-depth behavior scenarios definition. In this paper a device cloud platform architecture and a pilot implementation is presented. The device cloud allows arbitrary web, mobile or PC applications to connect to devices over the internet, to monitor and control their behavior and upgrade their firmware. It is also allowed to compose scripts and define cross-device and cross-event scenarios, which are maintained and executed within the cloud, thus allowing unprecedented level of automation and intelligent control. Main contributions of the presented device cloud are the interoperability with devices based on open technologies (TR-069), dynamic adaptation to device types, dynamic API layer towards user applications, intelligent bandwidth preservation and modular and scalable in-cloud scripting engine.

I. INTRODUCTION

With the increase of number of "smart" home devices, and with the increase of internet bandwidth and ubiquity of internet access, the need to control this ecosystem seamlessly has appeared. The establishment of a platform for device-user interaction regardless of user location would allow the development of "anywhere access" applications. Open technologies for application implementation (e.g. HTML-5 with Javascript), together with the ubiquity of the devices which are able to show applications (PCs, tablets, smart phones), would allow the clear leverage with the seamless API implementation provided by the Device Cloud platform. This platform builds upon the concept of Internet of Things (IoT), which envisages the possibility of unique identification and access to each device over the Internet. In recent years there is an increase in Device Cloud platform providers, which offer means of connectivity between devices, cloud and user applications.

In this paper a concept, architecture, implementation and demonstration use cases of a novel Device Cloud platform, called *Insight,* are given. The proposed device cloud tackles some of the most daunting problems in this area. Its architecture is adaptive so that it can accommodate any type of device with any data model specification, which maps directly to the cloud API and is accessible by apps in a semantically expectable manner. The connection of the device and the cloud is facilitated by the provision of cloud agents in various convenient forms, such as Android service, Linux daemon or pure-C software library which use IP connection for cloud access using open technologies, such as TR-069. Alternatively, low-performance devices can be connected to the cloud through an easy-to-install gateway supporting Zigbee and ZWave protocols. The device shadow continues to exist in the cloud even in the case the device is temporarily offline. Support for load balancing is designed in, so that all the device communication can be throttled if necessary and thus preserve bandwidth.

II. RELATED WORK

Since it was first coined in 1999, term Internet of Things was a promising paradigm that would resolve many problems of protocol convergence and ubiquitous access to all physical entities in the world [1]. The problem focus has recently shifted from wireless sensing network architectures (WSN) and device accessibility from the global network, towards the creation of cloud-centric frameworks for IoT usage and management. Gubbi et al. recently proposed a conceptual cloud framework for IoT access implemented as Aneka, which allows the creation of Sensor Applications which dynamically assigns computational resources and provides sensing data from the ubiquitous sensors [2]. However, this framework is sensing-centric, disregarding aspects such as configuration and maintainability of physical device ecosystem. The proposal of common application layer with thin server architecture pushing the operation logic and device accessibility to the cloud, was given by Kovatsch et al. [3]. However, this solution does not address security concerns, nor does it provide access to device nodes which may not be operational when requested by an application.

Commercial platforms related to the proposed device cloud are growing in number. Etherios (formerly iDigi) provides a hosted device cloud solution which enables apps creation which can control devices in real time, schedule operations and configure alarms [4]. Each device is exposed to the cloud through Cloud Connector libraries for Android or embedded systems. Many startups are also offering less mature M2M interfaces and cloud services such as Arkessa [5], Axeda [6], Carriots [7], Xively [8], Control4 Anywhere Access [9], etc. To the best of our knowledge, these solutions provide limited or no support for device identification, device model adaptability and device "shadow" image, which are issues handled in the proposed device cloud.

Motorola EDGE Manager [11] supports monitoring and configuration script written in JavaScript execution, however automatic script execution is allowed only for simple events in the system, where more complex events require operator management. This solution uses TR-069 Connection Request mechanism which does not exploit all TR-069 possibilities, especially data synchronization between end devices and monitoring platform.

III. DEVICE CLOUD ARCHITECTURE

Top level architecture of the proposed device cloud - Insight Device Cloud (IDC), is shown in **Figure 1**.

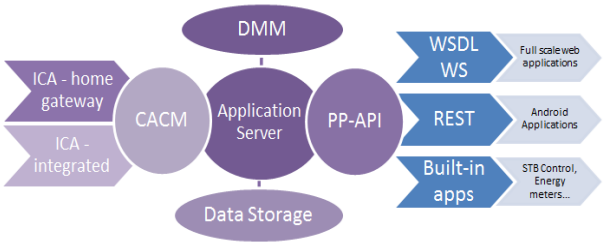


Figure 1. Top level architecture.

The *Application Server* is a scalable core of the IDC, providing the runtime environment for the *Data Manipulation Module* (DMM) and provides access to *Data Storage*. The *Cloud Agents Communication Module* (CACM) maintains the communication with the physical devices, supporting various device types and parameter sets through the concept of XML-based data model and dynamic API. *Insight Cloud Agent* (ICA) implements the device side of communication with IDC, available as Android service, Linux deamon or C library. The communication between CACM and ICA is based on TR-069 protocol. DMM controls end-to-end information flow while load-balancing access to storage units (database/cache). IDC exposes *Post-Processing APIs* (PP-API) for third party application development using standardized communication web-based interfaces such as web services based on web service description language (WSDL) or REST.

Several novel approaches are within the IDC architecture and functionality. Each device is represented by its *data model*, which is essentially a tree describing all device parameters, command hooks, objects etc (**Figure 2**). ICA adapts to this data model both in compile-time and run-time, making the device semantically close to the target application use cases. Based on the data model obtained by CACM, DMM dynamically adapts PP-API using WSDL to expose the API as close as possible to the device type and its functionalities.

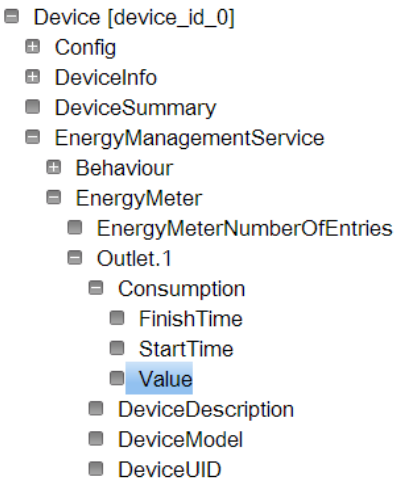


Figure 2 Example of data model specification

PP-API uses *security mechanisms* and reputation-based trust management to secure public domain data flow. Security in Cloud environment is a big issue, so messages on both ends are secured. Communication layers, towards ICA and third-party applications integrate HTTPS/SSL as the industry standard. WSDL based web-services endorse Web Service-Secure (WSS) protocol for message encryption. Federated Identity Management (FIdM) is integrated with Pluggable Authentication Module (PAM). This mechanism enables different high level application controllers with different access privileges. Besides specific security, IDC is applying Cloud-specific security through mechanisms through corrective and detective controls. *Load balancing* is provided by measuring the average duration of communication session with each device. If the threshold duration is reached, load is balanced by adjusting the interval between two consecutive Inform messages. Inform is a basic inter-entity message in TR-069 protocol. To trigger Inform asynchronously when immediate access to a device is required, ICA supports the use of STUN or XMPP to connection-request the device.

Each device is represented by its *cloud shadow;* this concept encompasses the usage of object-relational mapping libraries. Once the device is connected to the IDC, its shadow is created based on its data model. Cloud shadow provides advance features such as full device backup to mitigate risks of failure and device reconfiguration on start-up.

IDC allows automatic execution of Lua scripts through separate connected application. *Script Executor* (SE) interprets user defined scripts and executes actions based on device properties or events occurred in the system. Communication between IDC and SE is HTTP. Top level architecture of SE and IDC is shown in **Figure 3**.

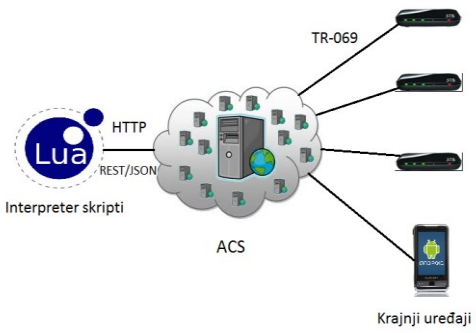


Figure 3. Script Executor and IDC top-level architecture.

IV. USE CASES AND EVALUATION

Evaluation was performed by the development of several web front-end applications using PP-API to provide monitoring and control of specific devices – set-top-boxes (STBs), home gateways, energy meters and light controls. Additionally, cross-device application for audience monitoring in TV networks was developed.

STB dashboard application (**Figure 4**) allows the user or service personnel to monitor current state of device, including active frontend connection, streaming protocols used, performance indicators (CPU load, memory consumption), user activity (currently watched channel/show) etc. Users are able to configure their device, perform channel scanning or even zap to a desired channel. The application can be accessed from web browser.

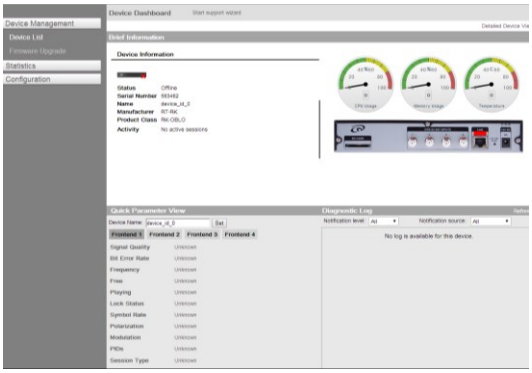


Figure 4 STB dashboard application

Energy meter dashboard application (**Figure 5**) shows the power usage statistics within a household. Power consumption can be disaggregated to specific consumption devices, such as power outlet and dimmer switches. Additionally, users are able to control lights or toggle devices within the home remotely.

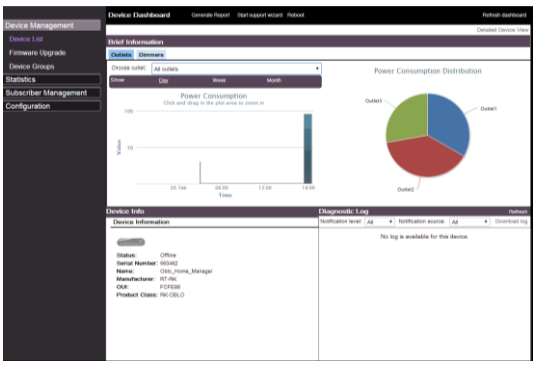


Figure 5 Energy meter dashboard application

Cross-device application for audience monitoring (**Figure 6**) uses PP-API to gather information about user behavior (channel watching statistics) from a group of STB devices. Application is implemented as a server enterprise back-end communicating via WSDL with the IDC. Front-end web application fetches aggregated statistics from the back-end part and populates the graphs accordingly.

The main selected evaluation metric was *response time.* Response time was measured for the following cases: (1) response to channel zapping in STB dashboard application split into (a) the time needed to process the request in the cloud, (b) the time needed for propagation through the Internet and (c) the time needed to perform actual zapping; and (2) response to light switch toggling in energy meter dashboard application when (a) instructed through IDC and (b) instructed through a Control4 Anywhere Access [9].

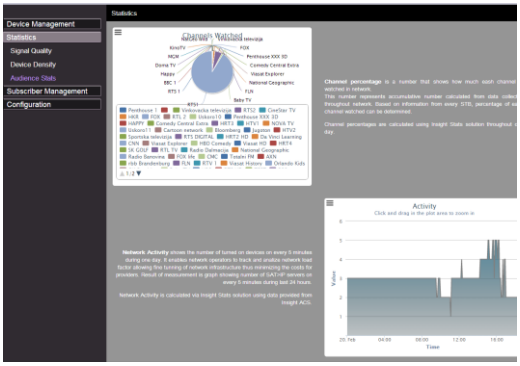


Figure 6. Cross-device application for audience monitoring

Figure 7 shows evaluation response in seconds for the test case 1. The time needed to perform the processing within the cloud is considerably smaller (~6%) when compared to propagation time and channel zapping time. For most applications, this overhead is not expected to be a drawback.

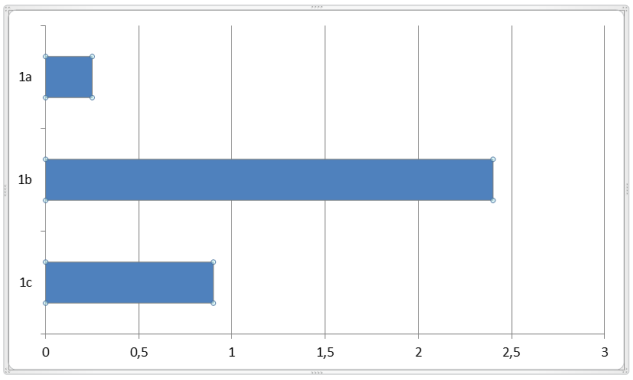


Figure 7. Test case 1: Average response time to channel zapping, time in seconds; (1a) processing the request in the cloud – IDC overhead; (1b) propagation through the Internet and (1c) actual zapping time on the STB.

Figure 8 shows evaluation response in seconds for the test case 2. Note that the actual dimmer device used when instructed both through the IDC and the Control4 Anywhere Access was the same (Oblo Smart Dimmer Switch [10]). The response time of IDC-based application (~2.9s) is lower than the response time of Control4 Anywhere Access (~3.4s). This shows that the proposed solution has comparable or better performance while keeping the architectural advantages that were introduced.

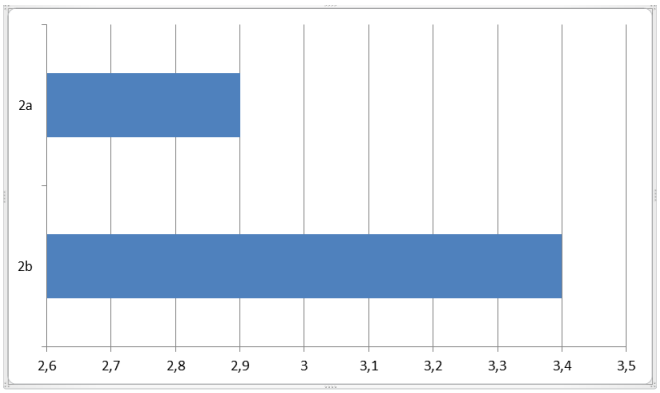


Figure 8. Test case 2: Response time in light switch toggling; (2a) using Insight Device Cloud, (2b) using Control4 Anywhere Access

Another evaluation was conducted using Script Executor web front-end application shown in **Figure 9** and comparing its speed and possibilities against a similar JavaScript based solution [11].

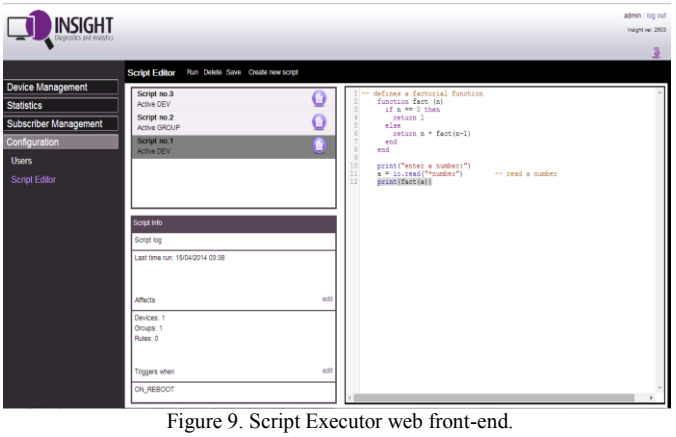


Figure 9. Script Executor web front-end.

Measurement results are shown in **Figure 10**.

	Lua	JavaScript
Memory consumption of interpreter	336 kB	739 kB
Script response time	268 msec	13 msec
Reaction to discreet events	Supported	Not supported
Complex event handling	Supported	Not supported
Configuration for a single device	Supported	Supported
Configuration for groups	Supported	Supported
Configuration according to certain criteria	Supported	Not supported
Stopping scripts	Supported	Not supported

Figure 10. Script measurement results.

V. CONCLUSION

Proposed device cloud brings several contributions to the state-of-the-art, including the ability to automatically identify new devices, adapt API dynamically, use automatic script execution and provide device representation in the cloud (*device shadow*) as long as it is required by applications, regardless of the device actual connection state. The overhead when using the proposed cloud is up to 200 ms what is acceptable for Internet of Things applications for device monitoring and control. Our initial evaluation shows that overall response time outperforms similar cloud solutions both in communication with Cloud Agents and in script execution. Future work will include proliferation of the proposed device cloud with the goal to increase the number of applications for various new devices.

ACKNOWLEDGMENT

This work is partially supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia, grant TR32041, 2011.

REFERENCES

[1] K. Ashton, "That 'Internet of Things' Thing," RFiD Journal, 2009.

[2] J. Gubbi, R. Buyya, S. Marusic, M. Planiswami, "Internet of Things (IoT): A vision, architectural elements, and future directions," Future Generation Computer Systems, vol. 29, 2013, pp. 1645-1660.

[3] M. Kovatsch, S. Mayer, B. Ostermaier, "Moving Application Logic from the Firmware to the Cloud: Towards the Thin Server Architecture for the Internet of Things," 6th Int. Conf. on Innovative Mobile and Internet Services in Ubiquitous Computing, 2012, pp. 751-756.

[4] Etherios, http://www.etherios.com/products/devicecloud/, 2014.

[5] Arkessa, http://www.arkessa.com/, 2014.

[6] Axeda, http://www.axeda.com/, 2014.

[7] Carriots, https://www.carriots.com/, 2014.

[8] Xively, https://xively.com/, 2014.

[9] Control4 Anywhere Access, http://www.control4.com/docs/product/4sight-with-anywhere-access/data-sheet/latest/4sight-with-anywhere-access-data-sheet-rev-a.pdf, 2013.

[10] Oblo Smart Dimmer Switch, http://oblo.rt-rk.com/products/smart-dimmer-switch, 2014.

[11] Motorola EDGE Manager [online]. Available: http://www.motorolasolutions.com/

Sound quality assessment in VOIP environment

Darko Lukša[1], Siniša Fajt[2], Miljenko Krhen[2]

[1]Polytechnic of Zagreb, Zagreb, Croatia
[2]Faculty of Electrical Engineering and Computing, University of Zagreb
e-mail address: darko.luksa@tvz.hr

Abstract

The need for measuring the sound quality in VOIP environment is a basic requirement of modern multimedia communication systems due to technical, commercial and legislative reasons. Measuring the sound quality using objective methods is not standardized up till now. Expected measurement results should be in correlation with subjective tests of sound quality. The usage of application for measurement of acoustic qualities of the space is explored, with an emphasis on evaluating speech intelligibility as the main parameter. In this work we present measurement method and influence of codec choice on different hardware platforms. Objectively measured small differences showed possibility to evaluate the sound quality of voice internet protocol.

I. INTRODUCTION

Nowadays, Voice Over IP (VoIP), IP telephony, audio streaming and videoconferencing becomes familiar for a wide range of users. Due to the nature of Internet communication there are still challenges to be overwhelmed to provide the end user with acceptable levels of service quality, especially because of unreliable connections quality and unexpected propagation delays. In the second chapter we presents main characteristic of main components in VOIP systems. Third chapter depicts test environment and methods, and discuss measured results.

II. THE OBJECTIVE OF THE PAPER

Assessing the quality of interactive VoIP streams is a difficult task which than that of assessing the quality of one-way streams. In this case more parameters need to be considered, such as the delay and jitter and various software configurations.

The VoIP tool employed is called Elastix [7], which is based on open-source tool Asterisk. Asterisk is the Open Source software PBX that's widely used in business telephony. It runs on Linux and provides a full-featured approach to voice and data transport over TDM, switched, and Ethernet architectures. Additionally, as one of the fastest growing Internet technologies, Voice over IP can provide a number of specific services compared to traditional telephony such as conferencing, events notification, presence notification, text-to-speech, instant messaging, video telephony and various multimedia transmissions. It could be expected that open source IP telephony products will replace a number of proprietary hardware and software with standard Linux servers and open call control over the coming decade.

Recent advances in embedded systems development qualify them for implementation of Asterisk based system. Such implementation could be found in small IP enabled embedded system boards such as Raspberry PI and various home routers.

Nowadays the most extended protocols for VoIP are SIP [4] and IAX protocol [5]. SIP protocol is widely used and represents the most popular standard until now in many VoIP applications. Before the apparition of IAX, the functionalities of call signaling and multimedia transport were implemented by different protocols acting in parallel (H.323 and SIP for signaling and RTP for the transport). However, IAX as a new concept in VoIP combines both functions in the same protocol. Thus IAX conveys all the signaling and multimedia information for the management of the voice streams into the same session which improves performance. Additionally, IAX also provides the possibility of using Trunk connections for several calls. Using this option, all the protocol overhead is common for all the calls between two IAX nodes, which clearly reduce the bandwidth usage.

During this conversation phase, SIP protocol rely the multimedia transport on Real-time Transport Protocol - RTP. On the other hand, IAX performs multimedia transport in a very different manner. Instead of use the services of RTP, IAX sends multimedia traffic into the one IAX session. Although this leads to the different impact on multimedia transport in this work we are focused on superior impact on bandwidth requirement by codec type choice.

At the transmitter side, the encoder takes an analog voice signal from an input such as a microphone and transforms the signal into a digital format that can be appended as the payload of an RTP packet. On the receiver, the decoder takes that payload content of the RTP packet and converts it back to an analog signal for play out to the receiving user. The broadest division among speech coding algorithms is between waveform quantization and parametric quantization, where the first group represents the signal by the time samples quantization. Second group denotes the signal by a binary representation of a speech model or spectral parameters.

The simplest nonparametric coding technique is Pulse Code Modulation (PCM), which is simply a quantizer of sampled amplitudes. ITU T Recommendation G.711 standardizes the u Law and A Law compression

algorithms, used in the United States and Europe, respectively. These algorithms use logarithmic scale quantization and are considered both uncompressed and reference for quality comparison with other algorithms. To enable uses in VOIP environment, ITU T recommendation G.711 incorporated Packet Loss Concealment and Discontinuous Transmission mechanisms for the G.711 codec.

Parametric codecs characteristically involve an analysis and synthesis method. In the analysis phase, speech is represented by a compact set of parameters that are encoded efficiently. At the synthesis phase, these parameters are decoded, and used in conjunction with a reconstruction algorithm to recreate a speech signal. The analysis phase can be open loop or closed loop. In the case of close loop analysis, the parameters are extracted by minimizing an objective metric, typically the square of the error between the original signal and the reconstructed speech. Due to the fact that close loop analysis usually involves the reconstructed speech, it is also called analysis by synthesis.

Some of the most commonly used codecs in the industry today which use Linear Prediction Coefficients are for the representation of the speech signals, and construct the signal using the analysis by synthesis ITU T G.729, ITU T G.723.1 and iLBC. For example, in the case of G.729 codec the autocorrelation coefficients are computed on 10 ms windows, and then converted to LPC coefficients.

Generally, packet networks have some inherent unsuitability with the transport of real time services, such as voice or video. The total delay on the network is compounded by several independent sources, which can be broadly divided in two categories: fixed delay components and variable delay components.
Variable delay are commonly referenced as delay jitter or simply jitter, is caused by the queuing that occurs in the different egress trunks that interface the Local Area Networks (LANs) to the Wide Area Networks (WANs), and add a variable and unpredictable amount of delay.

Coder delay refers to the time taken by the processor to compress a block of sound sample. This factor is manifestly affected by the processing capability of the device performing the selected compression algorithm, and the number of concurrent processes this device should handle. In this process the greatest importance has the algorithmic complexity and the frame size of the coder.

There are two other important factors that add to the overall delay are the decompression delay and the algorithmic delay. Although decompression delay is roughly 10% of the compression delay, but it should be noticed that it is multiplied by the number of frames sent in an RTP packet. Algorithmic delay refers to the unavoidable delay that some coders required in order to acquire future samples of the signal and depends on the number of future frames the coder or decoder needs to look ahead for proper operation.

Unavoidable packetization delay represents the time it takes to accumulate sufficient frames as needed by the designed number of frames in a single RTP packet. This delay is the product of the window size chosen for the compression algorithm (in ms) and the number of frames included in each RTP packet.

Queuing delay is a variable delay which depends on the trunk speed and the state of the queue in the router. It refers specifically to the queue inside the transmitting endpoint or gateway, and should not be mixed up with the propagation or network switching delay.

Network Switching Delay represents the propagation portion of the switching delay which results from the physical distance that needs to be traveled by the transmitted packets. Network congestion may adjust the number of hops required to reach the destination points, consequently the physical distance can vary from packet to packet, making packets arrive in an order different to which they were transmitted.

Specifically, we used Elastix as a unified communication platform that combines many popular sets of software available for Asterisk-based PBXs into a single interface. As a result, Elastix provides a highly user-friendly configuration of different VOIP parameters, especially for those users who are not familiar with Linux and complicated VOIP protocols implemented in Asterisk.

In exploitation it is important to clearly specify all the parameters consuming the VOIP PBX's resources CPU, RAM and hard disk. These parameters include the number of concurrent calls, type of codec used for different trunks, call features such as voice mail, music on hold and call conference to be supported. Even though it is very difficult to specify how much each of these parameters consume the VOIP-PBX resources, this information can benefit us to approximately determine the specification of the hardware for developing the VOIP-PBX. Table 1 presents required resources for main functionality. Due to the nature of different end point devices VOIP PBX an important assignment is converting different coding between them, which is processor and memory consuming task.

Commonly, in practice are used two types of phone for VoIP system: (1) Softphone and (2) specialized VoIP phone. They share the common purpose of making and receiving phone calls that are routed over the Internet, but are quite different in various aspects.

Softphone refers to software that can be installed in a computer device such as PC, notebook, smartphone, or tablet and utilizes its resources. Nowadays, there are plentiful softphones available to download, for example, X-lite, 3CX, Zoiper and many more which could be

Table 1 VOIP functionality and its resource consumption

Functionality	CPU	RAM	Hard disk
Concurrent calls	x	x	
Transcoding	**x**	**x**	
Voice mail, Call log, Backup			x

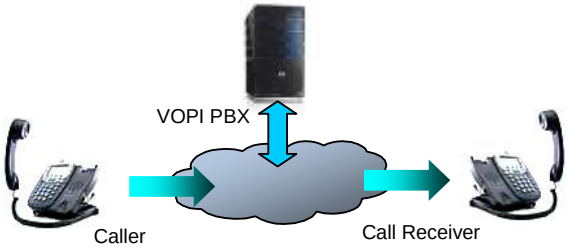


Figure 1 System diagram

Table 2 Codec properties

Codec	Bit rate [kbps]	Frame size [ms]	Data size per frame [bytes]
g.711 (A Law)	**64**	**20**	**160**
g.726	32	20	80
gsm	**13**	**20**	**33**
lpc10	2.4	22.5	7

found in different versions. The call features they provide typically are more flexible when comparing with what VOIP phones can offer.

On the other hand, VOIP hard phone refers to physical equipment that is devoted for telephony services. It is usually implemented using dedicated SOC processors for networking a wide variety of voice coders, jitter buffer management and packet loss concealment techniques, and a library of standard telephony algorithms required for IP phone operations. Comparing with the use of softphone, it is considered to be more user-friendly and due to usage of dedicated speaker and microphone provides better voice quality.

In our study we analyses influence of codec choice on different hardware devices ranging from small embedded systems to high performance computer systems.

III. DESCRIPTION OF SYSTEM

The functional diagram of the system deployed for the purpose of sound quality assessment is presented in Fig. 1. It basically includes three components: an Asterisk server, a Caller, and a Call Receiver.

For the Asterisk server we have used Elastix 2.4 with CentOS 64-bit Linux system on a Pentium 4 (2.4 GHz) with 2 GB RAM Memory. Used engine is Asterisk 1.8.20 using a standard configuration.

Asterisk server support different codec type for transcoding in a case of when receiving endpoints doesn't support used codec.

As a Caller and Call Receiver softphone was used. Choice of softphone over hard phone provides a flexibility of controlling and choosing codec on both endpoint sides. In our test environment free Zoiper softphone was used [6]. In Table 2 are presented main properties of exploited codecs.

III. THE MATERIALS AND RESULTS OF RESEACH

To determine the quality of a VoIP system, it is necessary to perform a number of objective measurements ("classical" methods for signal distortion and SNR) and subjective tests (overall quality score - user opinions).

In this study the performance and quality of a VoIP system dependent on a variety of codecs and parameters was measured, with the emphasis on the objective measurements. Measurements were performed in an anechoic chamber. The measurement system is shown in Fig. 2.

Asterisk were managed through Elastix interface. The software package Easera Pro (Electronic & Acoustic System Evaluation & Response Analysis) installed on Lenovo ThinkPad L512, Windows 7 Professional SP1 64 bit was used

Through Elastix interface more extensions or user (receiver and transmitter) were created. To place the call using the software IP phone (Softphone) ZoIPer. ZoIPer

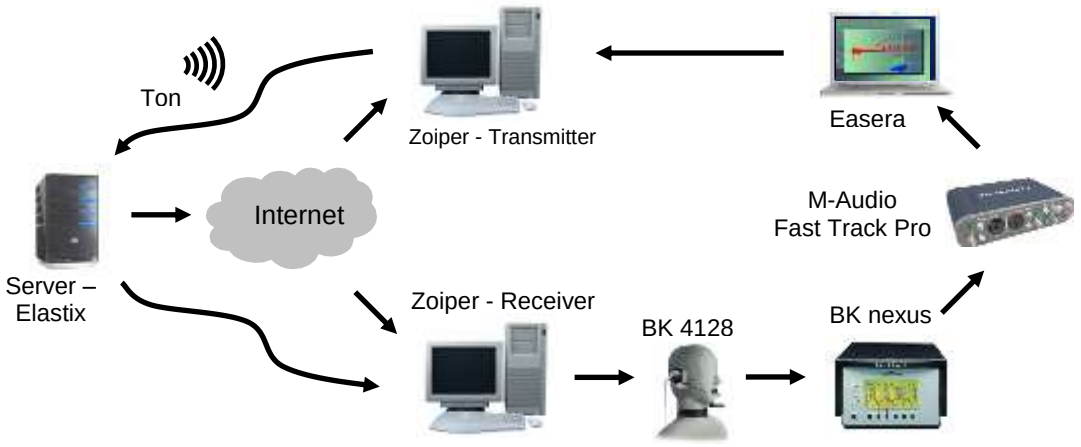


Figure 2 The Measurement System

Table 3 Speech intelligibility

Parameter	gsm_gsm default	gsm_alaw default	alaw_alaw default
STI	0,77	0,78	0,804
AlCons (%)	2,624	2,492	2,18
STI (male)	0,776	0,778	0,807
STI (female)	0,753	0,759	0,792
RASTI	0,855	0,828	0,855
equv. STIPa (male)	0,791	0,781	0,811
equv. STIPa (female)	0,77	0,765	0,8

is a software IP phone that supports both SIP and IAX2 signaling protocols and is the most commonly used software with Asterisk VoIP telephone switchboard. ZoIPer comes in several versions: free, business and OEM. The program has a built-in following audio codecs:

- G.711 μlaw
- G.711 alaw
- GSM
- Speex
- iLBC 20
- iLBC 30
- G.729 (at extra charge)

In this research the parameters by default mode were used or with the included parameters Echo cancellation, AGC and noise suppression for different combinations of codecs such as alaw - alaw, GSM - GSM and alaw - GSM. The amplitude frequency response is measured broadband with the followind codec combinations: alaw - alaw, GSM - GSM and alaw – GSM. It is noticed that there was a difference in the characteristics which can be seen in Figures 3-5.

However, in speech intelligibility by changing combinations of codecs differences were not observed. STI (Speech Transmission Index) during the measurements is ranged from 0.70 to 0.90. Index of speech intelligibility STI is a parameter for the objective evaluation of the quality of voice transmission (Table 3).

However, although in speech intelligibility differences were not observed, subjectively are perceived differences in quality. Specifically in order to examine how the system behaves by octaves, 1/48 octave analysis was made. The review of octave analysis and octave analysis of amplitude characteristics of the combination of codecs a fundamental difference in the color tone or sound is spotted. Octave amplitude characteristics of the combination of codecs GSM-GSM, GSM-alaw and alaw alaw is shown in Figure 6.

IV. CONCLUSIONS

On the basis of measurements it can be concluded that there is no objectively influence the choice of codec and their combinations on the intelligibility of voice, but it was objectively confirmed subjective aural impression of sound color change depending on the combination used codecs.

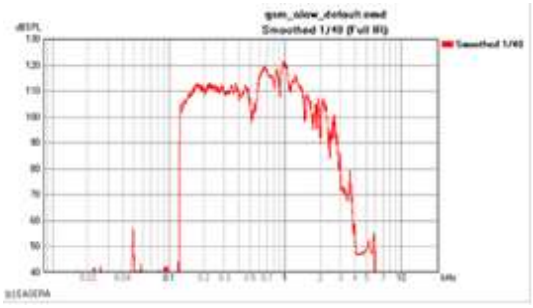


Figure 3 Amplitude characteristic of gsm/alaw codec

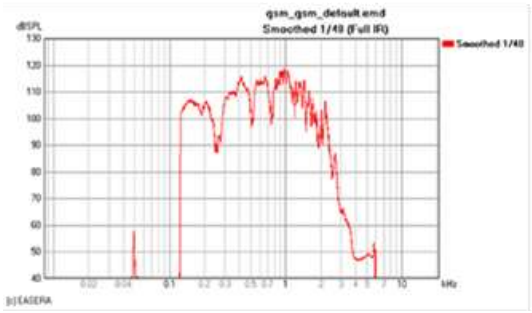


Figure 4 Amplitude characteristic of gsm/gsm codec

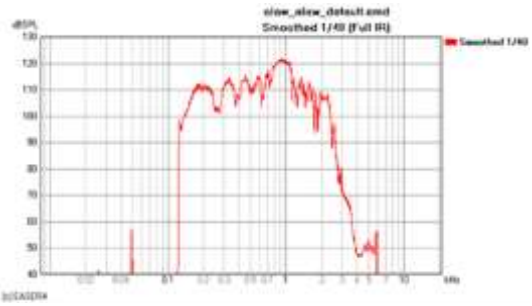


Figure 5 Amplitude characteristic of alaw/alaw codec

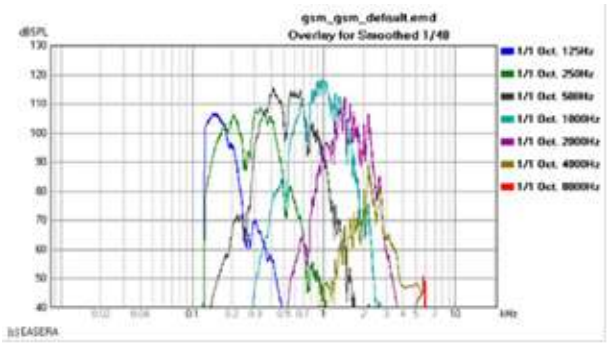


Figure 6 Codecs octal analysis of amplitude characteristics

REFERENCES

[1] Sruk, Vlado; Gašparović, Darko; Fajt, Siniša: Measurements of Speech Quality on VoIP Network, Proceedings of 2nd Congress of Alps-Adria Acoustics Association and 1st Congress of Acoustical Society of Croatia, Opatija,Croatia, 2005. pp. 317 – 320.

[2] Tomaš Goran, Furdek Marija, Fajt Siniša, Olujić Vladimir, Bruči Karlo: Perceptibility Of Aggressive Dynamics Processing In Digital Audio Broadcasting, Graz, Austrija: Joanneum Research Forschungsgeseellschaft mbH, 2007, pp. 1-4.

[3] Jim Van Meggelen, Leif Madsen, Jared Smith, Asterisk: The Future of Telephony, Second Edition, O'Reilly Media, Inc., Sebastopol, 2007.

[4] J.Rosenberg, H.Schulzrinne, G.Camarillo, A.Johnston, J.Peterson, and R.Sparks, M.Handley, E.Schooler, "SIP: Session Initiation Protocol", Internet Engineering Task Force (IETF), Request for Comments (RFC) 3261, Jun.2002.Available: http://www.ietf.org/rfc/rfc3261.txt.

[5] M.Spencer, B.Capouch, E.Guy, F.Miller, and K.Shumard, "IAX: Inter-Asterisk eXchange Version 2, draft-guy-iax-04", Internet Engineering Task Force (IETF), Internet-Draft, Mar.2008.Available: http://www.ietf.org/internet-drafts/draft-guyiax-04.txt.

[6] Zoiper softphone, http://www.zoiper.com/

[7] Elastix Open Source Unified Communications Server, http://www.elastixorg/

Automatic Interpretation of Aircraft Piston Engine Monitor Log

Dubravko Miljković
Hrvatska elektroprivreda, Zagreb, Croatia
dubravko.miljkovic@hep.hr

Abstract - Digital engine monitor during each flight produces large amount of data containing numerous engine parameters. These logs can be later downloaded to a personal computer and graphically presented with accompanying specialized software as sets of curves in a time domain. Potential problems are spotted when parameters exceed specified predetermined operating limits and when conditions for a particular diagnostic pattern are fulfilled. Catalogued diagnostic patterns consisting of EGT-CHT bars, one for each cylinder, are complex and are not easily detected from representation in form of curves on a display. Rule based pattern recognition of catalogued graphic engine fault patterns can provide necessary fault detection and isolation functions.

I. INTRODUCTION

Engine monitor is an advanced instrument for monitoring of piston engine parameters and produces sizable engine logs after each flight. After the flight logs can be downloaded to a PC where engine parameters can be presented in a graphical form suitable for the analysis by the maintenance personnel. Aside from spotting large deviations of a particular engine parameter, spotting of most diagnostic patterns corresponding to potential fault conditions is not so easy and requires considerable experience. By the automatic interpretation of data contained in the aircraft piston engine monitor log this process can be accomplished even with the less experienced personnel (possibly even the owner) and with greater speed.

II. ENGINE MONITOR

Engine monitor, called also Graphic Engine Monitor (GEM) or Engine Data Management (EDM) system is advanced and accurate piston engine-monitoring instrument that helps the pilots at managing of engine operation, [1-3]. Digital bar graph and numeric display is capable of displaying simultaneously exact Exhaust Gas Temperature (EGT) and Cylinder Head Temperature (CHT) temperature for each cylinder with the addition of Turbine Inlet Temperature (TIT) on turbocharged engines as vertical bars on the engine monitor's display. These bars form various EGT-CHT patterns that change during engine operation and may help detect abnormal engine conditions, [1-8]. Depending on engine monitor type the same bar can be used for both EGT and CHT (missing segment of a bar depicting CHT value), or, as is common on newer products, there are separate bars for EGT and CHT values, Fig. 1.

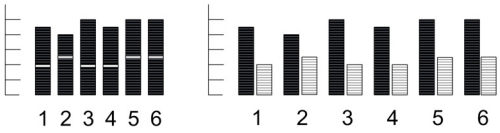


Figure 1. CHT-EGT values shown on the same and separate bars

III. ENGINE MONITOR LOG

Typical engine monitor records and stores all engine parameters to a log once every six seconds (default value) or at a user selected rate. After the flight, at a later time, this data can be transferred to a laptop PC. Engine parameters recorded by the one popular engine monitor (JPI EDM830) are shown in Table I. Logging of important engine temperature data (EGT, CHT, TITT) on a regular basis allows the creation of a engine-operation history. It remains a detailed record documenting each hour of an engine's operating life. Example of part of one engine log is shown in Fig. 2. After the record number, time data following by engine parameters and (optional) GPS data if a GPS receiver is connected to the engine monitor. If during a flight the pilot notices something unusual, it is possible to closely examine the log file belonging to that flight later.

TABLE I. MONITORED ENGINE PARAMETERS

Parameter	Description
EGT	Exhaust Gas Temperature
CHT	Cylinder Head temperature
OIL TEMP	Oil Temperature [1]
OIL PRES	Oil Pressure [1]
TIT 1	Turbine Inlet Temperature 1 [1]
TIT 2	Turbine Inlet Temperature 2 [1]
OAT	Outside Air Temperature
CDT	Compressor Discharge Temperature [1]
IAT	Intercooler Air Temperature [1]
CRB	Carburetor Air Temperature [1]
CDT - IAT	Intercooler cooling
RPM	Rotations Per Minute
MAP	Manifold Pressure
% HP	% Horse Power
CLD	CHT Cooling Rate [2]
DIF	EGT Span [3]
FF	Fuel Flow [1]

[1]optional, [2]fastest cooling cylinder, [3]difference between the hottest and coolest *EGT*

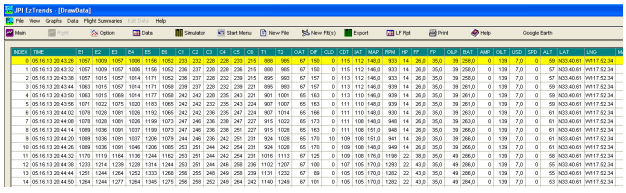


Figure 2. Example from the engine monitor log, include engine parameters and GPS data

IV. GRAPHICAL ANALYSIS SOFTWARE

EzTrends is the utility program that accompanies JPI engine monitors and enables the transfer of compressed data from the Engine Data Management (EDM) system to a personal computer (PC) for further detailed analysis, [9]. After decompression of the data software can plot the data on the screen as the graph where various engine parameters are presented as curves in different colors. EzTrends software allows great flexibility in how one can display engine monitor data. Graphs can present main and optional parameters either together, as shown in a Fig. 3, or separately, as in Figs. 4 and 5.

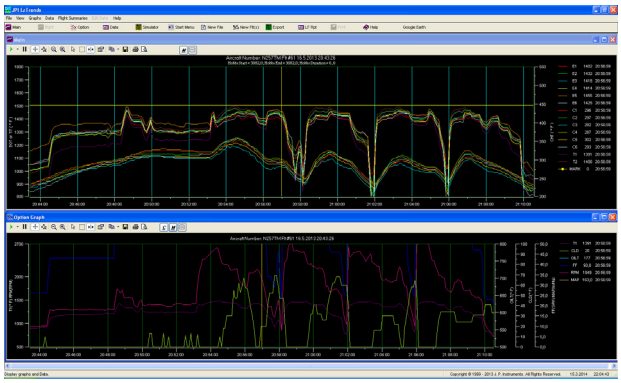


Figure 3. Main and optional parameters on the same screen

A. *Main Parameters*

Main parameters are Exhaust Gas Temperatures (EGTs) and Cylinder Head Temperatures (CHTs), The top six lines in the graph shown are EGTs (with the temperature scale on the left) and the bottom six lines show CHTs (with the temperature scale on the right), as shown in a Fig. 5.

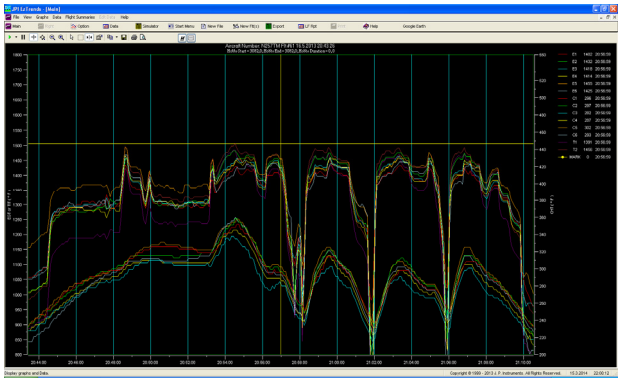


Figure 4. Main parameters (EGT, CHT, TIT)

Please note that the CHTs always lag behind the EGTs.

B. *Optional Parameters*

Optional parameters include Turbocharger Inlet Temperatures TIT1, (TIT2 if available), oil temperature OILT, fuel flow FF, engine speed RPM and manifold pressure MAP, Fig. 6.

C. *Flight Summary*

In a flight summary, Fig. 6, a schematic diagram of the engine is shown as small squares representing cylinders and connecting pipes representing exhaust pipes. Deviations from averages, for all cylinders, are color coded with the legend on the right side of the page. Operating the engine too hot is very damaging, [6,7].

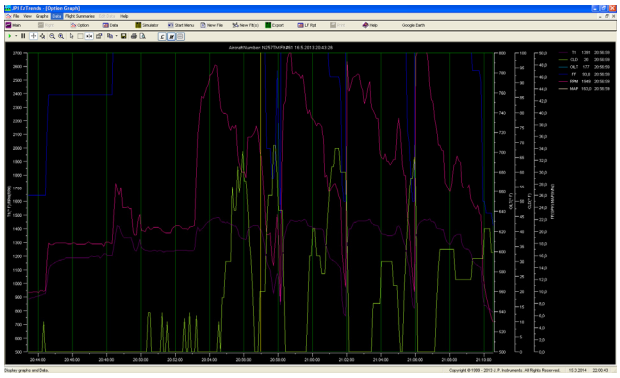


Figure 5. Optional parameters (CLD, OILT, FF, RPM, MAP)

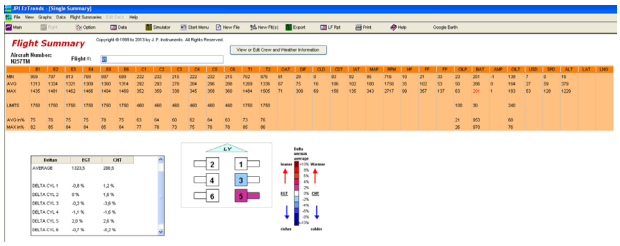


Figure 6. Flight summary as expressed by EzTrends software

The engine metal starts to soften when its temperature gets above 400°F. Hence, 450°F is chosen as the limit for engine CHTs. Serious damage to the engine may occur if CHTs rise above that value. Many turbochargers are limited to maximum turbine inlet temperature of 1650 °F. Exceeding that value may result in catastrophic failure of the turbocharger, [1]. When operating an engine one should respect limits listed in Table II, [2,3].

TABLE II. DEFAULT ENGINE MONITOR ALARM LIMITS

Measurement	Default Low Limit	Default High Limit
CHT		450 °F 230 °C
EGT [1]		1550 °F 843 °C [2]
OIL	90 °F 32 °C	230 °F 110 °C
TIT		1650 °F 900 °C
CLD		-60 °F/min -33 °C/min
DIF		500 °F 280 °C

[1]many engine monitors (e.g. JPI) don't have EGT alarm limit but DIF alarm limit instead, upper normal EGT value shown here, [2]on high performance aircraft, 1350 °F (732 °C) engines with less than 200 HP

As can be seen from previous description this software is graphically displaying log data together with some simple flight summary containing extreme values of engine parameters logged in previous flights. Maintenance personnel must carefully watch presented curve and with hindsight of experience spot abnormal engine behavior. By looking into parameters curves it is not difficult to spot large temperature deviations from default engine monitor alarm limits. However spotting fault patterns from curves is practically impossible. To spot fault patterns it is necessary to activate display simulator option, Fig. 7, showing CHTs and EGTs as vertical bars (showing patters while parsing a log). Again, one must remember and be proficient with various fault patterns.



Figure 7. Engine monitor display simulator for later log analysis

V. FAULT DETECTION AND ISOLATION

Fault detection is determination of faults present in a system and time of detection, [11]. On the other hand fault isolation is determination of kind, location and time of detection of a fault. Fault isolation follows the fault detection. Fault detection and isolation is identifying when a fault has occurred, and pinpointing the type of fault and its location, Fig. 8. Once the fault has been detected and isolated it is possible to advise appropriate corrective action. By elimination of the causes of a problem, it is possible to prevent its recurrence.

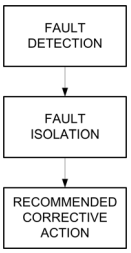


Figure 8. Fault detection, isolation and corrective action

For the automatic interpretation of engine monitor log files, limit checking is used for severe fault detection and pattern recognition of fault patterns is used for detection and isolation of various faults.

A. Limit Checkers for Engine Parameters

Limit checkers are very basic fault detection technique that can be applied using default engine alarm limits from table II. These limits contain allowable extreme EGTs (many monitors use DIF limit), CHTs, TITs and OIL temperatures during engine operation. Temperature extremes pose very wide limits that are suitable for detection of severe engine problems (1).

$$L_{L,i} < p_i < L_{H,i} \tag{1}$$

where

$L_{L,i}$ is low limit for parameter p_i
$L_{H,i}$ is high limit for parameter p_i
p_i is engine parameter i

EGTs, CHTs, TITs and OILT are primary parameters that should be considered for limit checking. Considering only upper limits for EGTs, CHTs, TITs, and both lower and upper limit for OILT (due to oil viscosity) following relations should be fulfilled, (2)-(5):

$$E_i < E_{MAX} \tag{2}$$

$$C_i < C_{MAX} \tag{3}$$

$$T_i < T_{MAX} \tag{4}$$

$$O_{MIN} < O < O_{MAX} \tag{5}$$

where

E_i is EGT of cylinder i i=1, …, N_c
C_i is CHT of cylinder i i=1, …, N_c
T_i is TIT of turbocharger i i = 1, 2
O is OILT
E_{MAX} is the maximal permissible value for E_i
C_{MAX} is the maximal permissible value for C_i
T_{MAX} is the maximal permissible value for T_i
O_{MIN} is the minimal permissible value for O
O_{MAX} is the maximal permissible value for O
Nc is the number of cylinders in engine

Despite the fact that limits from Table II consider only the extreme temperatures encountered in normal engine operation, these values are well known and documented in most piston engine and engine monitor documentation, i.e. one deals with well established facts, [1-3,10].

B. Recognition of Fault Patterns

Engine monitor is capable of displaying EGT-CHT patterns suitable for the fault detection. Patterns consist of bar graphs, darker bars represent EGT and lighter (or missing) bars CHT values. Each pattern corresponds to one or more engine problems. In pattern recognition it is possible to differentiate two classes of patterns:

1) Static Patterns

Static patterns don't change with time. There is no time evolution of pattern that depends of previous values from previous moments in time, Fig. 9. Such patterns can be described by one feature vector from a given moment.

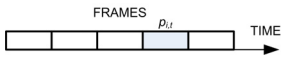


Figure 9. Static pattern described with feature vector of parameters at a particular moment

2) Dynamic Patterns

Dynamic patterns incorporates temporal dimension (evolution of parameters through the time). To detect such patterns it is necessary to consider values from the past. Dynamic patterns can be represented by more features vectors in the neighborhood that represent evolution of that pattern through the time, Fig. 10.

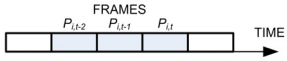


Figure 10. Dynamic pattern described with feature vector of parameters in the neighborhood of the particular moment (e.g. from a recent past)

In most simple case these patterns could be described by introduction of the simple rate of change of parameters that are elements of feature vector. This way it is also possible to avoid large feature vectors.

$$\Delta p_i = p_{i,t+1} - p_{i,t} \tag{6}$$

where

$p_{i,t}$ is the value of parameter p_i at the moment t
$p_{i,t+1}$ is the value of parameter p_i at the moment t+1

Rate of change may sometimes in engine monitor documentation be described in vague linguistic form. Terms like slow change and fast change require further clarification.

C. Integration Into the Method

For automatic parsing of engine logs limit checking is integrated with the pattern recognition for the recognition of engine fault patterns, Fig. 11.

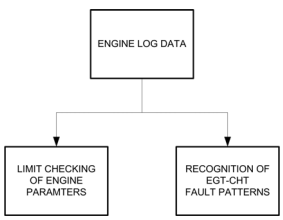


Figure 11. Log data analysis using engine parameters limits and recognition of various fault patterns

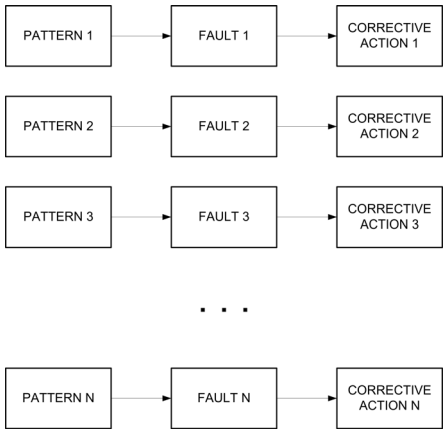


Figure 12. Fault pattern recognition for various faults

Limit checking is used for checking if engine parameters exceeded documented allowable extremes and pattern recognition for detection and isolation of various faults described by corresponding EGT-CHT fault patterns.

When using limit checking, there exists a possibility of using additional regime dependent limits based on engine speed - RPM, % of calculated maximal horse power - % HP or using flight phases, [12]. However, such limits are not widely discussed in comparison to use of limits corresponding to extreme values from Table II that practically represent common knowledge.

In a similar way measures for engine roughness and parameter fluctuations could be introduced in form of standard deviations of parameters in a recent period, however, no referent values are present in literature.

VI. RULE-BASED FAULT PATTERN RECOGNITION

Proposed pattern recognition technique employed in this method is rule based. This is due to rather precise fault pattern descriptions available in pilot's guides that accompany engine monitors, [1-3, 8]. There is also a lack of necessary numerous real world fault patterns that would otherwise justify use of some statistical pattern recognition techniques. Pattern recognition for the set of fault patterns is illustrated in Fig. 12.

A. Statistical Pattern Recognition

In statistical pattern recognition, a pattern is represented by a set of features, or attributes, viewed as a multidimensional feature vector. Various approaches in statistical pattern recognition are presented in [13]. Statistical pattern recognition requires large sample of normal and abnormal (faulty) patterns, Fig. 13, that are still difficult to acquire due to very reliable aircraft engine operation (MTBF values of an order of 5,000-30,000 hours) and consequently limited number of available recorded engine faults (and particularly various types of faults). For reliable pattern recognition each fault should be represented with at least several fault samples. It is not impossible to collect large sample base of this faults, but this would require collaboration of large owner base of aircrafts equipped with engine monitors willing to share such data. Producers of piston engines don't disclose engine failure data and statistics.

B. Rule Based Pattern Recognition

Rule based classifier represent the model as a set of

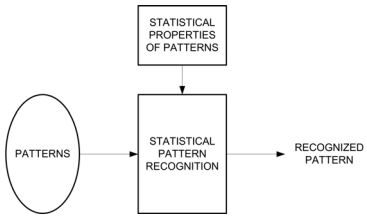


Figure 13. Statistical pattern recognition

IF-THEN rules. A rule covers a pattern x if the antecedent is satisfied (the rule is said to be triggered), [14].

Rule n: IF condition THEN conclusion (7)

A rule set may be not exhaustive (existence of patterns not covered by any rule) and may not be mutually exclusive (several rules may be triggered by the same pattern).

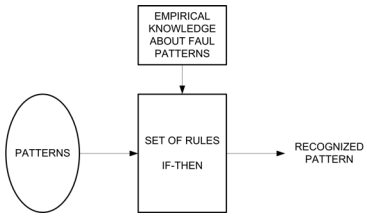


Figure 14. Rule based pattern recognition with knowledge coded into set of IF-THEN rules

Rule based pattern recognition, Fig. 14, as a kind of syntactic pattern recognition can be used instead of statistical pattern recognition if there is clear structure in the patterns. However, the development of a rule base can take great effort to complete. Mathematical descriptions for most fault patterns are produced from linguistic descriptions and expert opinion. Such mathematical descriptions in terms of necessary conditions for patterns are suitable for the rule-based pattern recognition.

Empirical knowledge about fault patterns is used for creation of simple rules for detection of numerous faults. Original rules intended for maintenance personnel are defined as descriptions in English language, usually in just one sentence and catalogued in manuals, [2, 3].

For detection of patterns it is necessary to know EGTs, CHTs, MAP, RPM and rate of change for EGT and CHT, (8) and (9):

$$\Delta E_i = E_{i,t+1} - E_{i,t} \tag{8}$$

$$\Delta C_i = C_{i,t+1} - C_{i,t} \tag{9}$$

where $E_{i,t}$, $E_{i,t+1}$, C_i, $C_{i,t+1}$ are values of E_i and C_i at moments t and t+1. Extreme values E_{max}, C_{max} that are maximal EGT and CHT values within particular record in a log are also needed, (10)-(11):

$$E_{\max} = \max\left(E_1,...,E_{N_C}\right) \tag{10}$$

$$C_{\max} = \max\left(C_1,...,C_{N_C}\right) \tag{11}$$

Beside EGTs and CHTs, the prerequisite for some patterns is previous ignition maintenance (before current flight) and GPS altitude h_{GPS}. Hence, the flag is introduced, F_{Maint} - recent ignition maintenance (before the flight). On installation with no available GPS data altitude information is not present and detection of particular pattern is omitted.

IF (recent ignition maintenance) F_{Maint}=1, else F_{Maint}=0 (12)

Following parameters are used in pattern recognition:

$$E_i,\ E_{max},\ \Delta E_i,\ C_i\ ,\ C_{max},\ \Delta C_i,\ \Delta E_{SPAN},\ M,\ R,\ F_{Maint},\ h_{GPS}$$

ΔE_{SPAN} denotes EGT span (DIFF)
M denotes manifold pressure (MAP)
R denotes engine speed (RPM)

The engine is normally operated with a slightly richer mixture. During leaning process, [1-4,8], EGTs could (temporarily) rise above limits from Table II. Leaning process may interfere with the fault detection producing false alerts. Hopefully leaning is performed only during very short periods in comparison with the duration of total flight and hence this is not a great disadvantage.

C. *Fault Patterns and Rules for Detection*

Total of 17 patterns are considered for inclusion and based on its description in [2,3] appropriate rules are formulated. Patterns in Fig. 15-29 are shown both for same and separate bars versions of CHT-EGT display. Probable causes are also described in more detail in [2,3]. Rules for detection of patterns are upgraded and revised version of rules from [12] that now include temporal changes. Bellow is the list of catalogued patterns (six cylinder engine, N_c=6), but now with derived simple mathematical description (conditions) suitable for program implementation. Resulting conditions are described in (13)-(27). "For one" means "for just one".

1. 75° to 100° EGT rise for one cylinder during flight
   - faulty spark plug, plug, wire or distributor

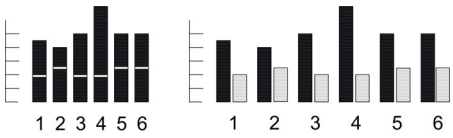


Figure 15. Pattern 1

$$\text{IF } (((E_{max}\text{-}E_i) > 75) \text{ AND } (R > 1500)) \quad \text{for one } i, \quad i=1,...,N_c \quad (13)$$

RPM greater of 1500 (as in cruise flight) is introduce to differ from Pattern 6.

2. EGT increase or decrease after ignition system maintenance
   - improper timing

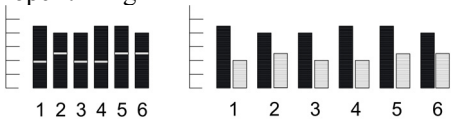


Figure 16. Pattern 2

$$\text{IF } ((F_{Maint}=1) \text{ AND } (M>20) \text{ AND } (((E_i>600) \text{ AND } (E_i<1000)) \text{ OR } (E_i>1500))) \quad \text{for all } i, \quad i=1,...,N_c \quad (14)$$

3. Loss of EGT for one cylinder
   - stuck valve

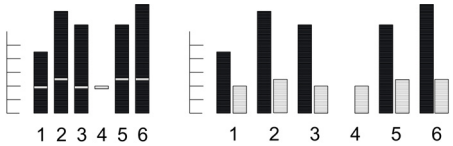


Figure 17. Pattern 3

$$\text{IF } ((M>20) \text{ AND } (E_i>0) \text{ AND } (E_i<600)) \quad \text{for one } i, \ i=1,...,N_c \quad (15)$$

EGT greater of 0 is introduced to differ from pattern 4.

4. Loss of EGT for one cylinder; no digital EGT
   - failed probe or wire harness

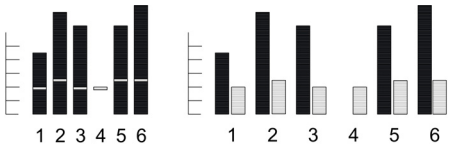


Figure 18. Pattern 4 (same as 3)

$$\text{IF } (E_i = 0) \text{ for any } i, \ i=1,\ldots, N_c \quad (16)$$

5. Decrease of EGT for one cylinder
   - faulty valve lifter

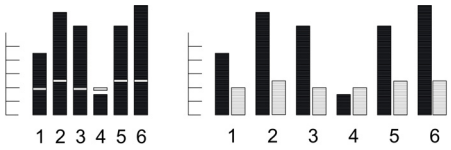


Figure 19. Pattern 5

$$\text{IF } ((M>20) \text{ AND } (E_i>600) \text{ AND } (E_i<1000)) \text{ for one } i, \ i=1,\ldots,N_c \quad (17)$$

6. Decrease of EGT for one cylinder at low RPM
   - low compression

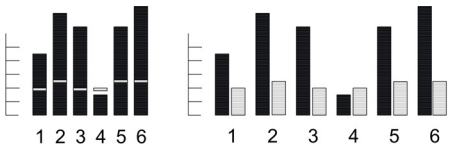


Figure 20. Pattern 6

$$\text{IF } (((E_{max}\text{-}E_i)>500) \text{ AND } (R<1500)) \text{ for one } i, \quad i=1,\ldots, N_c \quad (18)$$

7. EGT and CHT not uniform (injection engines only)
   - dirty fuel injector or fouled plugs

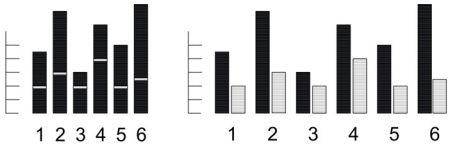


Figure 21. Pattern 7

Degree of uniformity is determined from the E_i/C_i ratio

$$\text{IF } ((E_i/C_i)<3 \text{ OR } (E_i/C_i)>6) \quad \text{for any } i, \quad i=1,\ldots, N_c \quad (19)$$

8. Decrease in EGT for all cylinders
   - decreased airflow into the induction system (ice?!)

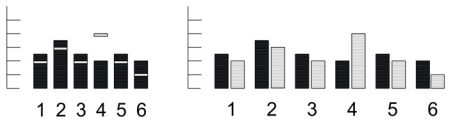


Figure 22. Pattern 8

$$\text{IF } ((M>24) \text{ AND } (E_i>1000) \text{ AND } (E_i<1200)) \text{ for all } i, \ i=1,\ldots,N_c \quad (20)$$

9. Slow rise in EGT, low CHT
   - burned exhaust valve

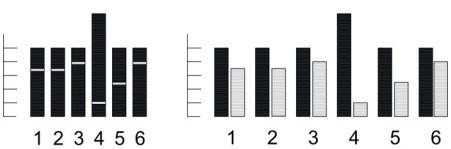


Figure 23. Pattern 9

$$\text{IF } ((E_i > 1550) \text{ AND } (C_i < 300)) \text{ for one } i, \ i=1,\ldots, N_c \quad (21)$$

Introduction of ΔE_i is not quite appropriate as rule refers to very slow change (trend) in large number of logs from successive flights

10. High CHT on cylinders on one side of engine
    - obstruction under the cowling

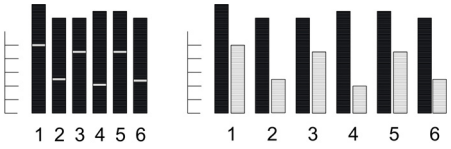


Figure 24. Pattern 10

$$\text{IF } (\text{abs}(C_{\text{left}} - C_{\text{right}}) > 100) \qquad \textit{left}=2n \ \textit{right}=2n+1, \ n=0,1,2 \quad (22)$$

Cylinders on the left side of the engine are even numbered, and on right side are odd numbered.

11. Rapid rise in EGT/CHT of one cylinder
    - detonation

With the introduction of ΔC_i it is expected that C_i rise rapidly with temperature gradient greater than 5 °F/s (30 °F in six seconds – default record time separation).

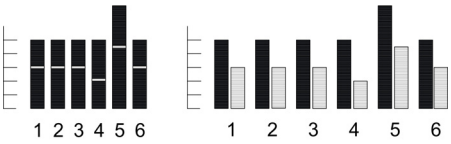


Figure 25. Pattern 11

$$\text{IF } (((E_i > 1550) \text{ AND } (C_i > 450)) \text{ OR } ((C_i > 350) \text{ AND } (\Delta C_i > 30))) \quad \text{for one i, } i=1,\ldots,N_c \quad (23)$$

12. Sudden off scale rise for any or all cylinders
    - preignition or failed probe

With the introduction of ΔE_i it is expected that E_i rise rapidly with temperature gradient greater than 10 °F/s (60 °F in six seconds – default record time separation).

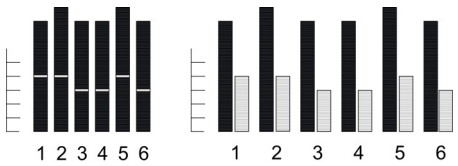


Figure 26. Pattern 12

$$\text{IF } ((E_i > 1550) \text{ OR } ((E_i > 1400) \text{ AND } (\Delta E_i > 60)) \text{ for any } i, \ i=1,\ldots, N_c \quad (24)$$

13. Loss of peak EGT-not implemented, leaning process
14. Decrease in peak or flat EGT response to leaning process-not implemented, leaning process (mixt. adj.)
15. Bellow 10,000 ft full throttle causes EGTs to rise
    - defective mechanical fuel pump

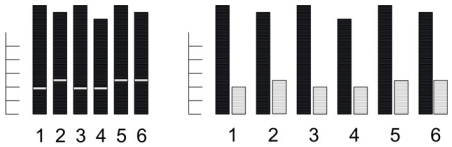


Figure 27. Pattern 15

$$\text{IF } (((h_{GPS} < 10000) \text{ AND } (M > 29)) \text{ AND } (((E_i > 1550) \text{ AND } (C_i < 450)) \text{ OR } ((E_i > 1400) \text{ AND } (\Delta E_i > 30))) \quad \text{for all } i, \quad i=1,\ldots,N_c \quad (25)$$

Manifold pressure MAP>29 due to full throttle requirement.

16. CHT more than 500º, EGT normal. Adjacent EGT may be low
    - leaking exhaust gasket to CHT probe

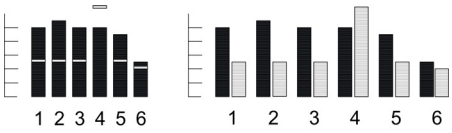


Figure 28. Pattern 16

$$\text{IF } ((C_{max} > 500) \text{ AND } (E_{max} < 1550)) \quad (26)$$

17. Large DIFF at low RPM
    - blow by in cylinder rings

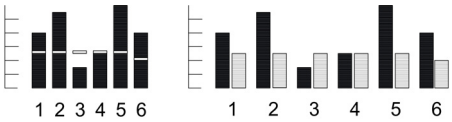


Figure 29. Pattern 17

$$\text{IF } ((\Delta E_{SPAN} > 500) \text{ AND } (R < 1500)) \quad (27)$$

D. *Parsing the Log File*

All records ($N_{records}$) in the log are checked for fault patterns ($N_{patterns}$) using procedure and derived rules:

```
for record=1 to N_records
{
 find E_i, E_max, ΔE_i, C_i, C_max, ΔC_i, ΔE_span, M, R, h_GPS, F_Maint
 for i = 1 to N_patterns
   check rule for pattern(i)
}
```

VII. CONCLUSION

Method for automatic interpretation of engine monitor log is proposed. It parses a log and is watching for engine parameter exceedances and potential fault patterns. Fault pattern recognition is realized using rule based pattern recognition where to each fault pattern correspond one or more rules. Such rules are developed by carefully analyzing decryptions of fault patterns intended for maintenance personnel and expressed in normal English language. Some rules are considered only below the particular altitude (acquire from GPS altitude if available in log) or after recent ignition maintenance and additional maintenance flag is manually determined before pattern recognition. This simple but effective approach may provide significant help to maintenance personnel.

REFERENCES

[1] *Pilot's Guide: Graphic Engine Monitor Data Logging System*, Insight Avionics, USA, 1995

[2] *Pilot's Guide: Engine Data Management, EDM-700, EDM-800, EDM-711 Primary*, J. P. Instruments, California, USA, 2007

[3] *Pilot's Guide EDM-730, EDM-830, EDM-740*, JPI, California, USA, 2005

[4] Bush M., "EGT Myths Debunked", *Cirrus Pilot*, Vol. 5, No. 7, July/August 2010

[5] Bush M., "Understanding CHT and EGT", *Cessna Pilots Association Magazine*, April 2009

[6] S. Mullender, Observations from an Engine Monitor Readout, Feb. 2008, http://www.diversiorum.org/sape/pilotage/Engines/edm-obs.html

[7] S, Mullender, If we take care of the Engine, the Engine will take care of us, Feb. 2005, http://www.huygens.org/sape/pilotage/Engines/index.html

[8] Roberts R., *The Pilot's Manual for Leaning and Diagnosing Engine Problems*, Electronic International, Oregon, USA, 2004

[9] *Pilot's Guide, EzTrends*, J. P. Instruments, 2006

[10] *Lycoming Operators Manual*, Lycoming, USA, March 1973

[11] F. Gustafsson, *Adaptive Filtering and Change Detection*, John Wiley & Sons, 2000

[12] D. Miljković, "Engine Fault Detection for Piston Engine Aircraft", MIPRO 2013, CTS, 20-24 May 2013, Opatija, Croatia

[13] Anil K. Jain, Robert P.W. Duin, and Jianchang Mao, "Statistical Pattern Recognition: A Review", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, No. 1, January 2000

[14] A. Webb and K. Copsey, *Statistical Pattern Recognition*, Willey, 2011

Regime Dependent Aircraft Piston Engine Monitoring

Dubravko Miljković
Hrvatska elektroprivreda, Zagreb, Croatia
dubravko.miljkovic@hep.hr

Abstract - Modern digital engine monitors with separate probes for each cylinder provide useful diagnostic information that precedes large class of engine problems. Limits for alarm alerts in such monitors are set to encompass all phases of engine operation. Method for monitoring engine parameters that considers multiple engine regimes corresponding to a current flight phase (engine run-up, taxi, take-off, climb, cruise and descent) and individual cylinder differences is proposed. Engine operation is presented graphically on the engine monitor display as patterns consisting of CHT/EGT bars. They can be compared to previously stored patterns corresponding to a particular flight phase possibly indicating deviations from previous flights. Flight phases are determined from engine RPM parameter together with speed and vertical speed derived from geometric altitude present in GPS data.

I. INTRODUCTION

Most of general aviation aircrafts (popularly known as small private airplanes) with the exception of turboprop and business jet aircrafts are powered by gasoline (Avgas) piston engines. Piston engine is a heat engine designed to convert energy contained in fuel into mechanical energy. It uses reciprocating pistons to convert pressure into a rotating motion. Aircraft piston engine is not very efficient at converting energy contained in a fuel to a physical motion. Only about one third of energy from fuel is converted to mechanical energy, [1]. Monitoring of engine temperatures make it possible to asses the efficiency of combustion process and indirectly detect most of engine problems, [2]. The main source of engine information traditionally available to pilot are several traditional gauges indicating cylinder head temperature (CHT), exhaust gas temperature (EGT), engine rotational speed (RPM, tachometer), fuel flow, oil temperature and oil pressure, [1,3]. These classical gauges give very basic information about engine condition. A single CHT and EGT gauge gives an average of each cylinder's head and exhaust gas temperature. Graphic engine monitor, [3-6], replaces this older method of viewing of one temperature with precise multi-cylinder engine monitoring of EGT and CHT temperatures plus myriad of other parameters. Piston engines used in general aviation aircrafts have a MTBF (considering total loss of power) in a range of 5000-30000 hours depending on engine complexity, operation and maintenance. With monitoring of engine parameters it is possible to detect impending engine problems and increase engine operational reliability. Engine monitors usually have one set of parameter limits for all flight phases that when exceeded issue an alarm. Introduction of separate parameter limits for various flight phases may improve fault detection capability of engine monitor.

II. ENGINE MONITOR

Engine monitor is advanced and accurate piston engine-monitoring instrument that monitors engine parameters and assist the pilot to operate engine in correct and optimal way, [1-6]. Temperatures are shown graphically as bars on the display of an engine monitor, Fig. 1. Each column in the bar on the display is composed of a stack of segments. The total height of each column represents the EGT while the missing segment in the column represents the CHT. In addition to graphically displaying EGTs and CHTs, on turbocharged engines the instrument continuously displays Turbine Inlet Temperature(s) (TIT). New products often have color display with separate columns for EGTs and CHTs, [6], Fig. 2. Engine monitors cover much more engine data then basic gauges (about dozen of parameters that are also recorded and can be downloaded and analyzed later). Various engine problems appear on the monitor's display as bar graph fault patterns, with two examples shown in Figs. 3 and 4. More examples can be found in [2,4,6].



Figure 1. Engine monitor with bar graph display (Insight Avionics, older model GEM 603)

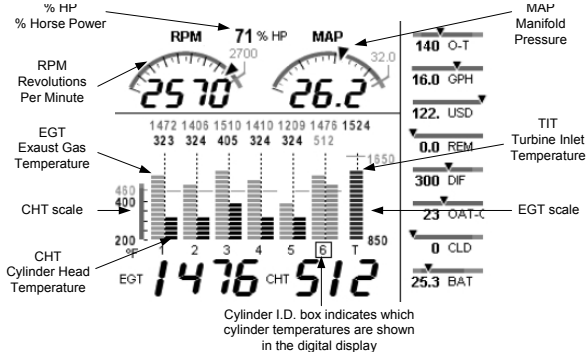


Figure 2. Engine monitor with separate bars for EGT and CHT (JPI EDM 830)

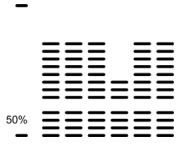


Figure 3. Low compression, Cyl. 4

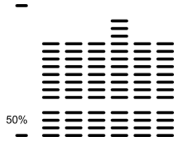


Figure 4. Detonation, Cyl. 4.

TABLE I. MONITORED ENGINE PARAMETERS

Parameter	Description
EGT	Exhaust Gas Temperature
CHT	Cylinder Head temperature
OIL TEMP	Oil Temperature [1]
OIL PRES	Oil Pressure [1]
TIT 1	Turbine Inlet Temperature 1[1]
TIT 2	Turbine Inlet Temperature 2 [1]
OAT	Outside Air Temperature
CDT	Compressor Discharge Temperature [1]
IAT	Intercooler Air Temperature [1]
CRB	Carburetor Air Temperature [1]
CDT - IAT	Intercooler cooling
RPM	Rotations Per Minute
MAP	Manifold Pressure
% HP	% Horse Power
CLD	CHT Cooling Rate [2]
DIF	EGT Span [3]
FF	Fuel Flow [1]

[1]optional, [2]fastest cooling cylinder, [3]difference between the hottest and coolest *EGT*

Monitored engine parameters (available in JPI EDM 830) are shown in Table I. The similar set of parameters is also available in other modern engine monitors. Engine monitor automatically records engine parameters during each flight into a log, Fig. 5. New log record is typically recorded every 6 seconds giving 600 records per hour. Recorded data can be downloaded with cable, wireless connection or memory card for later analysis using software, [7], installed on a PC for sophisticated graphical analysis. Suspicious data logs can be sent to a mechanic or engine manufacturer for further clarification.

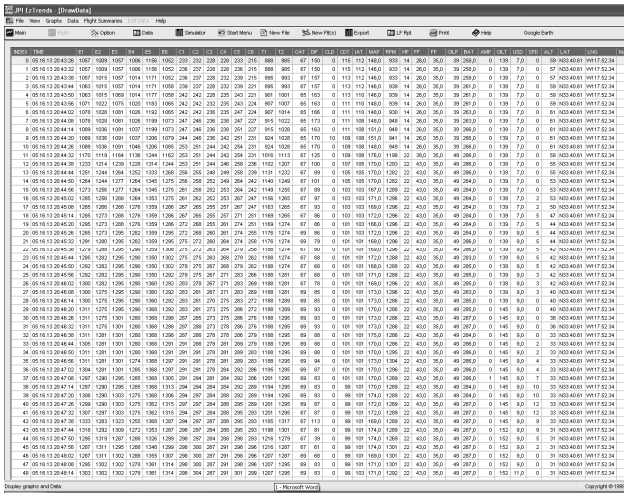


Figure 5. Example from engine monitor log, include engine parameters and GPS data

III. ENGINE PARAMETERS STATISTICS

Graphical representations of the probability distributions of key engine parameters CHT, EGT and TIT in form of frequency histograms are shown in Figs. 6-8, [8]. As can be seen in figures, the probability distributions have several peaks and as such are too complicated for simple application of common mathematical distributions. One solution to this problem is use of the statistical summary in form of percentiles, [8,9]. Percentile is the value of a variable below which a certain percent of observations fall. The 99th percentile is the value below which 99% of the observations may be found. Statistical analysis was performed on three engine log files supplied with the EzTrends2 software, [7],

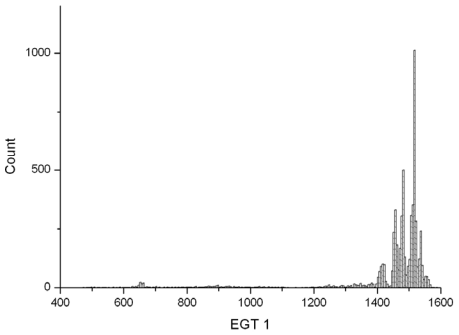


Figure 6. EGT 1 histogram

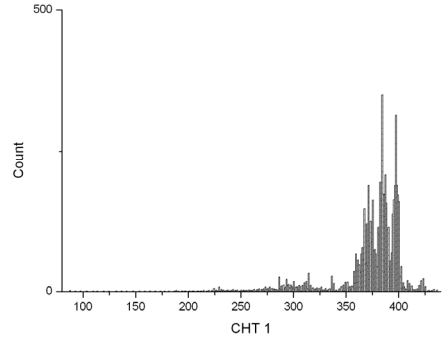


Figure 7. CHT1 histogram

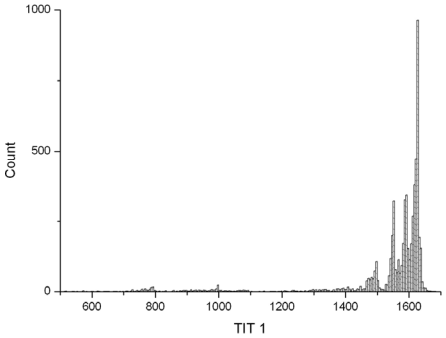


Figure 8. TIT 1 histogram

Flt#49 of duration 2.57 hours, Flt#56 of duration 0.43 hours and Flt#61 of duration 0.45 hours, belonging to a six cylinder, 280 HP turbonormalized Continental engine TSIO-550-G, from 2007 Mooney M20TN Acclaim aircraft, N257TM, with no known faults present. Resulting statistical summary of engine parameters encompassing all engine regimes is shown in form of Box-Wiskers plots in Fig. 9 with corresponding data given in Table II. Engine operation during a whole flight could be additionally summarized with flight summary generated by EzTrends2 software (average, max, min value of engine parameters) as shown in Fig. 10. Temperature deviations form average (see color legend) for each cylinder are shown in small picture. As can be seen, extreme values (minimums and maximums) of engine parameters during normal operation for the period of whole flight, shown in Table II and Fig. 10, are

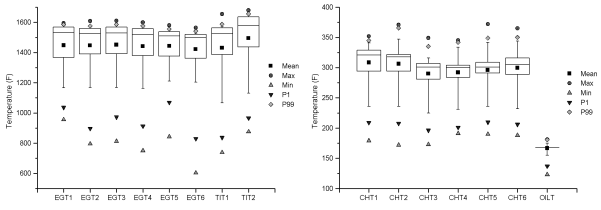


Figure 9. EGT1-6, TIT1, TIT2, CHT1-6 and OILT for all engine operating regimes

TABLE II. STATISTICAL SUMMARY OF SELECTED ENGINE PARAMETERS (TEMPERATURES) ENCOMPASSING ALL ENGINE REGIMES (LIMITS USEFUL FOR ALERTS ARE SHADED)

Param.	Mean μ	Std.Dev.σ	Min	Max	P_1	P_{99}
EGT1	1447,5	143,9	956	1595	1036,6	1586
EGT2	1446,8	149,0	796	1608	897	1576,2
EGT3	1451,3	148,0	813	1610	972	1586,2
EGT4	1441,4	150,6	751	1599	912,8	1577
EGT5	1443,3	118,6	843	1580	1069	1556
EGT6	1422,2	134,6	604	1564	829,2	1542,2
TIT1	1430,7	179,1	739	1655	837	1588,2
TIT2	1494,4	177,5	876	1679	966	1656
CHT1	308,3	31,1	179	352	209	345
CHT2	306,2	31,8	172	371	207,6	365,6
CHT3	290,0	28,0	173	349	196,4	335,2
CHT4	292,0	27,7	191	345	201	342
CHT5	296,0	27,5	190	372	209,6	349
CHT6	299,4	29,9	188	365	206	350,2
OILT	166,7	7,43	123	181	137	181

close to the default (but adjustable) engine monitor alarm limits shown in Table III, [6]. Alarm values are quite close match for EGTs (many monitors use DIF alarm limit instead) and TITs, and somewhat higher for CHTs.

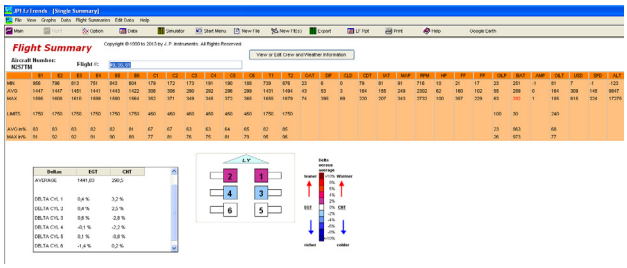


Figure 10. Flight summary as generated by EzTrends2 software

TABLE III. DEFAULT ENGINE MONITOR ALARM LIMITS

Measurement	Default Low Limit	Default High Limit
CHT		450 °F 230 °C
EGT [1]		1550 °F 843 °C [2]
OIL	90 °F 32 °C	230 °F 110 °C
TIT		1650 °F 900 °C
CLD		-60 °F/min -33 °C/min
DIF		500 °F 280 °C

[1]many engine monitors (e.g. JPI) don't have EGT alarm limit but DIF alarm limit instead, upper normal EGT value shown here, [2]on high performance aircraft, 1350 °F (732 °C) engines with less than 200 HP

IV. FLIGHT PHASE DEPENDENT REGIMES

During a flight engine operates through various regimes. In aircraft with a constant speed propeller engine is controlled with the throttle and prop lever. The throttle lever controls the target manifold pressure (MAP) while the prop lever controls the desired rotation speed at the propeller governor expressed in Revolutions Per Minute (RPM). The mixture lever control manages the richness of mixture. Table IV lists several regimes of engine operation corresponding to various phases of flight. Engine parameters during a particular regime fall within a narrower range compared to a wider range corresponding to all regimes, [8,9]. This way one can detect minor deviations for a particular regime. Separating engine operation in regimes can be done in many ways. Some of them are proposed bellow.

A. *Revolutions Per Minute - RPM*

Separation based on the engine operating speed expressed in Revolutions Per Minute (RPM) is quite simple solution, but it has some shortcomings. It doesn't consider altitude, load and manifold pressure (MAP).

TABLE IV. REGIME DETERMINATION FOR CONSTANT SPEED PROPELLER

Phase	Speed kn	Alt.ft QFE[1]	Vertical speed ft/min	RPM	MAP[3]
Engine run-up	0[2]	0[2]	0[2]	800-1000	
Taxi	< 20	0[2]	0[2]	800-1500	
Take-off	<45	0[2]	0[2]	2300+	
Climb	> 45	> 0	> 200	2000+	33-35
Cruise	> 45	> 0	>-200,<200	1800-1500	32
Descent	> 45	> 0	VS <-200	< 1800	18-24
Landing	< 45	0[2]	0[2]	< 1200	

[1]height above a specific aerodrome or ground level, [2]in practical considerations some tolerance must be allowed instead of precise value of 0, [3]approximative values

Same RPM has different meaning in term of engine load during climb, cruise and descent. Relation between the speed, altitude and RPM is shown in Fig. 11. High RPM is present booth in climb and cruise. Engine load is generally much greater during a climb. Example of regime switching variable r that separates operating regimes in 20% RPM brackets is defined by (1), where RPM_{MAX} is maximal value RPM can achieve (≈2850).

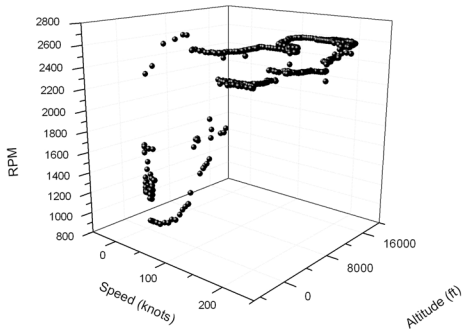


Figure 11. Figure Speed, altitude, RPM combinations during a flight

$$r = \left\lceil \frac{100RPM}{20RPM_{MAX}} \right\rceil \quad (1)$$

B. *Calculated Percent of HP - %HP*

This is more sophisticated choice that considers Revolutions Per Minute RPM, Manifold pressure MAP, fuel flow FF, outside air temperature OAT and sometimes pressure altitude. Horsepower is the product of torque (roughly proportional to MAP) and RPM. At a given torque, HP is directly proportional to RPM. % HP as a function of RPM and MAP is shown in Fig. 12. Regime switching variable r that separates operating regimes in 20% brackets is given with (2), [8,9].

$$r = \left\lceil \frac{\%HP}{20} \right\rceil \quad (2)$$

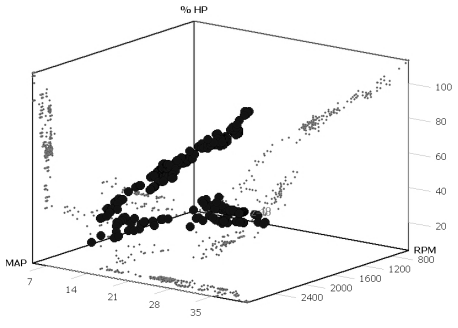


Figure 12. % HP as a function of RPM and MAP (manifold pressure)

C. Flight Phase

Full glass cockpit is still rare in general aviation aircraft. However, GPS receiver is commonly available. Many engine monitors can accept GPS data, Fig. 13. By incorporating GPS speed, GPS geometric altitude and calculated vertical speed to engine monitor RPM parameter it is possible to determine actual flight phase.

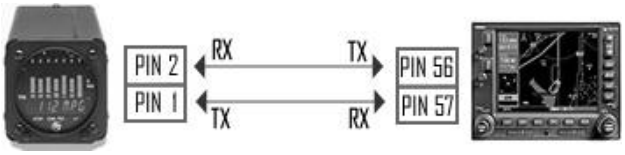


Figure 13. Communication between graphic engine monitor and GPS receiver (JPI EDM 700/80 and Garmin GNS 530) - JPI Instruments

The engine is under different load in climb, cruise and descent that sometimes can't be distinguished from engine parameters alone. The use of speed, vertical speed and engine RPM enables distinguishing of climb, cruise and descent phase. Vertical speed is determined from successive altitudes data Alt_{t1} and Alt_{t2} at t_i and t_{i+1}, (3).

$$VS = \frac{Alt_{t_{i+1}} - Alt_{t_i}}{t_{i+1} - t_i} \tag{3}$$

GPS speed is a groundspeed and in presence of very strong winds may vary significantly from indicated speed (IAS). It would be better if one could get IAS once the aircraft is aloft. Flight phase dependent regimes are particular suitable because most engine monitor documentation and literature describe potential problems in relation to the specific phase of a flight (e.g. "compare temperatures to past climbs") with readily available recommendations about engine operation and corrective actions, [2,5,6]. Speed, altitude and vertical speed during one flight are shown in Figs. 14-16.

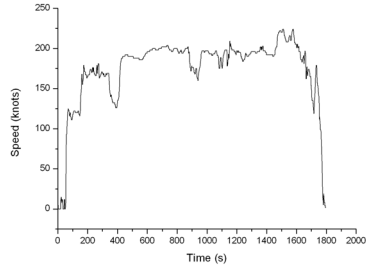


Figure 14. Speed of the aircraft during a typical flight

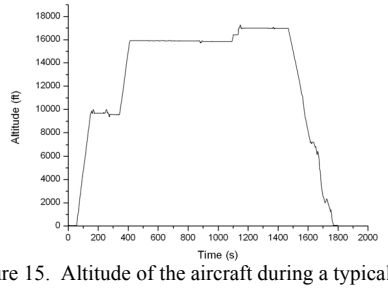


Figure 15. Altitude of the aircraft during a typical flight

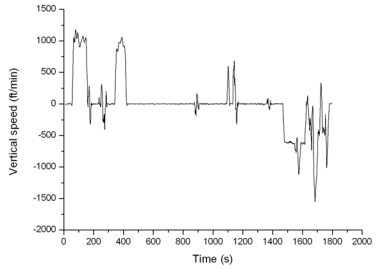


Figure 16. Vertical speed of the aircraft during a typical flight

D. Rules for Flight Phase Determination

Following flight phases are chosen: engine run-up, take-of, climb, cruise, descent. Landing is not included due to the short duration and low RPM that is not hard on the engine. Rules for the determination of the flight phase are given bellow (SPEED is GPS groundspeed, VS is calculated vertical speed from successive readings of GPS altitude and RPM is engine rotation speed):

IF SPEED < 3 AND 800 < RPM < 1700[1] Engine run-up
IF 3 < SPEED < 20 AND 30 < VS = 30[2] AND RPM < 1500 Taxi
IF 3 < SPEED < 45[3] AND 30 < VS = 30[2] AND RPM > 2300 Take-off
IF SPEED > 45[3] AND VS >200 AND RPM > 2000 (?) Climb
IF SPEED > 45[3] AND -200 < VS < 200 Cruise
IF SPEED > 45[3] AND VS < -200 Descent

[1]Recommended values for run-up are bellow 1500, value of 1700 is set after analyzing available logs, [2]In real world small margin of error (e.g. 30 ft/min due to GPS altitude error) must be allowed around value of 0 valid for ideal case, [3]Compromise value to allow for strong headwinds

With previously defined rules very small percentage of time engine parameters will not be labeled to any flight phase (≈1.5%). Disadvantage of such determination of flight phases is that it depends on availability of GPS signal. Under some rare occasions GPS data may be temporary unavailable and in that case only warning alerts not related to flight phase may be issued.

V. FLIGHT PHASE REGIME DEPENDENT STATISTICS

Statistical summaries for six previously defined engine regimes are shown in Figs. 17-22 and Table V**.** Please note that graphs vary in temperature scales (different temperature scale than used for all regimes shown in Fig. 9) and parameter spreads are smaller.

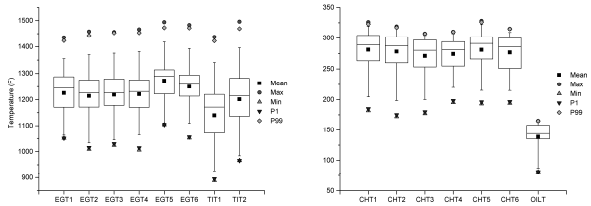


Figure 17. EGT1-6, TIT1, TIT2, CHT1-6 and OILT for engine run-up

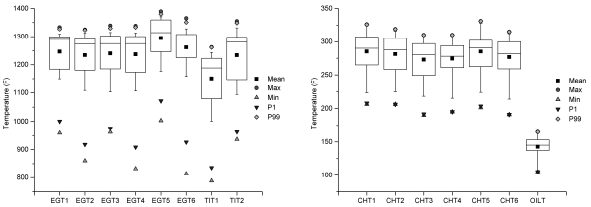


Figure 18. EGT1-6, TIT1, TIT2, CHT1-6 and OILT for taxi

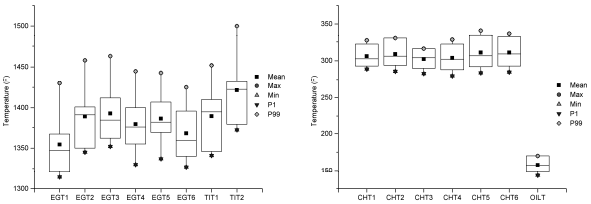


Figure 19. EGT1-6, TIT1, TIT2, CHT1-6 and OILT for take-off

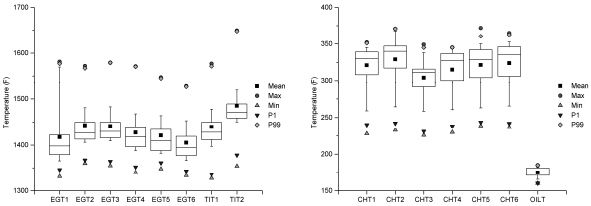


Figure 20. EGT1-6, TIT1, TIT2, CHT1-6 and OILT for climb

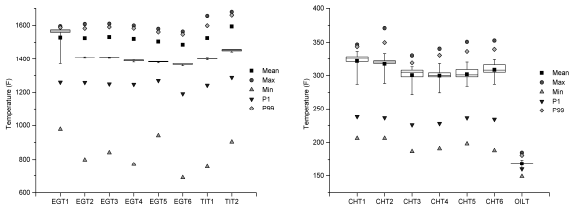


Figure 21. EGT1-6, TIT1, TIT2, CHT1-6 and OILT for cruise

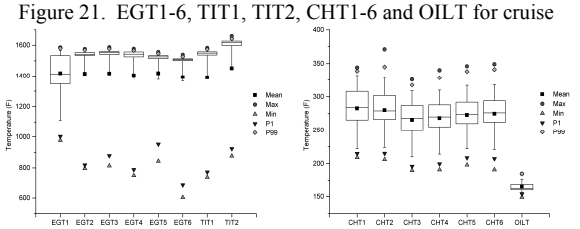


Figure 22. EGT1-6, TIT1, TIT2, CHT1-6 and OILT for descent

TABLE V. 1ST AND 99TH PERCENTILE VALUES FOR VARIOUS ENGINE PARAMETERS AND REGIMES (LIMITS USEFUL FOR ALERTS ARE SHADED)

Param. P_i	Regime											
	Run-up		Taxi		Take-off		Climb		Cruise		Descent	
	L_L	L_H	L_L	L_H	L_L	L_H	L_L	L_H	L_L	L_H	L_L	L_H
EGT1	1051	1423	1001	1327	1314	1430	1346	1577	1260	1588	1002	1581
EGT2	1015	1443	918	1323	1345	1458	1367	1563	1259	1583	818	1569
EGT3	1030	1453	976	1330	1352	1463	1364	1579	1250	1591	877	1578
EGT4	1014	1453	909	1333	1330	1445	1352	1570	1248	1584	788	1567
EGT5	1104	1473	1071	1380	1337	1443	1361	1545	1270	1562	952	1547
EGT6	1056	1472	926	1351	1327	1425	1343	1526	1191	1548	684	1527
TIT1	895	1422	836	1261	1341	1452	1336	1572	1242	1599	1396	1506
TIT2	966	1469	963	1349	1372	1500	1378	1648	1289	1660	1148	1293
CHT1	185	323	208	326	289	327	240	351	238	344	215	339
CHT2	175	317	206	319	286	330	242	371	236	350	215	345
CHT3	180	307	191	310	283	316	231	345	226	319	196	318
CHT4	198	310	195	310	280	328	238	345	228	332	199	329
CHT5	196	326	203	331	284	341	244	361	236	337	208	338
CHT6	196	315	191	315	285	337	242	363	234	340	207	341
OILT	81	165	105	165	143	170	160	185	160	181	154	185

VI. FLIGHT PHASE PROTOTYPE PATTERNS

Within the engine there exist small temperature differences between cylinders due to cylinder position (different air cooling), differences in distance traveled by the air from the intake and differences among injectors. This can be seen as slightly different bar heights belonging to various cylinders. Differences also exist with the change of flight phase. It may be beneficial to compare current pattern with previously stored prototype pattern corresponding to that particular flight phase.

A. Pattern Vector

Pattern vector $\boldsymbol{v}_T$ consist of components E_i and C_i that are EGT and CHT values for cylinder i at some record, N_C is the number of cylinders, (4).

$$v_T = [E_1, \ldots, E_{Nc}, C_1, \ldots, C_{Nc}] \tag{4}$$

B. Pattern Vector Normalization

To later facilitate better pattern comparison pattern normalization is performed, removing some influence of temperature differences within a flight phase to a shape of the pattern. Normalization is performed by determining highest *EGT* among all cylinders, E_{MAX}, (5). *EGT* and *CGT* values are then divided by the E_{MAX}.

$$E_{MAX} = \max(E_i) \qquad i=1,\ldots, N_C \tag{5}$$

Normalized *EGT* values, $E_{N,i}$ are, (6):

$$E_{N,i} = \frac{E_i}{E_{MAX}} \qquad i=1,\ldots, N_C \tag{6}$$

Similarly, normalized *CHT* values, $C_{N,i}$ are, (7):

$$C_{N,i} = \frac{C_i}{E_{MAX}} \qquad i=1,\ldots, N_C \tag{7}$$

All components are within the interval [0, 1], (8),(9):

$$0 <= E_{N,i} <= 1 \qquad i=1,\ldots, N_C \tag{8}$$

$$0 <= C_{N,i} <= 1 \qquad i=1,\ldots, N_C \tag{9}$$

Normalized test pattern is now vector $\boldsymbol{v}_{N,T}$ (10).

$$v_{N,T} = [E_{N,1}, \ldots, E_{N,Nc}, C_{N,1}, \ldots, C_{N,Nc}] \tag{10}$$

C. Prototype Pattern Vector

Prototype pattern vector for flight phase r is the vector $\boldsymbol{v}_{N,P,r}$, (11), that consists of averaged normalized values $\hat{E}_{N,i,r}$ and $\hat{C}_{N,i,r}$ determined by (12) and (13).

$$v_{N,P,r} = [\hat{E}_{N,1,r}, \ldots, \hat{E}_{N,Nc,r}, \hat{C}_{N,1,r}, \ldots, \hat{C}_{N,Nc,r}] \tag{11}$$

$$\hat{E}_{Ni,r} = \sum_{k=1}^{N_r} E_{N,i,r,k} \tag{12}$$

$$\hat{C}_{Ni,r} = \sum_{k=1}^{N_r} C_{N,i,r,k} \tag{13}$$

where $E_{N,i,r,k}$ and $C_{N,i.r,k}$ are EGT and CHT values of kth record and N_r is the total number of records for the flight phase r. Components of normalized prototype vectors for each of six flight phases are shown in Table VI.

Table VI. COMPONENTS OF NORMALIZED PROTOTYPE PATTERNS FOR EACH PHASE (BASED ON AVERAGE VALUES)

Param. P_i	Regime					
	Run-up	Taxi	Take-off	Climb	Cruise	Descent
EGT1	0,9612	0,9614	0,9692	0,9792	0,9952	0,9872
EGT2	0,9518	0,9511	0,9944	0,9974	0,9931	0,9832
EGT3	0,9557	0,9561	0,9970	0,9973	0,9973	0,9847
EGT4	0,9575	0,9533	0,9877	0,9903	0,9903	0,9747
EGT5	0,9985	0,9986	0,9925	0,9808	0,9808	0,9878
EGT6	0,9810	0,9728	0,9790	0,9682	0,9682	0,9670
CHT1	0,2202	0,2196	0,2192	0,2221	0,2097	0,1974
CHT2	0,2178	0,2166	0,2211	0,2279	0,2071	0,1958
CHT3	0,2123	0,2105	0,2164	0,2103	0,1959	0,1855
CHT4	0,2150	0,2114	0,2175	0,2181	0,1957	0,1874
CHT5	0,2201	0,2195	0,2226	0,2225	0,1971	0,1910
CHT6	0,2169	0,2131	0,2226	0,2242	0,2014	0,1920

VII. METHOD FOR FAULT DETECTION

Previous statistics and temperature profiles for each flight phase can be integrated in a fault detection method as illustrated in Fig. 23. Limit checkers may be applied to measured engine parameters, [10].

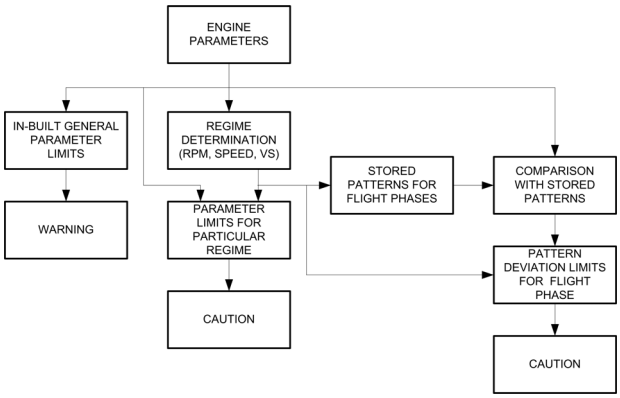


Figure 23. Integration of limit checkers with temperature profile in a method for fault detection

A. Limit Checkers for Global Limits

Limit checkers can be applied using engine parameters limits extracted from engine logs that encompass whole flights and are suitable for detection of severe engine problems issuing a warning alert, (14):

$$L_{L,i} < p_i < L_{H,i} \tag{14}$$

where

$L_{L,i}$ is low limit for parameter p_i
$L_{H,i}$ is high limit for parameter p_i
p_i is engine parameter i

B. Limit Checkers for Regime Dependent Limits

Same technique can also be applied to engine regime dependent limits, (15). This places narrower limits to engine parameters corresponding to a particular engine regime and is suitable for detection of smaller problems.

Such limits are beneficial as it may be of interest to detect smaller problems (issuing less severe caution alerts).

$$L_{L,i,r} < p_i < L_{H,i,r} \quad (15)$$

where

$L_{L,i,r}$ is low limit for regime r
$L_{H,i,r}$ is high limit for regime r

Percentiles from Table V can be used for setting limits.

C. Limit Checkers for Detection of Pattern Deviations

The Euclidian distance between vector of engine parameters represented in normalized engine pattern and stored normalized prototype pattern corresponding to current flight phase is determined. For vectors $\boldsymbol{p}$ and $\boldsymbol{q}$ the Euclidean distance between is given by (16).

$$d(p,q)=\sqrt{(p_1-q_1)^2+(p_2-q_2)^2+\ldots+(p_i-q_i)^2+\ldots+(p_n-q_n)^2} \quad (16)$$

Considering variance for EGTs and CHTs, σ^2_E and σ^2_C, weighted Euclidean distance between test vector $\boldsymbol{v}_{N,T}$ and prototype vector $\boldsymbol{v}_{N,P,r}$ is given by (17). Typical values for variances (cruise) are σ^2_E=0.000025 and σ^2_C= 0,000005.

$$d_{v_{N,T},v_{N,P,r}}=\sqrt{\frac{1}{\sigma_E^{\,2}}\sum_{i=1}^{i=N_E}\left(E_{N,i}-\hat{E}_{N,i,r}\right)^2+\frac{1}{\sigma_C^{\,2}}\sum_{i=1}^{i=N_C}\left(C_{N,i}-\hat{C}_{N,i,r}\right)^2} \quad (17)$$

Distribution of distances $d_{\boldsymbol{v}N,T,\boldsymbol{v}N,P,r}$ from all test patterns to prototype pattern for each regime r is shown in Fig. 24. Statistical summary of distance between prototype and patterns from the same regime are given in Table VII.

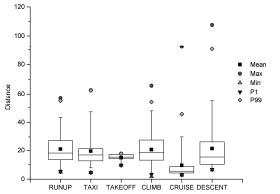


Figure 24. Difference between current and stored prototype patterns

TABLE VII. STATISTICAL SUMMARY: DISTANCES FROM PROTOTYPE

Regime	Mean μ	Std.Dev.σ	Min	Max	P_1	P_{99}
Run-up	20,87927	11,07503	4,89324	56,67822	5,2041	54,87364
Taxi	19,4386	11,36059	4,50276	62,60759	4,51575	62,50814
Take-off	14,88883	2,91856	9,60374	17,73574	9,60374	17,73574
Climb	20,49501	11,62248	1,69166	65,63849	3,2465	54,02237
Cruise	9,55232	10,01092	2,80137	92,11229	2,95024	45,60141
Descent	21,23669	16,6172	6,37109	107,54823	6,73234	90,53074

*Shaded values for P99 can be used for limits

Limit checker can be applied to distance as well, (18).

$$d_r < L_{H,r} \quad (18)$$

If distance is greater than 99th percentile limit caution is issued. This indicates that shape of a current pattern is too different from the stored prototype for a flight phase.

D. Percentiles and Detection Sensitivity

Failures in many technical systems follow a risk pyramid illustrated in Fig. 25, [8]. General idea behind this figure is that in a large sample number of minor events precede more severe and very severe events. Percentiles are fixed to 1% and 99% values, and this sensitivity may be set too high even for the smallest engine problems. Considering just upper limits of engine parameters (99th percentiles), 600 records of engine parameter per hour would translate to average of 6 caution alerts per hour

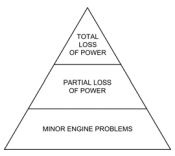


Figure 25. Pyramidal relationship of risk, tailored to engine problems



Figure 26. Permitted number of threshold exceedances within a specified time frame

for each parameter, or 72 caution alerts for all EGT and CHT values in 6 cylinder engine. To adjust the sensitivity of detection one may additionally specify the number of limit exceedances per time period (use of counter) that should be reason for concern (e.g. more than 5 exceedances per parameter in last 5 minutes). This number is adjusted to necessary fault detection sensitivity (fine tuning). Forgetting factor is needed (e.g. in form of resetting fault counter after a period of time). Combining thresholds and the number of threshold exceedances within specified time frame, Fig. 26, distinguishing between sporadic and frequent events, appropriate sensitivity could be achieved.

VIII. CONCLUSION

The graphic engine monitor is a tool for engine management that contributes to proactive maintenance. Exploiting statistical properties of engine parameters may be useful when devising acceptable parameter limits for fault detection. Statistical summary of engine parameters in forms of percentiles can be used for describing complex empirical distributions in a simple way. One fits all statistic of engine parameters acquired by analyzing the whole engine log is suitable for devising wide parameter limits for warning (severe) alert. For detection of smaller problems more precise engine parameter limits are needed that are related to a particular engine operating regime. Combining engine RPM with the additional GPS data, actual flight phase can be determined and used for separation of engine operation into regime corresponding to a flight phase. Within a particular regime engine parameters can be compared to much narrower limits, that if exceeded issue caution (less severe) alert. Use of flight phase for regime determination has an advantage of documented engine parameter values that are expected in various phases of flight. Temperature profile comparison with stored profiles (patterns) for each flight phase is included for addition detection of temperature anomalies. Percentile limits are often too sensitive for detection of potential faults; hence permitted number of threshold exceedances within a specific time period may be introduced to achieve desired fault detection sensitivity.

REFERENCES

[1] M. Bush, Understanding CHT and EGT, *Cessna Pilots Association Magazine*, April 2009

[2] *Pilot's Guide: Engine Data Management, EDM-700, EDM-800, EDM-711 Primary*, J. P. Instruments, California, USA, 2007

[3] D. Miljković, Engine Monitors for General Aviation Piston Engines, *CrSNDT Journal*, No. 10, 2013, pp. 19-23

[4] *Pilot's Guide: Graphic Engine Monitor Data Logging System*, Insight Avionics, USA, 1995

[5] *Ultimate Bar Graph Engine Analyzer (UBG-16) Operating Instructions*, Electronic International, Oregon, USA, 1997

[6] *Pilot's Guide: Engine Data Management, EDM-730, EDM-830, EDM-740 Experimental*, J. P. Instruments, California, USA, 2010

[7] *Pilot's Guide, EzTrends*, J. P. Instruments, 2006

[8] D. Miljković, "Statistical Properties of Aircraft Piston Engine Monitor CM Data", MATEST 2013, Nov. 2013, Zagreb, Croatia

[9] D. Miljković, "Engine Fault Detection for Piston Engine Aircraft", MIPRO 2013 – CTS 2013, 20-24 May 2013, Opatija, Croatia

[10] D. Miljković, "Fault Detection Methods: A Literature Survey", MIPRO 2011, Vol. III, CTS&CIS, 23-27 May 2011, Opatija, Croatia

Konfiguriranje Profinet I/O mreže u radu sa energetskim pretvaračem

Goran Malčić, Danijel Maršić, i Julije Čučuk
Tehničko veleučilište u Zagrebu
goran.malcic@tvz.hr

Sažetak - Rad u realnom vremenu je jedan od osnovnih uvjeta koji računalna mreža mora zadovoljiti da bi bila primjenjiva u industrijskom okruženju. Upravo taj uvjet uz određene adaptacije je moguće postići sa primjenom ethernet tehnologije, te se u tom dijelu vidi najveći napredak. U ovom radu prikazan je stručni istraživački rad na implementaciji industrijske računalne mreže Profinet I/O na primjeru povezivanja pretvarača napona i frekvencije sa dvije različite izvedbe nadređenih industrijskih računala tvrtke Siemens. Konfiguracija i upravljački program su izvedeni u programskim okruženjima STEP 7 te TIA Portal V12 kako bi se uočila razlika u implementaciji. Opisan je i prikazan na primjeru način upravljanja i konfiguracije komunikacije pretvarača ACS880 (ABB) sa nadređenim PLC uređajem u dva slučaja – PLC sa integriranim Ethernet portom te PLC uređaj sa zasebnim komunikacijskim modulom.

I. UVOD

Razloge zbog kojih je ethernet mrežna tehnologija važna za industriju i zbog kojih se zadnjih desetak godina probija u područje industrijskih računalnih mreža treba tražiti u njenoj relativno jednostavnoj primjeni i sveopćoj prisutnosti u društvu. Ethernet je sigurno najrašrenija mrežna tehnologija koja se danas aktivno koristi u svijetu - relativno stabilna, vrlo razvijena te dobro poznata sklopovska i programska infrastruktura koja se svakim danom nadograđuje novim proizvodima i idejama. Osim toga, postala je u širkom smislu riječi 'svakodnevna' mrežna infrastruktura jer osim što je svugdje prisutna, za konfiguraciju takve računalne mreže nisu potrebna potrebna velika stručna znanja i zato ih posjeduje veliki broj ljudi.

Naravno, ideja jedne mrežne infrastrukture na nivou cijele tvornice, počevši od pogonske do upravljačke razine, prisutna je odavno (npr. Siemens mrežna struktura ProfiBus tehnologije PA, DP, FMS [4]). Naravno, iznad tih razina bila je uvijek prisutna ethernet tehnologija koja se neovisno razvijala u odnosu na pogonske odnosno industrijske računalne mreže. Prije desetak godina na tržištu su se pojavili prvi proizvodi koji su nudili kompletna riješenja industrijskih računalnih mreža temeljenih na ethernet sklopovlju odnosno infrastrukturi, ali sa vlastitim protokolima (npr. AllenBradley Ethernet/IP tehnologija [5]). Naime, temeljna zapreka kod razvoja etherneta na pogonskom nivou je uvjet da mreža mora raditi unutar kritičnog vremena (eng. Real time). Realno vrijeme znači da sustav procesuira vanjske događaje unutar definiranog vremena [1][2]. Definirano vrijeme odziva je uglavnom ≤ 5ms za standardne aplikacije.

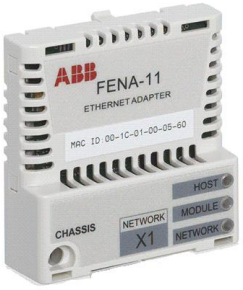
Tvrtka Siemens razvila je svoju tehnologiju Profinet (eng. *PROcess FIeld NET*) koja se bazira na Ethernet tehnologiji i na korištenju standarda. Profinet nije samo robusnija inačica Etherneta prilogođena za rad u industrijskim uvjetima. Inteligentni vanjski uređaji daju veliku količinu podataka za prijenos koja prelazi mogućnosti postojećih sabirničkih mreža i time ovoj tehnologiji čak daje prednost u odnosu na postojeće pogonske mreže. Profinet koristi optimizirani komunikacijski kanal za komunikaciju u realnom vremenu, što garantira prijenos vremensko kritičnih podataka između različitih stanica u mreži, unutar definiranog intervala. Real-time kanal može se implementirati na standardnim Ethernet kontrolerima primjenom softverskih riješenja ili u formi specijalnog hardverskog riješenja, a bazira se na 2 sloju ISO/OSI referentnog modela (Real-time protokol - RTC 1-3). Za potrebe Profinet real-time komunikacija kod procesa gdje je potrebna vrlo brza reakcija sustava upravljanja, nužan je bio razvoj ERTEC 400 i 200 (eng. *Enhanced Real Time Ethernet Controller*) ASIC procesora.

U literaturi se navodi da je rad sa Profinet tehnologijom moguć nabavom palete Siemens proizvoda te se dobiva dojam kako je to relativno zatvorena tehnologija. Zbog toga je ovom radu prikazano povezivanje pretvarača napona i frekvencije tvrtke ABB na PLC računalo serije S7-300 preko Profinet mreže. Prema uputama proizvođača ABB, takvo povezivanje i rad u realnom vremenu je moguć samo ako posjedujete PLC S7 sa integriranim Profinet komunikacijskim portom na procesoru. Takav sustav je objašnjen u prvom dijelu rada i svi problemi koji dolaze jer oprema nije od tvrtke Siemens, a ABB uputstva su bila nedostatna. Nadalje, ono što se nigdje ne navodi kao moguće, a niti se ne objašnjava slučaj kada posjedujete integrirani Profinet port na samom procesoru ili što učiniti ako već imate postojeći upravljački sustav kojemu uz procesor S7 serije dodajete zasebnu Profinet komunikacijsku karticu. Takav pristup je opisan u drugom dijelu rada i pokazuje da se može upogoniti i takav sustav.

Sustav je u oba slučaja izveden koristeći Profinet I/O komunikacijski koncept za implementaciju modularnih, distribuiranih aplikacija baziranih na industrijskom Ethernetu.

II. KOMUNIKACIJSKI MODUL FENA-11

Kako bi se pretvarač napona i frekvencije ACS 880 (ABB) mogao spojiti na ethernet mrežu potrebno je ugraditi komunikacijski modul FENA-11 za ethernet mrežu. Modul je kompatibilan s mrežnim standardima IEEE 802.3 i IEEE 802.3u i podržava tri protokola: *Modbus/TCP*, *EtherNet/IP* i *PROFINET IO*. FENA-11 modul kao medij podržava upletene parice čiji odsječak mora biti kraći od sto metara. Komunikacijski modul FENA-11 služi za slanje upravljačke riječi pretvaraču (start, stop), zadavanje reference pretvaraču, čitanje statusnih informacija i stvarne vrijednosti pretvarača te uklanjanje grešaka na pretvaraču. Komunikacijski modul podržava 10 Mb/s i 100 Mb/s brzine prijenosa podataka te automatski detektira korištenu brzinu u mreži. Na modulu se nalaze tri LED indikatora (HOST, MODULE, NETWORK) koji služe za dijagnostiku (Slika 1.).

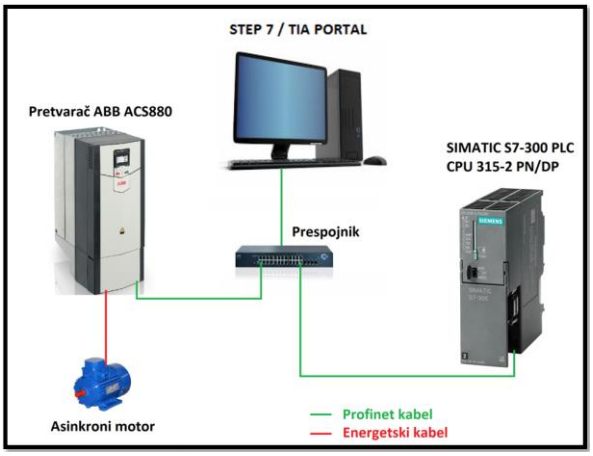


Slika 1. Komunikacijski modul FENA-11

III. S7-300 SA INTEGRIRANIM ETHERNET PORTOM

A. *Idejno rješenje*

U prvoj izvedbi sustava ostvareno je povezivanje nadređenog PLC uređaja sa integriranim Ethernet portom (CPU 315-2 PN/DP) te pretvarača napona i frekvencije ACS880 preko industrijske računalne mreže Profinet I/O. Razlog korištenja ovakve konfiguracije (CPU 315-2 PN/DP) u prvoj izvedbi sustava jest integrirani Ethernet port na samoj procesorskoj jedinici, što omogućava direktnu komunikaciju s podređenim uređajem (pretvaračem frekvencije) preko Profinet I/O protokola. To je ujedno i izvedba principa komunikacije predložena u priručniku proizvođača komunikacijskog modula FENA-11. Sve IP adrese u sustavu moraju biti statičke.



Slika 2. Idejno rješenje povezivanja prve izvedbe sustava

B. *Princip komunikacije*

Pretvarač frekvencije ACS880 može za komunikaciju sa nadređenim PLC uređajem koristiti više struktura procesnih podataka (eng. *Parameter Process data Object*, skraćeno *PPO*) definiranih prema PROFIdrive profilu 2.0. U ovom slučaju za prijenos podataka između PLC uređaja i pretvarača frekvencije koristi se *PPO Type 7* struktura podataka.

Odabrana struktura procesnih podataka (PPO) ima 12 procesnih riječi (PZD). Izlazne riječi služe za upravljanje radom, dok ulazne služe za nadgledanje statusa pretvarača frekvencije. Opis strukture procesnih podataka PPO TYPE 7 korištenih u ovoj izvedbi je prikazana na Slici 3.

	ULAZNI SMJER PODATAKA		IZLAZNI SMJER PODATAKA	
PROCESNA RIJEČ	OPIS	ADRESA	OPIS	ADRESA
PZD1	Statusna riječ	PIW256	Upravljačka riječ	PQW256
PZD2	Stvarna brzina vrtnje	PIW258	Referenca brzine	PQW258
PZD3 PZD4	Frekvencija motora	PIW262	Ne koristi se	
PZD5 PZD6	Struja motora	PIW266	Ne koristi se	
PZD7 PZD8	Moment motora	PIW270	Ne koristi se	
PZD9 PZD10	Snaga motora	PIW274	Ne koristi se	
PZD11 PZD12	Temperatura okoline	PIW278	Ne koristi se	

Slika 3. Korištena struktura procesnih podataka (PPO) Type 7

Upravljačka riječ je skupina od 16 internih binarnih signala, čija stanja određuje program PLC uređaja, koji moraju biti dostupni internoj sekvenci sustava upravljanja odnosno regulatoru pretvarača napona i frekvencije. U ovom rješenju adresa MW100 sadrži upravljačku riječ čiji pojedinačni bitovi imaju značenje prikazano na Slici 4.

Struktura upravljačke riječi PZD1				
bajt	bit	Opis značenja pojedinačnog bita	BIN	HEX
M101	0	Upravljački signal OFF1	0	E
	1	Upravljački signal OFF2	1	
	2	Upravljački signal OFF3	1	
	3	Omogući rad odabranog upravljačkog signala	1	7
	4	Omogući generiranje rampe	1	
	5	Pokreni generator rampe	1	
	6	Omogućavanje reference	1	
	7	Reset greške	0	
M100	0	Nije podržano	0	4
	1	Nije podržano	0	
	2	Omogući upravljanje putem mreže	1	
	3	Korisnički specificirani bitovi	0	
	4	Korisnički specificirani bitovi	0	0
	5	Korisnički specificirani bitovi	0	
	6	Korisnički specificirani bitovi	0	
	7	Korisnički specificirani bitovi	0	

Slika 4. Struktura upravljačke riječi PZD1

Za upravljanje pogonom putem Profinet I/O mreže koriste se digitalni upravljački signali OFF1 do OFF3. U ovom slučaju za pokretanje i zaustavljanje motora prema rampi koristit će se pozitivni brid upravljačkog signala ON/OFF1 što znači da je upravljačke signale OFF2 i OFF3 potrebno deaktivirati postavljanjem u logičko stanje 1. Jednako tako, potrebno je i neke od ostalih bitova upravljačke riječi postaviti u logičko 1 kako bi pretvarač bio spreman za pokretanje. Kako bi se to postiglo u upravljačku riječ mora biti upisana kombinacija $047E_{hex}$ što u stanje logičkog 1 postavlja pojedinačne bitove kako je prikazano na Slici 4.

U ovom rješenju se statusnu riječ može nadgledati i analizirati korištenjem adrese MW200 čiji pojedinačni bitovi imaju značenje prikazano na Slici 5.

Struktura ulazne statusne riječi PZD1		
bajt	bit	Opis značenja pojedinačnog bita
M201	0	Pogon spreman (Ready)
	1	Pogon spreman za rad
	2	Pogon u vrtnji (Running)
	3	Pogon u grešci
	4	OFF2 aktivan
	5	OFF3 aktivan
	6	Zabrana pokretanja aktivna
	7	Aktivno upozorenje pogona
M200	0	Referenca dostignuta
	1	Uključen daljinski rad pretvaračem (Remote ON)
	2	Max. Frekvencija dostignuta
	3	Korisnički specificirani bitovi
	4	Korisnički specificirani bitovi
	5	Korisnički specificirani bitovi
	6	Korisnički specificirani bitovi
	7	Korisnički specificirani bitovi

Slika 5. Struktura ulazne statusne riječi PZD1

C. *Implementacija u STEP 7/TIA Portal V12*

Prilikom testiranja sustava utvrđeno je da ne postoje značajnije razlike u postupku implementacije programskog rješenja sustava između programskih paketa *STEP 7* i *TIA Portal V12*. Zbog toga je postupak implementacije prikazan u razvojnoj okolini *TIA Portal V12* s obzirom da se radi o novijoj generaciji programske podrške za sustave automatizacije tvrtke *Siemens*.

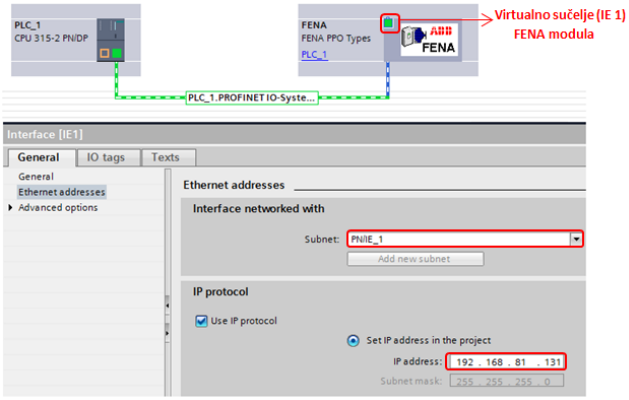
Prije izrade programske podrške u *TIA Portal-u* potrebno je prvo na energetskom pretvaraču podesiti parametre za komunikaciju preko Profinet I/O mreže. Na taj način će pretvarač biti vidljiv nadređenom PLC uređaju na mreži odnosno u upravljiv preko Profinet I/O mreže. Parametri pretvarača za mrežnu komunikaciju su podešeni sukladno uputama u priručniku proizvođača.

Nakon izrade novog projekta u *TIA Portal-u* potrebno je odabrati korištenu procesorsku jedinicu, a potom u prozoru za izradu sklopovske konfiguracije podesiti parametre ethernet porta. U ovom rješenju odabrana je IP adresa 192.168.81.130, maska podmreže 255.255.2550 i naziv uređaja na mreži *pn-io*.

Kako je PLC uređaj u stvarnosti jednim krajem spojen na (**X2**) ethernet sučelje procesorske jedinice a sa druge strane na komunikacijski modul pretvarača, potrebno je u prozoru sklopovske konfiguracije prvo definirati virtualnu Profinet I/O mrežu na koju se potom dodaje virtualni komunikacijski modul pretvarača FENA-11. Prije toga je potrebno instalirati upravljački program (GSD datoteku) za Profinet I/O komunikacijski modul FENA-01/11 preuzet od njegovog proizvođača (ABB).

Kako bi komunikacijski modul FENA-11 mogao komunicirati sa PLC uređajem potrebno je podesiti i njegove parametre. U ovom rješenju odabrana je IP adresa 192.168.81.131, maska podmreže 255.255.255 i naziv uređaja na mreži *fena*. Važno je napomenuti da IP adresa PLC uređaja i adresa komunikacijskog modula FENA-01/11 mora biti u istoj mreži, te da IP adresa u sklopovskoj konfiguraciji mora odgovarati IP adresi podešenoj u parametrima pretvarača.

Kako bi se Profinet I/O parametri dodijelili komunikacijskom modulu FENA-01/11 potrebno je pritisnuti na virtualno sučelje FENA modula (*IE 1*) te u *Ethernet addresses* postavkama odabrati izrađenu virtualnu Profinet I/O mrežu (PN/IE_1) i upisati IP adresu nadređenog kontrolera (192.168.81.130). Kao što se može primijetiti na Slici 6. maska podmreže mora biti ista kao i na procesorskoj jedinici (255.255.255.0) te se ne može mijenjati.



Slika 6. Dodjeljivanje Profinet I/O parametara modulu FENA-11

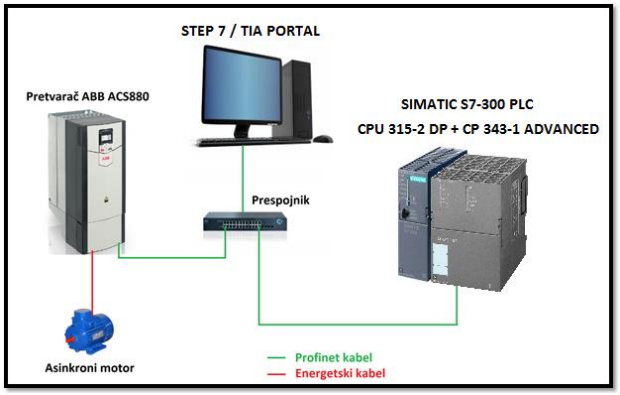
Posljednji korak u konfiguriranju pretvarača frekvencije za rad i slušanje zahtjeva sa Profinet I/O mreže je definiranje ulazno – izlaznih adresa pretvarača kojima će se slati odnosno primati poruke odabrane strukture podataka. Dvostrukim klikom na FENA modul otvara se ekran za pregled uređaja. Pomoću „povuci/ispusti“ metode umeće se već opisana PPO Type 7 struktura podataka u FENA komunikacijski modul.

Zadnji korak je izrada upravljačkog programa u organizacijskom bloku OB1. Upravljačka (PQW256) i statusna riječ (PIW256) se nalaze u perifernoj izlaznoj (PQ) odnosno ulaznoj memoriji (PI) PLC uređaja kod kojih nije moguće pristupati pojedinačnim bitovima u ladder dijagramu. Kako bi se to omogućilo, potrebno je uporabom naredbe za premještanje podataka (MOVE) prebaciti sadržaj upravljačke i statusne riječi iz periferne memorije u internu bit memoriju PLC uređaja kod koje je moguće pristupati pojedinačnim bitovima.

IV. S7-300 SA ZASEBNIM KOMUNIKACIJSKIM MODULOM

A. *Idejno rješenje*

U drugoj izvedbi sustava ostvareno je povezivanje nadređenog PLC uređaja sa zasebnim komunikacijskim modulom (CP 343-1 Advanced) i pretvarača napona i frekvencije ACS880 preko industrijske računalne mreže Profinet I/O (Slika 7.). Za centralnu procesorsku jedinicu odabran je CPU 315-2DP iz serije SIMATIC S7-300 serije koji ima integrirani PROFIBUS port. Kako bi CPU 315-2DP mogao komunicirati preko Profinet I/O mreže korišten je zasebni komunikacijski modul CP 343-1 Advanced koji služi kao Profinet I/O kontroler za operacije sa distribuiranim I/O uređajima na mreži.



Slika 7. Idejno rješenje povezivanja druge izvedbe sustava

B. *Princip komunikacije*

Kako bi PLC uređaj mogao komunicirati preko zasebnog komunikacijskog modula sa pretvaračem frekvencije potrebno je u upravljačkom programu pozvati komunikacijske blokove:

- FC11 „PNIO_SEND“ (za slanje podataka),
- FC12 „PNIO_RECV“ (za primanje podataka).

Karakteristike komunikacijskih blokova FC11 i FC12 su sljedeće:

- FC11 i FC12 su sinkroni komunikacijski blokovi,
- Za S7-300 kontrolere blokovi se nalaze u „SIMATIC_NET_CP“ biblioteci,
- Blokovi se moraju pozvati u glavnom organizacijskom bloku OB1,
- Završetak operacije prikazan je statusima „DONE“, „NDR“ ili „ERROR“

Zadaća komunikacijskoga bloka FC11 je prenositi podatke iz korisničkog programa centralne procesorske jedinice u modul CP343-1 Advanced. Blok prenosi podatke namijenjene izlaznim uređajima (eng. *IO devices*) iz zadanih izlaznih adresa CPU-a u CP modul. Kada se FC11 poziva u OB1 potrebno je unijeti parametre objašnjene na Slici 8.

Ime	Vrijednost	Komentar
CPLADDR	W#16#100	Početna izlazna adresa CP modula zapisana u hexadecimalnom obliku
MODE	B#16#0	Kako se CP modul koristi samo kao Profinet I/O kontroler MODE = 0
LEN	24	Broj bajtova koje treba poslati iz područja definiranog u varijabli SEND (**BYTE 24**)
SEND	P#**DB1**.**DBX0.0** **BYTE 24**	**DB1** – Podatkovno područje u koje se spremaju izlazne PZD riječi **DBX0.0** - Početna adresa izlaznih adresa pretvarača **BYTE 24** – Veličina izlaznih adresa pretvarača
IOCS	P#**DB2**.**DBX0.0** **BYTE 4**	**DB2** – Podatkovno područje u koje se spremaju informacije o stanju izlaznog područja **DBX0.0** – Početak adresa informacije **BYTE 4** – se dobiva iz formule (LEN/8+1)
DONE	M10.0	Bit koji prikazuje da li je slanje uspješno
ERROR	M10.1	Bit koji prikazuje da li je došlo do greške
STATUS	MW12	Statusna adresa za dijagnostiku i greške
CHECK_IOCS	M10.3	Pomoćni bit koji pokazuje treba li IOCS biti razmotren

Slika 8. Objašnjenje izrađenog komunikacijskog bloka FC11

Komunikacijski blok FC12 prima podatke iz modula CP343-1 Advanced i upisuje ih u korisnički program u CPU-u. Blok prihvaća procesne podatke ulaznih uređaja kao i informaciju o stanju od CP modula i pohranjuje u zadano ulazno područje CPU-a. Kada se FC12 poziva u OB1 potrebno je unijeti parametre objašnjene na Slici 9.

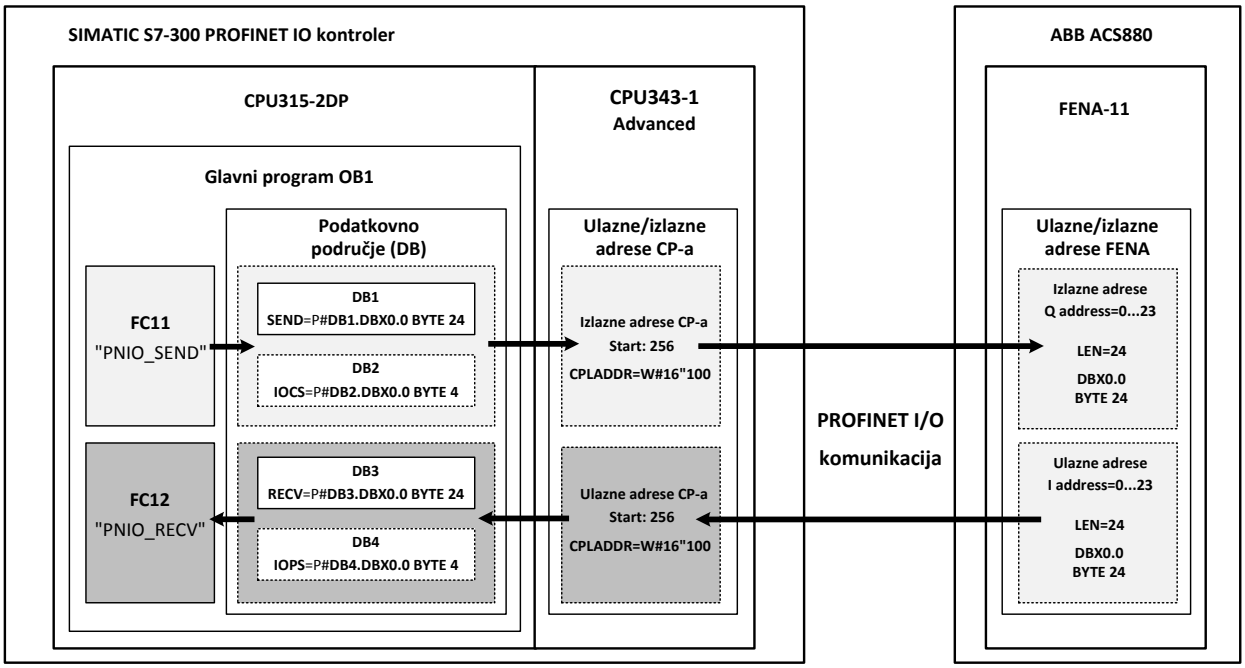
Ime	Vrijednost	Komentar
CPLADDR	W#16#100	Početna ulazna adresa CP modula zapisana u hexadecimalnom obliku
MODE	B#16#0	Kako se CP modul koristi samo kao Profinet I/O kontroler MODE = 0
LEN	24	Broj bajtova koje treba prihvatiti iz područja definiranog u varijabli RECV (**BYTE 24**)
RECV	P#**DB3**.**DBX0.0** **BYTE 24**	**DB3** – Podatkovno područje u koje se spremaju ulazne PZD riječi **DBX0.0** - Početna adresa ulaznih adresa pretvarača **BYTE 24** – Veličina ulaznih adresa pretvarača
IOPS	P#**DB4**.**DBX0.0** **BYTE 4**	**DB4** – Podatkovno područje u koje se spremaju informacije o stanju izlaznog područja **DBX0.0** – Početak adresa informacije **BYTE 4** – se dobiva iz formule (LEN/8+1)
NDR	M14.0	Bit koji prikazuje da li je slanje uspješno
ERROR	M14.1	Bit koji prikazuje da li je došlo do greške
STATUS	MW16	Statusna adresa za dijagnostiku i greške
CHECK_IOPS	M14.2	Pomoćni bit koji pokazuje treba li IOPS biti razmotren
ADD_INFO	MW18	Dodatna dijagnostička informacijska adresa

Slika 9. Objašnjenje izrađenog FC12 komunikacijskog bloka

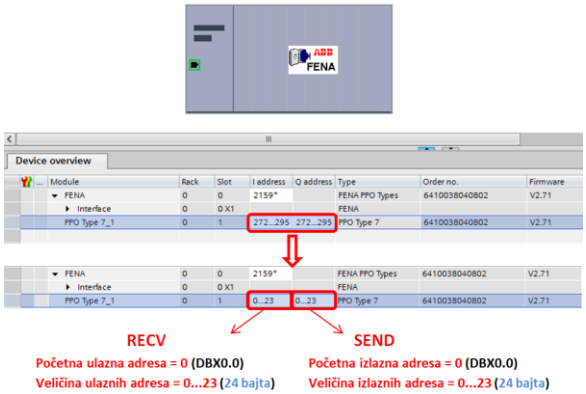
Za komunikaciju u sustavu koristi se *PPO Type 7* struktura podataka, isto kao u slučaju konfiguracije S7-300 sa integriranim portom, ali se razmjena podatka vrši preko komunikacijskih blokova FC11 i FC12 [6]. Slika 10. prikazuje principnu blok shemu PROFINET I/O komunikacije u sustavu. Kako bi pomoću komunikacijskog bloka FC11 iz glavnog programa OB1 mogli poslati podatak komunikacijskom modulu FENA-11 potrebna su dva podatkovna područja unutar procesorske jedinice. Podatkovni blok DB1 (SEND) služi za upisivanje procesnih PZD riječi dok podatkovni blok DB2 (IOCS) služi za spremanje informacija o stanju izlaznog područja. Prilikom upisivanja vrijednosti varijable SEND potrebno je definirati podatkovni blok (u ovom slučaju DB1), početak izlazne *Q* adrese FENA modula (DBX0.0) te njihovu ukupnu veličinu (BYTE 24) što se zajedno zapisuje u obliku pokazivača P#DB1.DBX0.0 BYTE 24. Komunikacijski procesor CP343-1 Advanced preuzima podatke iz podatkovnog bloka DB1 i preko svojih izlaznih adresa ih prosljeđuje FENA-11 komunikacijskom modulu pretvarača. Parametar CPLADDR označava početne izlazne adrese CP modula zapisane u heksadecimalnom obliku. Kako je početna izlazna adresa 256, potrebno ju je pretvoriti u heksadecimalni broj što se zapisuje kao W#16#100. Modul FENA-11 dobiveni podatak od strane CP modula obrađuje i šalje upravljačkoj jedinici pretvarača frekvencije ACS800. Komunikacijski blok FC12 za primanje podataka od strane CP modula djeluje na istom principu kao i FC11, jedino što se podaci šalju iz ulaznih *I* adresa FENA modula u ulazne adrese CP modula. Nakon toga se podaci prosljeđuju podatkovnom bloku (DB3) u procesorskoj jedinici, dok podatkovni blok DB4 služi samo za spremanje informacija o stanju ulaznog područja. Ostali parametri (NDR, ERROR, STATUS. Itd.) u obje funkcije služe samo nadgledanje grešaka u komunikaciji.

C. *Implementacija u STEP 7/TIA Portal V12*

Postupak sklopovske konfiguracije u programskim paketima *STEP 7/TIA Portal V12* sadrži sve korake koji su potrebni kod rada sa integriranim Profinet portom. Naravno, u ovom slučaju se sva podešenja vrše na zasebnom komunikacijskom modulu CP343-1 Advanced za koji se definira virtualna Profinet I/O mreža i dodaje virtualni komunikacijski modul pretvarača FENA-11. Kako se koristi zasebni komunikacijski modul početne ulazne i izlazne adrese pretvarača frekvencije moraju biti postavljene na način da kreću od 0 (Slika 11.).



Slika 10. Principna blok shema Profinet I/O komunikacije u sustavu



Slika 11. Definiranje ulazno – izlaznih adresa pretvarača (PPO 7)

Upravljački program sadrži komunikacijski blok FC11 koji koristi (DB1) podatkovno područje za spremanje izlaznih PZD riječi sa pretvarača veličine 24 bajta i (DB2) za spremanje informacije o stanju izlaznog područja pretvarača veličine 4 bajta. Blok FC12 blok koristi (DB3) za spremanje ulaznih PZD riječi sa pretvarača veličine 24 bajta i (DB4) za spremanje informacije o stanju ulaznog područja pretvarača veličine 4 bajta.

V. ZAKLJUČAK

S obzirom na u uvodu navedene prednosti etherneta, primjena Profinet mrežne tehnologije u izgradnji industrijske računalne mreže na pogonskom nivou trebao bi biti relativno jednostavan postupak primjenjujući norme i poštivajuću uputstva proizvođača opreme. U ovom radu je prikazano kako podrška, odnosno uputstva za rad sa ovakvim sustavima su više nego nedostatna. Zbog toga je provedeno istraživanje Profinet tehnologije i njezinog načina rada tražilo je desetke sati rada u traženju riješenja koje niti jedan proizvođač ne daje.

U prvom dijelu rada je prikazan sustav koji bi se trebao moći konfigurirati i upogoniti u vremenu od najviše par sati, što se pokazalo nemoguće. U drugom dijelu rada je prikazano koliko je za naizgled jednostavan sklop, koji se razlikuje od prvoga samo po odvojenoj jedinici za komunikaciju, potrebno uložiti vremena i znanja da bi sustav proradio. Takva rješenja bi trebala biti jednostavna, nadogradnje sustava omogućene i objašnjene od strane proizvođača opreme, no ovime se prikazalo da to nije tako. No ipak, i jedan i drugi proizvođač su izradili mrežne sklopove i programsku podršku koja je definirana Profinet normom što dokazujemo ispravnim radom sustava. Dakle poštivanje normativa neminovno mora dati riješenje, koje se na kraju pronašlo te je u radu objašnjeno. Sama primjena u budućim aplikacijama se može referencirati na ovaj rad što je naš doprinos.

VI. ZAKLJUČAK

[1] Pigan Raimond; Metter Mark: Automatisieren mit PROFINET: Industrielle Kommunikation auf Basis von Industrial Ethernet; Publicis Corporate Publishing; Auflage: 2. überarbeitete Auflage (15. Februar 2008)

[2] pi_white_paper_profinet_it_en_v1_0.pdf;PROFIBUS Nutzerorganisation e.V.; December 2008

[3] B01_PROFINET_system_en.pdf; PROFINET System Description Version April 2009; PNO 4.132

[4] Berger, Hans; Automating with SIMATIC; Publicis Corporate Publishing; 2nd revised edition 2003

[5] Goran Malčić, Ethernet/IP - Industrijski Ethernet, članak za stručni časopis Elektro, Zagreb 2004.

[6] Julije Čučuk: *Upravljanje pretvaračem ACS880 preko ProfiNet I/O mreže sa PLC uređajima tvrtke Siemens*, završni rad - diplomski/integralni studij, Elektrotehnički odjel Tehničkog veleučilišta u Zagrebu, 18.12. 2013

[7] B01_PROFINET_system_en.pdf; PROFINET System Description Version April 2009; PNO 4.132

[8] SIMATIC NET NCM for Industrial Ethernet, Siemens AG, 2005

Primjena SoftPLC koncepta procesnog računala prilagođene izvedbe s IEC61131-3 normom

Ivica Vlašić , Danijel Maršić i Goran Malčić
Tehničko veleučilište u Zagrebu/Elektrotehnički odjel, Zagreb, Hrvatska
ivica.vlasic@tvz.hr

Sažetak - Pristup projektiranju i izgradnji upravljačkih sustava u industriji se kontinuirano mijenja i poboljšava. Primjena IEC 61131-3 norme za programiranje PLC uređaja omogućila je razvoj novih vrsta upravljačkih sustava. Istovremeno proizvođači PLC-a prestaju s podrškom svojih starih sustava te prisiljavaju proizvođače opreme da koncepte svojih strojeva prilagođavaju novim sustavima. U radu je prikazana primjena SoftPLC rješenja tvrtke KW Software nazvano ProConOS® (eng. Programmable Controller Operating System) u tiskarskom postrojenju novije generacije. Osnovni koncept automatizacije tiskarskog postrojenja se nije mijenjao osim primjene prilagođenog (eng. custom) SoftPLC-a koji je naslijedio SIMATIC S5® sustav zadržavajući kompatibilnost u načinu primjene i održavanja.

I. UVOD

Pristup projektiranju i izgradnji upravljačkih sustava u industriji se kontinuirano mijenja i poboljšava. Paralelno s napretkom raznih tehnologija povećavaju se i dodatni zahtjevi kod razvoja upravljačkih sustava čime njihove mogućnosti postaju sve veće i veće. Kao odgovor na ove zahtjeve posljednjih dvadesetak godina je razvijeno nekoliko otvorenih međunarodnih standarda za industriju od kojih je najvažniji *IEC 61131-3* koji predstavlja skup smjernica za programiranje *PLC* (eng. ***P**rogrammable **L**ogic **C**ontroller*) uređaja. Primjena ovog standarda omogućila je razvoj novih vrsta upravljačkih sustava kao što je *SoftPLC* tehnologija.

Jedan od niza problema koji se javlja kod primjene *PLC* uređaja u industrijskim postrojenjima je nemogućnost miješanja uređaja različitih proizvođača zbog njihove međusobne nekompatibilnosti u pogledu izrade upravljačke logike. Iako su osnovne programske funkcije svih kontrolera praktički jednake, pojedine instrukcije su ipak bitno različite zbog čega se pravila pisanja programa mogu razlikovati. Ova činjenica bitno otežava razmjenu upravljačkih programa između uređaja odnosno sustava različitih proizvođača. Integratori sustava automatizacije zbog toga trebaju univerzalne alate za izgradnju svojih aplikacija da bi ih mogli koristiti na raznovrsnim PLC uređajima. Proizvođači rješenja za automatizaciju žele s jedne strane zadovoljiti potrebe kupaca (integratora sustava), dok s druge strane nastoje zadržati isplativu strukturu proizvoda u tehnološkom i poslovnom okruženju koje se mijenja iz dana u dan.

Navedeni problemi su tijekom 90-ih godina prošlog stoljeća potaknuli pojedine proizvođače programskih aplikacija na razvoj nove vrste kontrolera koji se naziva *SoftPLC*. Ovaj naziv predstavlja softver koji omogućuje izvršavanje upravljačka aplikacija u realnom vremenu na standardnoj računalnoj platformi. Pritom je upravljanje postrojenjem ostvareno preko standardiziranih modularnih *U/I* jedinica. Dakle, za implementaciju *SoftPLC* kontrolera potreban je hardver i softver kojima se postiže da standardna računalna platforma radi kao *PLC* uređaj. Osnova prednost *SoftPLC-a* je da sav potreban hardver nije ovisan o proizvođaču. Unatoč ovoj prednosti ni jedan proizvođač softvera nije uspio postaviti svoje rješenje *SoftPLC* softvera kao industrijski standard. To znači da programska podrška razvijena za jednu *SoftPLC* aplikaciju još uvijek ne može biti prenesena u *SoftPLC* nekog drugog proizvođača.

U radu je prikazan općeniti koncept *SoftPLC* tehnologije na primjeru rješenja *ProConOS* proizvođača *KW Software*. Također su prikazana višegodišnja iskustva u implementaciji i eksploataciji navedenog *SoftPLC* rješenja u tiskarskom postrojenju.

II. OSNOVE *IEC 61132-3* STANDARDA

IEC 61131-3 je naziv za prvi objavljeni globalni standard za programiranje PLC uređaja. Prvi puta je objavljen 1993. godine dok je trenutno aktualno drugo izdanje iz 2003. godine. Standardom su sažete potrebe industrijskih upravljačkih sustava kako u pogledu hardvera tako i softvera za programiranje. Smjernice i programski jezici *IEC 61131-3* standarda su primarno predviđeni za programiranje klasičnih PLC uređaja. Razvojem alternativnih rješenja industrijskih upravljačkih sustava pronašli su svoju primjenu u ugrađenim kontrolerima, programirljivim automatizacijskim kontrolerima (eng. ***P**rogrammable **A**utomation **C**ontrollers*, *PAC*) i upravljačkim sustavima baziranim na standardnim i specifičnim računalnim platformama (*SoftPLC*). Kako je standardom definiran ogroman broj detalja, od proizvođača upravljačkih sustava se ne može, u ovom trenutku, očekivati da će implementirati sve dijelove standarda, već samo određene dijelove. Dakle, standard osigurava mjerilo koje omogućuje kako proizvođačima tako i krajnjim korisnicima da procijene u kolikoj mjeri je svaki sustav za programiranje industrijskih kontrolera usklađen s *IEC 61131-3*.

Iako je većina proizvođača industrijskih upravljačkih sustava u velikoj mjeri implementirala određene smjernice *IEC* standarda, njihova hardverska i softverska rješenje se zbog komercijalnih i financijskih interesa u tolikoj mjeri razlikuju da je prijenos programa između

različitih upravljačkih platformi sve osim jednostavan. Ipak, primjenom standarda su elementi programskih jezika i struktura programa različitih proizvođača ipak postali slični što u konačnici integratorima sustava mnogo olakšava prebacivanje programa iz jednog u drugi upravljački sustav. Unatoč brojim preprekama koje stoje na putu provedbe osnovnih ciljeva *IEC 61131-3* standarda njegovim uvođenjem stvoreni su temelji za stvaranje jedinstvenog koncepta programiranja industrijskih kontrolera.

IEC 61131-3 standard se može općenito podijeliti na dva osnovna dijela:

A. *Zajednički elementi*

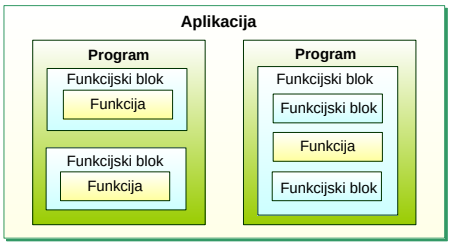
U ovom dijelu standarda su definirani sljedeći zajednički elemenati koji se mogu koristiti u svim programskim jezicima:

1) Oznake: služe za dodjeljivanje smislenih imena različitim elementima korištenih unutar IEC programskih jezika kao što su varijable, tipovi podataka, funkcijski blokovi i programi.

2) Tipovi podataka: predviđeno je nekoliko osnovnih i strukturiranih tipova podataka koje programer može definirati unutar programa.

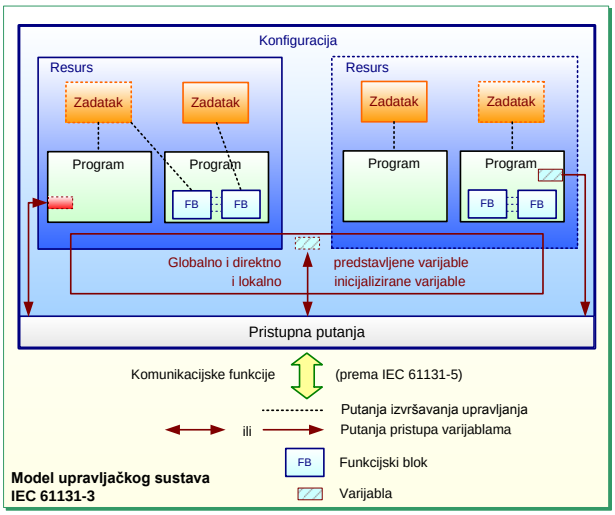
3) Varijable: su nazivi dodijeljeni podatkovnim elementima čiji se sadržaj može mijenjati tijekom izvođenja napisanog upravljačkog programa. Standardom su definirana tri tipa varijabli: lokalne, globalne i pristupne varijable.

4) Programske organizacijske jedinice: sve *IEC 61131-3* programske aplikacije se mogu razdijeliti u funkcionalne elemente nazvane organizacijske jedinice programa (eng. *Program Organisation Units, POU*) koje se mogu sastojati od funkcija, funkcijskih blokova ili cijelih programa (Slika 1.). Funkcije i funkcijski blokovi predstavljaju osnovu izgradnje kompleksnih upravljačkih programa jer sadrže strukture podataka i algoritme.



Slika 1. Primjeri upotrebe programskih organizacijskih jedinica

5) Softverski model upravljačkog sustava: je definiran kako bi se omogućilo bolje razumijevanje zajedničkih elemenata *IEC* programskih jezika (Slika 2.). Cjelokupna programska podrška koja služi za upravljanje određenim industrijskim postrojenjem se naziva konfiguracija (eng. *Configuration*). Konfiguracija je specifična za određeni upravljački uređaj jer su njezine mogućnosti temelje na njegovom hardveru o kojem ovisi brzina procesiranja, memorijske adrese ulaza i izlaza te ostale mogućnosti sustava. Unutar konfiguracije može se definirati jedan ili više resursa (eng. *Resource*) koji se mogu promatrati kao procesne jedinice sposobne izvršavati IEC programe. Svaki resurs može sadržati jedan ili više programa i/ili funkcijskih blokova koji se izvršavaju pod utjecajem upravljačkih zadataka (eng. *Tasks*). Zadatak se može izvršavati periodički ili pokretati pod utjecajem određenog okidača, primjerice uslijed pojave rastućeg brida neke varijable veličine bita. Program se može izgrađivati od niza različitih programskih elemenata napisanih u bilo kojem IEC programskom jeziku. Pristupna putanja definira simbolički naziv varijable dok direktno predstavljene varijable određuju fizičku lokaciju podataka (npr. hardverska adresa ulaznog signala) na način specifičan pojedinom proizvođaču. Globalne varijable su dostupne svugdje unutar resursa ili konfiguracije.



Slika 2. Softverski model industrijskog kontrolera

B. *Programski jezici*

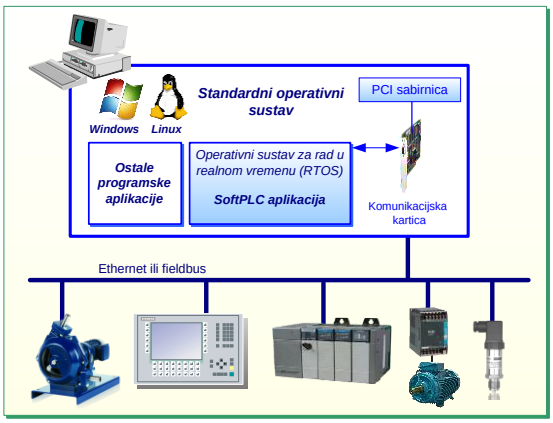
Standard podržava 5 programskih jezika što programerima PLC uređaja omogućava korištenje preferiranih jezika na osnovu prethodnih iskustava. Svaki od ovih programskih jezika ima prednosti u odnosu na ostale u specifičnim algoritmima upravljanja procesom. To su sljedeći programski jezici:

- Ljestvičasti dijagrami (eng. *Ladder Diagrams, LD*),
- Instrukcijska lista (eng. *Instruction List, IL*),
- Funkcijski blok dijagrami (eng. *Function Block Diagram, FBD*),
- Strukturirani tekst (eng. *Structured Text, ST*),
- Sekvencijalni funkcijski dijagrami (eng. *Sequential Function Charts, SFC*).

III. SOFTPLC TEHNOLOGIJA

Otvoreni standardi industrijskih upravljačkih sustava omogućili su primjenu *PC* (eng. ***P**ersonal **C**omputer*, dalje *PC*) računala za obavljanje raznih upravljačkih zadataka. U pojedinim industrijskim postrojenjima su paralelno uz primjenu klasičnih *PLC* uređaja pojavljuju koncepti upravljačkih sustava temeljenih na *PC* računalu. Rješenja takvih upravljačkih sustava razvijena su za razne operativne sustave (*Windows, Linux* i sl.) u kojima *PC* zamjenjuje klasični hardverski *PLC* uređaj. Upravljački sustavi temeljeni na *PC* računalima mogu biti izvedeni kao *SlotPLC* ili *SoftPLC*.

SoftPLC je naziv za programsku podršku kojom se postiže da standardna računalna platforma funkcionira kao *PLC* uređaj koji upravlja procesom u realnom vremenu (Slika 3.).



Slika 3. Koncept *SoftPLC* upravljačkog sustava

Komunikacija sa senzorima i aktuatorima u postrojenju je izvedena preko mreže odnosno njenih komunikacijskih sučelja spojenih na distribuiranu periferiju (fieldbus, ethernet i sl.). *SoftPLC* rješenje upravljačkog sustava je djelomično ovisno o pouzdanosti sklopovlja računala i nesmetanom funkcioniranju instaliranog operativnog sustava. *SoftPLC* aplikacija mora dijeliti resurse procesora s operativnim sustavom i eventualno s drugim instaliranim softverskim aplikacijama. Procesor PC računala mora izvršavati PLC program u realnom vremenu za što se koristi operativni sustav za rad u realnom vremenu (eng. *Real-Time Operating System, RTOS*) koji osigurava deterministički rad.

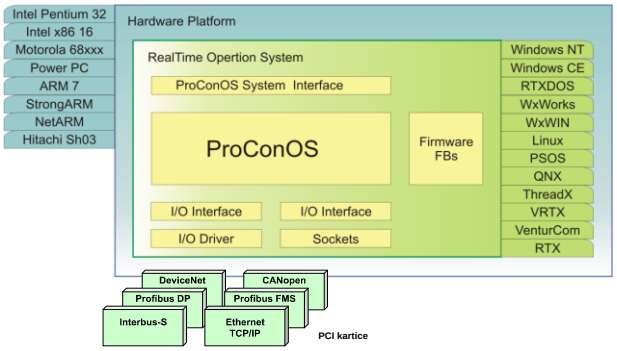
IV. *SoftPLC* RJEŠENJE *KWSoftware*

Na tržištu postoji nekoliko komercijalnih implemetacija *SoftPLC* tehnologije. Jedno od njih je prikazano rješenje *SoftPLC-a* tvrtke *Klöpper und Wiege Software*® (dalje *KW Software*) koja je razvila *PLC* operativni sustav nazvan *ProConOS*® (eng. ***Pro**grammable **Con**troller **O**perating **S**ystem*).

A. *Operativni sustav ProConOS*

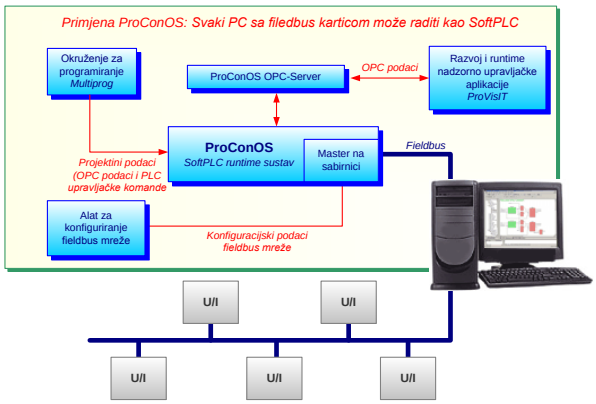
ProConOS je višezadaćni i vrlo brzi *PLC runtime*[1] sustav za kompleksne upravljačke aplikacije koji radi u realnom vremenu. Sukladan je s *IEC 61131-3* standardom i dizajniran tako da se može prilagoditi raznim standardnim ili specijalnim hardverskim platformama. Slika 4. prikazuje osnovni koncept *SoftPLC* rješenja proizvođača *KW Software*®.

KW Software® nudi samo odgovarajući *PLC runtime* softver (ProConOS®) dok je korisniku na raspolaganju velik broj hardverskih platformi na kojima može primijeniti taj softver. Dakle, posebnost ovog rješenja je u tome da korisnik može neovisno o proizvođaču softvera odabrati hardver koji će pretvoriti u softverski PLC uređaj, primjerice standardnu PC ili ugrađenu (eng. *embedded*) hardversku platformu. *ProConOS*® radi kao jedan od programa unutar jednog od mogućih operativnih sustava (npr. *Windows NT, Linux,* itd.) koji je instaliran na odabrani hardver. Kako bi *ProConOS*® mogao raditi deterministički potrebno je na operativni sustav instalirati odgovarajuću nadogradnju za rad u realnom vremenu. *KW Software* je razvio vlastita rješenja takvih dodataka kao primjerice *real-time* ekstenzija za Windows operativni sustav koja se naziva *KWRTK* (eng. *KW Realtime*[2] *Kernel*). Komunikacija s perifernim jedinicama se ostvaruje preko pripadajućih modula za podržane komunikacijske tehnologije (spojenih u PCI utore). *ProConOS*® podržava najčešće korištene sustave distribuiranih periferija kao što su *Interbus*®, *Profibus DP*®, *CANopen*®, *DeviceNet*®, *Modbus*® i sl. Otvoreno U/I sučelje nadalje omogućuje vezu preko raznih korisničko-specifičnih periferija.



Slika 4. Osnovni koncept *ProConOS* softverskog PLC-a

Slika 5. prikazuje funkcijski dijagram tipične implementacije *ProConOS*® softverskog *PLC-a* na standardnom računalu. Može se uočiti da svako računalo temeljeno na hardveru proizvođača Intel koje ima ugrađenu odgovarajuću fieldbus kraticu može raditi kao softverski PLC ako se na njega instalira *ProConOS runtime*® softver.



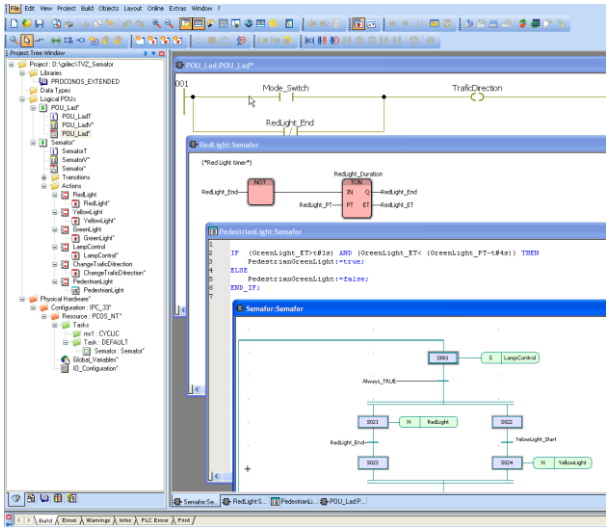
Slika 5. Primjer tipične implemetacije *ProConOS* softverskog PLC-a

Prednost ovog rješenja je u tome što se isto računalo može, osim za implementaciju softverskog PLC uređaja, koristi za ostale alate koji su obično potrebi za programiranje i rad PLC uređaja Na njemu se može nalaziti razvojno okruženje *KW Multiprog*® koje služi za razvoj i testiranje PLC programa. Jednako tako, isto

[1] *Runtime* softver koje je dizajniran kako bi podržavao izvršavanje računalnih programa napisanih u nekom programskom jeziku.

[2] eng. *Realtime* - u stvarnom vremenu

računalo može se koristi za razvoj i runtime izvođenje nadzorno upravljačkih aplikacija u programskom okruženju *ProVisIT®* koje sa softverskim PLC-om razmjenjuje podatke putem *ProConOS®* OPC Servera. Odgovarajuća fieldbus kartica (za *Profibus DP®*, *DeviceNet®* i sl.) ugrađena u PCI utore računala se konfigurira u alatu za konfiguriranje pripadajuće filedbus mreže koji se također može nalaziti unutar istog računala.



Slika 6. Izgled osnovnog ekrana okruženja *KW Multiprog*

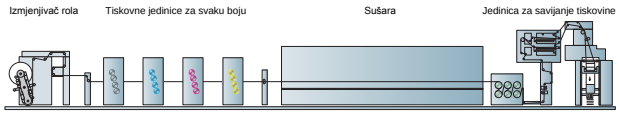
V. PRIMJENA *PROCONOS* U TISKARSKOM POSTROJENJU

A. *Zahtjevi automatizacije tiskarskog postrojenja*

Tiskarsko postrojenje *Rotoman®* za revijalni tisak je automat u koji kao sirovine ulaze papir i boja, a kao proizvod izlaze gotova tiskovina određenog tipa. U pogledu automatizacije možemo reći da se radi o kontinuiranom sustavu relativno velike brzine. Kapaciteti modernih postrojenja su od 30000 do 65000 primjeraka na sat što znači da se papirna traka prilikom tiskanja kreće brzinom od 8-15 m/s.

Slika 7. prikazuje blok shemu tiskarskog postrojenja koje se funkcionalno sastoji od sljedećih temeljnih cjelina:

- Opskrbljivač papirnom trakom (izmjenjivača rola) ima zadatak da osigura kontinuiranu opskrbu papirnom trakom tako da zamjeni istrošenu papirnu rolu novom bez zaustavljanja procesa. Dodatni zahtjevi su da se osigura varijabilna napetost papira kako bi se izbjeglo gužvanje.
- Tiskovne jedinice otiskuju sadržaj na papirnu traku i osiguravaju kvalitetu otiska koja se karakterizira kao precizan otisak pojedinačne boje (iz CYMK sustava komplementarnih boja) jedne preko druge. Osim primarnih postoje i niz sekundarnih zahtjeva kao što je opsluživanje prije tiska te zaštitni sustavi koji štite radni mehanizam od oštećenja, prilikom pucanja papira i sl., te zaštitu zdravlja i života osoblja.
- Sušara je dio postrojenja koji treba vrlo brzo osušiti boju nanesenu na papirnu traka te je isto tako brzo i kontrolirano ohladiti na izlazu. Toplinska energija se dobavlja gorenjem prirodnog plina. To podrazumijeva vrlo rigorozne zaštitne sustave budući da se razvijaju temperature od 800 ºC, a instalirana toplinska snaga je reda veličine 1 MW.
- Jedinica za savijanje (formiranje) tiskovine je sustav kompleksnih radnih mehanizama čija je zadaća da savijaju i odsijecaju papirnu traku tvoreći tako uobičajnu tiskovinu. Ovi radni mehanizmi su zbog relativno velikih brzina i složenih zahtjeva podložni pojačanom trošenju što donosi zahtjev za preciznim određivanje njihovog položaja i stanja da bi se ukupan sustav osigurao od lomova.



Slika 7. Blok shema tiskarskog stroja *Rotoman*

Zahtjevi automatizacije ovise o pojedinom segmentu postrojenja i oni se osiguravaju parcijalno dok je globalni zahtjev da se papirna traka kreće sinkronom brzinom kroz sve elemente. Ovaj zahtjev osigurava sofisticirani sustav motornog pogona kod kojeg nije samo zahtjev da se pojedini motori vrte sinkronom brzinom nego i da se osigura identični kut radnog mehanizma (eng. *Motion Control*).

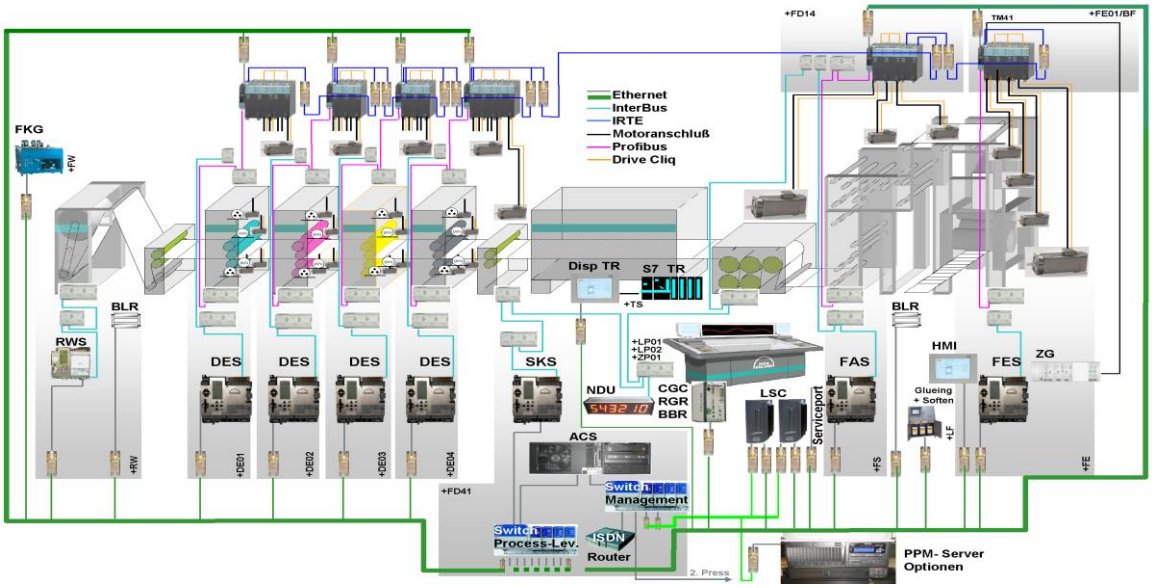
B. *Sustav automatizacije tiskarskog postrojenja*

Ako analiziramo navedene zahtjeva za automatizaciju tiskarskog postrojenja možemo kao primarne značajke primijetiti zahtjev za brzinu i preciznost te je logičan zaključak da se za ispunjenje istih primjeni sustav s više procesnih računala povezanih suvremenim industrijskim mrežama u procesnu cjelinu. Sustav automatizacije tiskarskog postrojenja *Manroland Rotoman* zadovoljava teorijske osnove modernih industrijskih mreža što se ogleda u strukturi umrežavanja na tri osnovne razine (Slika 8.).

Upravljačka razina se ostvaruje korištenjem standardne Ethernet mrežne infrastrukture koja spaja upravljačke konzole s procesnom razinom na jednoj strani te s radnom stanicom za organizaciju posla na drugoj. Ova razina omogućava implementaciju ukupnog postrojenja u informatički sustav poduzeća.

Procesna razina osigurava razmjenu podataka između bazičnih procesnih jedinica (izmjenjivač rola, tiskovne jedinice, sušara, jedinica savijanja i motorni pogon), a bazirana je na industrijskoj Ethernet mreži.

Razina polja zadužena je za procesnu sliku postrojenja. Ova razina izvedena je kao distribuirani sustav upravljanja svih aktuatora i senzora. Upotrijebljena je *Interbus-S* industrijska mreža s kojom se uspijeva osvježavati kompletna procesna slika brzinom od 500 kBd/s. *Interbus-S* je otvoreni industrijski mrežni standard koji je koncipiran kao serijski podatkovni prsten (eng. *Summation Frame Protokol*) koji radi isključivo u realnom vremenu u okviru master/slave metode kontrole prijenosa, a strukturiran je kao pozadinski posmični registar. Tijekom jednog mrežnog ciklusa serijski se isporučuju izlazni podaci članovima mreže ali se istovremeno od njih zaprimaju ulazni podaci.



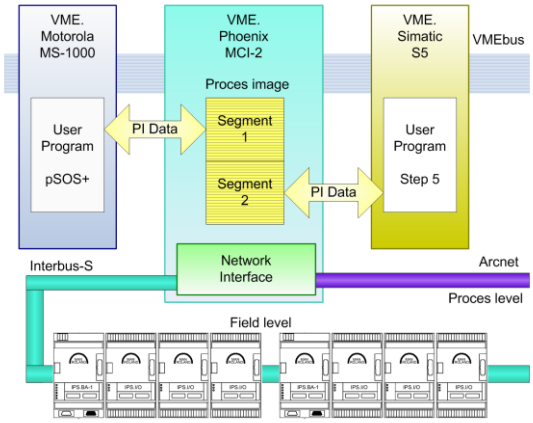
Slika 8. Industrijska mreža tiskarskog postrojenja *Rotoman*

Na kraju ciklusa svi izlazni podaci su distribuirani a ulazni upisani u ulaznu mapu slike procesa Koristeći prstenastu strukturu moguće je ostvariti istovremenu predaju i zaprimanje podataka (*full duplex*) a predvidljivo vrijeme pristupa za sve članove mreže je zajamčeno. Za potrebe svojih postrojenja tvrtka *Manroland* je konstruirala vlastite *Interbus-S®* inteligentne mrežne članove specifičnih karakteristika (*IPS-Interbus Press System*).

Motorni pogon je dosta zahtjevan dio automatiziranog tiskarskog postrojenja i izveden je korištenjem suvremenog sustava *SINAMICS*, tvrtke *Siemens*. Osnovu motornog pogona čine pretvarači frekvencije serije *SIMOTION D* koji su istovremeno i pogoni za elektromotore i posebno dizajnirana procesna računala s vlastitim programom sposobnim za održavanjem virtualne osovine tiskarskog postrojenja. S ostatkom postrojenja komuniciraju Siemensovom varijantom industrijske Ethernet mreža (***IRTE*** - ***I****socron* ***R****eal* ***T****ime* ***E****thernet*), a međusobno sustavom *Drive-CLiQ*.

C. SoftPLC rješenje modernizacije sustava

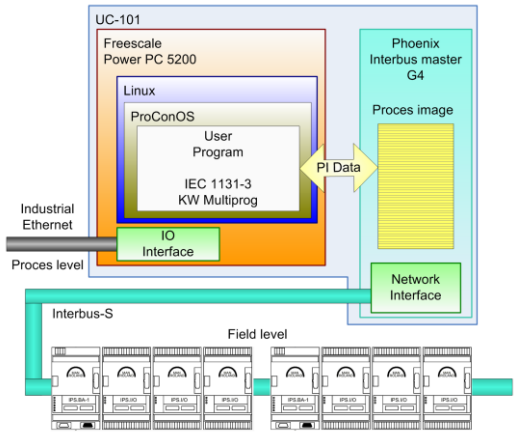
Razvoj tiskarskih strojeva u segmentima automatizacije je brži nego u području bazične mehanike radnih mehanizama što je logično gledajući tempo razvoja elektroničke industrije. Mnoga rješenja u segmentu automatizacije zastarijevaju s obzirom na performanse, ali i zbog prestanka proizvodnje i podrške osnovnih elemenata sustava automatizacije. Takav slučaj je i kod opisanog tiskarskog postrojenja čija je prethodna generacija bila automatizirana na bazi Siemensove PLC generacije *Simatic Step 5*® integriranih u VMEbus sustav sa Motorola 68000 CPU jedinicom, te *Phonix Interbus S*® kontrolerom distribuirane periferije (slika 9.). Programsko rješenje sastojalo se od dva segmenta. Prvi program upravljao je s negrafički procesnim elementima dok je drugi bio zadužen za procese grafičke obrade i sadržavao je relativne brze algoritme bazirane na matematičkim operacijama. Ovakva konfiguracija osiguravala je fleksibilna i napredna rješenje ali je problem nastao kad je obustavljana proizvodnja većine elementa.



Slika 9. Koncept procesnog računala UC-1 bazirano na VMEbus sustavu

Kako bi se zadržao kontinuitet u razvojnoj strategiji i smanjio budući utjecaj gore navedenih trendova kompanija je odlučila novu generaciju tiskarskih postrojenja automatizirati takozvanim *custom* rješenjima. Koristeći se pozitivnim efektima takvog pristupa pri razvoju vlastitih *Interbus-S*® modula odlučeno je da će se napraviti vlastito procesno računalo bazirano na otvorenom standardu. Rezultat toga je bio procesno računalo bazirano na *SoftPLC* rješenju *ProConOS*® nazvano *UnitControler UC*-101® (Slika 10.).

UC-101 je procesno računalo sklopovski koncipirano na *Power PC 5200* arhitekturi tvrtke *Freescale*® (Slika 11.). Kao *realtime* operativni sustav izabran je *Linux* kao još jedan element otvorenog koda. Ulazno-izlazni signali se kontroliraju preko *Interbus-S*® modula kao distribuirane periferije i stoga je u računalu integriran nadređeni Interbus-S član *(master)*. Računalo je smješteno u standardizirano kućište podešeno za montažu uz ostale IPS module.



Slika 10. Koncept procesno računalo UC-101 bazirano na custom SoftPLC sustavu

Programska oprema (software) uključuje *Xenomai/Linux* kao operativni sustav, *ProConOS®* kao *SoftPLC*, *KW Multiprog®* kao razvojno okruženje te *Phoenix Interbus CMD* program za konfiguriranje i dijagnostiku industrijske mreže.



Slika 11. Procesno računalo tiskarskog postrojenja UC-101

Pozitivni učinci primjene opisane tehnologije u prvom redu su, zahvaljujući konceptu distribuirane periferije, ne mijenjane osnovnog koncepta automatizacije stroja. U pogledu sklopovske implementacije dovoljna je samo zamjena procesnih računala što ne zahtijeva velike troškove redizajna sklopovske opreme. Ova činjenica ima veliki značaj kod rekonstrukcije i modernizacije postojećih tiskarskih postrojenje kada je uz cijenu vrlo bitno i vrijeme izvršenja radova.

VI. ZAKLJUČAK

SoftPLC rješenje tvrtke KW Software prikazano u radu je nazvano *ProConOS®* (**Pro**grammable **Con**troller **O**perating **S**ystem) jer predstavlja programsko bazirani PLC sustav koji se može implementirati na nizu standardnih ili specijalnih hardverskih platformi. Posebnost ovog rješenja je u tome da korisnik može neovisno o proizvođaču softvera odabrati hardver koji će pretvoriti u softverski PLC uređaj. Korisniku su, uz softverski PLC, na istoj platformi također dostupni ostali programski alati: *KW Multiprog®* za programiranje, *ProVisIT®* za izradu nadzorno-upravljačke aplikacije, softver za konfiguriranje i dijagnostiku korištene industrijske mreže, itd. U prikazanom tiskarskom postrojenju je zastarjeli PLC sustav *Simatic Step 5* zamijenjen softverskom PLC aplikacijom *ProConOS* implementiranoj na procesnom računalu sklopovske arhitekture *Power PC 5200®* firme *Freescale*. Kao realtime operativni sustav izabran je *Linux* dok se ulazno-izlazni signali kontroliraju preko Interbus-S modula kao distribuirane periferije.

LITERATURA

[1] ***: ProConOS Manual (English release 3.3), KW Software GmbH, Lageschestr. 32D - 32657 Lemgo, Mai 2003

[2] ***. SoftPLC Quick start guide, KW Software GmbH, Lageschestr. 32D - 32657 Lemgo, July 2003

[3] ***. ProConOS 3.3 WinRT Manual, KW Software GmbH, Lageschestr. 32D - 32657 Lemgo, February 2006

[4] ***. Operating Instructions UC100/101, manroland AG, Stadbachstrasse 1, Augsburg, Decembar 2006

[5] ***. PECOM Service manual, UC1-Hardware, manroland AG, Stadbachstrasse 1, Augsburg,

[6] ***. Configuring and installing Interbus, User manual, Phoenix Contact, November 2001

[7] Klöpper und Wiege Software GmbH, Michael Petig, Bernd Wüstenbecker: *PC-based programmable logic controllers*, Real-Time Magazine 97-4

[8] Teemu Tommila, Juhani Hirvonen, Lauri Jaakkola, Jyrki Peltoniemi, Jukka Peltola, Seppo Sierla & Kari Koskinen: *Next generation of industrial automation - Concepts and architecture of a component-based control system*, VTT Industrial Systems & Helsinki University of Technology, 2005.

[9] ***. ABB: *IEC 61131 Control Languages – Introduction*, 3BSE 021 358 R201 Rev B, October 2001.

CIS

International Conference on

INTELLIGENT SYSTEMS

Steering Committee

Chairs: Slobodan Ribarić, University of Zagreb, Croatia
Andrea Budin, Ericsson Nikola Tesla, Zagreb, Croatia

PAPERS

Intelligent support used for providing a pleasant user experience

J. Kaljun*, G. Harih* and B. Dolšak*
* University of Maribor, Faculty of Mechanical Engineering, Laboratory for intelligent CAD systems, Maribor, Slovenia
jasmin.kaljun@um.si

Abstract - This paper presents a prototype of intelligent advisory system based on the aesthetic factors regarding product design, with the emphasis on appurtenant design recommendations. Aesthetics certainly belong to the more complex design factors. Although some literature can be found about the aesthetic design of consumer goods, the designer still has to amass quite a lot of experience and knowledge in the field of aesthetics, in order to choose and carry out appropriate design actions for improving the aesthetic value of the product. The aim of the paper is not to discuss aesthetics from philosophical point of view, but to present problematic of aesthetics trough eyes of the mechanical design engineer, who is responsible for outer shape of the product already in conceptual design phase, where various influential factors determine the design.

I. INTRODUCTION

Desire is the fundamental motivation of all human action (Thomas Hobbes, 1588 – 1679) [1]. When using a "desired" product, user "feels" "pleasure" and "comfort" and it can be said that user has a "pleasant user experience". Today desire is the main marketing weapon. Stimulating a desire for a certain product by conventional advertising has to be supported by products' "self-expression". With the growth and globalization of the market, the future users of the specific product, in this case consumer products (household appliances, hand tools…), increasingly search, choose and buy desired product in virtual environment e.g. online stores [2]. In such a case, when selecting the product, users rely on:

a) the description of the product or. technical system, which is usually prepared by the manufacturer,
b) the comments of existing users,
c) test reports carried out by the relevant institutions and
d) virtual product presentation. It is a virtual visual perception of the product that is crucial, as it must support the aforementioned descriptions.

Products' form has to reflect the correct combination of, for the user, important characteristics such as functionality, ergonomics, practicality, robustness, elegance... In the virtual world a visual perception of products' form is a point of contact between the product and the prospective users. On the other hand also in the real world, the prospective user, when selecting a product, among wide range of products of same category located on the shelf, primary chooses that or those, whose form most clearly express the aforementioned properties, which are important to the user [3].

The modern process of designing technical systems, including consumer products, specifically household appliances and hand tools, is a relatively well-defined and relatively well-supported by modern computer tools. Support offered by many commercial and specially developed computer tools, mainly refers to the graphic representation, simulation, engineering analysis (e.g. finite element method) and the animation of the future operation of the technical system in a virtual world. In fact, these tools do not affect the scope of defining the form or design [4].

Despite the wide range of possibilities provided by the modern computer tools, the designer has to evaluate the information, find interdependences and end up with a broad list of possible partial solutions, among which an optimum combination has to be selected. Alongside this, the designer has to take care of a very complex and to a certain degree "subjective" area of aesthetic products. For this purpose, the designer has to know all the spheres of influence, their possible overlapping or exclusions and of course all of the available design (visual) elements. With its true/false decision seriously affects the attractiveness and usefulness of the product in practice, as well as technical feasibility, lifetime, economic viability, recyclability... In taking such decisions many different experts from various areas of expertise should cooperate, which often results in difficulties in coordinating the decisions that need to be in line with sustainable development.

A. The aim of the paper

The paper will discuss a framework of intelligent support system for the definition of the relevant visual perception of the product on the basis of functional, ergonomic and technological criteria, which will be able to offer support to the designer in the decision making process of selecting the most appropriate aesthetic (design) characteristics - features for a specific product or technical system.

The aim of the paper is not to discuss aesthetics from philosophical point of view, but to present problematic of aesthetics trough eyes of the mechanical design engineer, who is responsible for outer shape of the product already in conceptual design phase, where various influential factors determine the design.

B. Exsys Corvid

One of the development tools for the construction of expert systems is also Exysy Corvid 5.2.1 [5], a tool which basically allows object-oriented build of the expert system. The end result is web based expert system.

Usage of Exsys Corvid is based on defining "variables", which, due to their different characteristics allow a wide range of applications. Variables define not only the range of possible values, but also certain parameters, which can be controlled by the developer or modify the course of decision-making (reasoning) mechanism. At the same time variables define the behavior of the user interface.

On the basis of these variables "Logical Block(s)" are built to encode the knowledge in the form of production rules. Logical blocks define the knowledge base of the system. With the help of "Command Block(s)", a developer can optimize workflow of inference engine. In real application this means that a developer can predict a forward chaining in logical block A and backward chaining in logical block B using suitable commands in a command block

User interface, of Exsys Corvid application is set in a Java environment and can act as a Java applet or Java servlet. The entire system is mounted on an appropriate server, where it can be maintained using a basic software package. Design of user interface is partly determined within Exsys Corvid, the final form can be encoded in a website where the system is implemented.

Among various types of already mentioned variables, "confidence variables" represent a powerful feature of Exsys Corvid. Along with specific recommendation, "confidence variable" includes a "confidence factor" as a level of certainty. This enables expert systems to make multiple recommendations with differing degrees of confidence to reach a "best fit" in its conclusion. While in some cases, it is possible to give a specific recommendation with absolute precision; the real world is not often so clear-cut. Often multiple recommendations are simultaneously possible and the system ranks them based on relative merit. Applying of "confidence variables" enables the emulation of the real world and gives the type of recommendations that human experts would. Exsys Corvid provides many different ways to handle and utilize confidence values among other various formulas can be used to combine the values assigned to an overall confidence for the variable.

C. Oscar

In continuation of this paper an insight to knowledge base of prototype of an intelligent advisory system Oscar [6], based on expert design knowledge management using Exsys Corvid will be presented.

Oscar covers a field of hand manipulated power tools. In the proposed frame the aesthetic appearance of the product is not considered as an influential parameter, which however is the case in our system. Oscar is namely composed out of two sub-systems (Figure 1) that can be applied in two different modes. We can use them independently from each other, or simultaneously and interdependent on the same design project.

The intelligent decision support system Oscar is still a research prototype, and as such, a subject of intensive development, especially more subjective part of the system dealing with engineering aesthetics and aesthetic ergonomics.

Basically, the intelligent advisory system Oscar is a computer system that emulates the decision-making ability of a human expert. Such intelligent systems are designed to solve complex problems by reasoning about knowledge, like a human expert, and not by following the procedure of a developer as is the case in conventional programming.

The structure of the intelligent advisory system Oscar is unique, and significantly different from traditional programs. In foundation, the system is divided into three parts. The first part is fixed, independent of the rest of the system, called the inference engine. The second part is variable, called the knowledge base. The third part, which serves as communication module is called user interface. When running Oscar, the inference engine reasons about the knowledge base like a human expert and communicates with user through user interface.

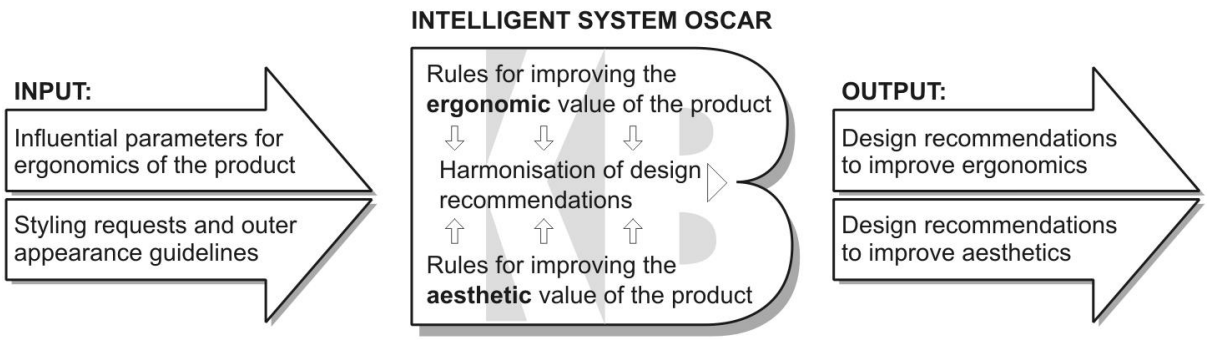


Figure 1. Basic structure of the intelligent advisory system Oscar

In the intelligent advisory system technology, knowledge base is expressed with natural language IF-THEN styled rules, called production rules.

Oscars' inference engine runs in so called conversational mode. This simply means that the inference engine requests initial data from the user at the start. Since problems are usually complex, system often cannot find a suitable solution using only initial data. Therefore the system has to find the way to solve the problem by requesting the missing data form the user.

With additional data system gradually approaches to the solution. The user feels like a participant in a dialogue led by an expert. To control a dialogue, the inference engine applies three different techniques: forward chaining, backward chaining and mixed chaining [7]. Forward chaining is a technique where the inference engine searches the solution (goal) by using specified procedure and asking questions. In backward chaining, the inference engine knows the solution and attempts to find production rules to support the solution. In mixed chaining the inference engine has an "idea" of the solution but cannot simply confirm it.

Therefore the inference engine deduces in forward chaining technique using information from previous user responses before asking the next question. So quite often the inference engine deduces the answer to the next question before asking it.

II. PRODUCT DESIGN BACKGROUND

Pleasant sensation is an accompanying product of the industrialized society and modern life and the objective fruit of the commercialization of the desire of consumption. The culture of pleasant sensation in modern life is, in reality, the demonstration of the instinctive desire of human beings. Product Design is one form of expression of this sensational culture. The product has become a carrier of the culture of pleasant sensation for people's modern material life [8].

The embodiment of pleasure culture in the design of industry goods manifests itself in two forms:

- concreteness in form, and
- symbolized metaphor.

Concreteness of the patterns of goods refers to visualized shaping of the goods, which directly incarnate some parts or organs of human body and give prominence to the shape in order to achieve physical pleasure. Symbolized metaphor means that the sexual symbol features are expressed by means of association and symbolization. Namely, the female features are represented by curvaceous-ness, tenuousness, tenderness, pink color, gentleness, while male features by linear-ness, coarseness, hardness, plain color, and dynamism.

Modern product design conveys two messages:

- rational (the function, material, and workmanship of the product), and
- sensible (the shaping, color, mentality, and way of sue), the former is the basis for the existence of the product, and the latter is more related to the formation and shaping of the product.

Apart from displaying the objectiveness of the product, the shaping of the product also reflects the cultural connotation of the product by means of some signs. Because under most circumstances, the purchasing behavior not merely involves the acquisition of the physical product itself, but also the search for something resonating with their inner desire [9].

III. THE THEORY OF AESTHETICS IN DESIGN

In the process of designing a product the aesthetic component of the product in most cases is not the main objective, as would it be in an artistic creation, which aims to achieve a high aesthetic value. The aesthetic component is a resultant of design parameters that determine the basic characteristics of the designed product [10]. The designer must develop the aesthetic component of a product included into the basic product concept through the entire creative process. Aesthetic value is formed by the elements (design solutions), which characterize the perception of the product, while challenging the subjective feelings of the user - the observer, which can be called aesthetic. The (appropriate) creation of aesthetic component is influenced by various influential factors, which can basically be divided (combined) into two major groups [11]:

- Subjective factors; the basic concept of aesthetics (philosophy) is bound to subjective perception. Subjective factors are closely linked to potential users and therefore relatively difficult to measure. They are formed by personal and cultural norms that apply in the environment originating the potential users. A large part of successful design depends on the designer's knowledge of the environment in which is the product going to be introduced. Different psychological and philosophical studies, which the designer can use in their work, can be very useful.

- Objective factors are a kind of "universal" measure of aesthetic value of each product. Objective factors dictate the aesthetic elements and relations between them, which can be called the building blocks of harmony. This group of factors includes, among others, functional factors such as clear shape, which is adapted to the main function of the product, to increase the aesthetic value of it.

A. The impact of the aesthetics to product development

Naturally, the impact of aesthetics as one of the design goals (a kind of added value) cannot be simply left to chance, but has to be considered already in the concept stage of product development. Required aesthetic value and guides to achieve these values must be, with other requirements in the areas of ergonomics, the mechanism of the structure and functions, well known and clearly defined already in the definition of development task, to allow designers to simultaneously satisfy all the requirements. Designers' task is to develop aesthetic elements in accordance with already presented factors.

The relationship between "technical" and "aesthetic" value of a product is defined by the purpose or mission of the product. The link between the technical - economic value and aesthetic - artistic value can be presented as the field of aesthetic value, which grows with the sidestepping of the basic machine elements towards consumer products, where products can be found which, together with the main function, provide even prestigious function.

IV. DEVELOPMENT OF KNOWLEDGE BASE

When analyzing the market and the anticipated requirements of customers, including estimates of the target groups regarding the developed product, the designer must have clear aims regarding the aesthetic component, and must define the semantic message of the product's visual identity. Defining the semantic forms of communication is very important, if not the key to mass produced consumer products, where the main function also complements the so-called prestigious function. The duty of the designer is to create such a product shape that best reflects the desired message, and follows the requirements of the relevant components regarding the aesthetic quality of the product. Aesthetic recommendations for the designs of products can also be an assistance and support to the designer. Aesthetic recommendations can be collected in the form of those aesthetics objectives in design, which bring together art, technical and even semantic elements. These objectives are [12]:

1. shape is consistent with the function (form follows function)
2. shape has the desired compositional consistency
3. shape reflects the planned semantic message

A. *Functionality of the product*

At the end of the 19th century, the phrase "Form follows function" was mentioned for the first time and became the 20th century leading thought in the field of industrial design. Indeed, it is in the "first plan" when developing industrially manufactured products regarding implementation of the main function, while the shape of the product mostly serves its ergonomic suitability and visual appearance. In practice this means that, the coffee grinder is intended primarily for "user friendly" coffee milling, and not decorating kitchen shelves, while expressing its identity and brand affiliation, among other things with their visual image [13].

Design recommendations in the field of aesthetics, in this case are:

- the form indicates how to use the product,
- the form of the product allows easy and safe use,
- the control areas are easily accessible,
- the use of different materials can add further functional highlight.

A representative rule in this group defines basic design goals for a tool for holding and pressing:

```
IF:
    Select the group of tools most
    suitable for tool in consideration:
    tools for holding and squeezing
THEN:
    Important safety issues are
    predicted - consider ergonomic part
    of the system.: Confidence = 5
    Shape (form) of the product should
    be consistent to products'
    function.: Confidence = 5
```

B. *Compositional consistency*

Combined with the known main function of the product, the designer has to compose a visual image of the product in such a way that it generates pleasant feelings and desires when used or merely possessed by the user, as well as the observer. The principles of visual composition are also present, in certain respects, in the designs of industrial products.

Thus, the objective compositional consistency can be archived by:

- the individual elements that constitute to the whole (in the form of a product) yet are mutually dependent so that the established rhythm, expresses the desired message (movement, resting ...).
- tendencies derived through individual elements must conform to the overall composition of the product.
- the proportions used should be in accordance with the principal function of the product:
- where possible, usage of pleasant (real) proportions is desirable,
- the use of different proportions on the different elements of one product is undesirable.

C. *Semantic consistency*

While the first two objectives of aesthetic design the care about subconsciously receiving of the product, semantic consistency carries, in concept stage already known, message of the product. Semantic message is reached with applying certain values of selected technical elements to the shape of the product. Some recommendations are presented in Table 1.

TABLE 1. DESIGN RECOMMENDATIONS FOR SEMANTIC CONSISTENCY [KALJUN ET AL. 2006]

The technical element	communication form
acceleration	aggressive, speed sports
hollow, crown	power, aggression, tension
convex, concave	tense
sharpness	tense, aggressive
tenderness	tense, natural, organic, warm
crispiness	cold, aggressive
tension	tense, wild
lead in	aggressive, threatening

Representative of production rules for semantic consistency is:

```
IF:
    How it the tool being used:
    professional use OR hobby use
AND:How to determine appropriate
    semantic meaning for the product?
    System shell derive semantic meaning
AND:Select the group of tools most
    suitable for tool in consideration:
```

```
    smashing tools
THEN:
    Tool Semantics aggression AND
    dominance
```

Presented rule recommends semantic message taking into consideration for a known usage of the product.

Different production rules are defined to translate semantic message to technical elements (Table 1). Representative rule cam be seen here:

```
IF:
    Tool Semantics aggression
THEN:
    Consider use of adequate technical
    realization of characteristic
lines.: Confidence = 5
    [Aest_Rec.ADD]  Lines/surfaces
    should express acceleration
```

V. CONLUSION

Some aspects of product design, such as aesthetics, which were in the past often unduly neglected in comparison with functional and economic efficiency of a new product, are now becoming increasingly essential.

Designers have the obligation to fashion the product to specific guidelines, regulations, standards and other elements in early developments stage, so that at the end, a product will be successful on the market. This is a difficult task for people whose job should be oriented into creative development of new goods rather than constantly studying certain rules.

Aesthetic design is a vital and complex part of product design process. In view of the complexity, multi-factors coupling and fuzziness of aesthetic knowledge, an intelligent support to aesthetic design in form of advisory decision support system is proposed.

In this context, the knowledge related to aesthetic design has be to collected, organized and encoded in form of production rules, which were found to be the most appropriate formalism due to their transparency and closeness with the human way of decision making process.

The knowledge built in the prototype of the intelligent system named Oscar is structured in form of different classes interconnected with various attributes and their values at the input side, while as the output of the system the user can expect (re)design recommendations leading to achievement of certain design goals that can improve the aesthetic value of the product being developed.

VI. REFERENCES

[1] T.M.N. Hobbes, Leviathan, Oxford: Clarendon Press, 2012.

[2] J. Lee, H.S. Lee, P. Wang, "An Interactive Visual Interface for Online Product Catalogs", Electronic Commerce Research, 4 (2004) 335-358.

[3] R. Coleman, "Fractal analysis of pathfinding aesthetics", Int. J. Simul. Model. International Journal of Simulation Modelling, 7 (2008) 71-80.

[4] M. Nagamachi, Kansei engineering, Boca Raton, Fla.; London: CRC ; Taylor & Francis [distributor], 2010.

[5] E.M. Awad, Building knowledge automation expert systems : with Exsys CORVID, Albuquerque, N.M.: EXSYS Inc., 2003.

[6] J. Kaljun, B. Dolsak, "Improving Products' Ergonomic Value Using Intelligent Decision Support System", Strojniski Vestn.-J. Mech. Eng., 58 (2012) 271-280.

[7] M.M.M.-L. Chein, "Graph-based knowledge representation computational foundations of conceptual graphs", in, Springer, [London], 2009.

[8] K.T. Ulrich, S.D. Eppinger, Product design and development, New York, NY: McGraw-Hill/Irwin, 2011.

[9] N. Crilly, D. Good, D. Matravers, P.J. Clarkson, "Design as communication: exploring the validity and utility of relating intention to interpretation", Des Stud Design Studies, 29 (2008) 425-457.

[10] V.J. Papanek, u. Galerija suvremene, Victor Papanek : dizajn za stvarni svijet : [Galerija suvremene umjetnosti, Zagreb, 14.3-31.3. 1974], Zagreb: Galerija, 1974.

[11] J. Kaljun, B. Dolsak, th, Design, "COMPUTER AIDED ERGONOMIC AND AESTHETIC DESIGN", (2004).

[12] J. Kaljun, U. Sancin, G. Harih, B. Dolšak, "Aesthetics as parameter of intelligent design support", in: Design 2012, Design Society, Dubrovnik, 2012, pp. 1293-1300.

[13] D. Quarante, B. Lapaine, Osnove industrijskog dizajna, Zagreb: Arhitektonski fakultet, 1991.

Regular Equivalence in Informed Network Search

Denis Helic
Knowledge Management Institute
Graz University of Technology
Graz, Austria
Email: dhelic@tugraz.at

***Abstract*—Search in networks is defined as a process in which an agent hops from one network node to another by traversing network links in search for given nodes. The simplest example of network search is a random walk where the agent selects a link uniformly at random from all outgoing links of the current node. On contrary, in an informed search the agent possesses (partial) background knowledge of the network. This background knowledge steers the agent's decisions when selecting the next link to traverse. The background knowledge of the network can be represented as a similarity matrix with similarities between pairs of nodes known to an agent. This matrix can be calculated in various ways in order to model various search scenarios or to best fit needs of an application. For example, similarities based on node degrees or some external information about the nodes have been commonly used in the past. In this paper we evaluate the measures that capture regular equivalence of nodes in a network with respect to their suitability as a similarity metric to inform search in networks. In particular we are interested in the properties of Katz similarity for this task.**

I. Introduction

Search in a network is a dynamical process in which an agent moves along the links between the nodes in search for a given node or a given set of nodes. Depending on the link selection mechanism we distinguish between a random walk and an informed search. In a random walk an agent selects the next link uniformly at random from the available links (the links going out from the agent's current node). In an informed search the agent possesses some background knowledge of the network, which guides it towards selecting specific links. In other words, the agent exhibits a certain bias towards selecting some of the links – the bias is induced by the background information that the agent has about the network. Another distinction between a random walk and an informed search is the stochasticity of their link selection mechanism. By definition a random walk is a stochastic dynamical process, whereas in an informed search the link selection mechanism can be (i) either deterministic in cases where the agent e.g. acts greedy and always follows an optimal link (according to a given criterion), (ii) or stochastic in cases where the agent draws a link from a given probability distribution (induced by the background knowledge).

Various studies investigated the nature of the background knowledge in e.g. informed search in social networks. For example, in studies of human search behavior that have been inspired by the famous small-world experiment by Milgram [1] the background knowledge about the network has been represented as a hierarchy of network nodes [2], [3]. The distance of nodes in the hierarchy has been used to inform deterministic decentralized search. Another type of the background knowledge has been analyzed in [4]. In this study a deterministic search is informed by the node degree (the number of nodes adjacent to a given node), a quantity that is always available locally as compared to a node hierarchy which assumes a global knowledge about the network. In [5] the authors built on this previous work and designed a stochastic algorithm that takes into account the node degree, as well as the nodes' homophily, i.e. the tendency of nodes to connect to other similar nodes. The presented algorithm probabilistically weighted the influence of the node degree and the node homophily on the link selection probability. Moreover, the algorithm modeled a situation in which an agent does not have information on the node similarities, in which case the decision was based solely on the node degree.

Similarly to social networks search in information networks is always local, i.e. users have at their disposal only the links going out from their current Web page. Human search behavior in information networks has been investigated in the work of West and Leskovec [6]. They analyzed click paths of users playing a navigation game on Wikipedia and in their subsequent work [7] they compared decentralized search algorithms using various distance functions and benchmarked them against the human navigation paths. The authors found that automatic deterministic search strategies with a complete background knowledge typically outperform human information seeking.

One of the most interesting findings was that human navigation in information networks exhibits two phases: (i) Zoom-out phase in which the users strive to reach the network core, or a hub in the network core, e.g. an overview Wikipedia page with many links to different parts of the Wikipedia. In this first phase the humans tend to base their search decisions on the node degrees, i.e. they tend to select high-degree nodes. (ii) Zoom-in phase in which the users leave the core and close in on the topic of their interest, i.e. they dive in into a Wikipedia cluster (community) dealing with a given topic. In this second phase the humans base their decisions primarily on the similarity between network nodes such as textual similarity between Wikipedia nodes.

Thus, similarly to [5] where the authors engineered an algorithm that interchangeably bases its decisions on the node degree and the similarity between nodes West et al. [6], [7] found empirically that humans base their search decisions on

those two quantities – alternating between the node degree in the first phase and the node similarity in the second phase.

Although the previous research has identified the importance of informing search in social as well as in information networks with the node degree and the similarity between nodes not much work has been done on the analysis of various network internal similarity measures and their ability to capture those two network properties. For example, in those previous studies [5]–[7] the similarity has been calculated from the external information about the network nodes: in [5] the authors calculated standard vector-based similarity on the titles of the scientific articles in the citation networks and in [6], [7] the authors applied the same similarity measure but calculated it on the complete text of Wikipedia articles.

However, it is an interesting research question to analyze the properties of the similarity measures that are based on the internal network linking patterns since the presence of the external information is not always guaranteed and its quality may greatly vary from one network to another. Moreover, measures based on the linking patterns can in a general case capture both node properties: its degree and its similarity to other nodes in a single quantity. Therefore, it seems important to investigate if and how such measures are able to inform search in networks. Thereby, we can concentrate on the suitability of these measures to reflect e.g. the two phase process of the navigation in social and information networks, where in the first phase the node degree plays the most important role and in the second phase that role is taken by the node similarity.

In this paper, we turn our attention to the analysis of internal linking patterns to serve as a similarity measure for informing search in networks. To that end, we concentrate on simple algebraic measures that capture so-called regular equivalence and analyze the distribution of such similarities over a range of synthetic networks. In particular we concentrate on Katz similarity as an example of a measure of regular equivalence in networks. Thereby, we are interested how Katz similarity captures heterogeneity in node degrees and the existence of highly interlinked clusters of similar nodes (network communities).

The paper is organized as follows. In the next Section we give a short overview of the related work in this field. In Section III we describe our methodology and Section IV gives the details about our experiments with synthetic networks. In Section V we present the experimental results and shortly discuss them. Finally, we conclude the paper and provide directions for further research.

II. Related Work

Search has been an important concept in theoretical and empirical studies of networks. For example, in Web search the famous PageRank calculation [8], which determines the importance of Web pages is based on a random walk on the Web network. In a so-called Random Surfer model a hypothetical user navigates the Web by clicking randomly on the Web links. In the limit of large number of clicks the fraction of visits to each Web page gives its PageRank. Another example of an application of the random walk includes the problem of detecting communities (groups of highly connected and similar nodes) in networks. For example, it has been shown that designing minimal codes for encoding paths of a random walker in a network is equivalent to optimally partitioning of the network into communities [9]. Thus, we can design community detection algorithms that aim at optimally compressing a random walk paths in a network.

Research on decentralized search in social networks started with Milgram's seminal small world experiment [1] in which he aimed to study the connectedness of the US society. Milgram found that people in such a large social network are connected by short chains of acquaintances. Moreover, he found that people are able to navigate large social networks efficiently, i.e. that they are able to find those short chains even if they only possess the local knowledge of the network. In the subsequent research, Kleinberg analyzed the second result of the Milgram's experiment – the ability of humans to find a short path when there is such a path between two nodes [2], [10], [11]. Kleinberg concluded that social networks possess certain latent properties that humans are aware of. This background knowledge of network structure allows humans to find a short path between two arbitrary network nodes efficiently. Kleinberg has also investigated the nature of background knowledge that is required for efficient decentralized search algorithms. In other words: What structural properties do efficiently navigable networks possess? To that end, Kleinberg designed a number of network models such as the 2D-grid model [10], hierarchical model [11], and group model [11]. Independently, Watts [3] introduced the notion of social identity as a membership in a number of social groups organized in hierarchies and showed the existence of efficient decentralized search algorithms by simulation. Finally, Adamic [12] and others applied such a model to explain search in e.g. professional social networks.

All of these models have been based on the assumption that an agent navigating the network possesses a complete background knowledge on distances or similarities between pairs of network nodes. Contrary to that, in [4] the authors argued that distances between nodes are typically available only for some of the node pairs and that the assumption that an agent possesses a complete background knowledge is unrealistic. Thus, the authors analyzed an informed search algorithm that bases its decisions on the node degree (the number of nodes adjacent to a given node), a quantity that is always available locally. The algorithm performed better in comparison with a random walk but could not reach the performance of an agent having a complete background knowledge.

In an information network decentralized search can be used to model user navigation in an information network, e.g. the Web or Wikipedia [6]. In that model the user navigates from a Web page to a Web page by following her own intuition and her own knowledge about the topics of various Web pages.

III. METHODOLOGY

A. Node Similarity

A crucial element of an informed search is the notion of distance or similarity between the pairs of network nodes. Mathematically, we may represent the background knowledge with a distance or a similarity matrix. The rows and columns of the similarity matrix represent the network nodes and the elements of the matrix give the similarity score between corresponding network nodes. If the similarity function is symmetric we obtain a symmetric similarity matrix. Partial background knowledge may be easily represented by blank elements in the matrix for those node pairs where the agent's knowledge is missing. In this paper we denote the similarity matrix with $\boldsymbol{\sigma}$ – the element σ_{ij} of the matrix gives the similarity between nodes i and j from the network.

Different similarity measures have been previously discussed in the literature for various tasks. For example, [13] gives an overview and evaluation of various similarity measures for the link prediction task. In general, similarities in networks can be calculated from external or exogenous information about the networks (e.g. textual content in an information networks), or from internal or endogenous information (internal linking patterns). There are two possible similarity measures based on the network linking patterns: structural equivalence measures and regular equivalence measures. Structural equivalence is based on the simple idea that two nodes are similar if they share many of the same network neighbors, whereas regular equivalence between two nodes is given if they have neighbors who themselves are similar. Since absolute numbers of common neighbors are hard to interpret, this number is typically normalized. Different normalizing constants give rise to different similarity measures such as cosine similarity, Pearson correlation coefficients, or Euclidean distance ([14] gives a nice overview and comparison of various normalization methods). In contrast, regular equivalence between nodes is established when two nodes do not necessarily share same neighbors but rather they have neighbors who themselves are similar, or in a more simple case two network nodes i and j are similar if j is similar to other neighbors of i. In past, several simple algebraic measures have been developed in the past to measure regular equivalence [15], [16].

Previous research showed that node degree as well as node similarity play an important role in informed search in many networks. Therefore, we concentrate in this paper on the analysis of regular equivalence measures since structural equivalence can not capture both of these aspects in a single quantity. These measures assign a zero similarity to all nodes that do not have neighbors in common – they only capture local similarity between nodes but can not reflect any global or long range dependences between nodes. They are suitable to inform search if the agent is already in the proximity of its goal, but they can not lead that agent from a distant part of the network to the cluster or community where the target is situated. In other words, structural equivalence would not be able to instruct an agent in its zoom-out phase but only in its zoom-in phase. On the other hand, because regular equivalence measures are defined recursively over node neighbors, they are able to capture global node similarities that go beyond immediate node neighbors. They assign a non-zero similarity value to each node pair that exhibits a structural dependence in the network, e.g. the nodes are connected by at least one path in the network.

B. Katz Similarity

In this paper we concentrate on a particular measure of regular equivalence called Katz similarity [13]. Let us first introduce the basic quantities relevant for measuring this similarity. Let us denote the adjacency matrix of a graph $G(V, E)$ with $\boldsymbol{A}$:

$$A_{ij} = \begin{cases} 1 \text{ if nodes } i \text{ and } j \text{ are connected by a link} \\ 0 \text{ otherwise.} \end{cases} \quad (1)$$

We define Katz similarity recursively as [17]:

$$\sigma_{ij} = \alpha \sum_k A_{ik}\sigma_{kj} + \delta_{ij}, \quad (2)$$

where δ_{ij} is Kronecker delta ($\delta_{ij} = 1$ if and only if $i = j$, otherwise $\delta_{ij} = 0$).

The idea is that the term σ_{kj} is large if the neighbor k of i is similar to j. We then sum over all neighbors of i, which is ensured by the product $A_{ik}\sigma_{kj}$ since A_{ik} equals 1 only if i and k are neighbors. Finally, the term δ_{ij} assigns a high value of "self-similarity" of a node to itself. The constant α weights the influence of neighbors as compared to the influence of the self-similarity to the nodes similarity.

We can write the expression for the Katz similarity in the matrix form as:

$$\boldsymbol{\sigma} = \alpha \boldsymbol{A}\boldsymbol{\sigma} + \boldsymbol{I}, \quad (3)$$

where $\boldsymbol{I}$ is the identity matrix.

We can now solve this equation for $\boldsymbol{\sigma}$, which gives us the following:

$$\boldsymbol{\sigma} = (\boldsymbol{I} - \alpha \boldsymbol{A})^{-1}. \quad (4)$$

In case of large networks with e.g. millions of nodes, inverting a matrix is computationally expensive. Therefore, we may want to evaluate the Equation 3 recursively. We would typically start by assigning initial values for similarity to zero, e.g. $\boldsymbol{\sigma}^0 = 0$. We then obtain (for the first three iterations):

$$\boldsymbol{\sigma}^0 = 0, \boldsymbol{\sigma}^1 = \boldsymbol{I}, \boldsymbol{\sigma}^2 = \alpha \boldsymbol{A} + \boldsymbol{I}, \boldsymbol{\sigma}^3 = \alpha^2 \boldsymbol{A}^2 + \alpha \boldsymbol{A} + \boldsymbol{I}. \quad (5)$$

Thus, in the limit of large number of iterations m we have:

$$\boldsymbol{\sigma} = \sum_{m=0}^{\infty} \alpha^m \boldsymbol{A}^m. \quad (6)$$

The elements of matrix $\boldsymbol{A}^m$ at index i and j, i.e. $[\boldsymbol{A}^m]_{ij}$ give the total number of paths of length m between nodes i and j. The Katz similarity can be seen as a weighted count of all paths between the nodes i and j with paths of length r getting weight α^r. As long as $\alpha < 1$ longer paths obtain

less weight than shorter ones.Thus, two nodes will be similar if they are connected either by few short paths or if they are connected by many longer paths.

The inverse of $(\boldsymbol{I}-\alpha\boldsymbol{A})$ does not exist for $det(\boldsymbol{I}-\alpha\boldsymbol{A})=0$, which is exactly the characteristic equation for eigenvalues of $\boldsymbol{A}$, i.e. in the matrix form $\boldsymbol{\lambda}=\mathbf{1}/\boldsymbol{\alpha}$. Thus, the inverse does not exist whenever $1/\alpha$ equals λ_i. If we start with $\alpha=0$ and gradually increase it, we first hit the equality $1/\alpha=\lambda_i$ for the largest eigenvalue λ_1, and then again every time when we approach the next eigenvalue. Thus, if we pick an $\alpha<1/\lambda_1$ we ensure that the inverse exists (or that $\sum_{m=0}^{\infty}\alpha^m\boldsymbol{A}^m$ converges).

Therefore, we may set the value of parameter α for each network as:

$$\alpha=\frac{1}{\lambda_1}f, \tag{7}$$

where λ_1 is the largest eigenvalue of the adjacency matrix $\boldsymbol{A}$, and f is a parameter in the interval $[0.01, 0.99]$. We call f the α-parameter.

C. Synthetic networks

We carry our experiments on synthetic networks. The goal is to generate networks that structurally closely reflect a typical social or an information network. Although the real networks exhibit a wide range of interesting structural properties we concentrate on two which have been observed in the majority of networks from different domains: (i) degree heterogeneity (where we typically have only a few of high degree and mid degree nodes that keep the network connected), (ii) and modular or community structure (with various groups of nodes that are tightly connected with each other and only loosely connected to other communities – typically via high and mid degree nodes). Such networks reflect also typical search scenarios such as search for a similar node within the same community and search for a distant target node in another community. The first scenario is simpler and consists only of the zoom-in phase where an agent needs only to close in on the target node, whereas the second scenario is a more complicated one and contains both search phases: zoom-out phase where the agent exits the starting community and enters the network core and zoom-in phase where the agent enters the target community and finally closes in on the target node.

In our experiments we want to analyze if Katz similarity can capture those two phases in a search process. Thus, for a given target node i:

1) The most similar nodes to i should be the nodes from its community – this would instruct an agent that entered the i's community and is already in the proximity of its target to explore that community in its search for the target.
2) The nodes from the other communities should have low similarity to the given target node i – this would prohibit an agent to enter a distant community.
3) The nodes that keep the network connected such as high and mid degree nodes should have high to moderate similarity values to a given node i – this would guide an agent to leave a distant community in the zoom-out phase. Combined with the high similarity to the nodes from the own community it would further instruct the agent to enter the i's community and explore it until it reaches its goal.

Thus, we generate networks with injected communities and heterogeneous node degrees. A standard generative model for such networks are so-called stochastic block models, which we describe next.

D. Stochastic Block Models

Stochastic Block Models (SBM) are generative models for networks with clusters or communities [18]–[23]. They are simple extensions of the standard Erdős–Rényi [24] or Bernoulli graph model. In Bernoulli model any pair of nodes from the network is connected with a constant probability p. In addition, each link is independent on any other link in the network. This model generates networks with rather homogeneous (Poisson) degree distributions. In a stochastic block model nodes are grouped into k communities and the probability of a link between two nodes is conditioned on their membership in different communities. Thus, a stochastic block model is a mixture of k^2 Bernoulli models with different linking probabilities.

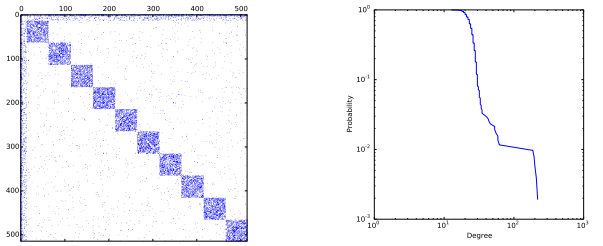
Mathematically, a stochastic block models is specified by: (i) the number of the communities k, (ii) a vector $\boldsymbol{s}$, which defines the membership of nodes in communities, and (iii) a $k\times k$ matrix $\boldsymbol{P}$, which defines probabilities of links between nodes from different groups.

The element P_{uv} of the matrix $\boldsymbol{P}$ defines the linking probability of nodes belonging to communities u and v. For the elements on the diagonal P_{uu} we obtain the linking probabilities within communities and the off-diagonal elements represent the probabilities of links that run between communities. To obtain a network with tightly connected communities with only a few links that connect the nodes from different communities we set higher probabilities in the diagonal elements P_{uu} and lower probabilities in the off-diagonal elements (see Equation 8).

To obtain a heterogeneous degree distribution we need a special community (or a couple of such communities) which are smaller than the other communities and which reverse the linking probabilities: the nodes from this community have low probability for connecting to other members of their community, but in turn they have high probability of connecting to all other communities. As a consequence these nodes obtain a larger number of links than average and thus represent high-degree or hubs in the network. The same modeling concept may be further applied to e.g. obtain mid-degree nodes, which also have a degree larger than average but still smaller than the hubs – the probabilities for connecting to the nodes from other communities should be larger than for typical nodes but smaller than the hub probabilities.

IV. Experiments

In our experiments we generate networks with 12 communities with one high-degree community and one mid-degree



(a) Adjacency matrix (b) Cumulative degree distribution

Fig. 1. **Left: Community structure.** The community structure is clearly visible in the visualization of the adjacency matrix. The thin blue line at the top and to the left of the matrix is the community of high-degree nodes. A blueish area beneath the top and to the right of the thin blue line is the mid-degree community. **Right: Heterogeneous degree distribution.** Cumulative degree distribution exhibits the existence of a small number of highly connected nodes (hubs).

community, which have 5 and 10 nodes respectively. All other 10 communities have 50 nodes. In the probability matrix for the high-degree community we assign a high probability of those nodes connecting to all other nodes from other communities. For mid-degree community we apply the same mechanism but assign smaller probabilities than for the high-degree community. All other 10 communities have a small probability of connecting to a different community (low off-diagonal entries) but a high probability of connecting to itself (high diagonal entries):

$$\boldsymbol{s} = \begin{pmatrix} 5 \\ 10 \\ 50 \\ 50 \\ \vdots \end{pmatrix} \boldsymbol{P} = \begin{pmatrix} 0.000 & 0.400 & 0.400 & 0.400 & \dots \\ 0.400 & 0.000 & 0.100 & 0.100 & \dots \\ 0.400 & 0.100 & 0.400 & 0.005 & \dots \\ 0.400 & 0.100 & 0.005 & 0.400 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \tag{8}$$

A typical adjacency matrix and a cumulative degree distribution are shown in Figure 1.

Now, we are interested in the average similarity of nodes to:

1) the nodes from their own community (Own),
2) the nodes from the other communities (Other),
3) the nodes from a moderate size community of mid-degree nodes (Mid),
4) the nodes from a small size community of high-degree nodes (High).

We investigate how these average similarities change as a function of the α-parameter f from Equation 7. Thus, we iterate over the interval $[0.01, 0.99]$ in steps of 0.01 for a total of 99 different parameter settings. For each setting we generate $1,000$ random networks (for a total of $99,000$ generated networks for all parameter configurations) using a stochastic block model with the group vector $\boldsymbol{s}$ and the probability matrix $\boldsymbol{P}$ from Equation 8. For each parameter value we plot the average similarities and the standard deviation over $1,000$ generated networks.

V. Results and Discussion

Figure 2 shows the results of our experiments. For smaller values (e.g. f less than 0.1) of the α-parameter the most similar nodes to arbitrary nodes are, on average, high-degree nodes and the nodes from their own community. The mid-degree nodes have moderate average similarity and the nodes from communities other than the own community have a small average similarity. Thus, these distributions of similarity are suitable for informing search in networks. For example, these configurations can support zoom-out phase, where an agent starting in a distant community needs first to leave that community and access nodes that provide shortcuts to the target community. High-degree nodes connect different parts of the network and an agent needs to visit them first to close in on the target community. Since the high-degree nodes have a large average similarity to an arbitrary node in a network an agent following this similarity intuition would first visit one of the high-degree nodes. On the other hand, once when the agent reaches the target community the most similar nodes to the target node are the nodes from its own community – thus, an agent would enter the target community and remain within it to explore the community nodes and eventually reach its target.

For mid range of values of α-parameter (greater than 0.1 and less than 0.9) the high-degree nodes start to dominate other node types and become the most similar nodes to arbitrary nodes regardless of their community. Thus, an agent would correctly leave a distant community by visiting one of the high-degree nodes. However, the agent would remain in the network core indefinitely long. Even if the agent succeeds in reaching the target community it would in most cases try to reach the target node by leaving the target community and revisiting the high-degree nodes. The exploration of the target community would not take place at all.

For even larger values of the α-parameter (greater than 0.9) even the mid-degree nodes become more similar to arbitrary nodes than the nodes from their communities. Thus, an agent would in most cases traverse links connecting high-degree and mid-degree nodes hoping in the network core and failing even to reach the target community.

Summarizing, regular equivalence measures in general and Katz similarity in particular can provide the background knowledge and inform search in networks efficiently. They are able to capture both phases of the search process: exiting a distant community and entering the network core with many long range links to the target community, as well as the exploration of the target community. However, it is necessary to configure the parameters for calculating the regular equivalence measures appropriately. Our experiments show that the desired behavior of these measures can be only obtained by lower values of the control α-parameter, i.e. for values less than 0.1.

VI. Conclusions

In this paper we provide a short overview of various search process on structurally complex networks. We then analyze the

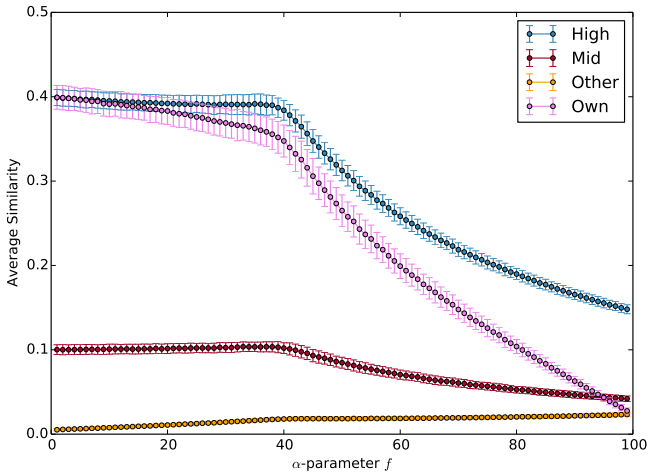


Fig. 2. **Average similarity as a function of α-parameter** For smaller values of the path weights the most similar nodes are those from the own community, closely followed by the hubs. The average similarity of the mid degrees is also significantly higher than the average similarity to nodes in the other communities. For large values of the path weights the similarity of high degree nodes dominates all other communities. The similarity to the nodes from the own community rapidly drops and for the values above 0.9 even the mid degree nodes become more similar on the average than the nodes from the own community. From the informed search point of view, the smaller values of α-parameter (i.e. values less than 0.1) are more suitable for providing guidance in both phases of search: zoom-out phase where high-degree nodes must be quickly reached, and the zoom-in phase where the nodes from the target communities need to be the most similar nodes.

suitability of regular equivalence measures to inform search in such networks. In particular, we analyze simple algebraic measures such as Katz similarity which capture several network structural properties such as node degree and node similarity into a single similarity score. Our experiments on synthetic networks with heterogeneous node degree distributions and injected community structure show that Katz similarity can efficiently inform search in networks. However, the parameters used to calculate the similarity need to be chosen with care and appropriately.

The main limitation of our work is the theoretical nature of the analysis. Specifically, we estimated the theoretical suitability of Katz similarity to inform search in networks. An experimental proof of concept would be also needed. For example, one can simulate an agent performing greedy search in networks by using the Katz similarity as the background knowledge and measure the agent's success rate and its search efficiency. Furthermore, the experiments should be repeated on empirical social and information networks that exhibit typical structural properties such as heterogeneous node degrees and community structure. However, we leave these experiments for the future work.

ACKNOWLEDGMENT

This research was in part funded by the FWF Austrian Science Fund research project "Navigability of Decentralized Information Networks" (P 24866-N15).

REFERENCES

[1] S. Milgram, "The small world problem," *Psychology Today*, vol. 1, pp. 60–67, 1967.
[2] J. M. Kleinberg, "Navigation in a small world," *Nature*, vol. 406, no. 6798, p. 845, August 2000.
[3] D. J. Watts, P. S. Dodds, and M. E. J. Newman, "Identity and search in social networks," *Science*, vol. 296, pp. 1302–1305, 2002.
[4] L. A. Adamic, R. M. Lukose, A. R. Puniyani, and B. A. Huberman, "Search in power-law networks," *Physical Review E*, vol. 64, no. 4, pp. 046 135 1–8, Sep 2001.
[5] O. z. Simsek and D. Jensen, "Decentralized search in networks using homophily and degree disparity," in *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, ser. IJCAI'05. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2005, pp. 304–310.
[6] R. West and J. Leskovec, "Human Wayfinding in Information Networks," in *Proceedings of the 21th international conference on World Wide Web - WWW '12*. New York, New York, USA: ACM Press, 2012.
[7] ——, "Automatic versus human navigation in information networks," in *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media, ICWSM*, 2012.
[8] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," in *Proceedings of the seventh international conference on World Wide Web 7*, ser. WWW7. Amsterdam, The Netherlands, The Netherlands: Elsevier Science Publishers B. V., 1998, pp. 107–117.
[9] M. Rosvall, D. Axelsson, and C. T. Bergstrom, "The map equation," *The European Physical Journal Special Topics*, vol. 178, no. 1, pp. 13–23, 2009.
[10] J. Kleinberg, "The small-world phenomenon: an algorithm perspective," in *Proceedings of the thirty-second annual ACM symposium on Theory of computing*, ser. STOC '00. New York, NY, USA: ACM, 2000, pp. 163–170.
[11] ——, "Small-world phenomena and the dynamics of information," in *Advances in Neural Information Processing Systems (NIPS) 14*. Cambridge, MA, USA: MIT Press, 2001, pp. 431–438.
[12] L. Adamic and E. Adar, "How to search a social network," *Social Networks*, vol. 27, no. 3, pp. 187 – 203, 2005.
[13] D. Liben-Nowell and J. Kleinberg, "The link prediction problem for social networks," in *Proceedings of the twelfth international conference on Information and knowledge management*, ser. CIKM '03. New York, NY, USA: ACM, 2003, pp. 556–559.
[14] E. A. Leicht, P. Holme, and M. E. J. Newman, "Vertex similarity in networks," *Phys. Rev. E*, vol. 73, p. 026120, Feb 2006.
[15] V. D. Blondel, A. Gajardo, M. Heymans, P. Senellart, and P. V. Dooren, "A measure of similarity between graph vertices: Applications to synonym extraction and web searching," *SIAM Rev.*, vol. 46, no. 4, pp. 647–666, Apr. 2004.
[16] G. Jeh and J. Widom, "Simrank: A measure of structural-context similarity," in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '02. New York, NY, USA: ACM, 2002, pp. 538–543.
[17] M. Newman, *Networks: An Introduction*. New York, NY, USA: Oxford University Press, Inc., 2010.
[18] P. W. Holland, K. B. Laskey, and S. Leinhardt, "Stochastic blockmodels: first steps," *Social Networks*, vol. 5, no. 2, pp. 109–137, 1983.
[19] K. Faust and S. Wasserman, "Blockmodels: Interpretation and evaluation," *Social Networks*, vol. 14, no. 1–2, pp. 5 – 61, 1992, special Issue on Blockmodels.
[20] C. J. Anderson, S. Wasserman, and K. Faust, "Building stochastic blockmodels," *Social Networks*, vol. 14, no. 1–2, pp. 137 – 161, 1992, special Issue on Blockmodels.
[21] T. A. Snijders and K. Nowicki, "Estimation and prediction for stochastic blockmodels for graphs with latent block structure," *Journal of Classification*, vol. 14, no. 1, pp. 75–100, 1997.
[22] Y. J. Wang and G. Y. Wong, "Stochastic blockmodels for directed graphs," *Journal of the American Statistical Association*, vol. 82, no. 397, pp. pp. 8–19, 1987.
[23] A. Goldenberg, A. X. Zheng, S. E. Fienberg, and E. M. Airoldi, "A survey of statistical network models," *Found. Trends Mach. Learn.*, vol. 2, no. 2, pp. 129–233, Feb. 2010.
[24] P. Erdos and A. Renyi, "On the evolution of random graphs," *Publ. Math. Inst. Hung. Acad. Sci*, vol. 5, pp. 17–61, 1960.

Algorithms for 2D Nesting Problem Based on the No-Fit Polygon

D. Domović, T. Rolich, D. Grundler and S. Bogović
University of Zagreb, Faculty of Textile Technology, Zagreb, Croatia
{daniel.domovic, tomislav.rolich, darko.grundler, slavica.bogovic}@ttf.hr

Summary - Strip-packing problem is apparent in textile industry where a set of items, i.e. cutting parts (2D convex or non-convex polygons) need to be placed on a rectangular container (fabric with an *m*×*n* area) so that cutting parts do not overlap and do not exceed the boundaries of the container. The goal is to find a placement that utilizes the area of a container. In this paper three methods (random search, greedy algorithm and genetic algorithm) are tested on sets of regular (convex polygons) and irregular (cutting parts) items. The goal is to find the optimal items placement that minimizes the cover area. In this paper a no-fit polygon (NFP) is used to assure two items touch without overlapping. NFP is constructed by rotating polygon *B* around a static polygon *A* in a way their edges always touch and never overlap. The result is a polygonal area enclosed by trajectory of rotating polygon's reference point which represents the overlapping area of *A* and *B*. Items touch if polygon *B* is placed on the NFP's border. Non-convex cutting parts are approximated with their convex hull since a NFP version for convex polygons is used in this paper.

I. INTRODUCTION

Nesting problems are a subfield of packing problems where a set of smaller elements i.e. items (convex or non-convex polygons) needs to be placed in one or more larger containers so that the following conditions are met: none of the two elements should overlap and elements need to be placed within the boundaries of a container [1].

The problem is applicable in variety of science branches and industry. That is why packing problems consist of wide variety of subproblems specifically modified according to the field of research. Two typologies have been made to systematize these problems: Dyckhoffs in 1990 which was revised by Wäsher in 2007 [2][3].

Packing problems can be researched from three perspectives. One can either try to: (1) find an optimal subset of items to pack in a container, (2) try to minimize the number of containers, or (3) try to find minimal container area/size.

In this paper a minimal container area is researched and applied on strip-packing problem commonly known as nesting problem since the cutting parts have to be closely nested to produce dense packing. Strip-packing problem is a common problem that arises in textile industry in the process of automated marker making. Problem goal is to find optimal placement i.e. optimal positions of cutting parts on material so the following conditions are met: each pair of cutting parts must not overlap and their surface area needs to be within the boundary of rectangular material area. An optimal placement that utilizes material usage needs to be created. Since material is formed in roles one dimension, usually width has a fixed size while the length needs to be minimized. In this paper, both dimensions can be various size so the goal is to find minimal area of a rectangle that serves as a bounding-box of nested cutting parts.

One of the first problems a researcher faces while nesting items is overlapping detection. Although this is a simple task for a human, computationally it is a hard problem that is getting more complex as the complexity of polygons visualization increases. The success of an overlapping detection method greatly depends on a way polygons are represented. There are several methods in the literature that are used to deal with overlapping detection: raster method, no-fit polygon, phi-functions and constraint graphs [4]. In this paper a no-fit polygon is used.

This paper is organized as follows. In section II methods and some previous research is described, including random search, greedy algorithm and genetic algorithm. In section III results of the implementation and a discussion is given. In section IV conclusions are made.

II. METHODS

A. No-fit polygon

A no-fit polygon (NFP) is a popular method for calculating an area in which two polygons will not overlap. Calculating the NFP is faster than using direct trigonometry methods and its advantage is that it allows the polygons to be represented with vertices and edges without the necessity for it to be approximated [4].

```
Define a static polygon A and a sliding polygon B
Perform initial trasformation
NFP = [first point in B]
newPoint, prevPoint = inf;
Until newPoint == prevPoint
        Calculate a touching point of A and B
        Find next start point
```

Figure 1. Algorithm for NFP construction for two convex polygons

NFP_{AB} of two polygons *A* and *B* is a resulting polygon of sliding movement of a polygon *B* around a fixed polygon *A*. During the sliding both polygons have fixed orientations. One point on polygon *B* should be chosen as a reference point. During the sliding a trajectory of a reference point represents the edges of NFP. During the sliding polygons *A* and *B* should always touch and never overlap.

To check if two polygons overlap a relationship between a reference point of polygon *B* and a NFP needs to be viewed. If a reference point is inside the NFP two polygons overlap. If a reference point is on a boundary of a NFP, two polygons touch. Finally, if a reference point is outside of NFP two polygons do not overlap.

The described NFP is an outer-fit NFP which is used when finding optimal non-overlapping position of two items. The other NFP version is an inner-fit NFP which is generated by sliding an item around the boundary of an inner side of a container. Using inner-fit NFP a set of feasible positions within a container is calculated in which a sliding polygon can be placed so that the overlap does not occur.

A no-fit polygon can be constructed using a sliding algorithm whose basic principle has been described above, using Minkowski sum and by decomposing items into star shaped or convex polygons.

Within the scope of this paper a NFP for convex polygons has been implemented in MATLAB environment based on an algorithm in [5] and is depicted in Fig. 1.

B. Convex hull

A convex hull of a planar set of points *P* is a smallest convex region enclosing *P* [6]. It can also be defined as a smallest convex set that contains *P*. The reason a polygon is approximated with a convex hull is that usually a convex hull has fewer points than an original polygon has so it becomes a simpler object to manipulate. Therefore computation is faster.

In this paper, cutting parts which are convex and non-convex polygons are approximated with their convex hulls i.e. transformed to convex polygons. All calculations are done with these approximations. Convex hull of cutting parts can be seen in Fig. 4. The yellow parts represent real shapes of cutting parts, while red parts are their convex hull approximations.

```
While maximum number of iterations is not met
    RND_perm = permutation of indexes of cutting parts
    Choose first polygon A as RND_perm[1]
    For an element i  in RND_perm
      Choose second polygon B as RND_perm[i]
      Calculate NFP of A and B
      In every point of NFP
          Calculate the area of rectangular bounding-box of
          polygon A and B
          Remeber in which point the bounding-box are is
          minimal
      Place polygon B in a point with minimal bounding-box
      area
      Set polygon A to a convex hull of polygons A and
      translated polygon B
```

Figure 2. Random search algorithm

C. Random search

One of the algoritms used in this paper is the random algorithm described in Fig. 2. Cutting parts are saved in a list therefore each cutting part can be accessed via its index. In each new iteration a new random permutation of indexes is created (RND_perm). Permutation list represents an order, i.e. polygons are taken from a list to form a placement in order their index appears in the permutation. The number of different versions of placement is *n*! where *n* is the number of cutting parts.

After two polygons are chosen, a NFP is calculated. A set of points is given in which two polygons can be placed to abut. Polygon *B* is placed in each NFP point. Then a bounding-box of a polygon *A* and *B* is calculated. A bounding-box is a minimal rectangular container that surrounds polygons, or in this case polygons *A* and *B*. For each NFP point a polygon *B* is placed in, an area of a bounding-box is recored. Final placement point for polygon *B* is chosen based on minimal bounding-box area criterium. Finally, a convex hull of polygon *A* with polygon *B* located in its best position is created. This convex hull serves as a static polygon *A* for the next iteration. A new polygon *B* is chosen form RND_perm and the procedure is continued until all cutting parts are used. Then a new iteration *i* starts.

D. Greedy algorithm

Greedy algorithm is an algorithm that builds up a solution piece by piece by choosing the currently best option that would bring immediate benefit to the solution

```
Choose initial cutting part from parts = [1,..,n]
    For each of the remaining n-1 cutting parts
      Calculate the NFP polygon points
      Calculate the bounding-box area when B is placed in
      each NFP point
      Place the polygon B in a position with minimal
      bounding box
    Place B in a NFP point that results in minimal bounding-
    box area
    Create a convex hull from A and B
    Set the convex hull as polygon A
```

Figure 3. Greedy algorithm

[7]. By creating a locally optimal solution the algorithm hopes to find the globally best solution. Although it may fail in producing an optimal solution, greedy algorithms often provide fast heuristics and are used in finding good approximations [8].

Greedy algorithms are used to solve many combinatorial optimization problems such as: Kruskal's algorithm for minimum spanning trees, Dijkstra's algorithm for a finding shortest paths etc.

In this paper a greedy algorithm for solving a 2D irregular strip-packing nesting problem has been created.

A greedy algorithm used in this paper is described in Fig. 3. An algorithm was inspired by Dijkstra's algorithm for solving the travelling salesman problem. A total of *n* iterations is calculated, and *n* different solutions are produced, where *n* is the number of cutting parts. The best one is declared the globally best solution produced by a greedy algorithm for this set of cutting parts. An iteration is equivalent to finding a shortest path from a town *i* in the travelling salesman problem (TSP).

First, an initial cutting part is chosen as a static polygon *A*. Then each of the remaining *n*-1 polygons are used as a sliding polygon *B* to test which gives a minimum bounding box area with polygon *A* while abutting in the NFP points. A position resulting in minimum bounding-box area is chosen. Optimal polygon *B* is the one with minimal bounding-box area and it is placed in a temporary solution. A convex hull is created from polygons *A* and *B*. Then a new polygon from the remaining *n-2* polygons is chosen. The procedure is continued until all polygons are used.

E. Genetic algorithm

Genetic algorithm (GA) is a heuristic algorithm based on evolutionary computation. GA mimics nature's evolution principle that describes the survival of the fittest. GA author, John H. Holland, presented the algorithm in the 1970s motivated by Darwin's theory of evolution. GA uses evolutionary methods to search the solution space in order to find a solution to the optimization problem [9].

The algorithm starts with initializing a population i.e. a set of individuals called chromosomes. Initial population is usually generated randomly within the solution space to enable searching the entire solution space. Population size depends on the type of problem, but usually the size of a few hundred to a few thousand individuals is used.

The algorithm continues with the selection process where a certain number of individuals from initial populations are selected based on their fitness to enter the reproductive process (crossover and mutation). Fitness of an individual is calculated using a fitness function *f(x)*. Good individuals are considered to pass good properties to their children in the process of crossover, and improve the quality of the population in the coming generations.

Crossover is a process of exchanging genetic material between parents. Individuals are crossing in order to maintain the good properties of the population. The higher the fitness of an individual is, the greater the likelihood of its survival and crossover. Crossover can create one or two children. The purpose is to get children chromosomes that will be just as good as or better than their parents [9].

After crossover some individuals mutate. Upon completion of the mutation, the new individuals are evaluated and copied into the population. Poor specimens are deleted from the population. The process of selection and reproduction is repeated until the termination condition is met. Termination condition may include maximum number of iteration achieved, maximum number of fitness function evaluations, the desired fitness function value has been achieved etc. [7].

1) Stochastic universal sampling

In a selection process in this paper a stochastic universal sampling is used [10]. Individuals are mapped to contiguous line segments. Size of each individual's segment corresponds to its fitness. Selection process is conducted by placing the equally spaced pointers over the line. Number of pointers corresponds to the number of individuals that need to be selected.

Consider number of individuals that need to be selected as NPointer. The distance between pointers equals to 1/NPointer. Position of a first pointer is given by a random number in the range [0, 1/NPointer].

2) Partially matching crossover

A partial matching crossover (PMX) has been used for recombination in this paper. A matrix of individuals from a current population is taken, a recombination is done on consecutive pair with a certain probability. After recombination a new population is used [11].

After two chromosomes *P* and *Q* are selected, two randomly chosen crossover points are created splitting the chromosomes to three parts: $P_1P_2P_3$ and $Q_1Q_2Q_3$. The middle labels between two cutting points are exchanged giving the new offspring: $P_1Q_2P_3$ and $Q_1P_2Q_3$. Duplicate labels may appear after the exchange. To resolve that problem it is considered that P_2 is consisted of $P_{21}P_{22}\ldots P_{2n}$ and Q_2 is consisted of $Q_{21}Q_{22}\ldots Q_{2n}$. Then, for a first child, a mapping of labels Q_{21} into P_{21}, Q_{22} into P_{22}, …, Q_{2n} into P_{2n} needs to be performed in the whole chromosome. Similar procedure is performed for the second child [12].

3) Move mutation

In this paper a move mutation has been used. A move mutation is a mutation type made for permutation/ordering/combinatorial optimization [10]. Individuals from the current population are taken, and each individual is mutated by moving variables with a given probability and afterwards returned to population [13].

III. RESULTS

A. Environment

All programs have been implemented in MATLAB environment following the algorithms in section 0. For genetic algorithm, a GEATbx (Genetic and Evolutionary Algorithms) toolbox has been used [14].

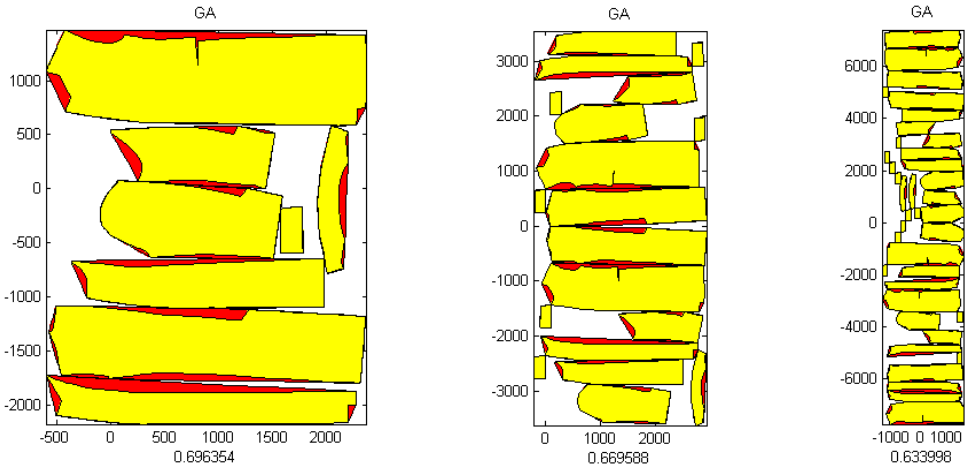


Figure 4. Best overall testing results

Additional programs have been downloaded from MATLAB Central repository and are enlisted in this section. The experiments have been run on personal computers with the following configuration: Intel Core i5-3470@3,2 GHz, 4GB DDR3, 500GB HDD.

B. Data sets

For the purpose of this research cutting parts of a male robe in garment production were used. Three sets with different number of cutting parts (8, 19 and 38) have been created.

Cutting parts were construted on paper in measure 1:5. Afterwards they were scanned and saved into bitmap format and converted to vector format using MATLAB. Since an approximate of 1200-7800 points were used to describe a polygon, using an application [15] these polygons have been simplified. Total number of points of a polygon is 1 percent of the original polygon points count (this parameter can be adjusted in [15]) which resulted in 13-78 points. The program tends to keep the aspect ratio between area and circumference as as close to 1 as possible before and after simplification.

This simplification has been done to convert curved edges of cutting parts into straight line edges. A transformation from cutting parts into convex hull has been done in the next step. Polygons created using this procedure are reffered to as version *V1*.

A different data set version *V2* is made of cutting parts from the version *V1* except *V2* is consisted of five times more points since five more points have been interpolated on a line between each two vertices of an original polygon from *V1*.

Cutting parts are saved in a list therefore each cutting part can be accessed via its index.

C. Hypothesis

The goal of this research was to find a dense packing using cutting parts aproximated with their convex hull.

Several hypothesis have been tested in the scope of this research. Hypothesis (1) included the premise that the data set version *V2* will be able to produce denser packing becauce in NFP, a sliding polygon could be placed directly on the edge of the NFP not just at the vertices which could lead to creating a convex hull of polygons *A* and *B* with a smaller area. That would mean a packing is denser.

Hypothesis (2) includes a premise that genetic algorithm is most likely to produce the best results.

In hypothesis (3) an assumption that greedy algotihm will not give the best solution.

In hypothesis (4) it is assumed that packing density depends on the number of cutting parts such that more cutting parts will produce lover cover area than less cutting parts.

D. Parameters

1) Random search parameter

In this research 10000 iterations has been used without rotation. Rotation has not been used since cutting parts in the real life techological process should not be rotated because of properties of material.

2) Greedy algorithm parameters

A total of *n* iterations has been made for greedy algorithm producing *n* different local best results since in every iteration a different cutting part has been chosen as a starting point. The solution with the best cover area has been declared as optimal.

3) Genetic algorithm parameters

Parameters described in Table I are used in the GEATbx toolbox in MATLAB. One population (parameter NumberPopulation in GEATbx) consisted of 50 individuals (parameter NumberIndividuals in GEATbx) is used in reproduction process of genetic algorithm in 200 generations. Best individual found upon this termination criterion is the best solution so far. Individuals are represented as permutations that correspond to variable format 5 in GEATbx.

TABLE I. PARAMETERS OF GENETIC ALGORITHM IN GEATBX

GA parameter		Value
Number of population		1
Number of individuals		50
Variable Format		5
Selection	Name	Selsus
	Pressure	2
	Generation Gap	1
Recombination	Name	recpm
	Rate	1
Mutation	Name	Mutmove
	Rate	1/m
Termination	Method	1
	MaxGenerations	200

In the selection process, a stochastic universal sampling method is used with pressure factor equal to 2 and generation gap parameter equal to 1. Generation gap parameter describes how many individuals will be produced in comparison to the number of individuals in the population [10]. If generation gap equals to 1, than the same amount of offspring as parents is produced and all parents are replaced with offspring.

Still the toolbox remembers the best individual found in all generations. If generation gap is less than 1, less offspring than parents is produced so parents with best solutions are able to survive to the next generation (elitism).

Partial matching crossover PMX (*recpm* in GEATbx) has been used for recombination with rate value 1.

A move mutation has been used. Mutation rate controls the number of changes per individual and equals to 1/*m* where *m* is the number of individuals.

The algorithm stops after the condition of maximum of 200 generations is met which equals to 10.000 objective function evaluations.

Fitness function is calculated using (1):

$$f(\mathrm{x}) = 1 - (areaCP / areaBB), \qquad (1)$$

where *areaCP* is the area of real non approximated cutting parts before approximation with convex hull, and *areaBB* is the area of a bounding box of two rectangles *A* and *B*. Note that, although approximations in form of convex hulls are used for nesting (so that NFP fox convex polygons can be used), real cutting parts area is used for calculating *f(x)*.

E. Discussion

The results in table II are used for hypothesis validation. Rows in table II illustrate the results for different data sets with 8 (CP8), 19 (CP19) and 38 (CP38) cutting parts. Each dataset comes in two versions: *V1* and *V2* as described in subsection Dataset. Results for cover area are shown in percentage for greedy algorithm (GRE), random search (RND) and genetic algorithm (GA). Cover area is the ratio between the area of cutting parts and minimum bounding-box of cutting parts. An average cover area of ten measurements is shown in column AVG and a standard deviation in column SD.

The results indicate that as the number of cutting parts increases cover area is getting lower for all *V2* experiments. *V2* dataset produced denser packing for 8 cutting parts using all algorithms. Only in GA *V2* gives better results than *V1* (if SD factor for CP38 is counted in). For random search algorithms *V1* always gave better results than *V2* in all datasets. Therefore, hypothesis (1) is not generally true. It is shown that datasets with less cutting parts and more points representing them will result in denser packing.

As the number of cutting parts increases genetic algorithm produces better results than the other two algorithms, because it searches the solution space diversely and with more guidance toward the best solution. Therefore, hypothesis (2) is justified. In Fig. 4 the best solutions obtained from all of the experiments are presented for 8, 19 and 38 cutting parts. All of the results have been obtained by GA. The yellow parts represent real shapes of cutting parts, while red parts are their convex hull approximations.

Greedy algorithm did not produce the best results compared to other algorithms which confirmed hypothesis (3). In the literature the results of greedy algorithm is usually used as the initial solution for some algorithms.

A relation between number of cutting parts and packing density has been detected for random search and GA for both *V1* and *V2* datasets. As number of cutting parts is increasing so is the size of the solution space. For 8 cutting parts the solution space is 8!, and for 38 cutting parts it is 38!.

TABLE II. COVER AREA

		GRE	RND		GA	
			AVG	SD	AVG	SD
CP8	**V1**	58.48	68.25	0.35	67.71	0.68
	V2	66.86	69.41	0.18	68.43	1.38
CP19	**V1**	45.16	64.7	0.82	64.81	0.9
	V2	44.58	44.54	8.02	64.89	0.62
CP38	**V1**	49.64	41.94	5.97	60.69	1.55
	V2	44.5	27.77	7.84	59.52	0.53

On the other hand, due to limiting computational resources, only 10.000 solutions for each data set is calculated in this research. This means that 24,8% of the solution space is searched for 8 cutting parts and only $5{,}23*10^{-44}$% for 38 cutting parts. The random search algorithm does not provide any guidance toward the best found solution as GA does so it produces worse results than GA for large number of cutting parts. Therefore, it can be assumed that the order in which cutting parts are nested is important and will have great influence on packing density. More points that represent a polygon do not guarantee the packing will be denser. The hypothesis (4) is valid.

It can also be noted that the shape of cutting parts influences the density of a packing because cutting parts, when optimal nesting position is found are converted to their convex hull which can lead to a lot of blank spaces that cannot be filled because a NFP for convex polygons is used.

IV. Conclusion

In this paper a research on finding dense packing using cutting parts of a male robe has been conducted. Three datasets with 8, 19 and 38 cutting parts have been used. Each data set came in two versions, one of which had five times more points representing each polygon. Experiments with random search, greedy and genetic algorithm have been performed. Results of an average of ten experiments and standard deviation have been presented in this paper.

It has been discovered that the shape and order of cutting parts influences the density of packing. If a polygon is represented with higher amount of points it does not mean the packing will be denser. Random search algorithm can produces best results for a small number of cutting parts. For large number of cutting parts it is best to use genetic algorithm.

References

[1] J. Egeblad, "Heuristics for Multidimensional Packing Problems," KøbenhavnsUniversitetKøbenhavns Universitet, Det Naturvidenskabelige FakultetFaculty of Science, Datalogisk InstitutDepartment of Computer Science, 2008.

[2] H. Dyckhoff, "A typology of cutting and packing problems," *Eur. J. Oper. Res.*, vol. 44, no. 2, pp. 145–159, Jan. 1990.

[3] G. Wäscher, H. Haußner, and H. Schumann, "An improved typology of cutting and packing problems," *Eur. J. Oper. Res.*, vol. 183, no. 3, pp. 1109–1130, Dec. 2007.

[4] J. A. Bennell and J. F. Oliveira, "The geometry of nesting problems: A tutorial," *Eur. J. Oper. Res.*, vol. 184, no. 2, pp. 397–415, Jan. 2008.

[5] E. K. Burke, R. S. R. Hellier, G. Kendall, and G. Whitwell, "Complete and robust no-fit polygon generation for the irregular stock cutting problem," *Eur. J. Oper. Res.*, vol. 179, no. 1, pp. 27–49, May 2007.

[6] R. L. Graham and F. F. Yao, "Finding the convex hull of a simple polygon," *J. Algorithms*, vol. 4, no. 4, pp. 324 – 331, 1983.

[7] S. Dasgupta, C. H. Papadimitriou, and U. V. Vazirani, *Algorithms*. 2006.

[8] D. M. Mount, "CMSC 451 Design and Analysis of Computer Algorithms," *Dept Comput. Sci. Univ. Md.*, pp. 1–135, 2003.

[9] M. Golub, "Genetski algoritam - prvi dio."

[10] H. Pohlheim, *Evolutionary Algorithms*. Springer, Heidelberg, 2003.

[11] "GEATbx for Matlab (v. 3.8): Function recpm." [Online]. Available: http://www.geatbx.com/docu/recpm.html. [Accessed: 18-Feb-2014].

[12] R. Baragona, F. Battaglia, and I. Poli, *Evolutionary Statistical Procedures: An Evolutionary Computation Approach to Statistical Procedures Designs and Applications*. Springer, 2011.

[13] "GEATbx for Matlab (v. 3.8): Function mutmove." [Online]. Available: http://www.geatbx.com/docu/mutmove.html. [Accessed: 18-Feb-2014].

[14] "GEATbx - Genetic and Evolutionary Algorithms Toolbox in Matlab - Main Page." [Online]. Available: http://www.geatbx.com/. [Accessed: 21-Feb-2014].

[15] "Decimate Polygon - File Exchange - MATLAB Central." [Online]. Available: http://www.mathworks.com/matlabcentral/fileexchange/34639-decimate-polygon. [Accessed: 18-Feb-2014].

An Optimal Control Module of Sustainable Natural Resources Consumption Control Synthesis for Decision Support Systems

E.Cherkashin [*] [**], S.Badmatsyrenova[*], I.Vladimirov[**], A.Popova[*], and A.Davydov[*]
[*] Institute of System Dynamics and Control Theory of SB of RAS, Irkutsk, Russia
[**] National Research Irkutsk State Technical University, Irkutsk, Russia
[***] V.B.Sochava Institute of Geography SB RAS, Irkutsk, Russia
eugeneai@icc.ru, kiana.wind@gmail.com, garisson@irigs.irk.ru, chudnenko@icc.ru, ethrik@gmail.com

Abstract - We consider optimal control problems solved in the process of decision support implemented as a submodule of a hybrid decision support system (HDSS). The HDSS is devoted to figure out a sustainable consumption of the forest resources in an industrial region. The proposed decision is presented to decision maker as possible scenarios of a modeled object development in the future. The scenarios are generated by a subsystem according to initial value combination obtained from various data sources. The mathematical model as an informational object representation is generated by means of automated reasoning that implements a technique of a model construction. The paper deals with the internals of the HDSS related to optimal control model implementation and its integration into the whole system. A HDSS is tested on the data of forest resources of Irkutsk Region.

I. INTRODUCTION

One of the approaches to decision support systems construction is the use of mathematical models and computer simulation to supply decision-making algorithms with additional information. This allows decision maker to manipulate more actual information and produce counterintuitive decisions. The additional information is a set of computer simulation scenarios of object under investigation and the analysis of the scenarios by means of a number of criteria. A set of Pareto-optimal scenarios is selected, and a multicriteria optimization is carried out over the set of the produced scenarios. Each stage of modeling must be supported with necessary data [1-3].

Within the above-described conceptual Hybrid approach to the Decision Support Systems (HDSS) environment and tools construction, we develop a HDSS for decision maker to support sustainable forest resource consumption [4]. The HDSS is based on computer simulation of the resource development in future according to the identified model, initial data, and a number of various scenarios of timber-harvesting and forest restoration. Identification of the model is carried out under control of a knowledge base, which constructs model state graph, identifies constants, and computes the initial conditions for the concrete scenario prescribed by a scenario generator. The user interface allows one to define criteria of a scenario assessment, initial conditions intervals for the generator and query computer simulation results as well as generated model elements. The HDSS construction environment contains also instrumental software for derivative work development.

The model construction is a creative problem due to variety of data granularity, properties of forest regions to be investigated, the structure of the object and the problem at hand. For the purpose, we use inference engine based on original Calculus of Positively Constructed Formulae (CPCF) [5-7], which have intuitionistic inferences in more broad class of formula with respect to popular resolution method of automatic theorem proving. The properties of the calculi allow us to construct logical control layers for the process of computer simulation [DSS] as well.

The HDSS environment is used to develop a prototype HDSS for management of forest resources of Irkutsk Region, Russia. Model identification algorithms are adjusted to construct a model for various source data and investigated object scales, *e.g.*, for the whole Irkutsk region and detailed forest species data, and for the set of forestry regions and raw species data. Sustainable scenarios of the forest resource development are obtained by means of computer simulation and solution of optimal control problems over the synthesized models.

Optimal control problem (OCP) for the model is defined over set of controllable variables, which in our case are amount of timber-harvesting volumes for all species (to be maximized), amount of tangible resources for forest fire protection and restoration, *etc*. In the last case, maximizing of the amount of the forest areas covered with forest to be of interest.

In this work, we use a simplified improvement procedure [8,9] for numerical solution of the OCP, which implements a second order improvement method based on the Krotov's sufficient optimality conditions. The procedure implements solving OCP of a discrete system, and it is originated from an approach for continuous systems. The procedure is still under development in the aspects of the efficiency of computation for various systems, application of nowadays methods for solving optimization tasks in the each step of improvement iteration.

Implementation of the solution of the OCP required coping with a number of technical problems such as figuring out derivatives and function compilation, result interpretation to the user with various visual diagrams, representation of the results as a digital map.

The paper describes the mathematical and software aspects of the implementation of HDSS.

II. General Design of HDSS

Decision support systems (DSS) are software components, which support decision maker with various methods (data and knowledge processing, objective and subjective models) for analysis and solution of weakly- and even nonstructured problems [4]. The idea of DSS appeared as an attempt of human behavior automation of data analysis at hand, problem solving, to elaborate a particular problem and create a decision. At present, there is no widely accepted definition of DSS, as its design sufficiently depends on a problem to be solved, available information and knowledge, and user abilities.

We suppose that DSS is a user application, which provides tools of data analysis, model identification, and computer simulation for decision maker to support his/her problems solution, make administrative and other decisions.

The main part of the DSS is Solution Generator (fig.1) that is controlled with source data on the object under investigation and the stated problem description. The generator uses knowledge based automated Reasoning Unit implemented on the base of Automatic Theorem Prover (ATP in fig.1). Reasoning Unit constructs a basic structure of the investigated object as a graph in computer

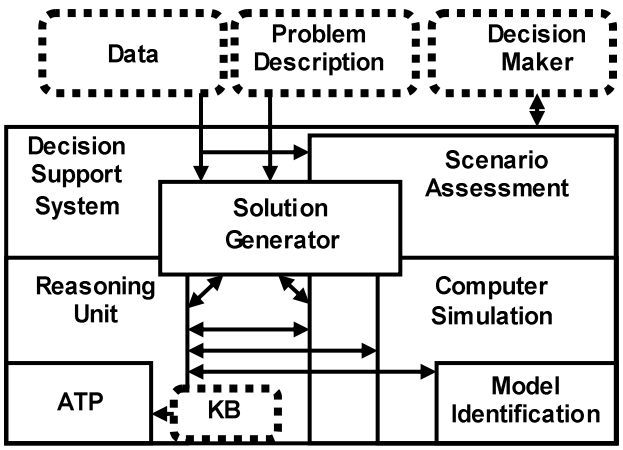


Figure 1. General design of HDSS

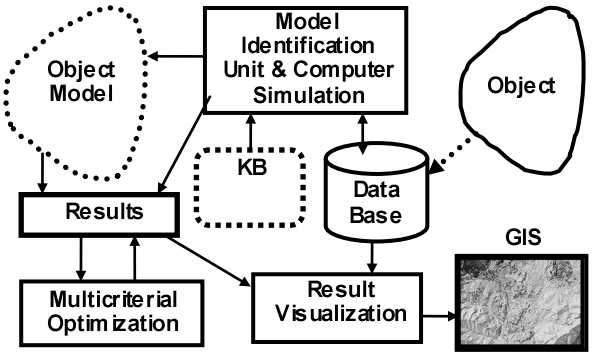


Figure 2. Information processes of HDSS

memory analyzing known data [18-asya]. Knowledge Base (KB in fig.1) includes the data about possible object change and heuristics used by mathematicians in the process of model construction; this allows us to adjust the model to particular properties of the object and the problem. Thus, the reasoning unit is used to structural and parametric model identification and assessment of the possible solutions [4].

On the identification stage of the model source data on the object under investigation is used; the data is stored in a database (DB). The data is used to figure out model parameters and initial conditions. After that, computer simulation is performed or an optimal control problem is solved. Each simulation is a scenario of forest dynamics; the scenario's condition depends on various combinations of the initial parameters controlled via user interface.

The scenarios are assessed according to a number of criteria, and, then, a multicriterial reduction of the scenario set is performed. If the results have spatial properties, they are visualized as digital map [4] (fig. 2) in a GIS module. The GIS module presents the simulation results at any time moment of the prognosis interval.

The theoretical aspect of the forest resources consumption domain is represented, in its general part, as model of forest dynamics, an optimal control problem, and a number of result assessment criteria. Let us briefly consider the representation.

III. Model of Forest Dynamics and Control

In V.B.Sochava Institute of Geography SB RAS, a system of models of forest resources has been developed [10]. A model knowledge base has been constructed, where the models are classified according to regional scale, modelled object and problem to be solved. Model coefficients identification techniques taking into account regional peculiarities has been developed [1-4]. Model of "Forest Dynamics and Control" (FDC) allows one to forecast of forest resources development of a region or a forestry for decades as proportions of forest areas of predominant species of an age group (it is described in English in [1]).

The model [1,10] corresponds to a graph structure of states (the area proportions) and edges between states, which correspond to intensities of transition of the area particles from one state to another. The graph describes forest growth, plantations, logging, burnout, and

segregation under capital construction, arable lands, reads, *etc*.

FDC is a regional-level model and uses dynamic variables. The total non-forest area S_N of the modeled territory; the total area uncovered with forest S_0; the areas covered with forest of *i*-th species predominating, including young forest of the first S_{1i} and second S_{2i} age classes, average-aged S_{3i}, maturing S_{4i}, mature S_{5i} and overmature S_{6i} forest stands. These states are all time-dependent. The dynamics of these area parcels is described by the following formulae:

$$dS_N / dt = -\lambda_{N0} S_N(t) + u_N(t); \quad dS_0 / dt =$$

$$= \lambda_{N0} S_N(t) - \lambda_{01} S_0(t) + u_L(t) + u_B(t) - u_{N0}(t) - u_C(t);$$

$$\frac{dS_{Ci}}{dt} = -\lambda_{C1i} S_{Ci}(t) + u_{Ci}(t) - u_{NCi}(t) - u_{BCi}(t);$$

$$\frac{dS_{1i}}{dt} = \lambda_{C1i} S_{Ci}(t) + \lambda_{01i} p_{0i} S_0(t) - \lambda_{12i} S_{1i}(t) - u_{L1i}(t) - u_{N1i}(t) - u_{B1i}(t);$$

$$\frac{dS_{2i}}{dt} = \lambda_{12i} S_{1i}(t) - \lambda_{23i} S_{2i}(t) - u_{L2i}(t) - u_{N2i}(t) - u_{B2i}(t);$$

$$\frac{dS_{3i}}{dt} = \lambda_{23i} S_{2i}(t) - \lambda_{34i} S_{3i}(t) - u_{L3i}(t) - u_{N3i}(t) - u_{B3i}(t) + I_i;$$

$$\frac{dS_{4i}}{dt} = \lambda_{34i} S_{3i}(t) - \lambda_{45i} S_{4i}(t) - u_{L4i}(t) - u_{N4i}(t) - u_{B4i}(t);$$

$$\frac{dS_{5i}}{dt} = \lambda_{45i} S_{4i}(t) - \lambda_{56i} S_{5i}(t) - u_{L5i}(t) - u_{N5i}(t) - u_{B5i}(t);$$

$$dS_{6i} / dt = \lambda_{56i} S_{5i}(t) - u_{L6i}(t) - u_{N6i}(t) - u_{B6i}(t) - I_i;$$

$$u_L(t) = \sum_{i,j} u_{Lji}(t); \quad u_C(t) = \sum_i u_{Ci}(t);$$

$$u_N(t) = \sum_{i,j} u_{Nji}(t) + \sum_i u_{NCi}(t) + u_{N0}(t).$$

Here, λ_{jki} is the intensity coefficients of transition of the area parcels from *j*-th category of lands or group of ages (state) to the other *k*-th forest state, with *i*-th species predominating, or to a definite category of lands (without the index *i*); p_{0i} is the share of the areas of uncovered with forest and populated with young forest of the first age group with *i*-th species predominating; $u_N(t)$ is an increase of the non-forest area due to capital construction on other categories of lands $u_{Nji}(t)$; $u_L(t)$, $u_{Li}(t)$, and $u_{Lji}(t)$, respectively, correspond to the annual area of logging as a whole, to crops of the *i*-th species, and in the forest stands of the *i*-th species and the *j*-th age group at time *t*; $u_B(t)$, $u_{Bi}(t)$, and $u_{Bji}(t)$, respectively, are the annually burnt-over areas as a whole, of crops of the *i*-th species, and forests of the *i*-th species of the *j*-th age group; I_i stands for the growth rate of the area of average-aged forests of young growth of the *i*-th species as a result of species succession, and I_i is the decrease rate of the area of overmature forests as a result of the species succession.

A. A Discrete-Time Variant of the Model

In the computer simulation we use a discrete-time variant of the model, where dynamic variables parameterized with *t*, are replaced with a discrete-time variables parameterized with the time moment number, so a general form of the forest growth equation become as follows:

$$S_{ji}^{l+1} = (\lambda_{j-1,ji} S_{j-1i}^{l} - \lambda_{jj+1i} S_{ji}^{l} + u_{ij}^{l})\Delta t.$$

Here, Δt is a differential approximation and *l* is a time moment number that varies from 0 to a finite number depending on value of Δt. If $\Delta t = 1$ the equation becomes a discrete-time vector equation (system of equations) of the following general form:

$$x(t+1) = f(t, x(t), u(t)), \quad x(t_0) = x_0,$$
$$t \in T = \{t_0, t_0 + 1, \ldots, t_1\}, \qquad (1)$$

where the function $x(t)$ takes values in the Euclidean space R^n, $u(t) \in R^m$. The function $x(t)$ is called a *state* (phase variable), $u(t)$ is called a *control*. The set of pairs $(x(t), u(t))$ satisfying (1) is called the *set of feasible pairs*, and it is denoted by *D*. *Cost function* is defined as follows:

$$I(x,u) = F(x(t_1)) + \sum_{t=t_0}^{t_1 - 1} f^0(t, x(t), u(t)) \to \min. \qquad (2)$$

The cost function realizes one of the assessment criteria. By the *optimal control problem* for discrete systems, we understand the problem of constructing a sequence $\{(x_s(t), u_s(t))\} \in D$, such that $I(x_s, u_s) \to \inf_D I$. This sequence is called a *minimizing sequence*. If for some pair $(x^*(t), u^*(t)) \in D$ the cost function $I = \inf_D I$, then $x^*(t)$ is called the *optimal trajectory* and $u^*(t)$ is called the *optimal control*.

B. An Improvement Algorithm for Discrete Systems

We will construct a minimizing sequence and an optimal control approximation by means of an iterative method. At each iteration, an improvement problem is solved. More precisely, at the beginning of iteration the pair (x^b, u^b) denotes the current feasible approximation of the optimal solution. Then, at the end of the iteration we are to obtain a new pair (x^e, u^e), such that $I(x^e, u^e) < I(x^b, u^b)$.

We use an approach implementing a second order improvement algorithm. This algorithm is obtained in [8]. Its properties of convergence and improvement have been proved in [8] as well. Iteration consists of the following steps:

Step 1. Compute $x^b(t)$ out of $u^b(t)$ from the system (1).

Step 2. Set a value of the parameter $\alpha = \alpha^b$.

Step 3. For α^b calculate $\psi(t,\alpha^b)$ and $\sigma(t,\alpha^b)$ from the system

$$\psi(t,\alpha) = -H_x^\alpha + (f_x^{'}\sigma(t+1,\alpha)f_u + H_{xu}^\alpha)\times$$
$$\times(f_u^{'}\sigma(t+1,\alpha)f_u + H_{uu}^\alpha - (1-\alpha)E)^{-1}H_u^\alpha,$$
$$\sigma(t,\alpha) = f_x^{'}\sigma(t+1,\alpha)f_x + H_{xx}^\alpha + (f_x^{'}\sigma(t+1,\alpha)f_u + H_{xu}^\alpha)\times$$
$$\times(f_u^{'}\sigma(t+1,\alpha)f_u + H_{uu}^\alpha - (1-\alpha)E)^{-1}(f_x^{'}\sigma(t+1,\alpha)f_u + H_{xu}^\alpha)^{'},$$
$$\psi(t_1,\alpha) = -\alpha F_x(x^b(t_1)), \quad \sigma(t_1,\alpha) = -\alpha F_{xx}(x^b(t_1)),$$
$$H^\alpha(t,x,\psi(\mathrm{t}+1,\alpha),u) = \psi^{'}(t+1,\alpha)f(t,x,u) - \alpha f^0(t,x,u),$$

where $\psi(t,\alpha^b)$ is a vector of n variables, $\sigma(t,\alpha^b)$ is a $n\times n$ symmetric matrix, and the derivatives of the function H^α are calculated at the point $(t,x^b(t),\psi(t+1,\alpha^b),u^b(t))$. Subscripts of H^α and other vector functions denotes the corresponding partial derivative on a vector variable, *e.g.* H_{xu}^α denotes the second order partial derivative on vector variables x, and then u. The prime sign denotes the transposition of a matrix.

Step 4. Find $\psi_\alpha(t,\alpha^b)$ and $\sigma_\alpha(t,\alpha^b)$ from

$$\psi_\alpha(t,\alpha) = H_x - (f_x^{'}\sigma_\alpha(t+1,\alpha)f_u + H_{xu})\times$$
$$\times(f_u^{'}\sigma(t+1,\alpha)f_u + H_{uu}^\alpha - (1-\alpha)E)^{-1}H_u^\alpha +$$
$$+(f_x^{'}\sigma(t+1,\alpha)f_u + H_{xu}^\alpha)(f_u^{'}\sigma(t+1,\alpha)f_u + H_{uu}^\alpha - (1-\alpha)E)^{-1}\times$$
$$\times(f_u^{'}\sigma_\alpha(t+1,\alpha)f_u + H_{uu} + E)\times$$
$$\times(f_u^{'}\sigma(t+1,\alpha)f_u + H_{uu}^\alpha - (1-\alpha)E)^{-1}H_u^\alpha -$$
$$-(f_x^{'}\sigma(t+1,\alpha)f_u + H_{xu}^\alpha)(f_u^{'}\sigma(t+1,\alpha)f_u + H_{uu}^\alpha - (1-\alpha)E)^{-1}H_u,$$
$$\sigma_\alpha(t,\alpha) = f_x^{'}\sigma_\alpha(t+1,\alpha)f_x + H_{xx} - (f_x^{'}\sigma_\alpha(t+1,\alpha)f_u +$$
$$+H_{xu}(f_u^{'}\sigma(t+1,\alpha)f_u + H_{uu}^\alpha - (1-\alpha)E)^{-1}\times$$
$$\times(f_x^{'}\sigma(t+1,\alpha)f_u + H_{xu}^\alpha)' + (f_x^{'}\sigma(t+1,\alpha)f_u + H_{xu}^\alpha)\times$$
$$\times(f_u^{'}\sigma(t+1,\alpha)f_u + H_{uu}^\alpha - (1-\alpha)E)^{-1}(f_u^{'}\sigma_\alpha(t+1,\alpha)f_u + H_{uu} + E)\times$$
$$\times(f_u^{'}\sigma(t+1,\alpha)f_u + H_{uu}^\alpha - (1-\alpha)E)^{-1}(f_x^{'}\sigma(t+1,\alpha)f_u + H_{xu}^\alpha)' -$$
$$-(f_x^{'}\sigma(t+1,\alpha)f_u + H_{xu}^\alpha)(f_u^{'}\sigma(t+1,\alpha)f_u + H_{uu}^\alpha - (1-\alpha)E)^{-1}\times$$
$$\times(f_x^{'}\sigma_\alpha(t+1,\alpha)f_u + H_{xu})',$$
$$\psi_\alpha(t_1,\alpha) = -F_x(x^b(t_1)), \quad \sigma_\alpha(t_1,\alpha) = -F_{xx}(x^b(t_1)),$$
$$H(t,x(t),\psi_\alpha(t+1,\alpha),u(t)) = \psi_\alpha^{'}(t+1,\alpha)f(t,x,u) - f^0(t,x,u).$$

The derivatives of the function H are evaluated at the point $(t,x^b(t),\psi_\alpha(t+1,\alpha^b),u^b(t))$.

Step 5. From the discrete chain

$$x(t+1) = f(t,x(t),u^b + \Delta u(t,\Delta x,\alpha)), x(t_0) = x_0,$$
$$\Delta u(t,\Delta x,\alpha) = -(f_u^{'}\sigma(t+1,\alpha)f_u + H_{uu}^\alpha - (1-\alpha)E)^{-1}\times$$
$$\times(H_u^\alpha + (H_{ux}^\alpha + f_u^{'}\sigma(t+1,\alpha)f_x)\Delta x), \quad \Delta x = x - x^b,$$

the derivatives of H^α are calculated at the point $(t,x^b(t),\psi(t+1,\alpha^b),u^b(t))$,

$$\psi(t,\alpha) = \psi(t,\alpha^b) + \psi_\alpha(t,\alpha^b)(\alpha-\alpha^b),$$
$$\sigma(t,\alpha) = \sigma(t,\alpha^b) + \sigma_\alpha(t,\alpha^b)(\alpha-\alpha^b),$$

and $x^e(t,\alpha)$ is determined, $u^e(t,\alpha) = u^b(t) + \Delta u(t,x^e(t,\alpha) - x^b(t,\alpha),\alpha)$ and the parameter α is figured out by solving the one-dimensional minimization problem over $\alpha \in 0,1$.

Step 6. If $I(x^e,u^e) < I(x^b,u^b)$, then let (x^e,u^e) be new approximation for the next iteration; otherwise, decrease α^b and repeat the process from the step 3.

The process stops if the improvement becomes less than a prescribed value.

C. Control Problem Module Implementation

Module for OCP solution has been implemented by means of Numpy, Scipy, and Simpy libraries of the Python programming language. The implementation consists of three classes: a Control Problem Base class (a stub for problem definition), and two classes corresponding to a method (the first order and the second order) of an approximation construction of the optimal solution.

To determine the model one must define a class by inheriting the stub class and defining the model f, the interval functional f^0, the terminal functional F as regular Python expressions (return values) and initial vector function of control u as the return value (an array of vectors for each discrete time moment) of a method of the class.

The model graph and its state variables are converted to vector function formula of the form (1) represented with symbols and expressions of the computer algebra library Simpy. The OCP solution implementation presumes that the formula (1) is of a general case, and no further heuristic implication asserted. Thanks to Simpy library symbolic manipulation routines, the vector function derivatives can be found, and then they are converted to a Python function of vector arguments.

The usage of the higher-order programming language and libraries instead of C or Fortran allowed us to check logical structure of the numerical methods with respect to basic matrix operations, disallowing programmer the implicit adjustment and optimizations of the algorithms "in place", *e.g.*, swapping vector dimensions in vector multiplications. As the result, we obtained shortest implementations of the algorithms that are very close to paper definitions. In further work, they can be converted (e.g., compiled) to a fast implementation on a compiled programming language with LAPACK-like libraries. The usage of Python also supplied us with higher order means of algorithms debugging, data printing, simple input-output, faster user interface implementation, integration with other modules, e.g., Matplotlib that emulates powerful graphical routines of Matlab.

The results of the computer simulation and OCP solution are stored in a database for further visualization. The visualization module allows comparing various scenarios by defining equations with simple scenario algebra.

D. The Modelling Scenarios and Multicriterial Assessment

Performing computer simulation in terms of the model, the user can access (to make relevant changes) the parameters of the main felling volume, tending fellings, burnouts and tree planting. Each of these parameters are

varied within 0% (none), 50% (a half of the volume), and 100% (total volume). Each scenario results of computer simulation are accounted using the following set of criteria:

1. $\max_j \min_i W^j(t_i)$, where $W^j(t_i)$ is standing crop (for all species and age groups) calculated in the *j*-th scenario at the year t_i (the greatest volume of felled forest);

2. $\min_j \max_i S_0^j(t_i)$, where $S_0^j(t_i)$, is the uncovered in forest area, it is calculated from *j*-th scenario in the year t_i (forest areas are better to be covered with forest);

3. $\max_j (S_5^j + S_6^j)$, where $S_5^j(t_i)$, and $S_6^j(t_i)$, respectively, correspond to the area of mature and overmature forests of all species calculated from *j*-th scenario for the forecasted period (it is better if ratios of the mature and overmature forest were greater);

4. $\min_j \max_i S_6^j(t_i)$, where $S_6^j(t_i)$ is the area of overmature forests of all species calculated from the *j*-th scenario in the year t_i (it is better if areas of the overmature forest are smaller).

Each calculated scenario is evaluated using this set of criteria. Due to the multicriteria nature of the best scenario choice problem, the principle of Edgeworth-Pareto is used to select scenarios as proposed variants to decision-making person.

In case of performing a solution of a control problem, a functional is also defined; the functional is added to the criteria or replaced one of them.

IV. APPLICATION OF CPCF IN HDSS

As we mentioned in the introduction section the logical inference in CPCF was used to deduce the model FDC graph and calculate parametric identification values of the model

The CPCF is not only machine-oriented, but it is also as a human-oriented automated reasoning method, mostly aimed at solving dynamic system control problems (DSCP), which include FDC model. The calculus has a number of properties that support the solution of DSCP. It has the only unique inference rule and simple scheme of axioms. CPCF supports modifiability of semantics of the derivative calculi (constructive, monotonic, temporal, etc.) [5,11], and, besides, it is possible to construct intuitionistic inferences of non-Horn formulas for a general class of formulas as compared to resolution method. The solemnization step is not required as CPCF supports both quantifiers. CPCF formulae can contain disjunctive branching in consequents, and the inference rule can be adjusted to control the inference branching after processing such formulae.

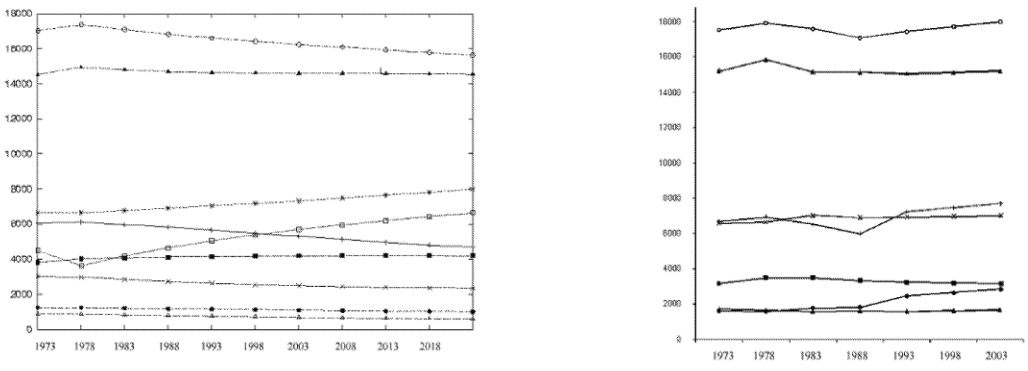
The inference of the model graph is constructed as satisfaction of a sequence of goals, which represent a general scenario of the model construction. In the process of the inference, the reasoning system issued queries to the database of forest inventory with special predicates having side effects. The results of the queries are the identified coefficients and initial values of the model. The metainformation on the database structure and attribute meanings is represented as a part of knowledge base and defined for each data source separately.

The model construction subsystem formed a basis of an experimental decision support system for sustainable resource consumption control of forest resources [8]. The usage of logical inference allowed us to raise level of generality of the model graph construction and its identifications. The expressive logical language allowed us to synthesize graphs for regional and forestry scales using the same knowledge bases but different databases and corresponding metainformation. Source data are also were processed by a special parts of the knowledge base.

The regional data is more specific and contain information on seven forest species and six age groups, whereas forestry data is more general: all species are joined to three classes (dark and light coniferous and deciduous forests). Age groups are also generalized; there are only three age groups (united young, middle-aged and maturing, and mature and overmature forest). The corresponding subset of rules reduced space of states of the graph to the sensible combinations and transition intensities are corrected in accordance.

A. Examples of Computer Simulation

The simulation software is being verified using a retrospective comparison. The year 1973 was chosen as initial time of computer simulation, as that year a reliable



a *b*

Figure 1. Verification of the Forest Dynamics and Control model

information was supplied by forest Inventory organizations.

The temporal dynamics of the forest resources (1973-2023) having regarded to natural species succession is sown in Fig. 3a., as well as fellings and burnouts (a constant increase in felling of coniferous forests is assumed), withdrawal of forests of the forest reserves as a result of capital construction, and the process of creation of forest crops and their conversion to young growth. The reference data of the verification is presented as graphics in Fig. 3b.

V. CONCLUSION

An approach to construction of hybrid decision support system on the base of integrated modules of computer simulation and optimal control problem solving has been considered in the paper. The usage of mathematical modeling in the process of decision elaboration in the HDSS allows decision maker to obtain additional objective information and produce counterintuitive decisions [1,4].

The construction of the HDSS and its environment is connected with the number of problem raging from technical to fundamental. Adaptation of the algorithms of model identification (structural and parametric) to the information at hand describing the domain and a concrete object under consideration is an example of the second case: most of the users that deal with economic domain experience difficulties in modeling and computer simulation in their unpredictable environment [12].

The support of the intermediate layer is provided by means of construction of a reasoning (inference) engine, which models the process of data analysis and model identification realized by mathematician that understands the problem domain. The engine itself is automatic theorem proving software based on original first-order calculi. The software successfully realized the process of identification of a forest resources dynamics and control model for an industrial region, namely, Irkutsk region, Russian Federation and its forestries. The usage of the logical engine allowed us, in particular, to generalize the process of database interaction by means of its metadata and to adapt general techniques of model construction to the database source data of various granularities.

Solving optimal control problems in the cycle of decision support system functioning allows decision maker to obtain more qualitative results of a criterion assessment, sacrificing the computational productivity as most of the optimal control procedures are based on iteration methods. In this case, it is necessary to develop new techniques of improvement on the base of higher order approximations and heuristic routines controlling the process of improvement.

We hope, when all will be elaborated and implemented, the usage of the HDSS will play the significant role as the decision support basis for sustainable forest resource consumption in Irkutsk and other world regions, whose ecological equilibriums depends greatly on the condition of their forest resources.

ACKNOWLEDGMENT

The research is carried on under support of Integration multidisciplinary project of Siberian Branch of Russian Academy of Sciences N 17 "Development of services and infrastructure of scientific spatial data for supporting complex multidisciplinary scientific research of Baikal nature territory" and Program of Siberian Branch of Russian Academy of Sciences IV.38.2.3. "New methods, technologies, and services for spatial and thematic data processing based on declarative specifications and knowledge". This work is also supported by Russian Foundation of Basic Research, grant No. 14-01-31449-mol-a.

REFERENCES

[1] I.N.Vladimirov, A.K.Chudnenko. Mathematical Modelling of Natural Phenomena. Issue "Multilevel modeling of the forest resource dynamics. Modelling of geographical processes and natural resources". V.4., No. 5, 2009, pp. 72-88. ISSN: 0973-5348.

[2] I.N.Vladimirov, A.K.Popova Modeling of the spatio-temporal dynamics of forest resources using the intelligent GIS // Geography and Natural Resources. 2009. № 1. pp. 26-31.[in Russian]

[3] E.I.Kuz'menko, Sh. Maksyutov, I.N. Vladimirov. Using the landscape map to assess the productivity of geosystems in the southern taiga of Western Siberia // Geography and Natural Resources. 2013. V.34, № 3. pp. 278-284. DOI: 10.1134/S187537281303013X.

[4] E.A.Cherkashin, A.K.Chudnenko. Hybrid GIS for Prognosis of Forest Stands. Vyestnik of Tomsk State University. Appendix No. 9(II). Reports of The V Allrussian Conference with International Participation «New Informational Technologies in Complex Systems Studies» - ICAM'04, Томск, 2004. – С. 69-72.

[5] S.N.Vassilyev, A.K.Zherlov, E.A.Fedunov, B.E.Fedosov. Intelligent Control of Dynamic Systems. Moscow. FIZMATLIT Publ., 2000 (in Russian).

[6] A.V.Davydov, A.A.Larionov, E.A.Cherkashin. On the calculus of positively constructed formulas for automated theorem proving // Automatic Control and Computer Sciences (AC&CS). 2011. Volume 45, Issue 7, pp. 402-407.

[7] A.A.Larionov, E.A.Cherkashin, A.V.Davydov. Theorem proving software, based on method of positively-constructed formulae. Proceedings of MIPRO 2011 - 34th International Convention on Information and Communication Technology, Electronics and Microelectronics, pp. 965-968.

[8] Badmatsyrenova S.B., Baturin V.A. Second-order successive improvement method for discrete control systems// Jornal of computer and systems sciences international. 2012, vol. 51, № 4. pp. 488-499. DOI: 10.1134/S1064230/12030033.

[9] S.B.Badmatsyrenova, V.A.Baturin. The control problem of the dynamics of the taiga landscape of Irkutsk region // OP&PM Surveys Appl. Industr. Math., 17, № 5, 2010, pp. 636-641. [in Russian].

[10] A.K.Cherkashin. A System of Mathematical models of Forest. Planning and Prognosys of Natural–Economic Systems. Novosibirsk. Nauka. 1984. pp. 46–57 (in Russian).

[11] A.Larionov, A.Davydov, E.Cherkashin. The calculus of positively constructed formulas, its features, strategies and implementation. Proceedings of 36-th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2013. pp. 1023-1028.

[12] A.Yu. Gornov, T.S. Zarodnyuk, T.I. Madzhara, A.V. Daneeva, I.A. Veyalko. "A Collection of Test Multiextremal Optimal Control Problems". Optimization, Simulation, and Control, Springer Optimization and Its Applications 76, A.Chinchuluun et al. (eds.). Springer Science+Business Media New York, 2013, pp. 257-274. DOI 10.1007/978-1-4614-5131-0_16.

Evolving Genetic Algorithms for Fault Injection Attacks

Stjepan Picek and Lejla Batina
Radboud University Nijmegen, Institute for Computing and Information Sciences (ICIS)
Postbus 9010, 6500 GL Nijmegen, The Netherlands
Email: stjepan@computer.org, lejla@cs.ru.nl
Domagoj Jakobović
Faculty of Electrical Engineering and Computing, University of Zagreb
Unska 3, Zagreb, Croatia
Email: domagoj.jakobovic@fer.hr
Rafael Boix Carpi
Riscure BV, The Netherlands
Email: BoixCarpi@riscure.com

***Abstract*—Genetic algorithms are used today to solve numerous difficult problems. However, it is often needed to specialize and adapt them further in order to successfully tackle some specific problem. One such example is the fault injection attack where the goal is to find a specific set of parameters that can lead to a successful cryptographic attack in a minimum amount of time. In this paper we address the process of the specialization of genetic algorithm from its standard form to the final, highly-specialized one. In this process we needed to customize crossover operator, add a mapping between the values in cryptographic domain and genetic algorithm domain and finally to adapt genetic algorithm to work on-the-fly. For the last phase of development we plan to go to the memetic algorithm by adding a local search strategy. Furthermore, we give a comparison between our algorithm and random search which is the mostly employed method for this problem at the moment. Our experiments show that our algorithm significantly outperforms the random search.**

I. Introduction

Genetic algorithms (GAs) are widely used to solve many different optimization problems. Since they are very general solvers, often it is necessary to customize the algorithm for some specific problem. Although the "No free Lunch" theorem states that when averaged over all problems all algorithms behave the same, that is true only when we posses no knowledge about the problem [1]. In this paper we start with standard genetic algorithm, and on the basis of experimental results we specialize it further to better relate to the problem at hand.

We are surrounded by small smart devices in our lives such as smart cards (contact and contactless), RFID tags, mobile phones etc. and their security is of utmost importance for a user. Considering the security and privacy of those devices there exists a special class of attacks that a malicious adversary can try, aiming at weaknesses of implementations. Those so called implementation attacks are divided into passive and active. The former explore some passive physical channel that is available while the device is processing secret data (such as power consumption or timing) and the latter is actively disturbing the computation aiming at finely tuned faults [2], [3]. The fault has to be effectively inserted such that the device does not break while an error computation will occur. This fact allows for the adversary, after inserting several "useful faults", to learn the secret key used. The famous Bellcore attack shows that having one faulty RSA signature and a correct one allows the adversary to factor the RSA modulus and hence completely break the system [4].

In this paper we consider so called glitching attacks where a sudden change in the supply voltage of a smart card can have the desired effect of a fault. However, inserting "useful faults" is considered to be very hard as there are many parameters that have to be tuned such as timing, glitch length, glitch voltage etc. The search in this multi-dimensional space is typically performed as a random search. Our work aims at improving this and deriving more efficient methods in finding the right ranges of all relevant parameters.

A. Related Work

As mentioned above, the research on fault attacks started around two decades ago with the seminal paper of Boneh, DeMillo and Lipton [4]. They showed the potential of this type of attacks assuming that they are possible. However, it took several years for a publication on an actual practical fault attack by Aumuller et al. [5]. The authors published in 2002 one of the first practical works on fault analysis, in which they describe a real-life scenario of the impact of injecting glitches in the clock and voltage lines of a cryptographic chip.

Van Woudenberg et al. describe a real attack scenario for an Optical Fault Injection attack where a laser is used to induce faults [6]. The authors also introduce the practical problem of setting the parameters for fault injection and they briefly discuss the lack of methodology.

Recently, a paper of Boix Carpi et al. [7] proposed a new algorithm that is a tailored search strategy but it is especially developed for finding the best choice of parameters for glitching. The authors also experiment with genetic algorithm and suggest investigating it further. We continue with this

recommendation aiming at setting common grounds for this new line research that could help security evaluators to assure that smart devices preserve user's privacy.

B. Our Contribution

In this paper we experiment with genetic algorithms when looking for parameters that can lead to successful fault injection. For the algorithm to behave as expected we needed to create a custom version of genetic algorithm and also interface it with a hardware-based system that inserts and evaluates fault attempts.

The rest of this paper is organized as follows: in Section 2 we present experimental setup, whereas in Section 3 we give experimental results of several search strategies as well as discussion about the results. Finally, in Section 4 we give a conclusion.

II. Experimental Setup

In this section we describe the fault injection scenario we follow, parameters of the problem and mappings between the solutions. On a more abstract level, the goal is to interface the system for injecting faults and evaluate them with the genetic algorithm framework. The role of genetic algorithm is to evolve new fault parameters and send those to the fault injection framework for an analysis. Motivation lies in the fact that finding such values is a difficult problem and there is no algorithm, except exhaustive search, that can do it. However, exhaustive search is often not a viable choice since there are constraints regarding the time that attacker has at his disposal. Furthermore, it is somewhat surprising that random search is regarded as the baseline strategy for this problem and as such is widely used.

Since random search is considered as a viable choice, that is also a pointer that research can go to the evolutionary computation area. We opted for genetic algorithms since they are directed random search method. In terms of an optimization problem, we want to minimize the time needed to find a set of parameters that can lead to the successful fault injection attack when we work in black-box scenario (no specific information about the device we attack).

A. Search Space Parameters

There are multiple parameters that need to be adjusted for a successful fault injection. Generally, they can be divided into those related with glitch shape, time and environment. In this work, we do not include environment or timing parameters in our experiments. The reason is the possibility to separate the search into two phases where in first phase we look for the adequate glitch shape and in the second phase for the correct timing periods to apply those glitches. Environmental parameters are not of interest since they can change even within the search and it is not clear how each of those changes affects the search.

Glitch shape parameters can be divided to glitch voltage [V] and glitch length [ns] parameters. With these assumptions, the goal is finding a suitable values for voltage and length that would provoke a successful fault injection. A pair of values for glitch voltage and length is considered a single point in search space (a possible solution).

B. Verdict Classes

Fault injection testing equipment and tool for the evaluation of introduced faults can output only verdict classes for each attempt. Therefore it is necessary to map each of those verdict classes to fitness values a GA can work with. There exist several possible classes for classifying a single measurement (attack attempt) [7]:

1) NORMAL: the target behaves as expected (the glitch is ignored)
2) RESET: the target resets as a result of the glitch
3) MUTE: the target stops all communication as a result of the glitch
4) CHANGING: multiple measurements in the same point yield responses that classify into different classes.
5) SUCCESS: the target response is a specific, predetermined value that does not happen under normal operation

If we define $F()$ as a function that scores the goodness of a point, and the higher $F()$ value is desired, then the formula for a single point classification is

$$F(SUCCESS) > F(CHANGING) > (F(NORMAL) == F(RESET) == F(MUTE)). \quad (1)$$

C. Interfacing Frameworks

For the fault injection (FI) framework we use Inspector program from Riscure BV. For genetic algorithms test suite we use the Evolutionary Computation Framework (ECF) [8]. ECF is a C++ framework intended for the application of any type of the evolutionary computation, developed at the University of Zagreb.

Since fault injections need to be done on the basis of the results from genetic algorithm, that makes it necessary for the fault injection system and genetic algorithm framework to communicate in real time. To be able to do that, we made an interface between those two systems.

The genetic algorithm is set to output one file for whole generation in each iteration of the evolution process. The fault injection software reads it and conducts the attacks with the values (search points) specified in the file. After each of the attacks, the FI software assigns a verdict class to each point and outputs appropriate parameter values and fitness values to a file. From that file the GA framework reads the fitness values and proceeds with the evolution.

The file system is chosen as a means of communication on the basis of the duration of candidate points evaluation; since each injection attempt can take up to several seconds, there is no need for faster process communication.

D. Target of Evaluation

We conduct all experiments with a smartcard based on an ATMega163+24C256 IC in CMOS technology. This card does not have any side-channel countermeasures or fault-injection

countermeasures. All processing on the card is performed in software, and the card is running on an external 4MHz clock frequency. The code we are attacking is the PIN authentication mechanism.

III. Search strategies and Results

Finding the correct parameter setting is a highly non-deterministic process, and countermeasures just add up to the complexity of this problem. As a result of this, security analysts usually set a value range for each parameter. Then, the fault injection setup is left testing thousands of different parameter configurations within given ranges for a-posteriori analysis. That search is typically done with strategies that involve a high search space complexity. Examples are Monte Carlo search for a successful setting within a set of value ranges, or test all possible value combinations within value ranges given a value step. Additionally, the parameter tuning (if performed) is usually done manually. As a result, the testing process can take days.

The analyst possesses some knowledge about the problem that is generally applicable, for instance, it can be assumed that two classes will always have the same relative positioning (NORMAL with low absolute values for glitch voltage/length, and RESET/MUTE class for high absolute values). It is not so interesting to search in the areas that are far from the decision boundary since it can be assumed that there are no outliers. However, there is no prior knowledge about the positions of those regions.

All the points are measured three times to enable the detection of changing behavior (the CHANGING class). Since a CHANGING class simply corresponds to different behavior in subsequent probes, a minimum of two is required for enabling detection of that class. The motivation for this lies in the fact that for some unstable protected targets more measurements per point are required to mitigate some side-effects during data acquisition.

A. Monte Carlo Search

This search strategy consists of performing measurements with randomly selected parameter combinations within the given initial search space. The random distribution for selecting values is considered to be uniform for each parameter present in the search space. This strategy is considered as the baseline search strategy.

A typical result for Monte Carlo search is given in Figure 1. All graphs are produced automatically by Inspector program and they depict only NORMAL, RESET/MUTE and SUCCESS verdict classes (CHANGING class is not shown even if found by the algorithm). Green dots represent solutions from NORMAL verdict class, BLUE represent solutions from RESET/MUTE and red dots represent solutions that belong to SUCCESS verdict class.

Two test runs are performed with the Monte Carlo strategy; a short one with 3072 measurements and a longer one with 76800 measurements. The short test completed with no SUCCESS measurement, and only a longer 76800 measurements test run produced 11 SUCCESS measurements. Because of the random nature of parameter values selection, there is also a significant number of repeating solutions.

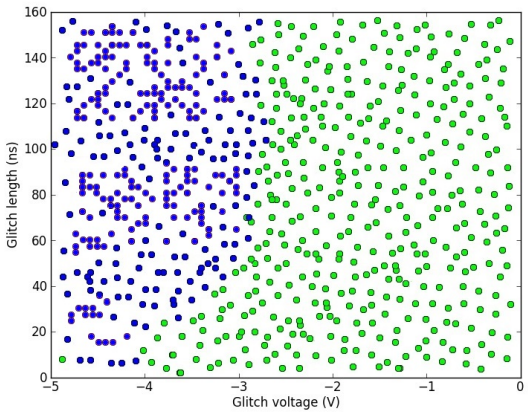


Fig. 1. Plot of measurement classification for Monte Carlo search.

TABLE I
Genetic algorithm parameters

Parameter	Parameter Value
Tournament size	3
Glitch length	[2, 150]
Glitch length resolution	2
Glitch voltage	[-5, -0.05]
Glitch voltage resolution	0.05
Population size	30
Stopping criterion	Stagnation in 10 generations
Mutation rate	0.1

B. Genetic Algorithm

For a genetic algorithm to work, we need to map verdict classes to distinct fitness values. Since we set the problem to be maximization, better classes will have higher fitness values. Mapping using in the initial version of the algorithm is NORMAL, RESET or MUTE verdict classes have value 1, CHANGING verdict class has value 8 and SUCCESS verdict class has value 10. Formally, the GA aims to find a (glitch length, glitch voltage) pair such that the fitness value F is maximal. Parameters of the search are given in Table I. Some of those values are set after the parameter tuning phase where we experimented with different population sizes and stopping criteria.

Since there are only a few different verdict classes, it means that a big portion of the search space will be impossible to differentiate. Therefore, we developed a version of tournament generational algorithm that is more appropriate for this specific problem. A single generation of the algorithm is described with the pseudocode as given in Algorithm 1.

In all the experiments, the population in the first generation is created randomly by the GA framework, where the initial fitness is zero for all individuals and then is sent to file to get the actual fitness values after the evaluation.

Algorithm 1 Genetic Algorithm - initial version

Require: count = 0, crx_count = 0
  repeat
    randomly select 3 individuals
    select best of the 3 individuals
    copy best to new population
    count = count + 1
  until $N \neq$ count
  replace old with the new generation
  repeat
    randomly select parents p_1, p_2
    $child_1$ = crossover (p_1, p_2)
    $child_2$ = crossover (p_2, p_1)
    replace p_1 with $child_1$
    replace p_2 with $child_2$
    crx_count = crx_count + 1
  until N * $p_c \neq crx_count$
  perform mutation on all individuals with p_m
  evaluate population

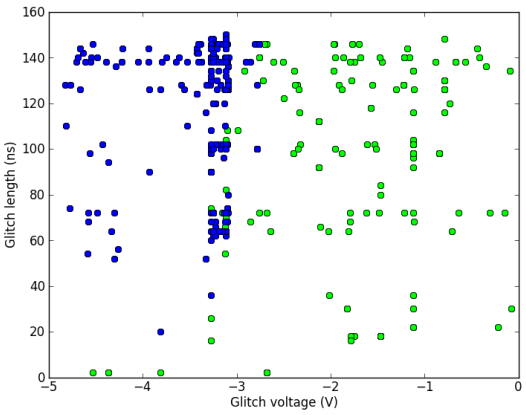


Fig. 2. Measurements for GA, initial algorithm.

Result from initial version of GA where the stopping criterion is 10 generations without improvement is depicted in Figure 2.

It can be observed that the algorithm outputs a valid point. By valid it is understood that the value is in the boundary between the two unimodal components. If speaking in terms of classifiers, the output point would be on top of the decision boundary between the two classes. The algorithm finished when no better fitness values could be found.

In Figure 3 we see a result of the same experiment but with 50 generations without improvement as a stopping criterion.

After running the first set of experiments we noticed some undesired effects. Once the algorithm finds a point with a high fitness value, it seems to fix one variable and explores the full range of the other. For this problem, spreading the search so much only in two directions is not so desirable in the sense that points too distant from each other that belong

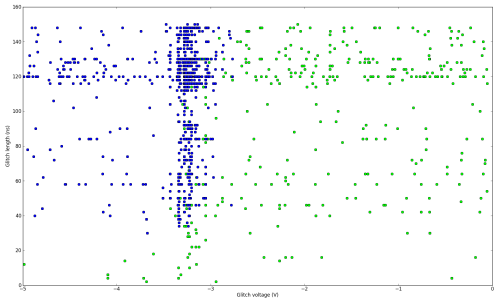


Fig. 3. Measurements for GA, initial algorithm, long run.

to the CHANGING class will likely classify to one of the low fitness valued classes (NORMAL or RESET/MUTE). To put it in other words, the decision boundary between classes will always have the same shape (convex, monotonically increasing curve and ideally it would mimic an exponential curve), so the direction that contains points belonging to the CHANGING class is not explored in depth by a vertical or horizontal line.

GA would be more successful in terms of finding more points that classify to CHANGING class by exploring the surroundings of a point that classifies into CHANGING inside a neighborhood of distance d (increasing d if needed) from it. Another effect that we observed is that once the GA finds a point belonging to the CHANGING class, the rest of the search space is ignored. This elitism is good in a sense that we want the search to be as short ass possible, but often it results in not exploring interesting parts of the search space sufficiently.

C. Improving Crossover and Parameter Tuning

Evolutionary algorithms in general (and GA in specific) do optimization by default, which means they always try to find points with a higher fitness value. The behavior of current algorithm is expected, especially with such small variances in fitness values. However, as we want to find as many points that belong to a certain class, we change the crossover operator to be more adequate.

New crossover operator crosses two points that should be of *different* classes, and it would choose a point between them that is closer to the better "parent" point. In this way we can point the algorithm to the desired region. Results for the new algorithm with improved crossover operator are shown in Figure 4. Although it may seem that the results depicted in Figure 4 are not better than those with the original algorithm, we remind that the graph does not show CHANGING class.

D. Optimization + Classification Algorithm

In a classical black-box optimization scenario, the best possible fitness value is not known in advance. The optimization algorithm will therefore always strive to find points with better values, regardless of how good the current solution is. If the best value is known, or if the goal is to find a solution of an acceptable predefined fitness value, then the algorithm is

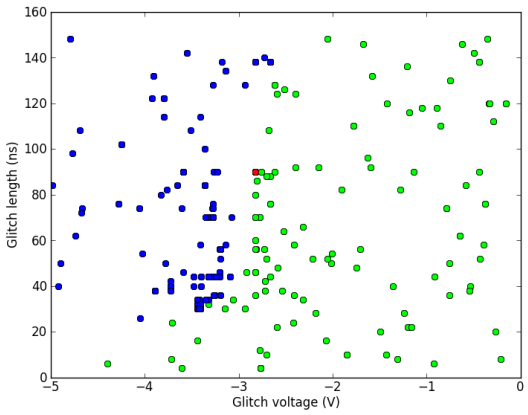


Fig. 4. Measurements for GA with modified crossover.

usually configured to stop as soon as it finds a single solution of the desired quality.

In this application we diverge from the described principle in the following ways:

1) we define a *target fitness* level where all the solutions having fitness equal or better than the target value are regarded as equally good and called *target points*;
2) instead of satisfying with a single target solution, the algorithm should find as many distinct target points as possible;
3) points worse than target value are divided into a number of *subclasses* corresponding to their fitness (we presume there is a small number of discrete fitness values).

The fitness mapping in this version is also not the same as in previous version of the algorithm, to allow distinction of different subclasses. Here we opted for a small change in values when mapping from verdict classes, so that NORMAL class retains the value of 1, RESET or MUTE classes have value 2, CHANGING class has value 8.5 and the target fitness level is set at 8. This allows the algorithm to perceive NORMAL and RESET/MUTE points as different subclasses, and all points that classify to CHANGING as target class points.

The next generation is produced in part by crossing the points from the current population of size N, and in part by selecting the better individuals by tournament selection operator. The proportion of crossover is controlled by the crossover probability (p_c), and the remaining individuals are selected with the tournament operator.

The selection of parents in the crossover operator is designed to promote diversity and explore boundaries between classes (points of different fitness values). The first parent is selected randomly from current generation; then, if the first parent is worse than the target level, the second parent is randomly selected from another class (point with a different fitness value). If, on the other hand, the first parent is of the target class, the algorithm tries to select another target class point as the second parent. If no such individual exists, a random parent is taken otherwise.

TABLE II
COMPARISON OF THE MONTE CARLO AND GA IN FINAL VERSION

Strategy	Measurements	#SUCCESS
Monte Carlo	3072	0
GA + classification	1560	8

The crossover operation also acts depending on the parent classes. If crossing points belong to different classes, the child is generated as a point between the parents in search space, but proportionally closer to the parent whose class is represented with a smaller number of individuals in current population. In this way, the search is guided towards the class that covers a smaller area of the search space. If crossing points are from the same class, the child is generated randomly between the parents in the search space.

After the crossover and tournament selection, a mutation operator is invoked on the new set of individuals, where the total number of mutations is determined with the mutation probability parameter (p_m). Algorithm 2 is a pseudocode for one iteration of the modified genetic algorithm.

Algorithm 2 Genetic Algorithm + classification

```
Require: crx_count = 0, mut_count = 0
  repeat
    select first parent
    if first parent of target class then
      try to find matching second parent
    else
      try to find second parent of different class
    end if
    perform crossover (depending on parent classes)
    copy child to new generation
    crx_count = crx_count + 1
  until p_c * N ≠ crx_count
  repeat
    select random individuals for tournament
    copy best of tournament to new generation
    mut_count = mut_count + 1
  until (1 − p_c) * N ≠ mut_count
  perform mutation on new generation with probability p_m
  evaluate population
```

A typical result of the modified GA version is shown in Figure5. In this case the algorithm succeeded in finding the most target class points in a limited number of measurements.

In Table II we give the comparison of the successfulness of Monte Carlo and genetic algorithm that uses modified crossover operator and classification.

IV. CONCLUSIONS

In this paper we explore a new option to successfully find the correct parameters for the fault injection. Currently employed methods can produce results, but are slow and often require more time than the security analyst is willing to spare.

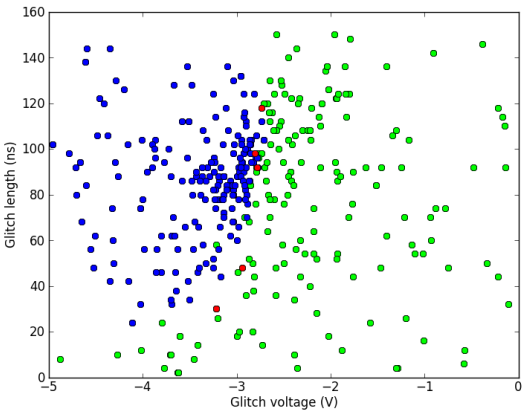


Fig. 5. Measurements for GA + classification.

We experimented with genetic algorithm where we started from its standard form and finished with a specialized algorithm. The results are promising and indicate that additional research is required in this area.

As a possible avenue for future research we plan to experiment with a memetic algorithm. That algorithm would comprise of the genetic algorithm in the last version plus an addition of a local search. When the algorithm finds a point that belongs to class INTERESTING (or better) from that point will inspect n points within distance d. If the new point belongs to a different class, the algorithm will keep that point and halve the distance value d for the next iteration of the algorithm. Additionally, since the genetic algorithm approach is working better than the baseline search strategy, it would be interesting to see how it behaves with fault injections that are based on electromagnetic radiation. The process would be similar to the scenario we describe here, only it would search for the parameters related with electromagnetic radiation.

ACKNOWLEDGEMENTS

This work was supported in part by the Technology Foundation STW (project 12624 - SIDES), The Netherlands Organization for Scientific Research NWO (project ProFIL 628.001.007) and the ICT COST action IC1204 TRUDEVICE.

REFERENCES

[1] D. H. Wolpert and W. G. Macready, "No Free Lunch Theorems for Optimization," *IEEE Transactions on Evolutionary Computation*, vol. 1, no. 1, pp. 67–82, April 1997.

[2] P. Kocher, J. Jaffe, and B. Jun, "Differential Power Analysis," in *Advances in Cryptology: Proceedings of CRYPTO'99*, ser. Lecture Notes in Computer Science, M. Wiener, Ed., no. 1666. Springer-Verlag, 1999, pp. 388–397.

[3] S. Mangard, E. Oswald, and T. Popp, *Power Analysis Attacks: Revealing the Secrets of Smart Cards (Advances in Information Security)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2007.

[4] R. D. D. Boneh and R. Lipton, "New threat model breaks crypto codes," *Bellcore 85 Press Release*, 1996.

[5] C. Aumüller, P. Bier, W. Fischer, P. Hofreiter, and J.-P. Seifert, "Fault attacks on RSA with CRT: Concrete results and practical countermeasures," in *Revised Papers from the 4th International Workshop on Cryptographic Hardware and Embedded Systems*, ser. CHES '02. London, UK, UK: Springer-Verlag, 2003, pp. 260–275.

[6] J. van Woudenberg, M. Witteman, and F. Menarini, "Practical optical fault injection on secure microcontrollers," in *Fault Diagnosis and Tolerance in Cryptography (FDTC), 2011 Workshop on*, 2011, pp. 91–99.

[7] R. Boix Carpi, S. Picek, L. Batina, F. Menarini, D. Jakobovic, and M. Golub, "Glitch it if you can: Novel parameter search strategies for successful fault injection," in *to appear in the proceedings of 12th Smart Card Research and Advanced Application Conference (CARDIS), Lecture Notes in Computer Science*, vol. xxxx, Berlin, Germany, Nov. 2013, pp. yyy–zzz.

[8] D. Jakobovic and et al., "Evolutionary computation framework," Jan. 2014. [Online]. Available: http://gp.zemris.fer.hr/ecf/

An overview of free software tools for general data mining

A. Jović*, K. Brkić* and N. Bogunović*
* Faculty of Electrical Engineering and Computing, University of Zagreb / Department of Electronics, Microelectronics, Computer and Intelligent Systems, Unska 3, 10 000 Zagreb, Croatia
{alan.jovic, karla.brkic, nikola.bogunovic}@fer.hr

Abstract - This expert paper describes the characteristics of six most used free software tools for general data mining that are available today: RapidMiner, R, Weka, KNIME, Orange, and scikit-learn. The goal is to provide the interested researcher with all the important pros and cons regarding the use of a particular tool. A comparison of the implemented algorithms covering all areas of data mining (classification, regression, clustering, associative rules, feature selection, evaluation criteria, visualization, etc.) is provided. In addition, the tools' support for the more advanced and specialized research topics (big data, data streams, text mining, etc.) is outlined, where applicable. The tools are also compared with respect to the community support, based on the available sources. This multidimensional overview in the form of expert paper on data mining tools emphasizes the quality of RapidMiner, R, Weka, and KNIME platforms, but also acknowledges the significant advancements made in the other tools.

I. INTRODUCTION

Data mining (DM) is the core step in knowledge discovery in datasets. It integrates all the analysis procedures that are required in order to reveal new and relevant information to an interested user. DM includes data preparation and data modeling. Datasets may be obtained from a variety of sources, including: traditional relational databases, data warehouses, web documents, or simple local textual files. It is important to prepare data in the most efficient way in order to extract as much information as possible [1]. After preparation, various models can be constructed, depending on the research goal. In order to properly interpret the models, standard evaluation and statistical procedures are pursued. The visualization of the results is also encouraged.

Free and publicly available software tools for DM have been in development for the past 20 years. The goal of these tools is to facilitate the rather complicated data analysis process and to offer all interested researchers a free alternative to commercial data analysis platforms. They do so mainly by proposing integrated environments or specialized packages on top of standard programming languages, which are often open source.

This paper focuses on the review of several such tools that have grown more efficient and useful over the years, some even comparable or better in certain aspects than their commercial counterparts. In particular, the main characteristics of RapidMiner [2], R [3], Weka [4], Orange [5], KNIME [6], and scikit-learn [7] will be outlined and compared. All of these free tools have implementations of general DM tasks. Other, more specialized tools such as Elki (clustering) or Anatella (big data), and tools with small DM community support (e.g. F#, GNU Octave) are not considered here.

The tools are compared in terms of general properties (language, license, etc.), core DM tasks (supported input, data preparation, data analysis, output), and support for some recent more advanced DM topics (big data, data streams, text mining, deep learning).

II. FREE DATA MINING TOOLS OVERVIEW

In a recent, 2013, poll published on the influential KDnuggets portal [8], regarding the use of DM tools in a real project, it is interesting to observe that in the top 5 tools there is only one commercial tool: Excel. The domination of free tools, mainly RapidMiner and R, probably stems from the maturity and availability of a large number of machine learning algorithm implementations. In Fig. 1, an adapted version of the poll is shown, with only free tools listed, for comparison.

Most of the modern DM tools are effectively software-based dataflow architectures. Some of the tools (e.g. RapidMiner, KNIME) are graphical integrated environments that enable visual component placement, connection, and dragging. The most common paradigm for such components is: pull the data in, transform, and push it further down the pipeline. The component only pulls the data in once all of its prerequisites are met. The under-the-hood implementation of such components is irrelevant for the average user, but more advanced users usually have the possibility of tackling with the code in order to improve it. Other tools (e.g. R) are just plain extensions of the underlying language in the form of specialized packages and/or GUI add-ons.

General characteristics of the six DM tools are listed in Table I. All of the tools have implementations for Windows, Linux, and Mac OS X operating systems. In the large Table II, the supported machine learning algorithms for each DM tool are summarized. The tools either implement an algorithm (+), use an external add-on (A) to support it, show some degree of support for the procedure (S), or do not implement it (-) at all. It must be noted here that the data in Table II should be considered temporary, because most of the tools are in constant state of upgrades. Nevertheless, it is important and useful to summarize their

TABLE I. GENERAL CHARACTERISTICS OF THE FREE DATA MINING TOOLS

Characteristic	RapidMiner	R	Weka	Orange	KNIME	scikit-learn
Developer:	RapidMiner, Germany	worldwide development	Univ. of Waikato, New Zealand	Univ. of Ljubljana, Slovenia	KNIME.com AG,Switzerland	multiple; support: INRIA, Google
Programming language:	Java	C, Fortran, R	Java	C++, Python, Qt framew.	Java	Python+NumPy+ SciPy+matplotlib
License:	open s. (v.5 or lower); closed s., free Starter ed. (v.6)	free software, GNU GPL 2+	open source, GNU GPL 3	open source, GNU GPL 3	open source, GNU GPL 3	FreeBSD
Current version:	6	3.02	3.6.10	2.7	2.9.1	0.14.1
GUI / command line:	GUI	both; (GUI for DM = Rattle)	both	both	GUI	command line
Main purpose:	general data mining	sci. computation and statistics	general data mining	general data mining	general data mining	machine learning package add-on
Community support (est.):	large (~200 000 users)	very large (~ 2 M users)	large	moderate	moderate (~ 15 000 users)	moderate

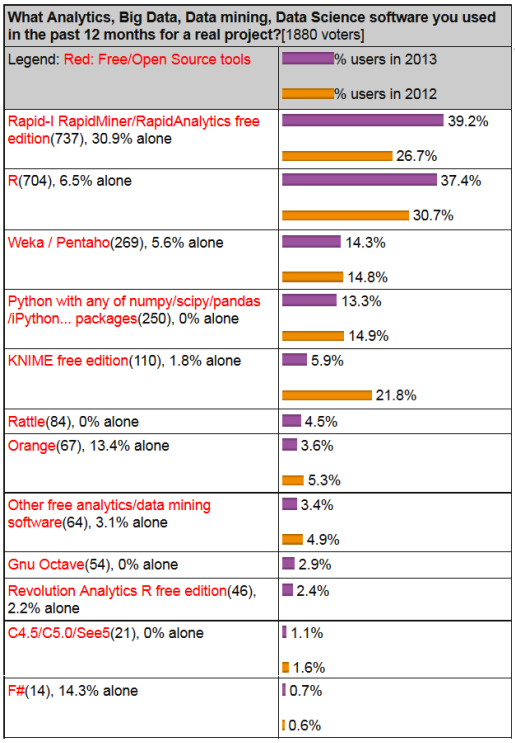


Figure 1. The community use of free DM tools, adapted from [8]

capabilities so that interested users can choose the appropriate environment for handling their problem. The algorithms shown in Table II were selected based on their significance and presence in most of the tools. Although the tools may contain some additional algorithms, not all of them could have been listed. Some of these were put in the appropriate rows with label "others". References to the work that describe the algorithms are not provided.

Table III lists the support for some of the more specialized topics in DM such as big data, text mining, etc.

III. TOOLS DESCRIPTION

A. *RapidMiner*

RapidMiner (previously: Rapid-I, YALE) is a mature, Java-based, general DM tool currently in development by the company RapidMiner, Germany. Previous versions (v. 5 or lower) were open source. The latest one (v. 6) is proprietary for now, with several license options (Starter, Personal, Professional, Enterprise). The Starter version is free with limitations only in respect to maximum allocated memory size (1 GB) and input files (.csv, Excel). The tool has become very popular in several recent years and has a large community support.

RapidMiner offers an integrating environment with visually appealing and user-friendly GUI. Everything in RapidMiner is focused on processes that may contain sub-processes. Processes contain operators in the form of visual components. Operators are implementations of DM algorithms, data sources, and data sinks. The dataflow is constructed by drag-and-drop of operators and by connecting the inputs and outputs of corresponding operators. RapidMiner also offers the option of application wizards that construct the process automatically based on the required project goals (e.g. direct marketing, churn analysis, sentiment analysis). There are tutorials available for many specific tasks so the tool has a stable learning curve.

Although RapidMiner is quite powerful with its basic set of operators, it is the extensions that make it even more useful. Popular extensions include sets of operators for text mining, web mining, time series analysis, etc. Most of the operators from Weka are also available through extension, which increases the number of implemented DM methods (Table II). The tool has very few shortcomings. The most important one is the transition to a novel model of business. It remains to be seen whether the transition to proprietary license will limit the number of its users, but it may not be helpful. The support for deep learning methods and some of the more advanced specific machine learning algorithms (e.g. extremely randomized trees, various inductive logic programming algorithms) is currently limited. However, big data analysis via Hadoop cluster (Radoop) is supported.

B. *Weka*

Weka is a Java-based, open-source DM platform developed at the University of Waikato, New Zealand. The software is free under GNU GPL 3 for non-commercial purposes. Weka has had mostly stable

TABLE II. DATA MINING ALGORITHMS AND PROCEDURES SUPPORTED BY THE TOOLS

Category	Name	RapidMiner	R	Weka	Orange	KNIME
Data import	textual files (.txt, .csv)	+	+	+	+	+
	specific input format files	+ (e.g..arff, .xrff)	A (foreign)	+ (.arff, .libsvm)	+	+
	Excel/ spreadsheet	+	A (xlsx)	-	-	A
	database table	+	A (RODBC)	+	+ (prototype)	+
	data from an URL	+	+	+	-	+
Feature selection	filters	+	A (FSelector)	+	+	+
	wrappers	+	A (FSelector)	+	+	+
Feature transformation	discretization	+	A (RWeka)	+	+	+
	normalization	+	A (RWeka)	+	+	+
	PCA	+	+	+	+	+
	ICA	+	+ (fastICA)	-	-	-
	MDS	-	+	+	+	+
	SVD	+	+	+	A	-
	random projections	-	+	-	-	+
Decision tree learner	ID3	A (Weka)	-	+	+	A (Weka)
	C4.5	A (Weka)	A (RWeka)	+	+	-
	CART	A (Weka)	A (RWeka)	+	+	A (Weka)
	others	+, A (own*, dec. stump)	+, A (own*, RWeka)	+ (dec. stump)	+ (own*)	+ (own*)
Classification rules	1Rule	+	A (RWeka)	+	-	A (Weka)
	PART	A (Weka)	A (RWeka)	+	-	A (Weka)
	RIPPER	+	A (RWeka)	+	-	A (Weka)
	others	+ (subgroup discovery)	A (RWeka)	+ (Ridor)	+ (CN2, subgroup discovery)	A (Weka)
Bayesian networks	Naive Bayes	+	+	+	+	+
	full bayesian network	A (Weka)	-	+	-	A (Weka)
	AODE	A (Weka)	A (AnDE)	+	-	A (Weka)
	others	A (Weka)	-	+ (DMNB, WAODE)	-	A (Weka)
Instance based learning	kNN	+	A (class, RWeka)	+	+	+
	LWL	A (Weka)	A (stats)	+	-	A (Weka)
	others	A (Weka) e.g. LBR	A (RWeka) e.g. LBR	+ (LBR)	+ (ML-kNN)	A (Weka)
Function based learning	regression analysis	+ (log., lin., polynomial)	+ (lin., nonlin.)	+ (log, lin.)	+ (log., lin., Lasso, PLS, trees, mean)	+(log., lin., poly, trees)
	ANN	+ (MLP)	A (nnet, RSNNS)	+ (MLP)	MLP	MLP, PNN
	SVM	+ (linear, evolut., PSO)	A (e1071, RWeka)	+ (SMO), A (libsvm)	+ (libsvm, liblinear)	+ (integr. libSVM)
	others	+ (Gaussian proc., RVM)	-	-	-	-
Hybrid learning methods	logistic model trees	A (Weka)	A (RWeka)	+	-	A (Weka)
	function trees	A (Weka)	-	+	-	A (Weka)
	Bayesian trees	A (Weka)	-	+	-	A (Weka)
	others	-	-	+ (DTNB), A (CBA)	-	A (Weka)
Ensemble	bagging	+	A (RWeka)	+	+	A (Weka)

Category	Name	RapidMiner	R	Weka	Orange	KNIME
learning	AdaBoost	+	A (RWeka)	+	+	A (Weka)
	random forest	+	A (random Forest)	+	+	+
	extremely randomized trees	-	+ (extra Trees)	-	-	-
	rotation forest	A (Weka)	S	+	-	A (Weka)
	LogitBoost	+	A (RWeka)	+	-	A (Weka)
	option tree	A (Weka)	-	+	-	A (Weka)
	stacking	+	A (RWeka)	+	-	A (Weka)
	other	+ (Bayesian boosting)	+ (boosted regression)	+ (Multi Boost)	-	A (Weka)
Hierarchical clustering	linkage-based	+	+	+	+	+
	AGNES	-	A (cluster)	-	-	-
	BIRCH	-	A (birch)	-	-	-
Centroid (partition) based clustering	*k*-means	+	+	+	+	+
	X-means	+	A (RWeka)	+	-	A (Weka)
	Partition Around Medoids	-	A (cluster)	-	A	A
	fuzzy *c*-means clustering	-	A (e1071)	-	A	+
Distribution based clustering	EM clustering	+	A (mclust)	+	-	A (Weka)
Density based clustering	DBSCAN	+	A (fpc)	+	-	A (Weka)
	OPTICS	A (Weka)	-	+	-	A (Weka)
ANN based clustering	Self-organising map	+	A(RSNNS)	A	+	A (Weka)
Association rules (unsupervised)	GSP	+	-	+	-	A (Weka)
	Apriori	A (Weka)	A (arules, RWeka)	+	+ (sparse, attr.-value)	A (Weka)
	FP-growth	+	-	+	-	A (Weka)
	Eclat	-	A (arules)	-	-	-
	Tertius	A (Weka)	A (RWeka)	+	-	A (Weka)
	others	-	A (arules Sequences)	+ (predictive Apriori)	+ (Apriori-SD)	A (Weka)
Evaluation methods and metrics	holdout	+	+	+	+	+
	cross-validation	+	A (cvTools)	+	+	+
	classif. accur., mean abs.err., MSE, κ	+	A (ROCR)	+	+	+
	TP, FP, FN, TN conf. matrix, precis., recall	+	A (ROCR)	+	+	+
	ROC, Lift chart, Cost-benefit	+	A (pROC)	+	+	+
Data visualization	histograms	+	+	+	+	+
	scatterplots	+	+	-	+	+
	other plots (e.g. box-whisker, mean/error)	+	+	-	+	+
	3D graphs (e.g. bivar. hist., surface plots)	S	A (lattice)	-	-	S (scatterplot3D)

* Own implementation of decision tree, mostly similar to C4.5 and CART

popularity over the years, which is mainly due to its user friendliness and the availability of a large number of implemented DM algorithms. It is still not as popular as RapidMiner or R, both in business and academic circles, mostly because of some slow and more resource demanding implementations of DM algorithms. Although it is not a single tool of choice in DM, Weka is still quite powerful and versatile, and has a large community support.

Weka offers four options for DM: command-line interface (CLI), Explorer, Experimenter, and Knowledge flow. The preferred option is the Explorer which allows

TABLE III. SUPPORT FOR SPECIALIZED AND ADVANCED DATA MINING TASKS

Name	RapidMiner	R	Weka	Orange	KNIME	scikit-learn
Big data	S (not free: Radoop)	A (ff, ffbase)	S (CLI, knowl. flow, distributedWekaHadoop)	-	A	S
Link, graph mining	-	A (igraph, sna)	A	-	A	-
Spatial data analysis	-	A (ggmap)	-	-	A	S
Time-series analysis	A	+, A(forecast)	S (several time series filters)	-	+	S (timeseries module has bugs)
Semi-super-vised learning	S	A (upclass)	S	-	S	+ (label propagation)
Data streams	+	A (stream)	A (massiveOnlineAnalysis)	-	+	S
Text mining	A	A (tm, RTextTools, qdap)	S	A	A	+
Paralelization	S (enterprise ed.)	A (snow, multicore)	S	-	+	A (joblib)
Deep learning	-	S (darch: incomplete)	-	-	-	S (Restricted Boltzmann Mach.)

the definition of data source, data preparation, machine learning algorithms, and visualization. The Experimenter is used mainly for comparison of the performance of different algorithms on the same dataset. Knowledge flow is akin to RapidMiner's operator paradigm in a sense that it allows one to specify the dataflow using appropriately connected visual components. Although not as visually appealing and extendable as RapidMiner, Weka's Knowledge flow still does the job.

Weka supports many model evaluation procedures and metrics, but lacks many data survey and visualization methods. It is also more oriented towards classification and regression problems and less towards descriptive statistics and clustering methods, although some improvements were made recently with respect to clustering. The support for big data, text mining, and semi-supervised learning is currently limited, while deep learning methods are still not considered.

C. R

The open-source tool and programming language of choice for statisticians, R, is also a strong option for DM tasks. R has been in development for the last 15 years and is the successor of S, a statistical language originally developed by Bell Labs in 1970s. The source code of R is written in C++, Fortran, and in R itself. It is an interpreted language and is mostly optimized for matrix based calculations, comparable in performance to commercially available Matlab and freely available GNU Octave.

The main language is extended by a myriad collection of packages for all sorts of computational tasks. Some of the packages are listed in Table II in parentheses. The tool offers only a simple GUI with command-line shell for input. It is certainly not a user-friendly environment because all commands need to be entered in the R language. The learning curve is steep, and although simple tasks such as drawing graphs and descriptive statistics can be learned rather easy, the language's full potential is difficult to master.

Some advanced users sometimes offer helpful reference cards with the list of the most significant functions [9]. From DM user's perspective, R offers very fast implementations of many machine learning algorithms, comparable in number to RapidMiner and Weka (from which a large number of algorithms is borrowed), and also the full prospect of statistical data visualizations methods. It has specific data types for handling big data, supports parallelization, data streams, web mining, graph mining, spatial mining, and many other advanced tasks, including a limited support for deep learning methods.

R's main problem is its language, which, although highly extendable, is also a difficult one to learn thoroughly enough to become productive in DM. Advancement for DM tasks in that direction is the Rattle package (author Dr. Graham Williams and other contributors) that offers a decent GUI for R. Rattle, which is in development from 2006, is similar to Weka's Explorer in the sense of user-friendliness. It loads separate packages from R upon request for a specific analysis. Rattle uses some of the R's standard implementations of DM methods and also additional packages. The only problem with Rattle is that it cannot use all of the R's algorithms or Weka's DM implementations. Nevertheless, Rattle is user-friendly and quite popular in DM community.

D. Orange

Orange is a Python-based tool for DM being developed at the Bioinformatics Laboratory of the Faculty of Computer and Information Science at the University of Ljubljana. It can be used either through Python scripting as a Python plug-in, or through visual programming. Its visual programming interface, Orange Canvas, offers a structured view of supported functionalities grouped into nine categories: data operations, visualization, classification, regression, evaluation, unsupervised learning, association, visualization using Qt, and prototype implementations.

Functionalities are visually represented by different widgets (e.g. read file, discretize, train SVM classifier etc). A short description of each widget is available within the interface. Programming is performed by placing

widgets on the canvas and connecting their inputs and outputs. The interface is very polished and visually appealing, offering a pleasant user experience.

One apparent downside of Orange is that the number of available widgets seems limited when compared to other tools such as RapidMiner or KNIME, especially because of the lack of integration with Weka. Still, the coverage of standard data mining techniques is quite good, as can be seen from Table II. Furthermore, there are a number of interesting widgets currently in development that can be found in the "Prototype" category, so it is reasonable to expect that the feature set will be expanded in the future.

E. scikit-learn

scikit-learn is a free package in Python that extends the functionality of NumPy and SciPy packages with numerous DM algorithms. It also uses the matplotlib package for plotting charts. The package keeps improving by accepting valuable contributions from many contributors and is supported by both INRIA and Google Summer of Code. One of its main strong points is a well-written online documentation for all of the implemented algorithms. Well-written documentation is a requirement for any contributor and is valued more than a lot of poorly documented algorithm implementations.

The package supports most of the core DM algorithms. However, several significant DM algorithm groups have been omitted currently, including classification rules and association rules. On the other hand, the package is strong in function-based methods including many general linear models and various types of SVM implementations. It is also quite fast despite being written in an interpreted language. This is mainly because the contributors are asked to optimize the code in various aspects, such as calling array based NumPy number crunching algorithm or writing wrappers for existing C/C++ implementations in Cython.

Despite its advantages, the use of scikit-learn still requires that one is a skilled programmer in Python because of its command-line interface. This will detract almost anyone not versed in this language because there are other tools that do not have this assumption.

F. KNIME

KNIME (Konstanz Information Miner) is a general-purpose DM tool based on the Eclipse platform, developed and maintained by the Swiss company KNIME.com AG. Its development started in 2004 at the University of Konstanz, Germany, and the initial version was released in 2006. KNIME is open-source, though commercial licenses exist for companies requiring professional technical support. According to the official website, KNIME is used by over 3000 organizations in more than 60 countries, and there seems to be a considerable community support.

The tool adheres to the visual programming paradigm present in most DM tools, where building blocks are placed on a canvas and connected to obtain a visual program. In KNIME, these building blocks are called nodes, and according to the official website, more than 1000 nodes are available through the core installation and various extensions. Nodes are organized in a hierarchy and can be searched by name within an intuitive interface. Each node is documented in detail, and the documentation is automatically shown within the interface once the node is selected.

A large repository of example workflows is available to facilitate quicker learning of the tool. One of the greatest strengths of KNIME is the integration with Weka and R. Although extensions have to be installed to enable the integration, the installation itself is trivial. Weka integration enables using almost all the functionality available in Weka as KNIME nodes, while R integration enables running R code as a step in the workflow, opening R views and learning models within R. Several other interesting free extensions are also available, e.g. JFreeChart extension that enables advanced charting, OpenStreetMap extension that enables working with geographical data, etc. There are also commercial extensions for more specific functionalities. Overall, KNIME seems to be one of the best choices for a user interested in a purely visual programming paradigm with a need for a large variety of nodes.

IV. CONCLUSION

Several DM tools were presented in this work. Overall conclusion is that there is no single best tool. Each tool has its strong points and weaknesses. Nevertheless, RapidMiner, R, Weka, and KNIME have most of the desired characteristics for a fully-functional DM platform and therefore their use can be recommended for most of the DM tasks.

REFERENCES

[1] D. Pyle, Data Preparation for Data Mining, San Diego: Academic Press, 1999.

[2] M. Hofmann and R. Klinkenberg, RapidMiner: Data Mining Use Cases and Business Analytics Applications, Boca Raton: CRC Press, 2013.

[3] Y. Zhao, R and Data Mining: Examples and Case Studies, San Diego: Academic Press, 2012.

[4] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," SIGKDD Explorations, vol. 11, no. 1, pp. 10–18, 2009.

[5] J. Demšar, T. Curk, and A. Erjavec, "Orange: Data Mining Toolbox in Python," Journal of Machine Learning Research, vol. 14, pp. 2349–2353, 2013.

[6] M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Kötter, T. Meinl, et al., "KNIME: The Konstanz Information Miner", in Data Analysis, Machine Learning and Applications (Studies in Classification, Data Analysis, and Knowledge Organization), Springer Berlin Heidelberg, pp. 319–326, 2008.

[7] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[8] G. Piatetsky, KDnuggets Annual Software Poll: RapidMiner and R vie for first place, 2013, Available at [last accessed 2014-02-22]: http://www.kdnuggets.com/2013/06/kdnuggets-annual-software-poll-rapidminer-r-vie-for-first-place.html

[9] Y. Zhao, R Reference Card for Data Mining, Available at: [last accessed 2014-02-23] http://www.rdatamining.com/docs/R-refcard-data-mining.pdf

Intelligent Multi Agent Systems for Decision Support in Insurance Industry

Ivan Markic*, Maja Stula**, Josip Maras**
*Croatia osiguranje d.d., Ljubuski, Bosnia and Herzegovina
**Faculty of Electrical Engineering, Mechanical Engineering and Naval Architecture in Split (FESB), Split, Croatia
ivan.markic@croatiaosiguranje.com, kiki@fesb.hr, josip.maras@fesb.hr

Abstract – MAS (Multi Agent Systems), belong to the field of AI (Artificial Intelligence), the study addressing the methods of construction of complex systems using a large number of agents and adjusting their behavior to a particular problem. MAS can be defined as groups of entities (agents) working together to solve problems not solvable by their individual skills and know-how. Business decisions are based on extraction of useful knowledge from different data sources. Those data sources, usually called the big data, might be huge digital data sets such as an internal data warehouse or external sources like the web. In order to process the big data, one needs to have smart solutions, analytics services, programming tools and applications. Intelligent agents can perform certain tasks such as big data processing, information retrieval and the creation of OLAP (On-Line Analytical Processing) cubes which represent multidimensional data. Given that volumes of data grow continuously, insurance companies have to be able manage them. The agent technology has the potential to become the key technology for management of ever-increasing amount of information. This paper describes how to develop MAS in order to gather useful knowledge from the data source and automate the specific process for better business decision.

Keywords: Multi Agent Systems, Artificial Intelligence, On-Line Analytical Processing, Decision Support, Big Data, Business Intelligence, JADE, OLAP, MAS, BI

I. INTRODUCTION

In the second half of the last century information sciences focused on data in general, and in particular on data collection, storage, transfer, sharing, analysis and, last but not least, data usage. Data volume increases on a daily level and we have become unable to store and analyze such amount. According to the Gartner's analysis, the amount of data grows by 50% every year. Gartner's analysis proves that information has become one of the most valuable assets that organizations have. Investors know how valuable information-centered businesses are and new information-based revenue streams are becoming a mainstream process across different industries. The "Big Data" technology innovation is challenging the existing information infrastructure. Information management is a top priority for organizations around the world. [1]

The term "Big Data" is used to describe large volumes of both unstructured and structured data, so big and complex that practically no one can use them with traditional management tools and applications for analysis. Text archive files, images, videos, social networks, e-commerce, medicine, the military, internet indexing, searching etc. account for the enormous data increase. Moreover, the information industry shifts to mobile computing, provision of cloud and it relies on analysis of large volumes of data.

Modern trends pose serious difficulties in management of large data volumes, and this is especially true of business analytics data because standard techniques cannot support it. New requirements expect from organizations to discover knowledge from their internal or external data sources which can help organizations to make better business decisions. The process of knowledge discovery is an integral part of the Business Intelligence (BI) technology approach that has been used increasingly, especially in last few decades, when a lot of new applications on that can help organization to improve business process and make better business decisions appeared on the market.

In order to improve their decision making process, organizations have created the right environment to respond to these new challenges. The environment which enables organizations to extract and store extracted information is called the Data Warehouse (DW). The DW covers a group of technologies and techniques which help employees to handle data more easily. The Figure 1 below shows typical architecture for supporting business decision systems within the organization. In order to have an effective system, appropriate methods have to be developed, which will gather and process data from multiple different data sources and in different data formats, such as the web, text documents, binary, etc. Such data should be cleansed and should be stored in the organization's data warehouse. Also, the data contained in the DW needs to be cataloged, maintained and corrected. Furthermore, the data warehouse needs to be updated as the new data comes in, for example, last month's sales. [2]

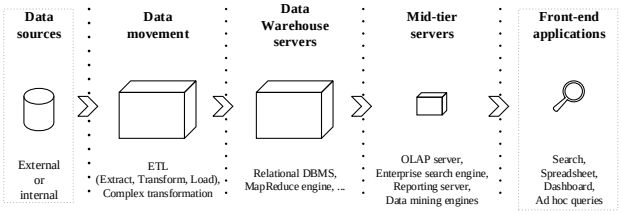


Figure 1 Typical decision support architecture

This paper presents information exchange between agents using a set of communication rules known as ontologies. The term "ontology" applies to business terms and rules that enable people to manipulate the real business problems and make business decisions based on intelligent agents. Introducing intelligent agents into big data processing, especially for business decision support, can reduce human workload. This is mainly due to agents'

independence, i.e. autonomous properties, which enable agents to act on behalf of humans in a business process, especially in the field of data acquisition and collection.

The following sections present an overview of the decision making technology process and agent-oriented technology programming approach using the JADE [1] technology (Java Agent Development Framework). The last section provides a simple example of exchange of business information between two agents that help the staff in an organization to make better business decision.

II. DECISION-MAKING TECHNOLOGY OVERVIEW

The Figure 1 presents typical architecture for supporting BI and knowledge discovery. The data comes from different sources, structured or unstructured. Data sources might be of different quality and in different formats. Sometime, the data might even be inconsistent. Problems of integrating, cleaning and standardizing such data are challenging but imperative because the necessary, successful environment exists.

The first step in establishment of architecture for a successful BI system as a basis for knowledge discovery and business decisions is collection of data from its source. The so-called ETL (Extract, Transform, Load) methods present the crucial procedure in the data gathering process. These methods help to establish and correct problems in data quality. They also facilitate efficient loading of large data volumes to the Data Warehouse (DW). [2]

The DW is a collection of data and collection of decision support technologies that assist an employee with sufficient knowledge (executive, manager and analyst) to make better and faster decisions. Data can be stored in many different types of databases; however, the data in DW is stored on a single site. In order to create a DW, data should be extracted from the source, cleansed (free of errors), transformed (summarized) and loaded into a data store – DW (ETL). DW servers are divided from operational databases and they have their own database which is used only for decision support. Data stored in a DW share the following characteristics: *Subject oriented* – data is logically organized around the main item, e.g. customers, sales, etc., *Integrated* – all of the data organized around such an item is combined and can be analyzed together, *Time variant* – historical data is maintained in historical form and *Nonvolatile* – data is read-only, not updated or changed by users. [3]

Once data is ready for querying, i.e. cleansed, filtered and corrected, employees should be able to query and perceive it from multiple dimensions. The Online Analytical Processing (OLAP) enables operations like filtering, aggregation, pivoting, rollup and drill-down of multidimensional data. The main difference between DWs and the OLAP is a different data view. DWs are based on relational technology while OLAP uses a multidimensional view of aggregated data. Use of a multidimensional view provides quick access to further analysis. A data analyst, manager or executive can analyze data from different views, in a fast and interactive way. The OLAP systems have the ability to answer to the "*who?*" and "*what?*", but it is their ability to answer the "*what if?*" and "*why?*" questions that sets them apart from Data Warehouses. The OLAP, with its capabilities, transforms the DW into strategic information, moving data access from information to knowledge. All the applications that use the OLAP have to be able to deliver information just-in-time. [4] The OLAP systems work with very large databases, up to petabytes, and thanks to that feature, the OLAP can be appropriate for intensive workloads. In addition, it is more tolerant to ACID[2] design philosophies [5]. The OLAP is nothing else but the *multidimensional* view of data in the DW. In multidimensional models a set of *numeric measures* exist that are subject to analysis. Such measures include sales, the budget, inventory etc. *Dimensions* represent the context for the measure. Examples of dimensions associated with the sale amount include the city, product name, date, etc. All dimensions uniquely determine the measure. Each dimension is further described by a set of attributes. For example, the *Product* dimension has the category of attributes and the industry of the product. The Figure 2 depicts the multidimensional view of the data.

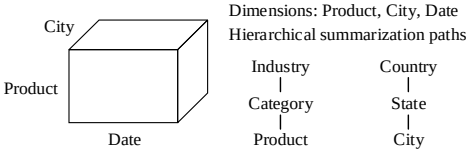


Figure 2 Multidimensional data

The OLAP conceptual model has features such as *aggregation* of measures by one or more dimensions. This feature is one of the key features in the OLAP. Computing and ranking of the total sales of a policy by a broker or within a city can be examples of aggregation. Another popular operation in the OLAP is *comparison* between two measures – for example, the sales and budget aggregated by certain dimensions. One of the most important dimensions in the decision support analysis is time, for example, the trend analysis. [7]

III. MULTI-AGENT SYSTEMS

The Agent Oriented Programming (AOP) is a software paradigm that brings concepts from the artificial intelligence theory to the center of the distributed systems. The AOP essentially shapes the applications as collections of components called agents, characterized as autonomous, with the possibility of communication. Being autonomous as they are, they have the capability to handle complex tasks. Agents must be able to communicate, negotiate and

[1] JADE is distributed agent platform created according to FIPA standards. It is developed in the programming language Java (J2SE), and it is being distributed via „Telecom Italia“, available as an open source platform with LGPL license. Source: http://jade.tilab.com/

[2] ACID represents design philosophies which focus on consistency and it is traditional approach of databases. ACID means: A-Atomicity, C-Consistency, I-Isolation and D-Durability. [6]

cooperate in "open" environment. Given that they have the ability of communication, they are able to interact with other entities, work for their own agenda and for agendas of other agents. The architecture of agent-oriented applications is exclusively peer-to-peer, implying that every agent is capable to initiate communication and to reply to requests at any time. A number of independent issues must be resolved when adopting an Agent system to solve specific domain problems - for example communication between agents. Multi Agent Systems should be on the top of agent-oriented infrastructure as independent, allowing developers to focus on the business logic. [8]

A. The JADE architecture

Applications developed on the JADE platform are comprised of a set of components called „*Agents*", each bearing a different name. Agents perform tasks and they communicate via message exchange. Agents live at the top of the platform that provides the basic information, such as message delivery. Just as the methods, objects etc., the programming agents represent one form of programming abstractions. The difference between an agent and object is the fact that the object is a programming abstraction describing the methods and attributes of the programming component, while the agent is a software abstraction at an even higher level of abstraction, enabling the simple and powerful tool for construction of complex programming entities. Unlike objects defined by attributes and methods, the programming agent is defined by its behavior. This fact is very significant because it reduces the programming of agent systems to the specification of the certain agents' behavior.

The JADE is completely developed in the Java programming language. It simplifies the implementation of Multi Agent Systems respecting the FIPA[3] (Foundation for Intelligent Physical Agents) rules, and via the graphical tools, it supports the development and testing phases. The agent platform can be distributed to several computers being controlled via graphical interface implemented in the JADE Framework. The programming agent needs to be: autonomous, mobile, purposeful, stabile, communicative and capable of learning. The JADE programming platform includes one or more containers. Containers can be executed on different hosts (computers, mobile devices etc.) making a distributed platform. Each container may contain one or more agents. Figure 3 [9] represents the basic elements of the JADE architecture. [8]

The Main container is a special form of container, capable of containing agents, but still very much different from other containers because the Main container is a place at which other agents register and it includes two different agents: the AMS (Agent Management System) represents an authoritative agent, the only one that is able to start and stop the platform, and the DF (Directory Facilitator) represents the "Yellow Pages", as we might call it, a place used by agents to advertise their services, that the other agents can browse through to search for information, depending on their needs [9]. The JADE platform may contain multiple containers like Container 1, on the Host 3, which contains the Agents A2 and A3.

Agents can communicate transparently, depending on their location; for example A2 or A3 can communicate; as well as A1 and A2 and although they belong to different containers, they are on the same platform. Communication is possible between A1 and A5, even though they belong to different platforms. Communication is based on asynchronous passage of messages. The message format is determined by the ACL (Agent Communication Language), a language defined by the FIPA (Figure 4 [11]). [9]

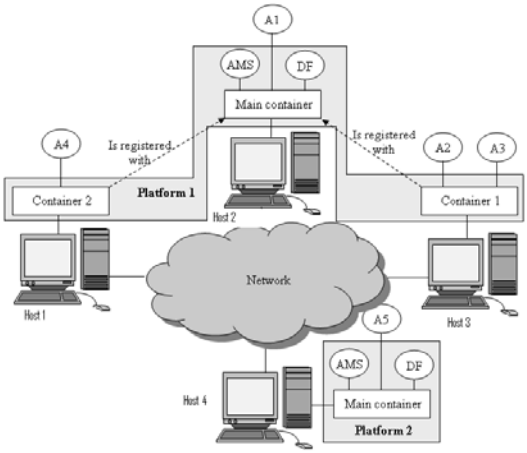


Figure 3 JADE architecture

A FIPA message includes the following data: the *Sender agent* - sending the message, the *Receiver agent* – receiving the message and the *Message* itself - intended for communication acts representing the sender's intention. For example, when the agent sends the INFORM message, it wants to inform the receiver of the fact (e.g. INFORM "It's raining today"). When the agent sends the REQUEST message, it demands an action from the receiver. The FIPA differentiates between 22 possible communication acts, covering 95% (it does not support the remaining 5%) of the possible situations. Content represents information contained in the message. [9]

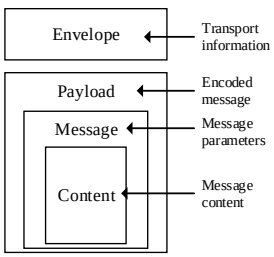


Figure 4 FIPA Message Structure

[3] FIPA (Foundation for Intelligent Physical Agents) is a nonprofit standardization organization dealing with the standardization of the programming agents and agent platforms. FIPA standardizes the architecture of agent systems, agent management, their communication etc.

An example of the FIPA-ACL (Agent Communication Language) message with the REQUEST performative:

```
(REQUEST
        :sender (agent-identifier :name Ivan)
        :receiver (agent-identifier :name Mark)
        :ontology travel-assistant
        :language FIPA-SL
        :protocol fipa-request
        :content
        ""((action
                (agent-identifier
                        :name Ivan)
                (book-hotel
                         :arrival 15/10/2006
                        :departure 05/07/2002 ... )
        ))""
)
```

[Part of code from [8]]

```
(PROPOSE
        :sender (agent-identifier
                :name Main@MY_IP/JADE
                :addresses (sequence MY_URL/acc ))
        :receiver (set
                (agent-identifier :name Seller@MY_IP/JADE
                :addresses (sequence MY_URL/acc )) )
                :content  "MAIN:
                        Number of policies: 8
                        Premium: 2900.0
                :reply-with
                        Seller@MY_IP/JADE136
                        :conversation-id  policy
)
```

[a part of the code from insurance example presented in the paper]

When a certain object is additionally described by a set of complex information, it is necessary to define the syntax for the receiver to be able to read the message content in order to obtain the necessary information. Ontologies represent abstract concepts and a relation for a specific domain. They can be shared and used many times across the internet by knowledge representation languages such as the Resource Description Framework (RDF) or Web Ontology Language (OWL).

According to the FIPA terminology, this syntax is known as the content language. The FIPA does not enforce these standards; they are recommendations how to communicate with AMS and DF agents –most significant in the JADE architecture.

An example of the content language recommended by the FIPA:

(Policy :number 123456 :seller Ivan :insured Mark)

When the agent receives this string, it has to be able to convert the string in the appropriate format to understand the information contained in the string. It also has to be able to share the meaning with the sender of the string and symbol concept („Policy, number, seller, insured") used for the structure expression. This set of concepts and symbols is known as ontology. The ontology is independent and inherently related to the corresponding research domain. Ontology-supported work makes the JADE capable for complex content operation. Software developers can choose to work with JADE classes, JAVA serialization, XML, or with some other technology, depending on their needs. The classification offered by the FIPA is the one separating the ACL message into two sections: *predicates* and *terms*, as indicated in the Figure 5. [8]

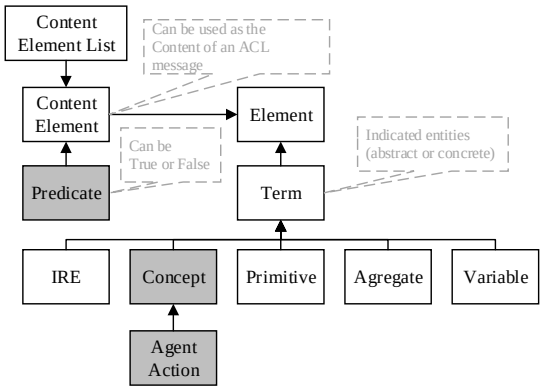


Figure 5 Reference model

Predicates are expressions describing the status of a certain object and they can be true or false:

(Insured (Person :name Ivan) (Municipality :name Split))

Description of the insured person, "person named Ivan" from "the Municipality of Split".

Predicates are used for the INFORM or QUERY-IF message content exchange, e.g. if two agents want to exchange some facts.

Terms are expressions identifying entities that exist in the real world. They can be abstract or concrete. They are classified into:

- Primitives – entities such as string and integers
- Concepts – entities with complex structure

(Person :name Ivan :Age 30)

Concepts are referred to inside predicates:

(Policy :number 1234 :insured(Person :name Ivan))

- Agents' actions: are special concepts that indicate action that can be perform by some agents.

(IssuingPolicy (Policy :number 12345) (Insured :name Ivan))

- Aggregations are entities or groups of other entities

(Sequence (Person :name Ivan) (Person :name Mark))

- IRE (Identifying referential expressions) are expressions that identify the entity (or entities) for which a given predicate is true
- Variables are terms frequently used in queries.

Agents have to be able to perform several competing tasks to respond to different requests. For this purpose, every agent has the possibility to implement different behavior types, ranging from those simple ones, performed only once to those performed until a certain condition is fulfilled, unless otherwise specified. The Figure 6 shows the JADE framework behavior hierarchy. [8] The behavioral state has to be contained in corresponding agents. Every agent cares about its behavioral state, such as the action completion and the action blockage.

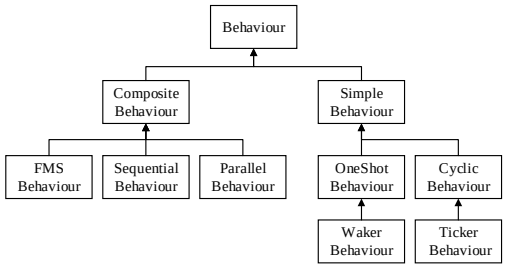


Figure 6 The Jade behavior hierarchy

The FIPA has specified interaction protocol stacks used for inter-agent communication. For every conversion, the JADE differentiates between the roles of the *initiator* and the agent replying (*Responder*) to the messages. The Figure 7 shows inter-agent communication structure. The Initiator sends the message, and the responder may reply with the message "not understood", refuse to reply with the message "refuse" or agree with the message via "agree". If the agent replies with "agree", the message is sent so quickly that it can be considered as an unnecessary burden. [10]

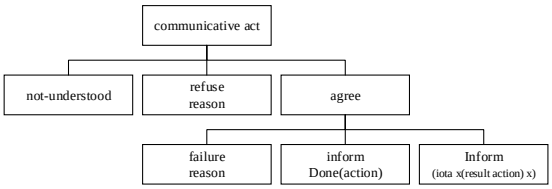


Figure 7 The Interaction Protocol Homogeneous Structure

IV. INSURANCE MAS EXAMPLE

This example comes from the Insurance domain and describes one of many situations in which employees need information to make a decision about their business actions. This is a problem not only for employees working in headquarters or static workers in other offices, but also for mobile employees. As the number of mobile and online users in the world continues to grow, so do the ways in which mobile users can gather the right information. All employees, ranging from a CEO to a seller, can request and obtain, automatically or on demand, a critical alert or information that is important for their business operations to their mobile phone or a smart phone. Every mobile phone equipped with appropriate software becomes a mobile agent capable of imitating communication with the headquarter agent who will respond to their queries.

To demonstrate this, we have developed two agents with different roles using the JADE architecture. One of them has a double role and represents the most important agent in the process since it discovers knowledge and upon discovering it, it sends it as information to mobile agents. Mobile agent is the second agent that only initiates communication.

The database includes two tables: one, titled "Policies" includes the concluded policies, and the second, titled "Persons" that contains data of the policy holders. The Figure 8 indicates the database structure:

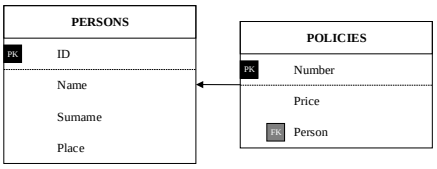


Figure 8 Database structure

The Agent that discovers knowledge, called the "Main" agent, obtains the statistical figures regarding the number of issued insurance policies and the realized premium. When the "Main" agent finds the right information it sends it to mobile agents, known as the "Seller" agents. The Main agent offers to the Seller agent available information that it has and offers the Seller agent to accept or refuse it, depending on the Seller's needs. The Seller agent performs the "Seller"behavior, sending the query on the schedule time passing through the corresponding communication steps (Figure 9). Every step requires corresponding behavior, depending on the message type.

The next code snippet covers a part of communication between the Main agent and the Seller agent.

```
if (messageRecieve.getPerformative() ==
ACLMessage.CFP){
messageAnswer.setPerformative(ACLMessage.INFORM);
messageAnswer.setConversationId("policy");
messageAnswer.setContent("MAIN: Available ...");
//Sent message
myAgent.send(messageAnswer);
} else if
(messageRecieve.getPerformative() ==
ACLMessage.ACCEPT_PROPOSAL){
messageAnswer.setPerformative(ACLMessage.PROPOSE);
messageAnswer.setConversationId("policy");
messageAnswer.setContent("MAIN: Number …");
//Send message
myAgent.send(messageAnswer);}
```

In the first step, the Seller agent sends the CFP (Call for Proposal) request for the available statistics. The behavior defining the CFP message is „SendCfpRequest". The Main agent grasps that this is a CFP message and sends back the INFORM message, informing the Seller agent about the available statistics.

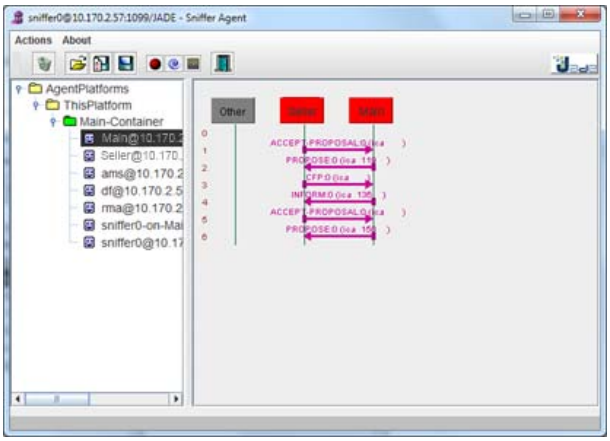


Figure 9 JADE Communication between Seller and Main Agent

The Seller agent accepts the proposal from the Main agent via "ReceiveCfpRequest" behavior and returns the message ACCEPT_PROPOSAL. The behavior defining the proposal is "CofirmRequest", whereby the Main agent, using the appropriate methods, contacts the distributed databases and, by performing the queries (SQL select), the Main agent sends the corresponding information via the message PROPOSE. Then, the Seller agent accepts the statistics via the "AcceptStatistics" behavior and prints them to the console.

After running the application, the console displays the following:

```
------------------------------------------------------------------------
Agent container Main-Container@10.170.2.57 is ready.
------------------------------------------------------------
SELLER: Request for statistic?
MAIN: Available statistics are number of sold policies and
premium!
SELLER: Accept! Give me all statistics.
MAIN:
        Number of sold policies: 8
        Premium: 2900.0
**************************************
```

V. CONCLUSION

Due to the large amount of information, intelligent software agents organized in the Multi Agent Systems represent a solution to the problem of gathering, processing and using information. Intelligent software agents are a collection of technology that ensure development of advanced services because it enables creation of intelligent services, tailor-made to the needs of a user. By using software agents, certain tasks, otherwise performed independently by users, can now be automated. Intelligent software agents support business processes due to their great potential, especially in terms of processing and maintenance of large amounts of information stored in heterogeneous, dislocated information sources. Implementation of a business process solution through the Multi Agent System based on the JADE framework, such as the one presented in the paper, is straightforward and intuitive. With the development and spread of new mobile hardware, problems such as detection of individual system entities and users and their communication are entering into a new phase. New trends and increase in the number of users of smart devices result in larger number of services these devices can provide. The fact that each user has a smart device in his/her immediate vicinity and that he/she can store new data at any moment, indicates the need to synchronize this data with other systems, as well as the need for various data access.

VI. REFERENCES

[1] Information 2020: Big Data and Beyond
http://www.gartner.com/, Business Intelligence & Information Management, Featured Webinar

[2] An Overview of Business Inteligence Technology
Communication of the ACM
August 2011, Vol. 54, DOI: 10.1145/1978542.1978562, Surajit Chaudhuri, Umeshwar Dayal, Vivek Narasayya

[3] Data Warehousing, Data mining, OLAP and OLTP technologies are essential elements to support decision-making process in industries, (IJCSE) International Journal on Computer Science and Engineering, Vol. 02, No. 09, 2010, 2865-2873, G.Satyanarayana Reddy, Rallabandi Srinivasu, M.Poorna Chander Rao, Srikanth Reddy Rikkula, ISSN: 0975-3397

[4] OLAP Council
http://www.olapcouncile.com

[5] Cloud-hosted databases: technologies, challenges and opportunities, Sherif Sakr, Springer Science+Business Media New York 2013

[6] CAP Twelwe Years Later: How the rules have changed, Eric Brewer, University of California, Berkeley, Computer magazine, February 2012.

[7] An Overview of Data Warehousing and OLAP Technology, Surajit Chaudhuri, Umeshwar Dayal

[8] Developing Multi-Agent Systems with JADE
Fabio Bellifemine, Telecom Italia, Italy
Giovanni Caire, Telecom Italia, Italy
Dominic Greenwood, Whitestein Technologies AG, Switzerland

[9] JADE Administration Tutorial
http://jade.tilab.com/doc/tutorials/JADEAdmin/index.html, Author: David Grimshaw (Ryerson University), March 26, 2010

[10] JADE Programmer's Guide
http://jade.tilab.com/doc/programmersguide.pdf
Fabio Bellifemine, Giovanni Caire, Tiziana Trucco (TILAB, formerly CSELT), Giovanni Rimassa (University of Parma)
08-April-2010. JADE 4.0

[11] http://www.fipa.org
Foundation for intelligent physical agents

Solving tourist trip planning problem via a Simulated Annealing Algorithm

Kadri Sylejmani*, Atdhe Muhaxhiri[x], Agni Dika* and Lule Ahmedi*
*Department of Computer Engineering, University of Prishtina, Prishtina, Kosova
[x]ProSolution, Prishtina, Kosova
kadri.sylejmani@uni-pr.edu, atdhe.muhaxhiri@prosolution-ks.com, agni.dika@uni-pr.edu, lule.ahmedi@uni-pr.edu

***Abstract*—Nowadays, by help of tourist trip planners, in many travel destinations, the trip itinerary could be prepared automatically. The trip planners enable customization of trip itinerary based on tourist preferences, available time and budget, where the itinerary is optimized by using optimization techniques from the field of metaheuristics. In this paper, we present a new approach based on simulated annealing algorithm for solving the tourist trip planning problem. The hard constraints include limited trip duration, working hours of points of interest (POIs) and tourist budget, while tourist preferences and traveling time comprise the soft constraints, which take part in a two component function for evaluation of quality of solutions. The search space is explored by using three types of operators, namely *insert*, *swap* and *shake*. The algorithm is tested against some real test data for the city of Prishtina in Kosova. In addition to fine tuning the values of algorithm parameters, further experiments are made for exploring different approaches of construction of initial solution, time duration of algorithm execution for various trip lengths, and influence of *shake* operator into the solution quality.**

I. Introduction

A tourist trip itinerary is plan for visiting a number of POIs in city (or a wider geographic region) during a limited period of time. In general, the itinerary contains a list of visits to POIs that are ordered based on the time of visit, where each visit might be annotated with a starting and ending time, visit duration, a short description of POI, etc. In addition, a trip itinerary might include a map with navigational information for traversing between individual POIs.

Usually, in a touristic city/region there might a high number of POIs which could be potentially visited, nevertheless the trip duration is usually limited to few days, and as result, only a subset of them can be actually visited. In addition, often it happens that the tourists preferences about types (e.g. monument, castle, statue, etc.) and categories (e.g. architecture, archeology, nature, etc.) of POIs are such that require visiting POIs that are located far from each other. Therefore, it is a complex task for the tourist to manually prepare the trip itinerary that would include the POIs that are the most preferred to her/him.

In general, the tourist trip planning problem is known as Tourist Trip Design Problem (TTDP) [1]. Given a number of POIs in a city/region, the goal is to prepare an itinerary that includes visits to POIs that meet, to a great extent, the preferences of the tourist. Depending on the kinds of features or constraints being considered for the problem, different problem modelings from the literature could be used to model the tourist trip planning problem. The Orienteering Problem (OP) [2] and the Team Orienteering Problem (TOP) [3] are used to model single and multiple day itineraries, respectively. The OP with Time Windows (OPTW) [4] and TOP with Time Windows (TOPTW) [5] enable modeling working hours of POIs, whereas Multi Constraint TOPTW (MCTOPTW) [6] allows enforcing additional constraints for the itinerary, such as maximum budget to spent, or maximum number of POIs of certain type (e.g. statue, castle, etc.) to be visited. The OP with Hotel Selection (OPHS) [7] enables selection of a hotel at the end of the trip of each tour.

Golden et al. [8] prove that OP is a NP hard problem, therefore many researcher tackle OP and its deriving extensions by using various approaches from the field of metaheuristics. Vansteenwegen et al. [9] discuss the most effective approaches that solve OP and its respective extensions.

In the process of trip itinerary planning, a method that estimates the interest of the tourist about POIs is required [10]. A such a method would consider the preferences of the tourist and features of the POIs, and based on their reciprocal match, it would generate the estimated satisfaction factors of the tourist with each POI. Nevertheless, introducing a new such a method is out of scope of this paper, hence, a method from the literature is used for that purpose.

The goal of this paper is to present a new algorithm that plans the tourist trip itinerary by considering multiple day tours, tourist budget, working hours of POIs and breaks during the trip of a day. In comparison to existing modelings in the literature for the trip planning problem, where optimization is done by only maximizing the satisfaction factor of a tourist with POIs, in our case, the optimization includes a second component, that is minimization of the total traveling time of the tourist.

The remaining of this paper is organized as follows: Section II discusses the state of the art algorithms, Section III presents the mathematical modeling of the problem, Section IV presents the proposed solution, Section V gives the computational experiments and Section VI outlines the main conclusions of the paper.

II. State of The Art

Vansteenwegen et al. [5] are the first to model TTDP as TOPTW problem. As result, they develop an approach based on the Iterated Local Search (ILS) framework. The neighborhood structure of the algorithm comprises of an insert step (for inserting new POIs) and a shake step (for escaping from local optima). The insert step adds consecutively new

points into the tours. The shake step removes one or more visits from each tour. The computation time of this algorithm is, with a factor several hundred, faster than the other presented approaches in the literature for TOPTW problem. This feature makes the algorithm suitable for application in the domain of tourist trip planning. Later, Garcia et al. [6] extend ILS to solve the Multi Constraint TOPTW, which allows modeling the constraints about maximum allowed budget to be spent and maximum allowed POIs of certain type or category to be visited.

Gavalas et al. [11] solve TOPTW by using two analogous approaches from cluster-based heuristics named as Cluster Search Cluster Ratio (CSCRatio) and Cluster Search Cluster Routes (CSCRoutes). The presented approaches aim at encouraging visits to dense areas that consist of a high density of "good" points (with high satisfaction factors). Both heuristics employ a clustering procedure to organize points into clusters based on distance from dense areas. This increases the probability that more visits would to take place inside individual clusters, and as result, two positive effects would be achieved, the decrease of duration of tours and minimization of number of transfers between different clusters. In terms of quality of solutions, this approach performs better than ILS [5], while the computation time remains in comparable level.

Garcia et al. [12] solve a variant of OPTW problem called Time Dependent Orienteering Problem with Time Windows (TDOPTW), which models the variability of traveling time between POIs. The TDOPTW is used to model different modes of traveling between any two POIs, such as walking, bicycle, public transportation or car. The ILS algorithm [5] is hybridized with Dijkstras Shortest Path algorithm to solve the TDOPTW problem. The Dijkstras algorithm is used to calculate the average travel times between all points available. Due to the constraint for a real time computation, the traveling times are calculated in an off-line-mode and saved for letter use by the ILS algorithm. The computation experiments performed with the data of city of San Sebastian in Spain, prove that the approach is able to generate valid routes in real time.

Souffriau et al. [13] extend MCTOPTW problem to the MCTOP with Multiple Time Windows (MCTOPMTW), where multiple working shifts of POIs can be modeled. The proposed approach for solving MCTOPMTW is a hybridization between ILS and Gready Randomized Adaptive Search Procedure (GRASP) algorithms. This approach is also based on the ILS approach of Vansteenwegen et al. [5] for TOPTW. The ILS-GRASP approach outperforms the approach of Garcia et al. [6].

In [14], we presented a tabu search method for solving the MCTOPTW problem. Oppositely to the approach of Garcia et al. [6] that tackles the same problem, we enforced an additional constraint to the problem that forbids any waiting time between consecutive visits. The method was shown to be competitive to the state of the art algorithms for solving the MCTOPTW problem.

Further, in [15], we extended MCTOPTW problem to MC Multiple TOPTW (MCMTOPTW) to model the trip itinerary for a group of tourists. Then, we solved the problem by devising an algorithm that is built on top of the approach presented in [14], which is used for construction of the starting solution. The proposed algorithm was again implemented under the framework of tabu search meta heuristic with a neighborhood structure that relays on three operators, namely *Separate* (a tourist from the group), *Join* (two tourists from two different subgroups) and *Insert* (a new POI into the trip of a tourist). The solution that is returned by the algorithm creates personalized itineraries for each tourist in the group, by allowing them to switch the company with different tourists during the trip, in accordance to their preferences.

III. Mathematical Modeling

We suppose that there are *N* POIs available in a city or region. Each POI *i* is characterized with a set of attributes, such as: coordinates (x_i, y_i), visit duration t_i, tourist satisfaction factor S_i, opening time o_i, closing time c_i and entry fee e_i. In addition, we suppose that traveling time t_{ij} between POI *i* and POI *j* is symmetric and it is known for all *m=1, ..., N* and *j=1, ..., N*, given that $i \neq j$. On the other hand, the tourists trip details, include: number of tours *M*, budget for the complete trip (all tours) B_{max} and duration of tour T_m, where *i=1, ..., M*. In general, individual tours can have different starting and ending times. Further, we assume that for each tour *m*, POI *1* and POI *N* are the starting and ending points, respectively. In practice, the starting and ending point could be the same, where in most cases that is a hotel (accommodation).

Next, based on [16], we formulate the envisioned trip planning problem as an integer programing problem, where we make use of two decision variables, namely x_{ijm} and y_{im} that are defined as in following. In addition, the starting time of a visit to POI *i* in tour *m* is denoted as s_{im}.

$$x_{ijm} = \begin{cases} 1, & \text{if visited point } i \text{ is followed with a visit to point } j \text{ in tour } m \\ 0, & \text{otherwise} \end{cases}$$

$$y_{im} = \begin{cases} 1, & \text{if point } i \text{ is visited in tour } m \\ 0, & \text{otherwise} \end{cases}$$

$$Max \sum_{m=1}^{M} \sum_{i=2}^{N-1} S_i y_{im}, \tag{1}$$

$$Min \sum_{m=1}^{M} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} t_{ij} x_{ijm}, \tag{2}$$

$$\sum_{m=1}^{M} \sum_{i=2}^{N-1} e_i y_{im} \leq B_{max}, \tag{3}$$

$$\sum_{m=1}^{M} \sum_{i=2}^{N-1} y_{im} \leq 1, \tag{4}$$

$$(y_{im} = 1) \implies (o_i \leq s_{im} \wedge s_{im} + t_i \leq c_i), \quad \forall i = 2, ..., N-1, \ \forall m = 1, ..., M \tag{5}$$

$$\sum_{i=1}^{N-1} \left(t_i y_{im} + \sum_{j=2}^{N} t_{ij} x_{ijm} \right) \leq T_m; \ \forall m = 1, ..., M \tag{6}$$

The first two constraints take part in a two component objective function for the envisioned problem, where Constraint (1) aims at maximizing the total collected score for tourist (by visiting POIs with the highest satisfaction factor), whereas Constraint (2) aims at reducing the total traveling time for the tourist throughout all tours. Constraint (3) limits the overall cost of all tours to the specified maximum budget B_{max}, whereas Constraint (4) makes sure that no POI is visited more then once. Constraint (5) ensures that the visits to individual POIs occur only during their respective opening hours. Finally, Constraint (6) limits the duration of individual tours to the respective maximal allowed durations.

The envisioned problem deviates from the original TOPTW problem [5] in two aspects: (1) it considers the entry fee of POIs by limiting the maximum allowed budget to a specified limit, and (2) it optimizes the solution not only based on the overall tourist satisfaction factor, but also based on the overall tourist traveling time.

IV. Solution Approach

The Simulated Annealing (SA) is a metaheuristic that is based on theories from thermodynamics where creation of the structure of crystals depends merely on the cooling function, which should neither be so quick or so slow, as defects in the crystal might appear. Analogical, in the field of optimization, SA uses a cooling function to determine whether the algorithm should also be able to select random solutions that might be worse then the current solution, or it should only accept better solutions, and thus act more as a hill climber. This feature enables SA algorithm to avoid premature convolution, thus it is able to escape from getting stuck in local optimum. The SA algorithm was initially presented by Kirkpatrick et al. in [17]. Recently, the SA algorithm was used by Lin & Yu [18] to successfully solve the related problem of TOPTW.

A. Solution Representation

The representation of a given current solution (denoted as S_c) is done by using an array of M members, where each member is a list, which itself contains the sequence of POIs that are part of a single tour (day). In addition, the POIs that are not included in any of the tours, are kept in separate list called "QueueList". Considering an example of a tourist visiting a city that has $N = 10$ POIs for a duration of $M = 2$ tours, an arbitrary candidate solution might be represented as in following:

$$S_c = \{[1, 8, 2, 5], [9, 7, 10]\}$$
$$QueueList = [3, 4, 6]$$

B. Neighborhood Exploration

There is a number of operators in the literature that could be used to solve the proposed problem. The *insert* operator [5] is used for insertion of new POIs in the itinerary, or *remove* operator to remove one or more POIs from the itinerary. In addition, the *swap/exchange* operator can be used to exchange POIs inside the itinerary with POIs outside of the itinerary, or also it can be applied to exchange POIs within itinerary itself (cross-exchange) [19].

In this implementation, the structure of the neighborhood exploration procedure consists of three operators, namely *insert*, *swap* and *shake*. The *swap* operator is only implemented to exchange POIs in the itinerary with POIs outside of the itinerary. Nevertheless, the cross exchange between the POIs within the itinerary occurs indirectly through *swap* operator, since when a POI goes outside of the itinerary at a given iteration, it has the chance of going back to the itinerary at a later iteration, and there exist probability that the same POI would go to some other position in the tour than it has been at the time when it was lastly removed.

C. Evaluation Function

The evaluation (fitness) of the solutions to the problem at hand are done by using a function that consists of two components, as given by Equation (7). The first component E_s represent the overall normalized satisfaction factor of the tourist with the POIs included into the itinerary, whereas the second component E_t represents the overall normalized satisfaction of the tourist with the overall traveling time during the tours. Both components are multiplied with their corresponding wight coefficients w_s and w_t, respectively, which are used to make a trade off between finding itineraries that include POIs with high satisfaction factors, or in finding itineraries with less travel time. In this implementation, the values of the weight coefficients are assumed to be set by the user, whereas in the future work, we aim to automatize them by taking a multi-objective approach, such as for instance the method of *Pareto front* of the space of candidate solutions.

In reference to Constraint (1) and Constraint (2), note that in order to have a maximizing fitness function, where both its components would have a maximization nature, the second component of Equation (7) needs to consider the complementary value of the normalized traveling time, since its original aim is to minimize the traveling time.

In order to make sure that individual components of the evaluation function do not predominate each other, their respective values need to be normalized based on their maximal possible values. In regard to the component of satisfaction factor with POIs, the normalization (Equation (9) is made by supposing a maximum number of POIs (denoted as τ) that can be visited during a tour (day), and by considering the satisfaction value of the POI with the highest satisfaction factor (denoted as ω). On the other hand, the normalization of the traveling time component is done by considering the total duration of all tours.

$$E(S) = w_s E_{s_{norm}} + w_t(1 - E_{t_{norm}}) \tag{7}$$

$$E_{s_{norm}} = \frac{\sum_{m=1}^{M} \sum_{i=2}^{N-1} S_i y_{im}}{M * \tau * \omega} \tag{8}$$

$$E_{t_{norm}} = \frac{\sum_{m=1}^{M} \sum_{i=1}^{N-1} \left(t_i y_{im} + \sum_{j=2}^{N} t_{ij} x_{ijm} \right)}{\sum_{m=1}^{M} T_m} \tag{9}$$

TABLE I. ALGORITHM PARAMETERS

Symbol	Description
α	Specifies the proportion of temperature decrease
β	Maximal temperature used to commence the search process
γ	Minimal temperature used to end the search process
δ	Maximal number of iterations without improvement
ϵ	Number of POIs to remove during the Shake step
ζ	Specifies one of the modes for generation of the initial solution

D. Simulated Annealing Implementation

The proposed SA algorithm can be fine tuned by using six different parameters. Table I gives more details about the utilized parameters, where the first column presents the respective symbols, whereas the second column explains the purpose for using the envisioned parameters.

As shown in the pseudo code (Figure 1), the parameters can be used to fine tune different mechanism within the algorithm. The ζ parameter sets one of five possible modes for generation of the initial solution, as described in following:

- *Mode 1* - inserts POIs into the itinerary in a random order
- *Mode 2* - initially orders POIs based on their satisfaction factor and then inserts them into the itinerary in a descending order, where POIs with highest satisfaction are inserted first
- *Mode 3* - orders POIs as *Mode 2*, except that it inserts them into the itinerary in a ascending order, where POIs with lowest satisfaction are considered first
- *Mode 4* - calculates the average distance of each POI with all other POIs, and then inserts them into the itinerary by giving priority to the POIs that have shorter average distance to the others
- *Mode 5* - inserts POIs into the itinerary by giving priority to POIs that have a shorter visit duration

The evaluation function described earlier (Subsection IV-C) is used to evaluate the candidate solutions during the search process, such as the initial solution at the start, and the neighbor solutions later. The iterative part of the algorithm depends on its main loop, which is controlled by the initial temperature, which, at the start, is set to its maximal value (parameter β), the temperature decreasing ratio (parameter α), the minimal temperature (parameter γ), and the cooling function. In this implementation, the cooling function follows a decreasing exponential law, which depends on current iteration t and parameter α. At a given iteration t, the temperature T_t is calculated by using Equation 10, where T_{t-1} represent the temperature in previous iteration.

$$T_t = T_{t-1} e^{-\alpha * t} \tag{10}$$

In a given iteration t, the neighborhood of a current solution consists of all neighbors (candidate solutions) that yield as results of trying to swap every POI p that is off the itinerary with every POI q that is currently in the itinerary. There are two mechanisms that are applied priory to evaluating the fitness of the neighbors. The first is elimination of non feasible

```
1:  procedure GENERATE ITINERARY(α, β, γ, δ, ϵ, ζ)
2:      S_c ← GenerateInitialSolution(ζ)
3:      Evaluate S_c
4:      S_b ← S_c
5:      T ← β
6:      t ← 0
7:      i ← 0
8:      while T > γ do
9:          for each POI p in QueueList do
10:             for each POI q in S_c do
11:                 S_n ← ApplySwap(p, q)
12:                 if S_n is a feasible solution then
13:                     Fill empty space in S_n
14:                     Evaluate S_n
15:                     if E(S_n) > E(S_c) then
16:                         S_c ← S_n
17:                         if E(S_c) > E(S_b) then
18:                             S_b ← S_c
19:                             i ← 0
20:                         else
21:                             i ← +1
22:                         end if
23:                     else
24:                         a ← GenerateRandomNumber(0, 1)
25:                         b ← (T−γ)/(β−γ) e^((E(S_n)−E(S_c))/T)
26:                         if a < b then
27:                             S_c ← S_n
28:                         end if
29:                     end if
30:                 end if
31:                 if i > δ then
32:                     S_c ← ApplyShake(ϵ)
33:                     Reorder POIs in S_c
34:                     Fill empty space in S_c
35:                 end if
36:             end for
37:         end for
38:         t ← +1
39:         T ← ApplyCoolingFunction(t, α)
40:     end while
41:     return S_b
42: end procedure
```

Fig. 1. The SA algorithm for generation of trip itinerary

neighbors, which do no meet the hard constraints, such as maximum budget, at most one visit to a POI and limited duration of the tours, whereas the second mechanism is about filling empty spaces in the itinerary that might have come out due to application of the envisioned operators. Depending on the size of the empty space, additional non included POIs are inserted into the itinerary.

This implementation follows the general guidelines of SA algorithm in reference to adaption of a neighbor solution as a next current solution. Therefore, a neighbor is always accepted as a next current solution when it is better then the actual current solution. Nevertheless, if a neighbor is worse then the actual current solution, it can still be accepted, with a certain probability, as next current solution. The probability of acceptance is higher if the corresponding quality of the

neighbor is not that worse then the quality of actual current solution, and if the running temperature is higher (this is usually the case in the early iterations of the SA algorithm when it undergoes its exploitative phase). In the pseudocode, variable a gets a random number between 0 and 1, whereas variable b also gets a value between 0 and 1 through the expression in Line 25. If the value of a happens to be smaller then the value of b, then the neighbor solution would be accepted as the next current solution.

If the algorithm undergoes a consecutive number of iterations without improvement (as defined by parameter δ) then the *shake* operator will be applied. Considering the ϵ parameter, the *shake* operator removes a number of POIs from the itinerary. Afterwards, at first, the POIs are reordered by being shifted as much as possible towards the beginning of the tours, so that the empty spaces are gathered at the end of the tours, then the mechanism for filling up the empty spaces (described above) is put into function, and lastly the modified solution is adapted as next current solution. The application of *shake* operator might be important as it might help the SA algorithm to escape from getting caught into the local optimum.

V. Computational Experiments

The algorithm is developed by using Java 1.7. The experiments are conducted by using a machine with an Intel Core Duo 2.5 GHz processor with a Windows 7 operating system and 3GB of RAM memory. In the following subsection, we describe the test set, and then in Subsection V-B, we present the experimental results.

A. Test Set

The test set is made of fifteen instances, where eleven of them consist of two tours (days), while the remaining four have 1, 3, 4 and 5 tours, respectively. In addition, the tourist interest about types (e.g. monument, park, castle, etc.) and categories (e.g. archeology, architecture, religious art, etc.) of POIs is modeled by using the mechanism of random number generation of Java programming language. Further, all fifteen instances of the test set are executed against fifty real POIs in city of Prishtina in Kosova. Each POI, besides its name and description, is also characterized with its estimated tourist satisfaction/interest factor, coordinates, opening and closing time, visit duration and entry fee. In addition, each of the fifteen instances is associated with an additional point (e.g. hotel, bus station, city down town, etc.), which serves, both as a starting and ending point for the tours. The distance between individual POIs is expressed in the unit of minutes.

The first ten instances, except for the tourist satisfaction factor, are characterized with same attribute values, as it follows: 2 tours, each tour starts at 08:00 hrs and ends at 17:00 hrs, break time 30 minutes, wight for satisfaction factor 60% and weight for travel time 40%. The other five instances have diverse attribute values.

B. Experimental results

The values of the algorithm parameters (except ζ) have been fine tuned through some preliminary experiments that have been conducted by using the complete test set. The fine tuned values are: $\alpha = 0.005, \beta = 4000, \gamma = 80, \delta = 210, \epsilon = 6$.

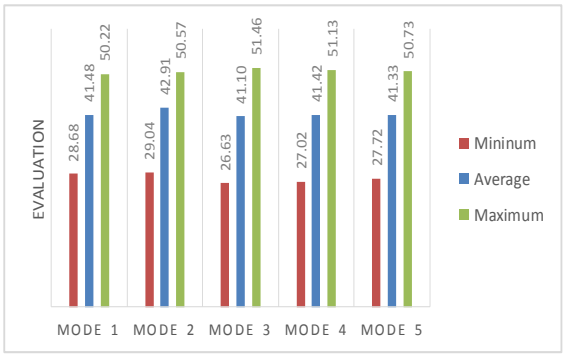


Fig. 2. Algorithm performance for different modes of initial solution

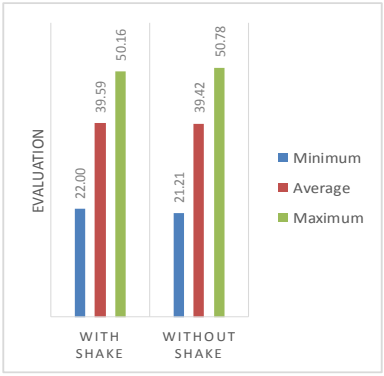


Fig. 3. Algorithm performance when *shake* operator is applied

In the following, we present the results about four types of experiments, such as: the performance of different modes of initial solution, the effect of *shake* operator into the search process, the variations between different executions of a given instance, and the time needed to execute instances with different number of tours. Each experiment includes ten executions for each instance of the test set. Therefore, the results presented below are averaged over the ten executions of individual instances.

Figure 2 shows that, in average, *Mode 2* of initial solution, where POIs with higher satisfaction factor are more common, performs better than the other considered modes. This result can be explained based on the indication that *Mode 2* of initial solution is more diverse, since most of its POIs have high satisfaction factor, hence they get more prone in being subject of exploration operators, and therefore make the algorithm more exploratory.

The diagram in Figure 3 depicts the averaged results over the complete test set, where the performance of the algorithm is presented in variance to the application of *shake* operator. By considering the average results, it can be noted down that the *shake* operator helps the algorithm to achieve only a slight improvement in the overall quality of solutions.

Figure 4 presents the variance in time over ten different executions of *Instance 12*, which consists of two tours. In terms of quality of solutions, for this particular instance, it is evident

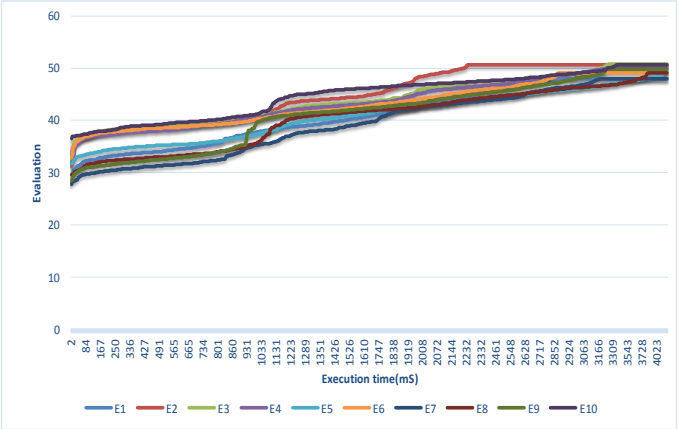


Fig. 4. Variations between different executions for instance *Instance 12*

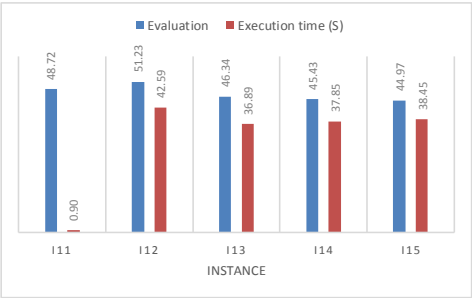


Fig. 5. Computation time for different number of tours

that the difference between different executions is at most a little less than 3 points, whereas the computation time, in the worst case scenario is a little less than 4 seconds. In addition, it can be noted down that, although, in some executions, the starting solutions are much worse, at the end of executions the algorithm manages to produce solutions with comparable quality to the cases where initial solution is better.

Figure 5 shows the performance of the algorithm when executed for instances 11 until 15, which consist of 1 to 5 tours, respectively. In terms of evaluation, the algorithm produces comparable results for all five considered instances, whereas the computation time for *Instance 11* (with one tour) is a little less than 1 second, and for other instances it is much higher.

The particular trip planning problem tackled in this paper differs from the related problems in the literature, therefore the results of the proposed algorithm can not be experimentally compared with the results of the state of the art algorithms.

VI. Conclusions

In this paper, we presented a new approach based on simulated annealing for planning the tourist trip itinerary. The computation time for itineraries with one tour is quite short, whereas for itineraries with two tours and more the average computation time is about 39 seconds. Although, the algorithm works slightly better by using *Mode 2* of initial solution (where POIs with higher satisfaction are given priority for insertion), we can conclude that the difference by using other modes is quite small. Despite our expectations, the application of *shake* operator does not prove to be very successful in improving the quality of solutions. Therefor, it remains an open question to investigate application of other operators that might be helpful in further improvement of the results.

References

[1] W. Souffriau, P. Vansteenwegen, J. Vertommen, G. V. Berghe, and D. V. Oudheusden, "A personalized tourist trip design algorithm for mobile tourist guides," *Applied Artificial Intelligence*, vol. 22, no. 10, pp. 964–985, 2008.

[2] T. Tsiligirides, "Heuristic methods applied to orienteering," *Journal of the Operational Research Society*, pp. 797–809, 1984.

[3] I. Chao, B. L. Golden, E. A. Wasil *et al.*, "The team orienteering problem," *European journal of operational research*, vol. 88, no. 3, pp. 464–474, 1996.

[4] M. G. Kantor and M. B. Rosenwein, "The orienteering problem with time windows," *Journal of the Operational Research Society*, pp. 629–635, 1992.

[5] P. Vansteenwegen, W. Souffriau, G. Vanden Berghe, and D. Van Oudheusden, "Iterated local search for the team orienteering problem with time windows," *Computers & Operations Research*, vol. 36, no. 12, pp. 3281–3290, 2009.

[6] A. Garcia, P. Vansteenwegen, W. Souffriau, O. Arbelaitz, and M. Linaza, "Solving multi constrained team orienteering problems to generate tourist routes," *status: published*, 2009.

[7] A. Divsalar, P. Vansteenwegen, and D. Cattrysse, "A variable neighborhood search method for the orienteering problem with hotel selection," *International Journal of Production Economics*, 2013.

[8] B. L. Golden, L. Levy, and R. Vohra, "The orienteering problem," *Naval Research Logistics (NRL)*, vol. 34, no. 3, pp. 307–318, 1987.

[9] P. Vansteenwegen, W. Souffriau, and D. V. Oudheusden, "The orienteering problem: A survey," *European Journal of Operational Research*, vol. 209, no. 1, pp. 1–10, 2011.

[10] W. Souffriau, J. Maervoet, P. Vansteenwegen, G. V. Berghe, and D. Van Oudheusden, "A mobile tourist decision support system for small footprint devices," in *Bio-Inspired Systems: Computational and Ambient Intelligence*. Springer, 2009, pp. 1248–1255.

[11] D. Gavalas, C. Konstantopoulos, K. Mastakas, G. Pantziou, and Y. Tasoulas, "Cluster-based heuristics for the team orienteering problem with time windows," in *Experimental Algorithms*. Springer, 2013, pp. 390–401.

[12] A. Garcia, O. Arbelaitz, P. Vansteenwegen, W. Souffriau, and M. T. Linaza, "Hybrid approach for the public transportation time dependent orienteering problem with time windows," in *Hybrid Artificial Intelligence Systems*. Springer, 2010, pp. 151–158.

[13] W. Souffriau, P. Vansteenwegen, G. Vanden Berghe, and D. Van Oudheusden, "The multi-constraint team orienteering problem with multiple time windows," *Transportation Science, Articles in Advance*, pp. 1–11, 2011.

[14] K. Sylejmani, J. Dorn, and N. Musliu, "A tabu search approach for multi constrained team orienteering problem and its application in touristic trip planning," in *Hybrid Intelligent Systems (HIS), 2012 12th International Conference on*. IEEE, 2012, pp. 300–305.

[15] K. Sylejmani, "Optimizing trip itinerary for tourist groups," Ph.D. dissertation, Vienna University of Technology, Faculty of Informatics, Austria, 2013.

[16] K. Sylejmani and A. Dika, "Solving touristic trip planning problem by using taboo search approach," *International Journal of Computer Science Issues(IJCSI)*, vol. 8, no. 5, 2011.

[17] S. Kirkpatrick, D. G. Jr., and M. P. Vecchi, "Optimization by simmulated annealing," *science*, vol. 220, no. 4598, pp. 671–680, 1983.

[18] S. W. Lin and V. F. Yu, "A simulated annealing heuristic for the team orienteering problem with time windows," *European Journal of Operational Research*, vol. 217, no. 1, pp. 94–107, 2012.

[19] F. Tricoire, M. Romauch, K. F. Doerner, and R. F. Hartl, "Heuristics for the multi-period orienteering problem with multiple time windows," *Computers & Operations Research*, vol. 37, no. 2, pp. 351–367, 2010.

Generating Spatio-Temporal Streaming Trajectories

E. Mešković*, D. Osmanović *, Z. Galić** and M. Baranović**
* University of Tuzla, Faculty of Electrical Engineering, Franjevačka 2,
75000 Tuzla, Bosnia and Herzegovina
emir.meskovic@untz.ba, dario.osmanovic90@gmail.com
** University of Zagreb, Faculty of Electrical Engineering and Computing
Unska 3, 10000 Zagreb, Croatia
{zdravko.galic, mirta.baranovic}@fer.hr

Abstract - Recent research efforts in data stream management systems (DSMS) focus mainly on processing continuous queries over traditional data streams, and only a few addressed spatio-temporal continuous queries. OCEANUS presents an effort to extend TelegraphCQ DSMS with spatial support providing a platform for spatio-temporal streaming applications. Data type system that represents the formal basis for modeling moving objects in data streams, as well as an approach for managing moving objects in spatio-temporal data streams based on user-defined aggregate functions (UDAF) are presented. In this paper we are concerned with improving proposed approach for extracting moving objects out of spatio-temporal data streams based on UDAF. It is based on two methods for detecting static locations on trajectories of moving objects and for reducing total number of units in sliced representation by introducing trajectory buffer. First method leads to more precise results of certain operations over moving objects such as retrieving information about speed and direction, but also enables introduction of few new operations such as retrieving static locations and overall stoppage time on moving object's trajectory. Second method leads to better memory usage during processing spatio-temporal continuous queries in DSMS.

I. INTRODUCTION

Moving objects databases actually represent spatio-temporal databases that manage spatial objects with continuously changing location and/or geometry and they represent relatively intense research area in recent years. These databases are overloaded with frequent updates of moving objects positions, and new approach of considering trajectory of a moving object as a continuous data stream which carries information about object's position at any instant has been proposed. To manage this data, several DSMSs have been developed, that support the management of data streams and standard relations as well. Mainly based on relational paradigm, they have very rudimentary support for spatio-temporal data types and operations that would enabled efficient management of moving objects.

The majority of published research has focused on moving objects databases or data streams separately, while only a few attempts to combine these two areas. The goal of our previous research [1][2] was to examine the possibility of combining data models and associated algebras developed for traditional spatio-temporal database management systems (DBMS) with spatio-temporal data streams in order to efficiently manage moving objects. We have proposed an approach for spatio-temporal data stream management based on extending relational-based DSMS with a full-fledged set of spatio-temporal data types and associated operations. Formal specification of operations for moving objects in the context of data streams is presented by using a conceptual framework of many sorted algebra consisting of data types and operations needed to support spatio-temporal data in data streams. Implementation of abstract data model at discrete level, including spatio-temporal predicates and operations, is based on the concept of extracting moving objects out of data streams using UDAFs.

Extracted moving object is formed of unit types each consisting of a time interval and a value describing simple function on the unit time interval as it is presented in [1][2]. The unit type that describes movement of a corresponding moving object as a linear function of time is created by using two successive spatio-temporal data stream elements with spatio-temporal references consisting of successive locations and time instants. Complete sliced representation presents a moving object as a set of temporal units whose time intervals are disjoint and, if adjacent, their values are different. Assembling unit types into a moving object based on the spatio-temporal references of data stream elements is achieved by using UDAFs.

In this paper we are concerned with improving proposed method for creating temporal units in sliced representation of a moving object. It consists of two extensions that would enable more precise results of certain operations over moving objects (such as *speed* and *mdirection*) and an optimization of memory usage during processing spatio-temporal continuous queries. First extension implies detecting locations on moving object's trajectory when it actually doesn't move (such as stations, semaphores, etc.). Second extension introduces trajectory buffer with required accuracy which at the end leads to reducing number of temporal units in sliced representation of a moving object.

II. RELATED WORK

The upgrade of DBMSs with the ability to manage moving objects, but also data streams and their DSMSs, are of great interest for scientific-research community in last two decades. However, researches done within these two areas were usually conducted separately, especially in their initial phase, and only recently they are trying to combine.

In [3] authors consider moving objects databases from the location based perspective and introduce model of trajectory of a moving object as a polyline in three-dimensional space, where two dimensions refer to space and third dimension to time. This model takes into account the temporal and regional uncertainty of moving objects by associating uncertainty threshold with each line segment of the trajectory. They also consider querying moving objects with uncertainty given by their trajectories and introduce the set of most essential operators for doing this.

Model of a trajectory presented in [4] captures the inherent parameter of uncertainty of the moving objects location, which impacts both the semantics of spatio-temporal queries and the algorithms for their processing. The operators which capture the spatial, temporal and uncertainty aspects of a moving object are implemented as UDAFs on top of existing object-relational DBMS and can be used for answering queries and generating notification triggers. Proposed model is implemented in DOMINO project.

Inherent imprecision of the information about moving objects location at a given time in moving objects databases is also addressed in [5]. This work deals with the issues of constructing, representing and querying the trajectories of moving objects. Uncertain trajectory is modelled as 3D cylindrical body and set of novel but natural spatio-temporal operators which capture the uncertainty is presented.

The geometry of the uncertain trajectories of the objects with movement constrained to networks and the interaction between such objects in terms of operations is presented in [6]. It is based on the framework in [7] which is extended with data types and operations supporting uncertainty.

Work presented in [8] introduces a flexible trajectory modelling approach that takes into account model-inferred actual positions, time-varying uncertainty, and nondeterministic uncertainty ranges. It also presents an efficient mechanism to evaluate probabilistic range queries on those evolving-density trajectories, as well as estimators that effectively infer evolving densities of trajectory data.

Method for trajectory segmentation and sampling based on the representativeness of the (sub)trajectories in the moving objects databases is presented in [9]. Most representative subtrajectories are find based on local density and trajectory similarity information of each segment of the trajectory.

Work presented in [10] introduces term trajectory stream for point objects and several new types of window operators that would enable more efficient management of those trajectories. As a result, TSV (Trajectory Stream Visualizer) based on TelegraphCQ and Java that enables posing and visualization of queries on moving objects trajectories is created. However, it has significant limitations in terms of supporting spatio-temporal data, since it only uses built in spatial data types and operations inherited from TelegraphCQ.

Our approach for generating spatio-temporal streaming trajectories is similar to one presented in [3], [4] and [5] in sense of inherited uncertainty but is based on actual location of moving object contained in data stream element rather than expected one calculated using motion vector. We should also keep in mind that our approach extracts moving object out of data stream using UDAF in terms of spatio-temporal data type, and that *trajectory* is one of supported operations over this data type that yields trajectory made by moving point as a line in 2D space. For this reason, our model of trajectory for point objects is closed to one presented in [10] but overcomes limitations of supporting spatio-temporal data with supporting data type system presented in [1].

III. MODELING MOVING OBJECTS IN DATA STREAM SYSTEMS

In next section we will introduce extensions of an abstract algorithm for iterative function used in UDAF for extracting moving objects out of data stream. These extensions use spatio-temporal data types for representing moving objects presented in our approach for modelling moving objects in data stream systems [1]. It is based on a system of data types on abstract level supporting historical and current queries, defined using many-sorted algebra and second-order signature, and build upon work presented in [11][12] and [13].

Modelling on abstract level allows us to make definitions in terms of infinite sets, without fixing any finite representations of these sets. Although the abstract level is the conceptual model we are interested in, it should be noted that it is not directly implementable and that additional step of fixing a concrete, finite representation is needed. To ensure support for spatio-temporal data in DSMS, in [1] we have defined abstract data types by extending abstract data type system introduced in [12], [13] and [3].

New sorts representing attribute in a relation or data stream that represent a basis for building spatio-temporal DBMS in extensible DBMS environment, such as SECONDO [14], with the goal to support modelling and querying history of movement, are presented. We have also approached modelling data streams by adding new sorts and type constructors with two significant differences compared to the corresponding sorts and type constructors for modelling relations:

- Relation tuple may contain data of temporal types, while for data stream tuple that is not the case. We consider that data stream tuple may contain data about current location or extent of moving object, but not the moving object entirely.

- Each data stream tuple must implicitly contain timestamp that represents time when it is generated at data source.

Based on two approaches for modelling moving objects, the one presented in [1] and the one presented in [3], it is obvious that in the context of spatio-temporal data streams data types obtained by using type constructors *moving* and *streaming* try to model similar phenomena from the real world, and that is unit of data that changes over time. However, there are important differences:

- Data types obtained with type constructor *moving* do not change without explicit intervention of user, while data types obtained with type constructor *streaming* change under the influence of external data source.
- Implementation of data types obtained with type constructor *moving* can be done in different ways, and sliced representation with different ways of interpolation between slices. On the other hand, data types obtained with type constructor *streaming* have discrete sliced representation, in which one value holds until next one arrives.

Since both sets of data types are related to the same moving objects from a real world, the most of operations that are defined for data types obtained with type constructor *moving* can also be applied to data types obtained with type constructor *streaming*. This is also introduced in our previous work [1].

IV. IMPROVING UDAF FOR EXTRACTING MOVING OBJECTS OUT OF DATA STREAMS

Spatio-temporal data stream encapsulate information about successive locations of moving objects that represent their sources, i.e. information on how these objects move around in space. A main task in extracting this information is that every single moving object that sends data into spatio-temporal data streams is presented in a form of sliced representation based on the spatio-temporal references of data stream elements (*intime*(*point*) type). If we consider spatio-temporal references of two successive data stream elements generated at same data source, it is easy to calculate a linear function that describes movement of a corresponding source between two points presented with spatial stamps during time interval presented with timestamps of data stream elements. If $p_1(x_{p1}, y_{p1})$ denotes spatial stamp of first data stream element and t_1 its timestamp, and $p_2(x_{p2}, y_{p2})$ denotes spatial stamp of second data stream element and t_2 its timestamp, linear movement of a point moving object along x and y axes can be presented with following equations:

$$\begin{aligned} x_0 &= (-t_1)(x_{p2} - x_{p1})/(t_2 - t_1) + x_{p1} . \\ x_1 &= (x_{p2} - x_{p1})/(t_2 - t_1) . \\ y_0 &= (-t_1)(y_{p2} - y_{p1})/(t_2 - t_1) + y_{p1} . \\ y_1 &= (y_{p2} - y_{p1})/(t_2 - t_1) . \end{aligned} \quad (1)$$

The coefficients x_0, x_1, y_0 and y_1 represent the value *MPoint* which describes the movement of a point 2D object as a linear function of time, i.e. its meaning is [3]:

$$f((x_0, x_1, y_0, y_1), t) = (x_0 + x_1 t, y_0 + y_1 t) . \quad (2)$$

To define the unit of sliced representation (*unit*(*point*) type), it is necessary to assign a time interval with a function presented with *MPoint* during which it holds. This time interval is defined with timestamps t_1 and t_2 of two data stream elements used for unit creation. To create moving object in a form of sliced representation (*moving*(*point*) type) it is necessary to assemble units obtained with procedure described above. And as we have already mentioned previously, assembling unit types into a moving object based on the spatio-temporal references of data stream elements can be achieved using UDAFs, as we presented in our previous research [1][2]. This method of adding new unit to the moving object can be represented by an abstract algorithm we introduced in [15]. In this algorithm we use *mpoint, upoint,* and *ipoint* to denote data types *moving*(*point*), *unit*(*point*) and *intime*(*point*), respectively.

A function that is used in iterative operation in UDAF implements presented abstract algorithm and is called once for each value aggregated, i.e. for each spatio-temporal reference of new data stream element generated at the same source.

A. Detecting static locations on moving object's trajectory

Static locations on a trajectory of moving object are the places where moving object stops for one short period of time. For vehicles moving on the road network these locations can be traffic lights or traffic congestions, for the buses are bus stations, and for taxis are places for entry and exit of passengers, etc. During these short stops, moving objects that represent data sources continue to send information into data stream, even though they represent the same spatial data. This means that a number of successive data stream elements have identical spatial references.

If we consider creating a new unit of sliced representation when such spatial stamp is arriving, this would mean that the linear function that describes the movement of a moving object during last unit will be the same, only the time interval of that unit during which the function is valid will be changed. However, this approach of defining units of a moving object would lead us to have incorrect results when applying specific operations on a moving object. For example, the *speed* operation would result with a lower value than the actual speed of a moving object during one unit, and the *mdirection* operation would result with a certain direction even though the moving object is not moving.

In order to avoid the occurrence of such incorrect results, static locations of a moving object need to be represented as separate units in a sliced representation. Creation of such unit begins when the spatial reference of a new arriving data stream element coincides with the moving object's location on the right end of the last unit's time interval. When data stream element whose spatial reference is different from static location presented in the last unit of sliced representation arrives, creation of that

unit ends and creation of a new unit of sliced representation begins.

These changes should be considered when creating linear function for new unit of a moving object presented with equations (2). Abstract algorithm for function performing creation of linear function is given in Fig. 1. Function call should be embedded into abstract algorithm for iterative function used in UDAF for extracting moving objects out of data stream in a fallowing way:

- If function call returns *null*, there is no need for creating new unit of sliced representation, since right end of time interval in last unit is set to time reference of new arriving data stream element (line 6 in abstract algorithm).
- Otherwise, new unit of sliced representation is created with linear function set to returned value of function call.

In [1] we have presented a subset of operations defined specifically for moving data types that are appropriate for streaming data types. With presented approach and changes made in UDAF for extracting moving objects out of data stream it is possible to extend this set of operations over moving objects. Here we introduce two new functions with signatures:

locations: *streaming*(*point*) → *points*

stoptime: *streaming*(*point*) → *int*

Function *locations* applied over type *streaming*(*point*) returns a set of points in which each element presents one static location on a trajectory of a moving object. This function simply scans sliced representation of a moving object and for each slice with $x_1 = 0$ and $y_1 = 0$ adds new point into resulting point set. Function *stoptime* returns a overall stoppage time of a moving object on its trajectory expressed in seconds. It simply summarizes time intervals of all slices (units) with $x_1 = 0$ and $y_1 = 0$.

B. Trajectory buffer

Total number of units in a sliced representation of a moving object can be reduced depending on the accuracy requested for representing moving objects. As we can see from a method for crating moving objects in form of sliced representation and method for detecting static locations of a moving object presented previously, the new unit of sliced representation of a moving object is not created with the arrival of each new data stream element. If the spatial reference of data stream element is lying in the same direction as the direction of trajectory segment obtained with projection of the last unit on the x-y plane, we just need to change the right end of the last unit's time interval. The same approach applies if data stream elements with the same spatial references are arriving.

If data stream elements arrive with high frequency, it often happens in practice that the spatial reference of new data stream element deviates very little from the direction of trajectory segment previously mentioned. Depending on the permitted deviation from the actual trajectory of a moving object, and on the required accuracy for presenting moving objects, we can introduce a *threshold* which represents the distance from the trajectory and which when exceeded implies a creation of a new unit of sliced representation. As long as the deviation is less than this threshold, already introduced approach of changing the right end of last unit's time interval applies.

In this way, we introduce the concept of *trajectory buffer*, which presents a wrapper around the trajectory of a moving object inside which moving object can be found in practice. Radius of a trajectory buffer around each trajectory segment has value of a requested accuracy threshold. This approach is illustrated on Fig. 2.

Since the most of aggregate functions in DSMSs does not support two parameters, the accuracy threshold should be defined in another way. As it now presents a feature of created moving object, the best way is to store that value into moving object itself. In this way, when using operations on moving objects you can also get information about result deviation from the actual value. Proposed method for creation a unit of sliced representation can also be embedded into abstract algorithm for iterative function used in UDAF, in a way to call function that implements abstract algorithm given in Fig. 3. If this function returns *true*, right end of time interval in last unit is set to time reference of new arriving data stream element. Otherwise, new unit of sliced representation is created.

We should keep in mind that the actual movement of a moving object is already approximated by a linear function between two spatio-temporal references of data stream, and that this approach represents an additional approximation of the movement. This concept is similar to the concept of uncertainty of moving object's trajectory presented in various works. Since it is related directly to

```
algorithm newMpoint(mp,ip)
input: moving object of type mpoint, new arriving spatio-temporal reference of type ipoint
output: linear function Mp of type MPoint for new unit of moving object
1: Mp = ∅
2: n := no_units(mp)
3: ip1 := atinstant(mp, mp.un.p.end)
4: if val(ip) = val(ip1) then
5:         if mp.upn.Mp.x1 = 0 and mp.upn.Mp.y1 = 0 then
6:                 mp.un.p.end := inst(ip)
7:         else
8:                 Mp := val(ip)
9:                 p.s := p.e := inst(ip)
10:                p.lc := true, p.rc := true
11:        end if
12: else
13:        compute linear function Mp between spatio-temporal references ip1 and ip by
           using equations (1)
14: end if
15: return Mp
end
```

Figure 1. Abstract algorithm for function performing creation of linear function for new unit of moving object

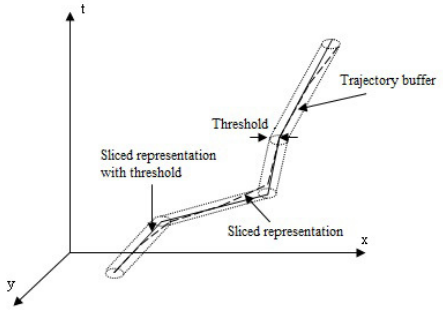


Figure 2. Trajectory buffer around sliced representation of a moving object formed with permitted accuracy threshold

algorithm *less_than_th*(*mp*,*ip*)
input: moving object of type *mpoint*, new arriving spatio-temporal reference of type *ipoint*
output: *bool* representing whether new spatial reference is within *th* distance from trajectory segment of moving object's last unit
method
1: $n := no_units(mp)$
2: $un := mp.u_n$
3: $un.p.end := inst(ip)$
4: $l := projection(un)$
5: $d := distance(l, val(ip))$
6: **if** $up.Mp = mp.u_n.Mp$ **or** $d < mp.th$ **then**
7: **return** *true*
8: **else**
9: **return** *false*
10: **end if**
end

Figure 3. Abstract algorithm for testing whether new spatial stamp is lying within permitted accuracy threshold

moving objects presented in a form of sliced representation it implies not only to operation *trajectory*, but also to other operations on moving objects. Defining new set of operations on moving objects that take into account this uncertainty is beyond scope of this work and one should use standard set of operations on moving objects but also be aware of this inherited uncertainty.

For this reason, it is necessary to determine very carefully the applications in which this approximation is allowed and appropriate, and also when it certainly should not be used. For example, it would not be appropriate for the movement of vehicles on the road network, but perhaps with certain accuracy it would be appropriate for the movement of ships at sea.

C. *Query examples and experimental results*

As we have already introduced in previous section, proposed method for detecting static locations enables introduction of two new functions: *locations* and *stoptime*. These two functions has been embedded into CSQL, an SQL-like query language that supports continuous queries over spatio-temporal data streams. Beside set of queries presented in our previous works [1][2] and [15] now we can pose additional queries that illustrate proposed approach and usage of aforementioned functions.

Q1: *For each taxi find all places on its trajectory when it was not moving during last hour.*

```
SELECT id, locations(moving(geometry))
FROM carStream <PAST '1 hour'>
WHERE carType = 'taxi'
GROUP BY id
```

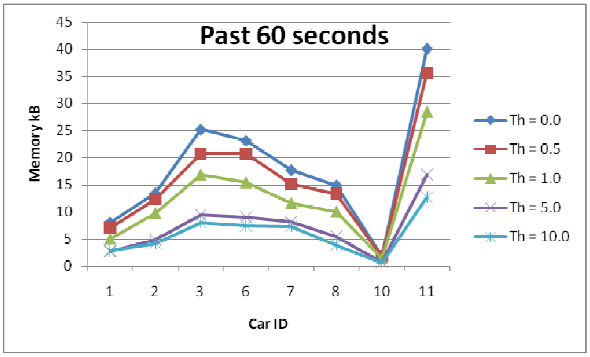


Figure 6. Memory usage in continuous query with past window of 60 seconds and different values of accuracy threshold

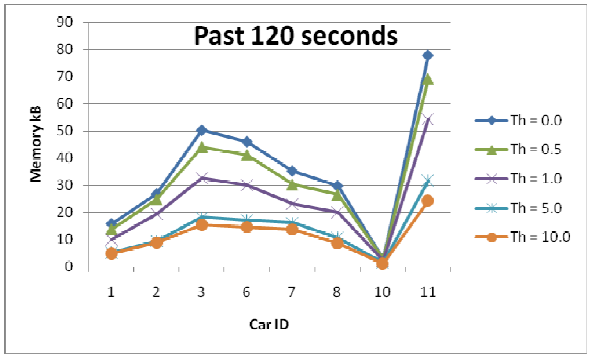


Figure 5. Memory usage in continuous query with past window of 120 seconds and different values of accuracy threshold

Q2: *Find each police car and its driver that didn't move for more than 10 minutes during last 90 minutes.*

```
SELECT id, driver
FROM carStream <PAST '90 minutes'>
WHERE carType = 'police'
GROUP BY id, driver
HAVING stoptime(moving(geometry)) > 600
```

Q3: *Find each bus and its driver that was staying at a particular location sometimes during last half an hour.*

```
SELECT id, driver
FROM carStream <PAST '30 minutes>
WHERE carType = 'bus'
GROUP BY id, driver
HAVING
ST_Contains(locations(moving(geometry)),
ST_GeomFromText('POINT(557234      322701)',
25832))
```

None of these queries were possible with existing set of functions supported in OCEANUS prototype. For example, it was possible to determine did some vehicle passed some location of interests (using function *passes*), but it wouldn't give as an answer whether it was staying at this location for one short period of time. These query examples illustrate improvements gained with proposed method for detecting static locations on moving object's trajectory.

Generating moving objects with trajectory buffer approach enables better memory usage during processing spatio-temporal continuous queries which can be illustrated through few simple experiments shown on Fig. 4 and Fig. 5. Both figures illustrate memory usage in continuous query with two different size of window *past* and accuracy threshold in range of 0.5 to 10 meters. Fragment of INFATI data set [16] consisting of around 4500 tuples was used in both experiments.

As one can see from both figures, memory saving is significant as the threshold value is higher, even for this relatively small set of data stream elements. For example, for car with id=11 it ranges from 11% with threshold value of 0.5 meter to 68% with threshold value of 10 meters. This saving also increases with window size since number of units reduced in each sliding window is higher when window size is larger. This is also illustrated with Fig. 5 representing results of window size two time bigger than the one on Fig. 4. With these experiments, we have

shown that careful and appropriate choice of accuracy threshold in combination with windows size can lead to significant memory saving during processing spatio-temporal continuous queries on moving objects.

V. CONCLUSIONS AND FUTURE WORK

In our previous work we have introduced a formal framework for managing moving objects in data streams, consisting of data types and operations needed to support spatio-temporal data in data streams. Implementation approach is based on UDAF for extracting moving objects out of spatio-temporal data streams and in this work we propose two methods for its extension. With first method we are able to detect all static locations on moving object's trajectory and also to get more precise results when using certain functions on moving objects such as *speed* and *mdirection*. Detecting static locations also enables introduction of new functions, such as *locations* and *stoptime*, that enable posing quite a new range of continuous queries some of which were illustrated in last section. Second method enables better memory usage in a case when data arrives in system with high frequency and accuracy of moving object's trajectory is not a priority. We introduce two new terms, trajectory buffer and accuracy threshold, which leads into concept of uncertainty when applying functions such as *trajectory* on moving objects.

Focus of our future work will be on defining new set of operations on moving objects that would consider uncertainty inherited with proposed approach. This operations should produce results with information about probability that moving object has been in the given range. They should take into account the temporal uncertainty and the regional uncertainty of moving objects as it has been proposed in [3].

REFERENCES

[1] Z. Galić, E. Mešković, K. Križanović, M. Baranović, "OCEANUS: A Spatio-Temporal Data Stream System Prototype," in Proc. IWGS 2012 3rd ACM SIGSPATIAL Int. Workshop on GeoStreaming, Redondo Beach, CA, USA, 2012, pp. 109-115

[2] E. Mešković, Z. Galić, M. Baranović, "Managing Moving Objects in Spatio-Temporal Data Streams," in Proc. MDM 2011 12th IEEE Int. Conf. on Mobile Data Management, Lulea, Sweden, 2011, pp. 15-18

[3] R.H. Gütting, M. Schneider, "Moving objects databases," Morgan Kaufmann, 2003

[4] G. Trajcevski, O. Wolfson, C. Hu, H. Lin, F. Zhang, "Managing uncertain trajectories of moving objects with DOMINO," in Proc. 4th International Conference on Enterprise Information Systems, Ciudad Real, Spain, 2002, pp. 218-225

[5] G. Trajcevski, O. Wolfson, K. Hinrichs, S. Chamberlain, "Managing uncertainty in moving objects databases," ACM Transactions on Database Systems, Vol. 29(3), 2004, pp. 463–507.

[6] V.T. de Almeida, R.H. Gütting, "Supporting uncertainty in moving objects in network databases," in Proc. 13th annual ACM international workshop on Geographic information systems, New York, NY, USA, 2005, pp. 31-40

[7] R.H. Güting, M.H. Böhlen, M. Erwig, C.S. Jensen, N.A. Lorentzos, M. Schneider, M. Vazirgiannis, "A foundation for representing and querying moving objects" ACM Transactions on Database Systems (TODS), Vol. 25(1), 2000, pp. 1–42

[8] H. Jeung, H. Lu, H. Wang, S. Sathe, M.L. Yiu, "Managing evolving uncertainty in trajectory databases," IEEE Transactions on Knowledge and Data Engineering, 2013, Aug. 22, PrePrint

[9] C. Panagiotakis, N. Pelekis, I. Kopanakis, E. Ramasso, Y. Theodoridis, "Segmentation and sampling of moving objects trajectories based on representativeness," IEEE Transactions on Knowlege and Data Engineering, Vol. 24, no. 7, 2012, pp. 1328-1343

[10] K. Patroumpas, T. Sellis, "Managing Trajectories of Moving Objects as Data Streams," Spatio-Temporal Database Management, 2nd Int. Workshop STDBM'04, Toronto, Canada, 2004, pp. 41-46

[11] Y. Huang, C. Zhang, "New data types and operations to support geo-streams," in GIScience, vol. 5266 of Lecture Notes in Computer Science, T. J. Cova, H. J. Miller, K. Beard, A. U. Frank, M. F. Goodchild, Eds., Springer, 2008, pp. 106–118.

[12] M. Erwig, R. H. Güting, M. Schneider, M. Vazirgiannis, "Abstract and discrete modeling of spatio-temporal data types," in ACM-GIS, 1998, pp. 131–136.

[13] R. H. Güting, M. H. Böhlen, M. Erwig, C.S. Jensen, N.A. Lorentzos, M. Schneider, M. Vazirgiannis, "A foundation for representing and quering moving objects," ACM Transactions on Database Systems, Vol. 25(1), 2000, pp. 1–42.

[14] V. T. de Almeida, R. H. Güting, T. Behr, "Querying moving objects in SECONDO," in Mobile Data Management, IEEE Computer Society, 2006, pp. 47–51.

[15] Z. Galić, M. Baranović, K. Križanović, E. Mešković, "Geospatial data streams: formal framework and implementation", unpublished

[16] C. S. Jensen, H. Lahrmann, S. Pakalnis, J. Runge "The INFATI Data", *TimeCenter Technical Report*, 2006.

Automatic recognition of handwritten corrections for multiple-choice exam answer sheets

Marko Čupić, Karla Brkić, Tomislav Hrkać, Željka Mihajlović and Zoran Kalafatić
University of Zagreb, Faculty of Electrical Engineering and Computing
Unska 3, 10 000 Zagreb, Croatia
{marko.cupic, karla.brkic, tomislav.hrkac, zeljka.mihajlovic, zoran.kalafatic}@fer.hr

***Abstract*—Automated grading of multiple-choice exams is of great interest in university courses with a large number of students. We consider an existing system in which exams are automatically graded using simple answer sheets that are annotated by the student. A sheet consists of a series of circles representing possible answers. As annotation errors are possible, a student is permitted to alter the annotated answer by annotating the"error" circle and handwriting the letter of the correct answer next to the appropriate row. During the scanning process, if an annotated"error" circle is detected, the system raises an alarm and requires intervention from a human operator to determine which answer to consider valid. We propose rather simple and effecive computer vision algorithm which enables automated reading of a limited set of handwritten answers and minimizes the need for a human intervention in the scanning process. We test our algorithm on a large dataset of real scanned answer sheets, and report encouraging performance rates.**

I. Introduction

Automated grading of multiple choice exams is a necessity in modern university courses taken by a large number of students (e.g. a few hundreds). Although in recent years there has been a notable shift towards taking multiple choice exams on computers [1], [2], written exams are still inevitable when the number of students taking the exam surpasses the number of computers the university can provide at a given point in time. Manually grading hundreds of multiple choice exams is tedious work prone to errors, and a task more suitable for a machine than for a human. Using automated grading saves time, enables faster processing of exams and reduces the probability for grading errors. If exams are graded by a machine, the results of an exam can be known in as little as a few hours after the exam has finished, with very little human effort, hence offering a considerable benefit both for students and for course staff.

In this paper, we consider an existing system for automated exam grading [3], [4] currently in use at the Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia. The system operates by scanning and recognizing answer sheets of a predefined format on which the desired answers are annotated by the student. Although offering a fair amount of automation, the system is not yet fully automated, as it requires operator input in cases when the student makes an annotation error, i.e. decides to change an already annotated answer. In these cases, the student is expected to annotate a special error column (or annotate more than one answer where only a single answer is expected) and handwrite the letter corresponding to the appropriate answer in a dedicated rectangle. When the system detects that more than one answer or the error column had been annotated, a prompt is raised to the human operator, who is required to read the handwritten letter and input it back into the system. Hence, a certain amount of supervision is required during the automated grading process. We address this problem by proposing a computer vision technique for handwritten letter recognition suitable for use in the context of automated exam grading. The recognition task is constrained, as a limited number of letters need to be supported, given that multiple choice exams rarely offer more than six choices. Still, the relative simplicity of the task balances itself with high performance requirements: the recognition should be as fast as possible and as reliable as possible, and generalize well over a large number of different handwritings. Errors are highly undesirable, so the recognition module should be able to assess its own certainty in the recognition result. For uncertain recognitions, human input should be requested.

II. Related work

Character recognition has been a subject of intensive research for a long period of time. While recognition of printed text is today mostly considered a solved problem, recognition of handwritten text remains a challenging task, primarily due to high variability in handwriting among different people. A large number of methods have been proposed for handwritten text recognition. These methods can generally be categorized either as online or offline. Online recognition involves the automatic and real-time conversion of a text as it is written on a special digitizing device, where a sensor tracks the movements of the pen and an algorithm processes this information, while offline methods start from a static image usually obtained by scanning a document. An overview focused on offline handwritten character recognition methods can be found in [5].

The task of character recognition generally involves the following processing steps [5], [6]:

(i) Preprocessing (for instance including size normalization, image binarization, etc.),
(ii) Feature extraction,
(iii) Classification.

It is presumed that an adequate method for character localization has been previously applied, in order to find parts of the image containing each of the characters.

The key step in the process is the selection of an appropriate feature extraction method (or representation), since features must at the same time be sufficiently discriminative among the

character classes, but also invariant to the expected variations inside the class. Furthermore, the size of feature vectors should be relatively small in order to reduce computational complexity and therefore the time needed to perform the classification. A variety of different feature representations have been proposed in the literature. A standard survey can be found in [7]. Commonly used feature representations can be classified in three broad classes [5], [7]:

(i) *Statistical representations*, that represent a character image by a statistical distibution of points. This class includes methods such as analyzing densities of points or other features in different regions of an image (zoning), projection histograms, counting the number of crossings of an image contour with line segments in specified directions, and distances of line segments from given boundaries.

(ii) *Structural representations*, based on geometrical and topological properties, such as aspect ratio, distances between characteristic points or strokes, cusps, openings in different directions, chain codes, graph and tree representations, etc. These kinds of features encode some knowledge of the structure of the character, i.e. what sort of components it is made of.

(iii) *Representations obtained by global transformations*, that describe an image by a linear combination of a series of simpler well-defined functions. The features are then formed by coefficients of a such linear combination. Fourier transforms, Gabor transforms, wavelets, moments, Karhunen-Loeve expansion, etc., can be counted in this category.

As already stated, using distances of the image elements from the predefined boundaries has been used and described in the literature. Kim et al. [8] used distances of image pixels from the minimum bounding box of a whole word, along a certain number of horizontal and vertical scan-lines, in combination with several other feature types combined in 120-dimensional feature vector, applied to much harder problem of recognition of a limited set of handwritten cursive words. By using a heterogeneous classifier consisting of two multi-layer perceptrons and a hidden Markov model, they achieved recognition rate of 92.7% for 32 classes of legal words. Similarly, Mohamed [9] represented cursive handwritten word images by computing the location and number of transitions from background to foreground pixels along vertical lines (image columns). Hidden Markov model was used for classification. The obtained recognition rate was 83.9% (for the first ranked class). Surinta et al. [10] use the distance values between the closest black pixels and predefined hotspots in four or eight different directions to represent characters. They report the recognition rate of 90.1%.

In this paper, we adhere to the general idea of using image distances, and propose an offline character recognition method particulary suitable for recognizing a limited set of characters in the context of automated grading. In contrast to previously described methods, our method is specifically optimized to the set of supported characters, i.e. it can be said that it incorporates a degree of prior knowledge. It is designed so as to maximize separability between supported character classes by measuring specific properties characteristic to individual classes (e.g. it assumes that the existence or absence of the middle dash is an important classification cue).

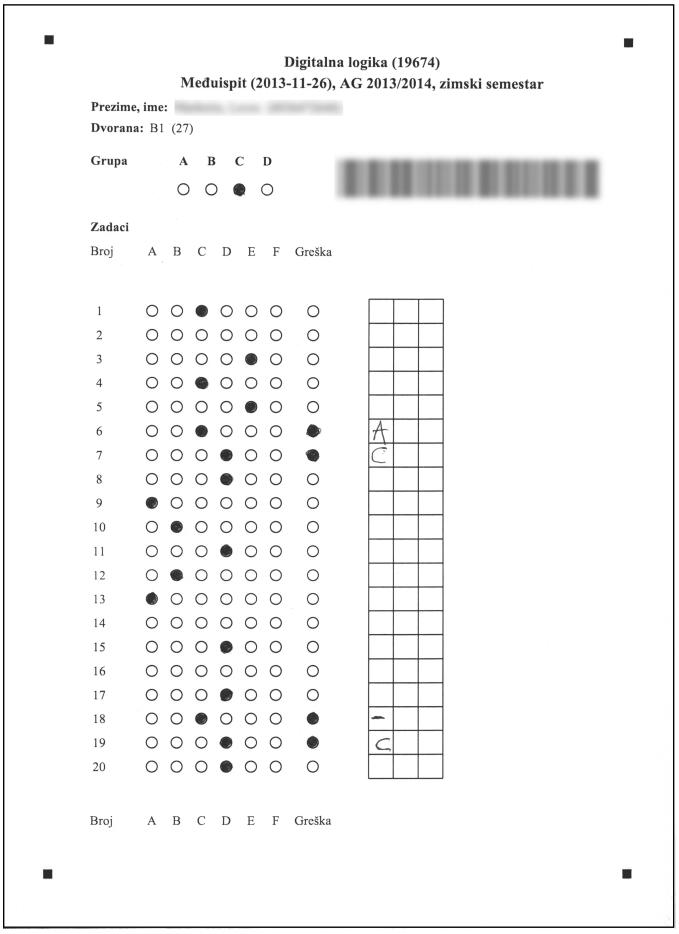


Fig. 1. An example of a standardized answer sheet.

III. THE EXISTING SYSTEM

The first step in developing any automated grading system is to introduce a standardized answer sheet that can be machine-processed. In this paper, we consider an automated grading system called *FreeFormReader*, currently in use at the Faculty of Electical Enginnering and Computing at University of Zagreb. An example of a standardized answer sheet that can be processed by the system is shown in Fig. 1. The answers are annotated by the student in an answer matrix that consists of an appropriate number of rows of empty circles. Each row corresponds to a single question, and each column to a single answer. To annotate an answer, the student darkens the appropriate circle. It is often the case that a student annotates an answer and later changes his mind. If that happens, the student is expected to annotate the "Error" cell ("Greska" in Croatian, as shown in Fig. 1) in the appropriate row or darken the correct answer in the same row, and write the letter corresponding to the final answer in the rectangle near the appropriate row, also illustrated in Fig. 1. When the system detects that the "Error" cell is marked, or that two or more answers are marked, an alarm is raised and a human operator must inspect and manually enter the correct answer.

Given that many exams at our university are graded in a way that wrong choices add negative points to the exam score

(usually -0.25 points), it sometimes happens that a student wishes to revoke his answer to a particular question, and leave the question unanswered. In that case, the same procedure is followed, except that instead of handwriting the letter of the appropriate answer, the student is expected to write the dash symbol ("-"). On the other hand, if the student is confident that for a certain question none of the offered answers are correct, or that there is more than one correct answer, the student is allowed to annotate the "Error" column in the corresponding row and write the letter X in the appropriate rectangle. This situation can occur when there has been a mistake in forming the exam. If an X is detected, the human operator should manually inspect the student's solution.

Currently, the system supports exams with up to 23 answer choices. However, the most frequent case are questions with four to six answers and for these cases we are trying to further automate the recognition process. Our practical experiences suggest that errors in annotation happen in about 20% of exams. The process of manually correcting annotation errors is tedious and it is easy to make a mistake, given the repetitive nature of the task.

IV. Representing and Classifying the Characters

In devising an appropriate representation and classification method for handwritten characters that could be used within our exisiting system, we adhere to the following requirements.

(i) The required computations for representation and classification should be very fast, so that current processing speed of the system is retained.
(ii) It can be assumed that the set of supported characters is limited to six answer letters (A, B, C, D, E, F) and two characters for revoking the answer and signaling a mistake in the question (dash and X).
(iii) The classification framework should be able to output confidence level. Predictions with low confidence level should be presented to the human operator.

In the following section we describe the method that we propose for the character recognition.

A. The edge distance representation

After scanning the document, an approximate location of the table containing the handwritten correction answers is manually annotated by the human operator. This has to be done only for the first exam sheet, since the approximate position of the table remains the same for all the other sheets. Then the precise location of the table and the locations of individual cells must be found. This can be done either by finding lines using Hough transform (and finding individual cells by finding their corners defined by intersections of found lines), or using a method described in [4], or any other. Currently, for experimentation purposes, we use (slow) Hough transform.

Once the individual cells have been found, a procedure for finding a minimal character bounding box is performed. Then the corresponding part of the image is further preprocessed by reducing its dimensions to 24×32 pixels, since we expect that in this resolution the distinguishing features of characters are still preserved. We then perform binarization by simple histogram-based adaptive thresholding. The result is a 24×32 binary image in which the character extends from top to bottom and from left to right margin of the image. The size normalization is essential, since the proposed features will only have low variability in case of size normalized images, thus lending themselves to character recognition.

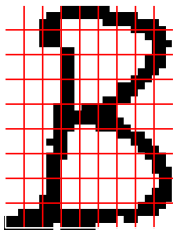


Fig. 2. The grid of $M \times N$ horizontal and vertical scan-lines

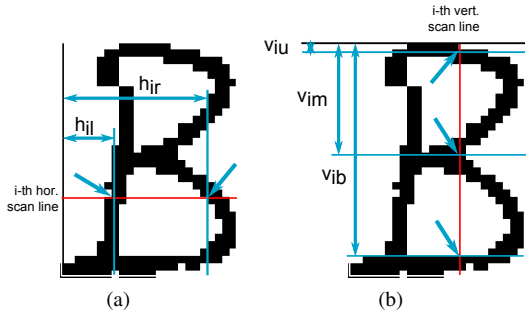


Fig. 3. The definition of features: (a) horizontal features; (b) vertical features

To facilitate the recognition, we propose a simple but effective feature extraction method resulting in an *edge distance* representation of characters (Figs. 2 and 3). In the image, M horizontal and N vertical equidistant scan-lines are defined (in our experiments, $M = N = 8$), as shown in Fig. 2. On each horizontal scan-line i, two features are defined (Fig. 3(a)): the relative distance of the leftmost black image pixel on that line from the left margin of the image (h_{il}) and the relative distance of the righmost black image pixel on the same line again from the left image margin (h_{ir}). On each vertical scan-line, however, three features are defined (Fig. 3(b)): the relative distance of the uppermost black pixel on the $i-th$ vertical scan-line from the upper image border (v_{iu}), the relative distance of the bottom-most black pixel on the same line from the top of the image (v_{ib}), and finally, the third feature (v_{im}) intended to reflect the position of the "middle stroke" that exists in some letters (such as B and E), i.e. the relative distance of the pixels of this "middle stroke" crossing the $i-th$ vertical line from the top of the image, if the "middle stroke" exists. More precisely, v_{im} is defined as the relative distance from the top of the image to the black pixel below the uppermost pixel that is at the same time disconnected both from the uppermost and from the bottommost pixels and is closest to the center of the letter. If it does not exist, i.e. if there are no pixels on the $i-th$ vertical scan-line disconnected from uppermost and bottommost black pixels, the feature v_{im} is set to 0. The described representation results in a feature vector of dimension $2M+3N$ (in our case the dimension was $2 \times 8 + 3 \times 8 = 40$).

Intuitively, the features defined in the described way should be sufficiently discriminative for characters belonging to dis-

tinct classes, while maintaining a relatively low variability inside a class, regardless of different handwriting styles. We emphasize here that our goal is not to define rotation invariant features, since we do not expect hanadwritten characters to have any larger degree of rotation. The features (h_{il}), (h_{il}), (v_{iu}) and (v_{ib}) describe in a rough way the outer shape of the letter, while (v_{im}) enables easier distinguishing between letters that could have similar outer shape in some handwritings (i.e. C and E; B and D).

B. Selecting the classifier

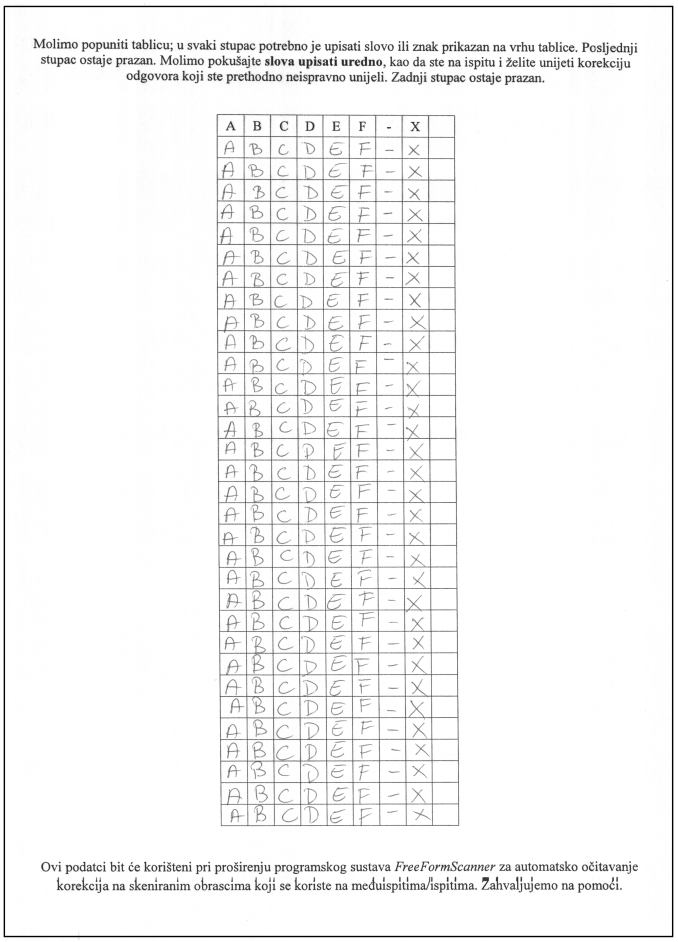
Given the desired performance requirements, we opted to use the random forest classifier proposed by Breiman [11]. The random forest classifier is reasonably fast to train, the classification results are fast to compute, and it offers performance comparable [12] to other popular classifiers such as support vector machines [13]. The main idea of the random forest approach is combining randomized decision trees into ensembles. When building a decision tree within a random forest, for each split at inner tree nodes, the algorithm randomly selects a subset of m attributes and then performs the split by the best attribute from the selected subset, based on the Gini impurity index. In contrast to SVM, which is posed as a binary classification algorithm, random forests inherently support multiple classes. Given a sample, each of the trees of the classifier casts a vote for one of the classes, and the final class of the sample is determined by the majority of votes. Tunable parameters of the random forest algorithm are the number of trees, the number of top candidates to consider (m), and the allowed depth of individual decision tree.

V. Experiments

For experimental validation of the proposed procedure, we collected two datasets. The training dataset contains handwriting samples from 289 individuals, and the test-dataset contains samples from 15 individuals. Each individual was asked to write all the considered characters (A, B, C, D, E, F, X, -) 32 times, using a specialized sheet as illustrated in Fig. 4. Hence, for each individual we collected a total of 264 samples plus the eight printed samples from the table header, resulting in the training dataset size of 76296 samples and the test set size of 3960 samples. There is no mixing of samples from the training and the test set, i.e. individuals contributing to the training set were not a part of the second group that contributed to the test set. Multiple samples per person were collected with the goal of improved learning, given the inevitable variations that occur within different instances of the same handwriting.

The edge distance representation was implemented in Java in a straightforward way, according to the described algorithm. For each scanned image, the table containing the handwritten letters was located using Hough transform; then the cells were located using intersections of appropriate border lines. The procedure for the minimum letter bounding box described in [4] was then applied, and the character image was extracted and further binarized. On the binarized image, the edge distance features were extracted.

For testing classification performance of the random forest classifier, we used the open source data mining tool Weka [14]. Weka offers a graphical user interface for experimenting


Molimo popuniti tablicu; u svaki stupac potrebno je upisati slovo ili znak prikazan na vrhu tablice. Posljednji stupac ostaje prazan. Molimo pokušajte **slova upisati uredno**, kao da ste na ispitu i želite unijeti korekciju odgovora koji ste prethodno neispravno unijeli. Zadnji stupac ostaje prazan.

A	B	C	D	E	F	-	X	

Ovi podatci bit će korišteni pri proširenju programskog sustava *FreeFormScanner* za automatsko očitavanje korekcija na skeniranim obrascima koji se koriste na međuispitima/ispitima. Zahvaljujemo na pomoći.

Fig. 4. An example of an annotation sheet used to collect handwritten character samples.

with various classifiers, but it is also available as a standalone Java library. Given that the existing automated grading system is written in Java, we expect that using Weka will ensure a seamless integration of the final classifier with the existing code.

A. Classification results

We experimented with several parameter settings, and found that the best performance is obtained by using a forest of 100 trees, considering $m = 6$ random attributes in each split, and not limiting the tree depth. Using this setup, the random forest classifier was able to correctly classify 96.5% of samples from the test set (3821 of 3960). The confusion matrix is shown in Table I. It can be seen that, as expected, most confusion occurs between visually similar letters. For example, 20 instances of the letter E were classified as C, although only four C's were classified as E's. It seems more common for the letter E to lack a middle dash, making it similar to C, than for the letter C to contain an additional dash that would make it similar to E. Other notable confusions include 18 instances of the letter B being classified as E, 16 instances of the letter E being classified as B, etc.

In order to better illustrate the amount of confusion among different classes, a visualization of the confusion matrix is

TABLE I. THE CONFUSION TABLE FOR THE TESTING SET. ROWS CORRESPOND TO CORRECT LABELS, AND COLUMNS TO THE NUMBER OF CLASSIFICATIONS AS A GIVEN SYMBOL OUTPUT BY THE RANDOM FOREST.

	A	B	C	D	E	F	-	X
A	483	2	1	5	0	1	3	0
B	5	454	0	14	18	3	0	1
C	0	1	487	0	4	1	2	0
D	1	9	3	479	1	0	2	0
E	1	16	20	0	454	3	0	1
F	2	2	0	0	2	489	0	0
-	0	0	0	0	0	1	493	1
X	11	0	1	0	0	0	1	482

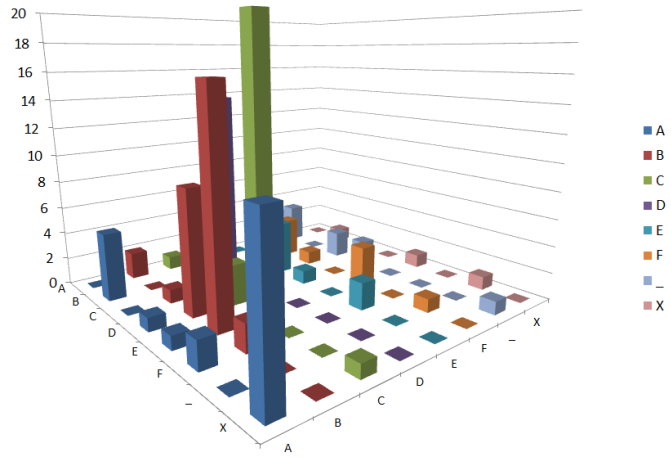


Fig. 5. Visualization of the confusion matrix without diagonal elements



(a) (b) (c) (d) (e) (f) (g)

Fig. 6. Examples of misclassifications. From left to right: A classified as B, B classified as D, C classified as E, D classified as B, E classified as B, F classified as E and X classified as A.

shown in Fig. 5. In order to emphasize misclassifications, we removed the bars corresponding to correctly classified examples so that only misclassifications are left. The height of each bar corresponds to the number of misclassified examples at the corresponding location of the confusion matrix.

A few misclassified samples from the test dataset are shown in Fig. 6. The correct labels from left to right are A, B, C, D, E, F, X. Note that most of these examples are hard to classify even for a human. For example, the examplified letter B reminds of a heart symbol and in a certain way to letter D, the letter E looks like a G, the letter F could easily be confused for an E, and the character X reminds of the Greek letter λ.

B. Varying the decision threshold

As discussed before, the proposed system should provide a confidence measure of the predicted label, so that labels with low confidence can be presented to the human operator. In order to find the optimal classification certainty threshold below which manual operator intervention will be required, we have considered two facts:

(i) The classifier will inevitably produce a certain number of misclassifications that will be misclassified with a certainty greater than the confidence threshold; these will initially pass undetected and will require later intervention based on student complaints.

(ii) There will be some correctly classified letters for which the system will ask the operator to manually confirm the classification, because the classification certainty will be smaller than the confidence threshold; this scenario costs operator time.

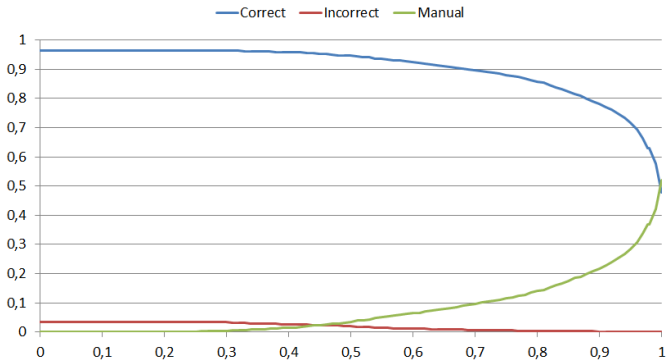


Fig. 7. The percentage of correct classifications, misclassifications and uncertain classifications requiring manual input, depending on confidence-level threshold (for the first method; for the second method the diagram is similar). X-axis corresponds to confidence-level thresholds, while y-axis represents the percentage of correct, incorrect and uncertain classifications.

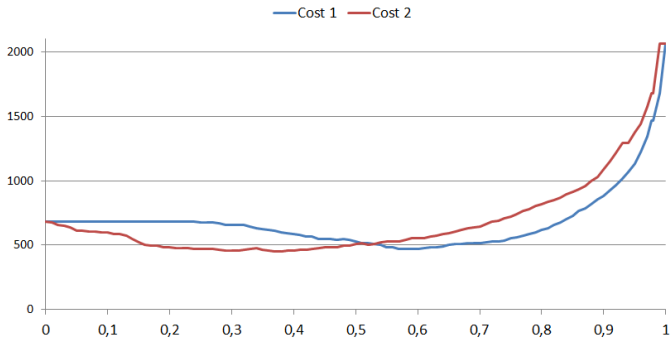


Fig. 8. Performance of penalty function. X-axis corresponds to confidence-level thresholds, while y-axis represents values of penalty functions for the two proposed approaches (less is better). *Cost 1* corresponds to method (i) and *Cost 2* to method (ii).

To tackle these two situations, we observe that it takes more time for course staff to manually inspect and correct missclasifications upon receiving a complaint by a student than at the time of sheet proccesing. Therefore, we define the following cost function: for each classification with insufficient confidence level we add a penalty of 1 point; for each undetected misclassification (misclassification with a confidence level greater than the threshold) we add a penalty of 5 points. Given this penalty function, we define two approaches to solving the problem of selecting the optimal threshold.

(i) In this approach we require that the absolute value of confidence for the predicted character class must be greater than θ. We evaluated the clasification performance for each distinct θ in $[0, 1]$ which occured as confidence of prediction for any character. With the change of threshold, the number of correct classifications, misclas-

sifications and uncertain classifications changes, and is illustrated in Fig. 7. The performance of penalty function is presented in Fig. 8. As can be seen, for this case the optimal certainty-level threshold is 0.59 at which we have 92.8% correct classifications, 1.2% misclassifications and 6.1% manual interventions.

(ii) In this approach we require that the relative value of confidence for the predicted character class must be greater than θ. This means that the difference between the probability of the most probable class and the second most probable class must be greater than θ. We evaluated the clasification for each distinct θ in $[0, 1]$ which occured as confidence of prediction for any character. With the change of threshold, the number of correct classifications, misclassifications and uncertain classifications changes, and is similar to the one illustrated in Fig. 7. The performance of the penalty function is presented in Fig. 8. As can be seen, for this case the optimal certainty-level threshold is 0.37 at which we have 92.6% correct classifications, 1.0% misclassifications and 6.4% manual interventions.

VI. Conclusion and outlook

We have presented a technique based on edge distances and random forest classifier that can be used to improve an existing automated grading system by machine-based recognition of handwritten characters. In the existing setup, these characters represent answer alterations in case a student annotates an answer in a predefined answer matrix and then changes his mind. Extensive experiments were performed on a large dataset containing samples of eight supported characters. Experiments indicate that by introducing the proposed technique into the existing automated grading system, more than 92% of such answers could be machine-read, hence saving valuable operator time and further improving the automated grading process.

Acknowledgements

This work has been financially supported by the University of Zagreb, Croatia, under grant Research support 2, contract 2013-ZUID-19, and by the Research Centre for Advanced Cooperative Systems ACROSS (EU FP7 #285939). This support is gratefully acknowledged.

References

[1] M. Čupić and Ž. Mihajlović, "Computer-based knowledge, self-assessment and training," *International Journal of Engineering Education*, vol. 26, no. 1, pp. 111–125, 2010.

[2] Ž. Mihajlović and M. Čupić, "Software environment for learning and knowledge assessment based on graphical gadgets," *International Journal of Engineering Education*, vol. 28, no. 5, pp. 1127–1140, 2012.

[3] M. Čupić, "A case study: using multiple-choice tests at university level courses – preparation and supporting infrastructure," *International Journal of Intelligent Defence Support Systems (IJIDSS)*, vol. 3, no. 1/2, pp. 90–100, 2010.

[4] M. Čupić, K. Brkić, T. Hrkać, and Z. Kalafatić, "Supporting automated grading of formative multiple choice exams by introducing student identifier matrices," in *MIPRO, 2011 Proceedings of the 34th International Convention*, May 2011, pp. 993–998.

[5] N. Arica and F. T. Yarman-Vural, "An overview of character recognition focused on off-line handwriting," *Trans. Sys. Man Cyber Part C*, vol. 31, no. 2, pp. 216–233, May 2001. [Online]. Available: http://dx.doi.org/10.1109/5326.941845

[6] G. Vamvakas, B. Gatos, and S. J. Perantonis, "Handwritten character recognition through two-stage foreground sub-sampling." *Pattern Recognition*, vol. 43, no. 8, pp. 2807–2816, 2010.

[7] Ø. D. Trier, A. K. Jain, and T. Taxt, "Feature extraction methods for character recognition - a survey," *Pattern Recognition (PR)*, vol. 29, no. 4, pp. 641–662, 1996.

[8] J. H. Kim, K. Kyung, Kim, and C. Y. Suen, "Hybrid schemes of homogeneous and heterogeneous classifiers for cursive word recognition," in *In Proc. 7th International Workshop on Frontiers in Handwriting Recognition*, 2000, pp. 433–442.

[9] M. Mohamed and P. Gader, "Handwritten word recognition using segmentation-free hidden markov modeling and segmentation-based dynamic programming techniques," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 5, pp. 548–554, May 1996. [Online]. Available: http://dx.doi.org/10.1109/34.494644

[10] O. Surinta, L. Schomaker, and M. Wiering, "Handwritten character classification using the hotspot feature extraction technique." in *ICPRAM (1)*, P. L. Carmona, J. S. Snchez, and A. L. N. Fred, Eds. SciTePress, 2012, pp. 261–264.

[11] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001. [Online]. Available: http://dx.doi.org/10.1023/A:1010933404324

[12] A. Jović, K. Brkić, and N. Bogunović, "Decision tree ensembles in biomedical time-series classification." in *DAGM/OAGM Symposium*, ser. Lecture Notes in Computer Science, A. Pinz, T. Pock, H. Bischof, and F. Leberl, Eds., vol. 7476. Springer, 2012, pp. 408–417.

[13] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995. [Online]. Available: http://dx.doi.org/10.1023/A:1022627411411

[14] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: An update," *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, Nov. 2009. [Online]. Available: http://doi.acm.org/10.1145/1656274.1656278

Discrimination Between Serbian and Slovenian Language by Texture Analysis

Darko Brodić*, Zoran N. Milivojević**, Čedomir A. Maluckov* and Milena Jevtić*
* University of Belgrade, Technical Faculty in Bor, Bor, Serbia
** Technical College Niš, Niš, Serbia
dbrodic@tf.bor.ac.rs

Abstract - The paper proposed the method for the language identification according to texture analysis. First, the algorithm encrypted the text given in different languages to the cipher based on the baseline status of each script element in text. Then, the cipher was subjected to the texture analysis. The aim of this analysis was the extraction of texture features. Obtained texture features showed significant diversity between languages. Hence, they were suitable for creating the discrimination and identification criteria between languages. The proposed method was tested on Serbian and Slovenian documents from custom oriented database. The experiments gave encouraging results.

I. INTRODUCTION

Language identification is defined as a task of determining the language in a certain document. It represents a preprocessing step, which is commonly included in the optical character recognition (OCR) systems, information retrieval systems (IRS) or web page encoding [1]. That's why, it used to be an initial step in a text processing system [2].

Language identification has become more and more important, as more and more textual data are present on the Internet. In recent years, the number of web pages and languages has increased tremendously. Currently, the number of languages which has been implemented in Google is around 150 [3].

Many studies of language identification on written text have been reported in the past [4-11]. They achieved a good accuracy. Consequently, a comparative survey of language identification methods for written text was reported in [12]. However, text from the Web is considerably different [13]. The difficulties presented by the Web domain are generally related to the noisier nature of the text, including things like spelling errors, small amounts of text or multilingual documents.

Two major stages in language recognition systems can be identified [14]: (i) the language modeling stage, where the distinctive features of languages are determined and stored in models, and (ii) the classification stage, in which the model of the (partial) input document is compared to the reference language models.

This work was partially supported by the Grant of the Ministry of Science from Republic of Serbia, as a part of the project TR33037 and III43011 within the framework of Technological development program.

The language model most similar to the input document represents the language of the document. As a basis for document modeling, we employ cryptography [15]. Cryptography converts data from a readable to an unreadable state. Readable data is called a plain text, while unreadable data represent a cipher [16]. This process is realized with a cryptographic algorithm. It is used to encrypt and/or decrypt data. Although there exist three different types of cryptographic algorithms, we used a hash algorithm (HF) which performs a one-way encryption. It means that once the data has been encrypted, there is no path to retrieve the original data [17]. However, the decryption process is an unessential for further texture processing.

The modeling is based on the baseline characteristics of each letter. This way, we perform the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups. Due to clustering the number of the active variables has been reduced. Then, the obtained model is replaced with the cipher [15]. Cipher is used as appropriate criteria for further texture analysis. From the above, it is clear the such a model greatly reduces the computation costs. Furthermore, the cipher text is subjected to the feature extraction. The gray level co-occurrence matrix (GLCM) is used as a basis for feature extraction [18-19]. Then, the obtained texture features are classified. According to these results, it is possible to make a strong margin in some feature characteristics, which can be used as criteria for characterization of the certain language as well as for distinction and dissemination between languages.

The rest of the paper is organized as follows. Section 2 describes the proposed algorithm, which is based on the extraction of texture features from the cipher text. Section 3 describes the elements of the experiment. Section 4 shows the results of the experiment. Furthermore, it discusses the obtained results and establishes the decision-making criteria for the language discrimination in web pages. Section 5 gives the conclusion with the future research plans.

II. PROPOSED ALGORITHM

The algorithm consists of the following stages:

- Cipher modeling,

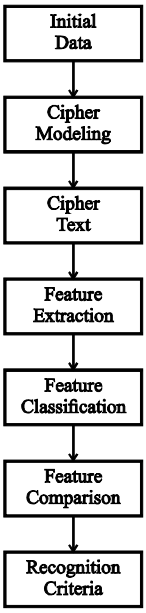


Figure 1. The flow of the proposed algorithm

- Feature extraction,
- Feature classification,
- Feature comparison, and
- Recognition criteria.

Figure 1 shows the flow of the proposed algorithm.

A. *Cipher Modeling*

Cipher modeling converted different scripts by HF according to the text-line structure definition. It is based on the assumption that each text-line in printed documents contains following three vertical zones [20]: (i) the upper zone, (ii) the middle zone, and (iii) the lower zone.

All letters, diacritics, ligatures or signs represent the constituents of any script. They are classified according to the zones herein. Figure 2 (a) shows the script characteristics in accordance to vertical zones. The base letter **a**, which occupies the middle zone, is mapped to B. The ascender letter **b** occupies both the upper and middle zones. It is mapped to A. The descender letter **p** occupies the lower and middle zones. It is mapped to D. Finally, the full letter **Lj** enlarges among the upper, middle and lower zones. It is mapped to F. Hence, the model is based on the following elements {B,A,D,F}. For further texture analysis, above set is encrypted to cipher set {0,1,2,3}, respectively. It is shown in Figure 2 (b).

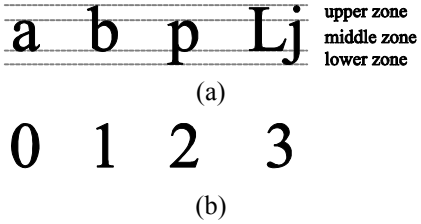


Figure 2. Illustration of the cipher modeling

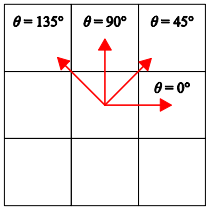


Figure 3. Illustration of the window of interest (the four directions of adjacency for calculating the Haralick texture features with $d = 1$).

B. *Feature extraction*

In our approach, the features extraction is established with texture analysis of the cipher text. Due to text characteristics the cipher text represents 1-D image given with vector **A** featuring M rows (in this special case $M = 1$) and N columns. Furthermore, the number of gray levels G is determined by the number of the levels in cipher text, which is set to 4 (using into account the elements from the set {0, 1, 2, 3}).

The method begins at the top left corner of the image **A** and counts the occurrences of each reference pixel in reference to neighbor pixel relationship, i.e. in the window of the interest (WOI). Each element (i, j) of GLCM represents the sum of the times the pixel with the value i was located at some distance d and angle θ from the pixel of intensity j. Accordingly, the element (i, j) gives the number how many times the gray levels i and j appears as a sequence of two neighbor pixels located at a defined distance d along a chosen direction θ. Figure 3 illustrated the WOI.

The GLCM for an image $A(x, y)$ featuring M rows and N columns is parameterized by the offset $(\Delta x, \Delta y)$. It is:

$$C(i,j) = \sum_{x=1}^{M}\sum_{y=1}^{N}\begin{cases}1, & \text{if } A(x,y) = i \wedge A(x+\Delta x, y+\Delta y)\\ 0, & \text{otherwise}\end{cases} \quad (1)$$

The offset $(\Delta x, \Delta y)$ represents the pixel displacement d and the orientation θ for which the co-occurrence matrix is calculated. Consequently, the feasible values of the parameters d and θ are narrowed to 1 and θ (for the 1-D image), respectively. The probability version of GLCM is:

$$P(i,j) = \frac{C(i,j)}{\sum_{i=1}^{M}\sum_{j=1}^{N} C(i,j)}. \quad (2)$$

Figure 4 shows the values of probability GLCM **P** for the Serbian and Slovenian language.

0.4105	0.2057	0.0694	0.0033
0.2115	0.0192	0.0017	0.0017
0.0619	0.0092	0.0008	0
0.0050	0	0	0

(a)

0.3929	0.2024	0.0708	0.0054
0.2015	0.0299	0.0082	0.0009
0.0726	0.0073	0.0018	0
0.0054	0	0.0009	0

(b)

Figure 4. The probability GLCM **P**: (a) for Serbian, (b) for Slovenian

C. Feature classification

The extracted fourteen texture features [18] can be used for the classification. In our case, the three features only can be used for dissemination between Serbian and Slovenian language. It is reasonable because both languages belong to the same group of South Slavic languages sharing similar characteristics like the corresponding dialect [21]. Hence, the Energy, Maximum (probability), and Entropy are the only features which can distinguish both languages.

Energy is a measure of textural uniformity of an image under consideration. It is defined as:

$$Energy = \sum_{i=1}^{M}\sum_{j=1}^{N} P^2(i,j). \quad (3)$$

Energy reaches its highest value when gray level distribution has either a constant or a periodic form. In contrast, if the image contains a large number of small entries, the energy will receive smaller value.

Maximum (probability) determines the most predominant pixel pair in the image. It is defined as:

$$Maximum = \max_{i,j} P(i,j). \quad (4)$$

Entropy is a feature which measures the randomness of gray level distribution of the image. It is defined as:

$$Entropy = \sum_{i=1}^{M}\sum_{j=1}^{N} P(i,j)\log P(i,j). \quad (5)$$

Entropy measures the disorder of an image. It achieves its smallest value when the image has mainly uniform level. However, when the image is not texturally uniform many GLCM elements have very small values, which imply that entropy is very large.

D. Feature comparison

All texture features (proposed in [18]) are unique for each language. To identify any of the languages, their comparison is mandatory. If two languages are compared, then the texture feature ratios is a solid way to start with. In our case, three texture feature ratios are of the interest:

- Energy ratio,
- Maximum ratio, and
- Entropy ratio.

E. Recognition criteria

To establish recognition criteria between two languages, the texture feature ratios are used to create a logical relation, which can be used for language distinction. This logical equation is the aim of the recognition process given by the proposed algorithm.

III. EXPERIMENTS

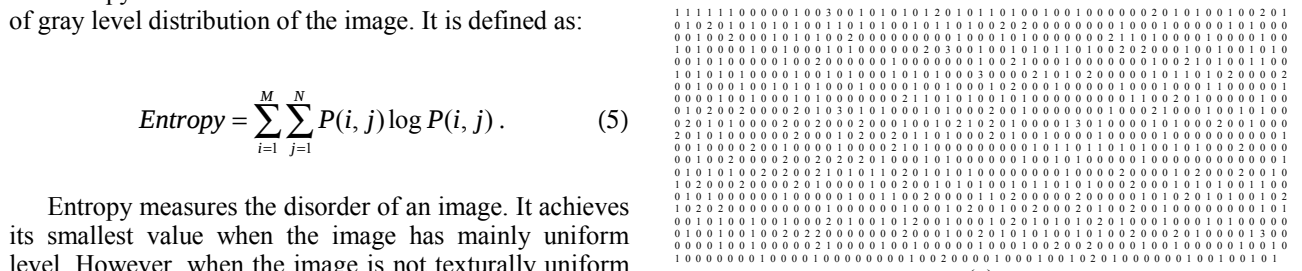
For the sake of the experiment, a custom-oriented database is created. It consists mainly of short texts which are extracted from different web sites. However, it includes few longer texts from PDF documents. The database includes 100 documents. This number of samples represents so-called "rule of 100", which can be used as rule of thumb in factor analysis [22].

The smallest document represents the text which length is roughly 500 characters only, while the longest one contains around 5000 characters. Figure 5 illustrates the initial data from document given in Serbian and Slovenian as well as the appropriate ciphers.

IV. RESULTS AND DISCUSSION

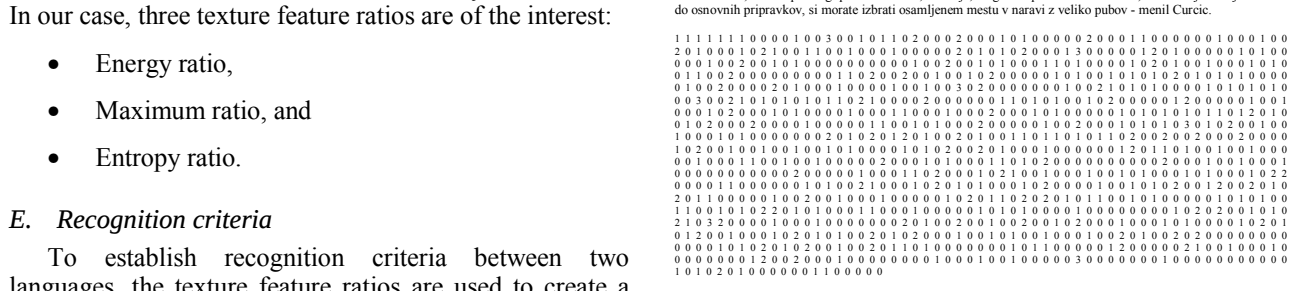
The obtained results can be divided into following categories: (i) texture features for Serbian text, (ii) texture features for Slovenian text, and (iii) texture features ratios between Serbian and Slovenian text.

LONDON, 30. novembar, (Njuz) – Nekadašnji fudbaler Partizana i pobednik rijalitija „Veliki brat“ Saša Ćurčić dobio je poziv menadžera Arsenala Arsena Vengera da do kraja sezone vodi omladince ovog kluba u nočne izlaske širom Londona.
Kako saznaje Njuz, bivši fudbaler je pristao da mlade nade Arsenala upozna sa čarima noćnog života, za mesečnu platu od 4.000 funti, koliko smatra da će mu biti dovoljno za plaćanje cena koktela, piva i garderobe u lokalima širom britanske prestonice.
Prvi čovek Arsenala smatra da omladinci ovog kluba treba da na vreme otkriju čari izlazaka, jer je mnogo bolje da kroz tu fazu prođu u ranom tinejdžerskom dobu, nego kada zaigraju za prvi tim Topdžija.
- Razmišljali smo ko bi mogao da im pokaže sve najznačajnije klubove po Londonu, a da im istovremeno za šankom ispriča i neku anegdotu vezanu za fudbal. Saša Ćurčić nam je svima bio prva opcija, pogotovu kada smo na Youtube na nekim snimcima videli kako sjajno pleše, što je takođe veoma važno za posao koji mu predstoji – izjavio je Arsen Venger.
Saša Ćurčić kaže da mu je velika čast da omladincima Arsenala otkrije ono što njemu nije imao ko da pokaže, zbog čega je morao sam da se snalazi.
- Najvažnije mi je da igrači na vreme dođu u dobru kondiciju. Poslednja stvar koju želite je da vaši omladinci već nakon drugog piva ne mogu više i požele da idu da spavaju. Razmišljao sam i da odemo negde van Londona na bazične pripreme, treba samo da odaberemo neko izolovano mesto u prirodi sa dosta pabova – smatra Ćurčić.



(a)

LONDON, 30. November (Njuz) - Nekdanji nogometaš in zmagovalka resničnostnega šova "Big Brother" Sasa Curcic je bila povabljena, vodja Arsenala Arsene Wenger do konca sezone traja mlade iz kluba v noči po Londonu.
Kako odkrije novice, nekdanji igralec dogovorili, da mladi hopefuls Arsenala spoznajo čarobno nočno življenje, za mesečno plačo 4.000 £, da misli, da bo dovolj za plačilo koktajli cen, pivo in oblačila v trgovinah po vsem britanski prestolnici.
Prvi človek verjame, da Arsenal mladi ta klub potrebuje za pravočasno odkrili čare gre ven, ker je veliko bolje, da gredo skozi to fazo, v svojih zgodnjih najstniških letih, ko pa igrajo za prvo ekipo strelec.
- Mislila sva, ki bi jim pokazati vse večje klube v Londonu, in da se hkrati v baru in povedal anekdoto, povezano z nogometom. Sasa Curcic vse nas prva možnost, še posebej, če ste na YouTubu videti nekaj posnetkov na velikem plesu, ki je tudi zelo pomembno delo pred njim - je dejal Arsene Wenger.
Sasa Curcic dejal, da je počaščen, da Arsenal mladi odkrili, kaj je bilo, nihče ne bi naredil vtis, zato sem moral spopasti.
- Najbolj pomembna stvar se mi je, da igralci pridejo na čas v dobrem stanju. Zadnja stvar, ki jo želim je, da je vaša mladost, vendar po drugi piva ne morem, in želijo, da gredo spat. Zato sem mislil, da se nekje zunaj Londona do osnovnih pripravkov, si morate izbrati osamljenem mestu v naravi z veliko pubov - menil Curcic.



(b)

Figure 5. Initial document from database and cipher text: (a) Serbian, (b) Slovenian

Table I shows the texture features obtained from the Serbian text.

TABLE I. TEXTURE FEATURES FOR SERBIAN TEXT

Serbian	*Energy*	*Entropy*	*Maxi*
Min	0.2444	-1.6801	0.3754
Max	0.2856	-1.5006	0.4441
Mean	0.2633	-1.5921	0.4156

From Table I the Serbian text is characterized with energy between 0.2444 and 0.2856, entropy between -1.6801 and -1.5006 and maximum between 0.3754 and 0.4441. Table II shows the texture features obtained from the Slovenian text.

TABLE II. TEXTURE FEATURES FOR SLOVENIAN TEXT

Slovenian	*Energy*	*Entropy*	*Maxi*
Min	0.2239	-1.7199	0.3341
Max	0.2755	-1.5701	0.4388
Mean	0.2422	-1.6556	0.3843

From Table II the Slovenian text is characterized with energy between 0.2239 and 0.2755, entropy between -1.7199 and -1.5701 and maximum between 0.3341 and 0.4388. Table III shows the texture feature ratios obtained from the Serbian/Slovenian texture feature ratios.

TABLE III. TEXTURE FEATURE RATIOS

Serbian/ Slovenian	*Energy Ratio*	*Entropy Ratio*	*Maxi Ratio*
Min	0.9781	0.9109	0.9699
Max	1.1821	0.9959	1.2501
Mean	1.0821	0.9566	1.0921
Criteria	> **1** (99%)	< **1** (100%)	> **1** (99%)

From Table III the comparison between Serbian and Slovenian text is characterized with energy ratio between 0.9781 and 1.1821, entropy ratio between 0.9109 and 0.9959 and maximum ratio between 0.9699 and 1.2501. Such close values are reasonable due to the language similarities [21]. However, the energy ratio > 1 as well as the maximum ratio > 1 are obtained in 99%. Furthermore, entropy ratio < 1 is obtained in all cases. Accordingly, this texture feature ratio can be used as a solid criteria in discrimination process between Serbian and Slovenian language. Hence, the logical equation that can be used for language recognition between Serbian and Slovenian language is given as `Energy ratio < 1`.

The proposed approach is characterized as computationally non-intensive. The algorithm is coded in Matlab without any programming parallelism. Testing shows that typical processing time is as low as 0.1 second per text including 2000 characters. Furthermore, the algorithm doesn't need any prior training like n-gram approach. At the end, it is suitable for short text samples of 300-400 characters.

V. CONCLUSIONS

The manuscript proposed a new approach to the language distinction between different languages. This approach establishes a model based on the baseline status of each letter in text. Then, it is encrypted making a cipher, which is subjected to the texture analysis. The analysis extracts the texture features of the cipher text. Obtained texture features showed significant diversity between languages. Accordingly, they are a starting point in the process of creating the language discrimination and identification criteria. The proposed method was tested on documents from custom oriented database. The algorithm achieved a good accuracy. At the end, it is manifested as computationally non-intensive algorithm.

Future research plans include the creating of database languages that can distinguish them mutually. Furthermore, the proposed algorithm will be accustomed to be used effectively in OCR.

REFERENCES

[1] S. Kranig, "Evaluation of language identification methods," Proceedings of ACM SAC, 2006.

[2] G. G. Chowdhury, "Natural language processing," Annual Review of Information Science and Technology, vol.37, no.1, pp. 51-89, 2003.

[3] C. C. Ng, A. Selamat, "Feature Preprocessing on Web Page Language Identification," Proceedings of the 5th Postgraduate Annual Research Seminar (PARS2009), Faculty of Computer Science and Information Systems (FSKSM), Universiti Teknologi Malaysia (UTM), pp. 238-243, 2008.

[4] E. M. Gold, "Language identification in the limit," Information and Control, vol.1967, no. 5, pp. 447-474, 1967.

[5] W. B. Cavnar, J. M. Trenkle, "N-Gram-Based Text Categorization," Proceedings of Third Annual Symposium on Document Analysis and Information Retrieval, pp. 161-175, 1994.

[6] T. Dunning, "Statistical Identification of Language," Technical report CRLMCCS-94-273, Computing Research Lab., New Mexico State University.

[7] C. Souter, G. Churcher, J. Hayes, J. Hughes, S. Johnson, "Natural language identification using corpus-based models," Hermes Journal of Linguistics, vol.13, pp.183–203, 1994.

[8] M. J. Martino, R. C. Paulsen Jr., "Natural language determination using partial words," United States Patent US 6.216.102 B1, April 10, 2001.

[9] I. Suzuki, Y. Mikami, A. Ohsato, Y. Chubachi, "A language and character set determination method based on N-gram statistics," ACM Transactions on Asian Language Information Processing (TALIP), pp. 269-278, 2002.

[10] T. Olvecky, "N-Gram Based Statistics Aimed at Language Identification," Proceedings of the Student Research Conference in Informatics and Information Technologies, pp. 1-7, 2005.

[11] B. Martins, M. J. Silva, "Language identification in web pages." 2005 ACM symposium on Applied computing, pp.764-768, 2005.

[12] L. Grothe, E. W. De Luca, A. Nürnberger, "A Comparative Study on Language Identification Methods, " Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), pp. 980-985, 2008.

[13] E. Amitay, "Hypertext - the importance of being different," Master's thesis, Centre for Cognitive Science, Edinburgh University, 1997.

[14] A. Poutsma, "Applying Monte Carlo Techniques to Language Identification," Proceedings of Computational Linguistics in the Netherlands (CLIN), pp.179-189, 2001.

[15] D. Brodic, Z. N. Milivojevic, C. A. Maluckov, ”Recognition of the Script in Serbian Documents using Frequency Occurrence and Co-occurrence Analysis,” The World Scientific Journal, vol. 2013, art. 896328, pp.1-14, 2013.

[16] C. Smith, ”Basic Cryptanalysis Techniques,” Tehnical Report, SANS Institute, 2009.

[17] V. O. C. Yaksic, ”A Study on Hash Functions for Cryptography,” Global Information Assurance Certification Paper, SANS Institute, 2003.

[18] R. M. Haralick, K. Shanmugam, and I. Dinstein, ”Textural Features for Image Classification,” IEEE Trans. on Systems, Man and Cybernetics, Vol. SMC-3, 1973, pp. 610-621.

[19] S. Cagnoni, ”Evolutionary Image Analysis and Signal Processing,” Studies in Computational Intelligence (2009), vol. 213, Springer-Verlag, Berlin Heidelberg.

[20] A. W. Zramdini, R. Ingold, ”Optical Font Recognition Using Typographical Features,” IEEE Transaction on Pattern Analysis and Machine Intelligence, vol.20, no.8, pp. 877-882, 1998.

[21] R. Sussex, P. Cubberley. ”The Slavic languages. (Cambridge Language Surveys),” Cambridge, Cambridge University Press, 2006.

[22] R. C. MacCallum, K. F. Widaman, S. Zhang, S. Hong, ”Sample size in factor analysis,” Psychological Methods, vol.4, no.1, 84-99, 1999.

Script Independent Feature Set for Handwritten Text Recognition

Deepti Khanduja
Department of Computer Engineering
MNIT Jaipur, India
Email: deepti.khanduja89@gmail.com

Neeta Nain
Department of Computer Engineering
MNIT Jaipur, India
Email: neetanain@yahoo.com

***Abstract*—The efficiency of any character recognition technique is directly dependent on the accuracy of the generated feature set which could uniquely represent a character and hence correctly recognize it. This paper proposes a hybrid approach combining the structural features of the character and a mathematical model of curve fitting to simulate the best features of a character. As a preprocessing step the character is binarized and transformed to a thinned skeleton and the spurious edges are removed. Then, a combination of structural features of the character like number of end points, loops and intersection points are calculated. Further, the thinned character image is statistically zoned into partitions and quadratic curve fitting model is applied on each partition forming a feature vector of coefficients of the curve. This vector is combined with the spatial distribution of the foreground pixels for each zone and hence script independent feature representation. The approach has been evaluated experimentally on English and Hindi scripts. The algorithm achieves as average recognition accuracy of 89% for any script without incorporating any script specific features.**

I. Introduction

In the field of pattern recognition, character recognition is regarded as one of the most interesting and challenging steps for script processing. Handwritten character recognition (HCR), a tool to recognize and convert scanned text script images into digitized form broadly consists of 3 major stages, i.e. preprocessing, segmentation and recognition. HCR initializes with pre-processing that includes binarization, skew and slant normalization of text document. In further steps preprocessed text image is divided into lines, lines into words and finally each word is partitioned into individual character. In the simplest case this character is fed to the recognition algorithm, however to avoid complexity of the recognition system, a set of features is extracted for each character in order to reduce its dimensionality and this reduced representation of character is fed into the recognition phase to classify the character into its appropriate class. To make this feature set extraction task more effective and refined a new preprocessing step applied on individual character named thinning has been introduced in this paper. Since feature vector is the direct input to the recognition step, accuracy of the system depends on the effectiveness of the feature vector obtained. A number of approaches are introduced earlier to simulate the feature extraction like character moment based [1], direction based[2], neural network based[3] and transform based [4]. In most of the approaches either structural features are estimated or spacial and statistical features are calculated. In this paper we have proposed an approach based on both structure and spacial distribution. The paper comprises of the details of proposed approaches in Section II. Section III defines the neural network

classifier applied. Results and conclusions are appended in Section IV and V respectively.

II. Proposed Approach

A. THINNING- a step further in preprocessing, a refinement

Before applying any feature extraction mechanism, as a preprocessing step, thinning has been introduced. Thinning is an image processing operation that refines a binary valued character image by approximating the center skeleton of the character. This step reduces chaos in the structural elements of the character and retains their generality more effectively. It is usually required that the resulting image after thinning is connected so as to infer the shape and topology of the original image. The algorithm for thinning is summarized in Algorithm 1.

The algorithm for thinning has been tested on several Devanagari and English script characters with the results presented in Table I.

B. Structural Feature Set Extraction

Structural features are used to acquire the topological information of the characters, assuring minimal affect of shape/font/size variation on the feature set. The proposed technique captures the endpoint(EndPts), intersection point(IntersectPts) and presence of hole(HasLoop) in the character image as shown in Fig. 1(a)-(c). EndPts and IntersectPts are composed of the number of pixels with single and three or more neighborhood respectively. The HasLoop feature is true if the character contains one/more closed loop or hole is/are present in the character image. The procedure for computing Structural Features is Procedure 2.

C. Statistical Feature Set Extraction

To derive the enhanced features from the character image that can significantly capture the trend of foreground pixel distribution, as a statistical measure in a processed image, a

Algorithm 1: Algorithm for thinning character image.

Input: 2-Dimensional binary segmented character I.

Output: Thinned character image.

1 Calculate the minimum stroke width *W* of character to be used as threshold of non-spurious edges.

2 Compute skeleton of the input character image using standard mechanism defined in terms of the hit-or-miss transform as described in (1).

$$I \oslash B = I - (I \otimes B) = I \cap (I \otimes B)^c \qquad (1)$$

$$B = \begin{bmatrix} 1 & 1 & 1 \\ 1 & x & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

where B is the structuring element.

3 The resulting character matrix consist of single element edge irregularities leading to distortion and spurious edges. Repeat steps 4-7 to remove such edges.

4 Find the end-point pixels with 1-neighborhood depicted by N_1.

5 Traverse the connecting pixels in clockwise direction using 8-neighborhood until the pixel with 3 or more neighborhood is found depicted by N_3.

6 Calculate the distance *D* between N_1 and N_3.

7 If $D < W$ Remove the edge as it is a spurious edge.

Procedure StructuralFeatures(Image).

1 Compute number of neighbors, *N*, in Image by using 8-connectivity mechanism.

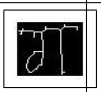
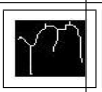
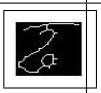
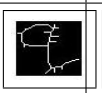
2 For each foreground pixel (i,j) do

$$\text{if} N_{(i,j)} = \begin{cases} 1 & EndPts = EndPts + 1, \\ > 2 & IntersectPts = IntersectPts + 1, \\ else & \text{do nothing.} \end{cases}$$

Set binary variable HasLoop=1 if exists holes exists in the Image using Flood Fill Algorithm [5].

3 **return** *[EndPts IntersectPts HasLoop]*

TABLE I. RESULT OF THINNING ALGORITHM.

Input Character	Skeleton	Thinned Character	Input Character	Skeleton	Thinned Character
					
					
					
					

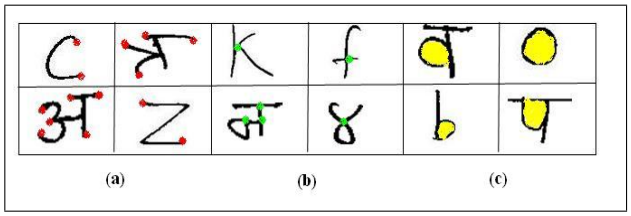


Fig. 1. (a) End points (b) Intersection points (c) Loop.

two way method is applied. First fit a curve on the character and determine the coefficients of the curve equation as features, secondly a count of the number of foreground pixels is used as a direct measure of distribution. To further improve the resolution of process output we have segmented the character image into fixed small blocks of size 5x5. Each such block will derive a curve and count of foreground pixels.

Curve fitting is defined as a way to capture the trend of spatial distribution of pixels by assigning a single function across the entire range possibly subject to constraints. Each constraint can be a point, angle, or curvature. To fit a curve on the specified skeleton pixels we postulate a function form f(x) (2), a polynomial function of 2nd order to describe the general trend in the data. Fig. 2 shows the results of quadratic curve fitting with the co-efficients labeled along.

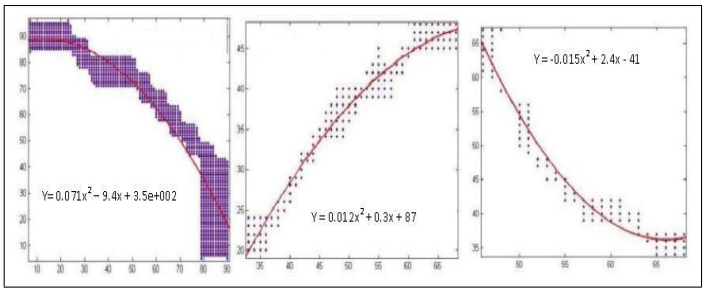


Fig. 2. Quadratic curve fitting along with the resulting polynomial equation.

$$f(x) = ax^2 + bx + c \tag{2}$$

While fitting the curve f(x) over the data pixels there will be inevitable measurement errors. Hence the best we can try to capture the trend of curve is to choose the parameters of the function as such that the curve fits as close as possible to the pixel distribution so as to minimize the fitting error. So ultimate goal of the algorithm is to minimize the Residuals r_i (3), difference between the observed y value and the value corresponding to fitted curve.

$$r_i = y_i - f(x_i) \quad \text{for 1= 1,2,3....n} \tag{3}$$

Equation (4) is used to calculate the standard measured error, i.e. mean square error(MSE), for all points. Hence by an iterative method we can minimize the residual r until it reaches a global minimum by adjusting the parameters a,b,c. The evolution of the normalized mean of the squares of the errors for a sample curve is shown in Fig. 3.

$$\begin{aligned} MSE &= (\frac{1}{\sqrt{n}})||r||_2 \\ &= \sqrt{(\frac{1}{n}\sum_{i=1}^{n}(r_i)^2)} \\ &= \sqrt{(\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2)} \end{aligned} \tag{4}$$

D. Complete Approach

Feature set generation is the method of converting highly redundant, variable and diverse data to a small size, robust,

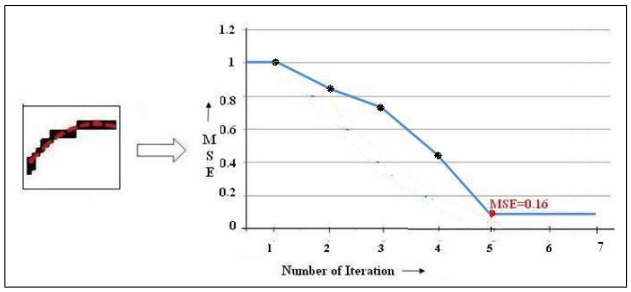


Fig. 3. Evolution of mse for a sample curve

Procedure StaisticalFeatures(Zone$_i$).

1 Find Sum$_i$, (5)the number of dark pixels for Zone$_i$. It constitutes 1 feature for each input zone.

$$Sum_i = Sum_i + Zone_{ij} \qquad \text{for all } \mathbf{Zone_{ij} = 1} \quad (5)$$

2 Using curve fitting mechanism over the skeleton of character to generate a quadratic polynomial QuadPolyCoeff$_i$ using (6).

$$f(x) = a_0 + a_1 x + a_2 x^2. \quad (6)$$

Using a_0, a_1 and a_2 compute standard error(4) and iterate until MSE$_j \approx MSE_{j+1}$.

3 Polynomial generated in previous step gives 3 values of corresponding coefficients $(a_0, a_1$ and $a_2)$ constituting 3 entries in the feature vector.

4 **return** *[Sum$_i$ QuadPolyCoeff$_i$]*

abstract and complete set that conveys all features of the original data. Generally data with greater size increases complexity and becomes time consuming for applications to process. Hence mapping to a corresponding small set of data makes task easier.

In character recognition task extracting those features that are essential to differentiate among characters is a tedious task. A simple method that generates most appropriate and complete set of features is always needed. In this paper we have tried to fabricate a simple mathematical approach Algorithm 2 based on a combination of Statistical and

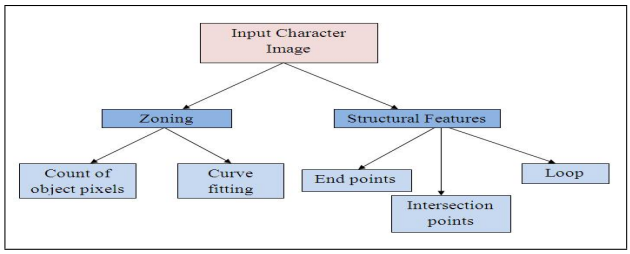


Fig. 4. Block diagram for the proposed Feature Extraction Approach

structural feature extraction techniques. The approach can also be represented by a block diagram as shown in Fig. 4. Fig. 5(a)-(d) presents the results of implementing Algorithm 2 on sample character images.

Algorithm 2: Algorithm to extract feature vector.

Input: 2-Dimensional binary segmented character.

Output: Feature vector consisting of 125 features to be fed to the neural network for further classification.

1 Apply thinning on the input image using Algorithm 1 to deduce a 1-pixel wide character skeleton.

2 Call Procedure StructuralFeatures() to compute structural features of Input Character resulting in a vector of 3 features. Re-size input image Img to dimensions 100x100 and partition Img in 5x5 block using equal partitioning, thus resulting in 25 blocks each of size 20x20.

3 Repeat Procedure StatisticalFeatures() for each zone.

4 Hence a complete feature vector consists of ((1*25) +(3*25)+3) i.e. 103 entries that give a wide range of features to differentiate among characters.

5 Finally this feature vector can be fed as input to the neutral network for training and for testing as well.

III. Neural Network Classifier

The classification stage uses the feature vector generated in the feature extraction stage to identify the text image. Classification is accomplished by a single layer feed forward neural

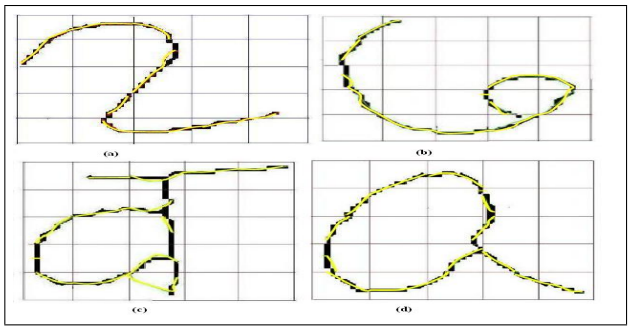


Fig. 5. Implementation of Algorithm 2

network with back propagation learning. The topology of the classifier consists of an input and output layers with neurons determined by the length of feature vector and number of classes respectively. The network employs gradient descendent rule in attempt to minimize the squared error between the network output values and the target values for these outputs. The trained network is used to test a input character image entered by the user and the classifier maps any input pattern to a number of classifications. The probability density function (pdf) of each of the class is generated, then an unknown image, X, belongs to class i if: $f_i(X) > f_j(X)$, $\forall j \neq i$, where f_k is the pdf for class k.

IV. Experimental Results

The proposed feature set generation technique has been trained and tested on $9,500$ scanned and standard data-set samples. The samples comprise of handwritten Hindi numeral, and character open source database *"Devanagari numeral and character database for offline handwritten character recognition"* [6] and a database obtained by CVPR Unit, ISI Kolkata [7]. Also a database of English numeral and characters is generated by scanning handwritten images written by 100 users. The results for testing has been summarized in Table II. Equation 7 and 8 presents the accuracy and error measure of the proposed method which achieves an average accuracy of 89% and false ratio of 10.29%. Also the accuracy of validation i.e. the ratio of correctly classified characters is 86% approximately.

$$Accuracy = \frac{^1TP + ^2TN}{TP + TN + ^3FP + ^4FN} \quad (7)$$

$$Errorratio = \frac{FP + FN}{TP + TN + FP + FN} \quad (8)$$

V. Conclusions

This paper proposes a novel and efficient feature set extraction technique using statistical and structural features of the text image for script independent character recognition. Results generated by the experiments on the test bed images conform to the greater level of efficiency with respect to the size of feature set, time utilized to train the network and its complexity. The approach of Quadratic Polynomial Curve Fitting used for generating feature set has been experimentally derived by testing various feature extraction approaches. For n data points the complexity of the algorithm is $O(kn)$, where k is the number of iterations. As the image is divided into 25 zones, hence the complexity achieves O(25*kn) $\approx$ O(kn). Whereas, the algorithms to compute transform [4] takes O(nlogn) and fuzzy model [8] has exponentially growing computational complexity. As a measure of statistical comparison, proposed approach gives accuracy of 89% in established test bed environment compared to the quoted accuracy of 80% and 83.5% in manuscripts of [2] and [3] respectively. These results uphold the application of the approach to generate feature vector.

References

[1] Dhir, R., *"Moment based invariant feature extraction techniques for bilingual character recognition"*, 2nd International Conference on Education Technology and Computer (ICETC), vol.4, june 2010, pp.80–84.

[2] M. Blumenstein; B. Verma; H. Basli, *"A novel feature extraction technique for the recognition of segmented handwritten characters"*, Proceedings of the Seventh International Conference on Document Analysis and Recognition, (ICDAR 03), 2003, pp.137–141.

[1] TP= True positive

[2] TN= True negative

[3] FP= False positive

[4] FN= False negative

TABLE II. EVALUATION OF FEATURE SET EXTRACTION AND CLASSIFICATION APPROACH.

Database	Number of images	TP	TN	FP	FN	Accuracy(in %) (7)	Error rate(in %)(8)
Hindi Character database [6]	4000	744	2778	367	111	88.05	11.95
Hindi Numeral database [7]	2000	431	1393	135	41	91.2	8.8
English Numeral database	1000	207	732	39	22	93.9	6.1
English Character database	2500	372	1769	502	57	85.64	14.34

[3] Pradeep, J.; Srinivasan, E.; Himavathi, S. *"Neural network based handwritten character recognition system without feature extraction,"* International Conference on Computer, Communication and Electrical Technology (ICCCET), 18-19 March 2011, pp.40–44.

[4] Mowlaei, A.; Faez, K.; Haghighat, A.T., *"Feature extraction with wavelet transform for recognition of isolated handwritten Farsi/Arabic characters and numerals"*, 14th International Conference on Digital Signal Processing, vol.2, 2002, pp.923–926.

[5] Hearn, D. and Baker, P., *"Computer Graphics: C Version"*, Prentice Hall International, isbn: 9780135786345, 1997.

[6] Vikas J Dongre, Vijay H Mankar, *"Development of comprehensive Devanagari numeral and character database for offline handwritten character recognition"*, Journal of Applied Computational Intelligence and Soft Computing (ACISC), 2012.

[7] U. Bhattacharya and B. B. Chaudhuri, *"Handwritten numeral databases of Indian scripts and multistage recognition of mixed numerals"*, IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 31(3), 2009, pp. 444–457.

[8] Hanmandlu, M.; Murthy, O. V Ramana; Madasu, V.K., *"Fuzzy Model Based Recognition of Handwritten Hindi Characters,"* 9th Biennial Conference of the Australian Pattern Recognition Society on Digital Image Computing Techniques and Applications, 3-5 Dec. 2007, pp.454–461.

An Efficient Feature Extraction Method for Segmented Cursive Characters Recognition

Subhash Panwar
Department of Computer Engineering
Malaviya National Institute of Technology
Jaipur, Rajasthan India.
Email: panwar.subhash@gmail.com

Neeta Nain
Department of Computer Engineering
Malaviya National Institute of Technology
Jaipur, Rajasthan India.
Email: neetanain@yahoo.com

***Abstract*—Handwritten document analysis is a persistent research area nowadays. It has large applications in various image processing domain as human computer interaction, machine translation and automation of reading various scanned images for handwritten fields in forms, postal address on envelopes and amounts in banks checks. The main factor which influence the performance of handwritten text recognition is the selection of an appropriate set of features for representing input samples. In this paper, we propose a efficient feature vector for handwritten character recognition, using a hybrid of the statistical and structural properties of a character to represent the particular character class. Experiments have been performed on standard database of handwritten digits and letters. The recognition accuracy is tested on a Neural Network classifier with different parameters. The results have been compared with existing features extraction algorithms. The comparative results shows the effectiveness of our approach.**

***Keywords*-Handwritten text segmentation; Connectivity strength parameter; Offline text Recognition;**

I. Introduction

Handwritten Text Recognition systems transform the handwritten documents into its electronic form for further transformation. Nowadays, although recognition of printed isolated characters is performed with high accuracy, recognition of cursive handwritten characters still remains an open problem in the research area and to increase the character recognition accuracy, novel features extraction techniques are indispensible [19].

Handwritten Text Recognition system involve several steps: preprocessing, segmentation, feature extraction and classification. In this paper, we address the feature extraction step.

Feature extraction step plays one of the most important roles in recognition system. In simplest case, the binary images are fed to a recognizer and a template matching is done. However, in most of the recognition system, in order to avoid extra complexity and to increase the accuracy of algorithms a more compact representation is required. For this purpose, a set of feature is extracted for each class that helps distinguish it from other classes.

So we can say that in this step of recognition, we extract discriminant information from an image of a segmented character and reduce its dimension of representation and produce a vector of feature to represent the character. Such suitable compact representation of the image by means of a vector of features is used for classification to classify the character efficiently.

A good survey on feature extraction methods for character recognition can be found in [2]. In the following, hundreds of documents image features extraction approaches are categorized in three major groups: a) statistical, b) structural and c) transformations based features.

The statistical features include histograms, moments and direction features. The structural features include registration, line element features, Fourier descriptors, topological features, and so on. The transformation methods include principal component analysis (PCA), linear discriminant analysis (LDA), kernel PCA, and so on.

A feature extraction algorithm must be robust enough such that for a variety of instances of the same symbol, similar feature sets are generated, thereby making the subsequent classification task less difficult [2].

The most common statistical features used for character representation are: a) zoning, where the character is divided into several zones and features are extracted from the densities in each zone [7] or from measuring the direction of the contour of the character by computing histograms of chain codes in each zone [10], b) projections [12] and c) crossings and distances [14].

In paper [11], dimensional features are extracted for isolated Greek handwritten character recognition based on efficient feature extraction followed by a suitable feature vector dimensionality reduction scheme. Extracted features are based on (i) horizontal and vertical zones, (ii) the projections of the character profiles, (iii) distances from the character boundaries and (iv) profiles from the character edges. The combination of these types of features leads to a 325-dimensional feature vector.

Structural features are based on topological and geometrical properties of the character, such as loops present in the character image, maxima and minima, reference lines, ascenders, descenders, cusps above and below a threshold, cross points, branch points, strokes and their directions, inflection between two points, horizontal curves at top or bottom, etc .

The structural approach for recognizing Greek handwritten characters is introduced in [15].They extracted a 280-dimensional vector consisting of histograms and profiles. The horizontal and vertical histograms are used

in combination with the radial histogram, out-in radial and in-out radial profiles.

Due to the fact that a continuous signal contains more information than needed representing this signal by a linear combination of a series of simpler well defined functions is easier. The coefficients of the linear combination provide a compact encoding known as transformation. Fourier Transforms (FT) is most popular among global transformation. Since the first n coefficients of the FT can be used in order to reconstruct the image [16], then these n coefficients are considered to be a n-dimensional feature vector that represents the character. Other transformations used are the Discrete Cosine Transform (DCT) [17], moments [18] etc.

Cursive characters present high variability in shapes, a feature extractor should have negligible sensitivity to local shifts and distortions. Therefore feature extractors that perform local averaging are more appropriate than others that yield an exact reconstruction of the pattern (e.g. Zernike polynomials, moments). The feature extractor, fed with the binary image of an isolated cursive character, generates local and global features.

The remaining of the paper is organized as follows. In Section II the proposed methodology is presented as preprocessing and feature extraction. Section III describes the configuration of the experimental data set and the classifier parameters. Experimental results are discussed in Section IV and finally conclusions are drawn in Section V.

II. Proposed Methodology

For uniformity of feature representation we first normalize all binary characters images to a $N \times N$ matrix with respect to the original aspect ratio. As we are interested in the relative positions of the strokes in the pattern but not in the size of the pattern we transform the character image into its skeleton using thinning operation. A skeleton shows the general shape of a pattern, and some important feature can be extracted from it. There are many approaches to find the skeleton of an object [3].

After finding the skeleton we divide the image into 5×5 zones as shown in Figure 4.

A. Feature Extraction

In our proposed approach, we employ three types of features. The first and second set of features are local features and are based on zoning. The image is divided into horizontal and vertical zones. In the first set of features we calculate the area of characters in each zone.

In second type of features, the curve fitting, also known as regression analysis is used to find the best fit curves for series of foreground pixels in each zone. We use the least square method of curve fitting which minimizes the square of the error between original data and the values predicted by the equation. It is a statistical method of fitting a function to a data set. There are five types of least squares fitting approaches such as Linear, Polynomial, Exponential, Logarithmic and Power. In term of

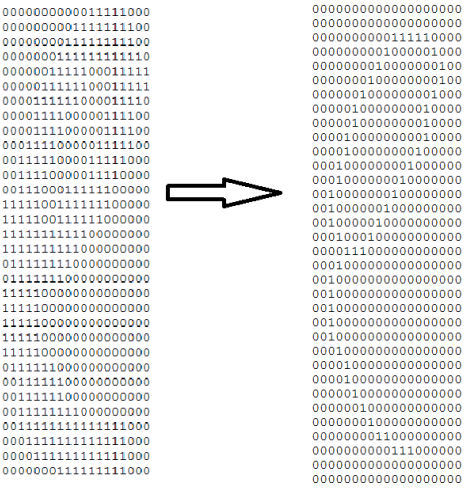


Figure 1. (a) Example matrix format of a handwritten character, (b) The corresponding skeleton matrix of the character.

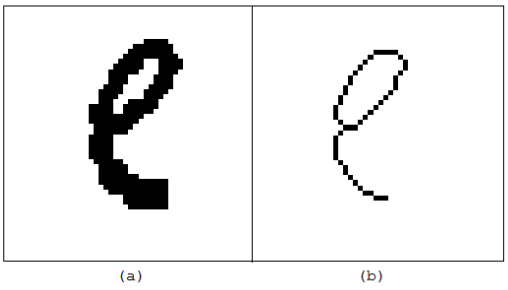


Figure 2. (a) Image of the handwritten character shown in 1, (b) The corresponding skeleton image.

required computing complexity and appearance of curve in character images , we use the second order polynomial function for best fitting. Quadratic polynomial functions are a reasonable choice owing to their low degree of complexity and enough flexibility to encode free hand curve in a small area (zone). The parametric representation of a polynomial of second order is given as

$$P(t) = c_0 + c_1 t + c_2 t^2$$

where $t \in [0, 1]$, is the parameter and the polynomial $P(t)$ goes through the points which satisfies $P(t)$.

Thus, compute $P(x_i, y_i)$ for $i = 1, 2, 3, ..., n$

$$c_0 + c_1 x_i + c_2 {x_i}^2$$

where coefficients c_0, c_1 and c_2 are the n unknowns, while the x_i and y_i are known as foreground pixels in the respective zone.

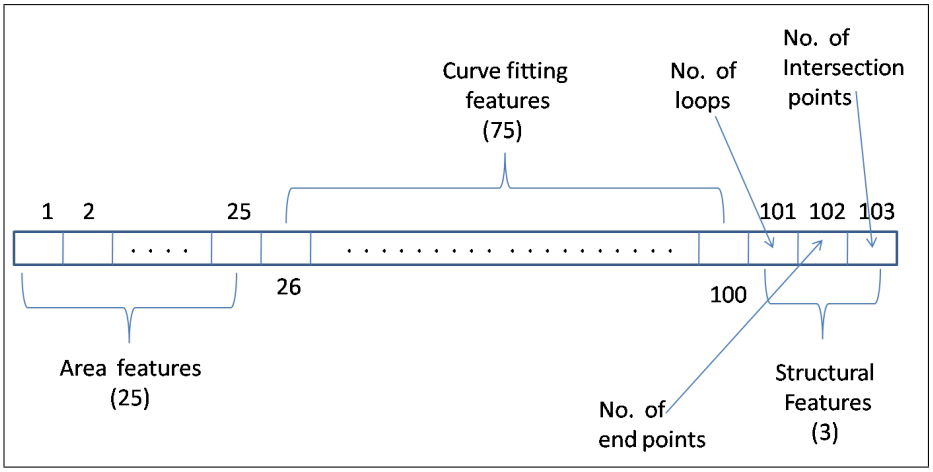


Figure 3. Structure of the proposed feature vector for character representation. It is a hybrid of features like: statistical (area), coefficients of a polynomial, and structural features like end points, intersections and loops.

$$\begin{bmatrix} 1 & x_1 & {x_1}^2 \\ 1 & x_2 & {x_2}^2 \\ 1 & x_3 & {x_3}^2 \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} P_1 \\ P_2 \\ P_3 \end{bmatrix}$$

We find the polynomial coefficients c_0, c_1, and c_2 and use them as features. Thus for each zone, three features are generated.

The third feature set is a global feature set and uses the structural information of characters like number of loops, end points and intersection points present in the character as shown in Figure 4.

The extraction of such global and structural features, is done by testing a 3x3 neighborhood of every foreground pixel of the skeleton. The following cases are determined.

- An end point is identified if there is only one foreground pixel in a 3×3 neighbourhood.
- An intersection point has more than two neighbours.
- Loop detection is based on the presence of circular or elliptical shapes within the character using the flood fill algorithm [6].

The complete features of our technique are shown as feature vector (FV) in Figure 3. Corresponding to (every character) 5×5 zones we have 25 area features listed at index $FV[1]$ to $FV[25]$. Every zone has three coefficients of the parametric curve, hence 75 quadratic curve coefficients listed at index $FV[26]$ to $FV[100]$, followed by the three structural features: end points, intersection point and loop; at $FV[101]$to $FV[103]$.

III. Experimental Design

For performance testing of our approach we have focused on universality and robustness. We attempt to be objective by using a database which covers maximum syntactic variations and fluctuations in cursive handwriting. For this, the test set is designed by people from diverse academic backgrounds and from different geographical areas.

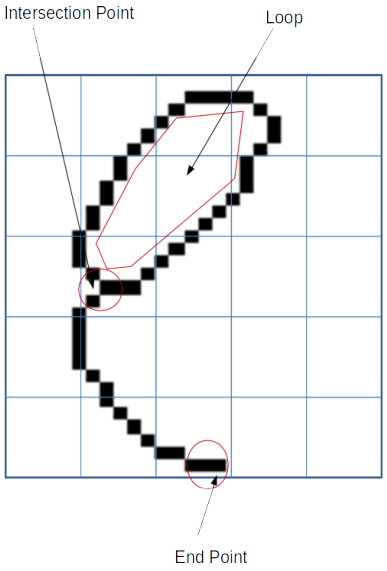


Figure 4. The example showing the three structural features of a character.

For robustness of our classifier extensive tests are conducted with different network training parameters like: varying the range of instances with different number of epochs and activation functions ranges with various thresholds of Mean square error range.

A. Organization of the database

As the cursive handwritten data set is not available, the database used for our experiments was generated with 100 writers from different academic background (primary school to UG college) and from different geographical regions (rural to urban). We have used the 10 digits, 26 lowercase and 26 uppercase English cursive characters. This concluded in 62 classes each one containing 100 patterns. The database has $62 \times 100 = 6200$ characters. Complete bifurcation of dataset in terms of number of

Table I
DATABASE USED IN OUR EXPERIMENTS.

	Number of classes	Training Size	Test size
Numerals	10	700	300
English capital letters	26	1820	780
English small letters	26	1820	780

training and testing instances is shown in Table I.

B. Configuration of NN classifier

The selection of a suitable classifier is an important part of any recognition system, because it highly affects the accuracy of results and speed. A well known classifier, feed forward neural network is designed for testing our proposed approach. The neural network was implemented using well known feed forward training algorithm with the mean square error function.

A constant learning rate 0.2, is used. The input samples are fed after shuffling their order at every epoch. The number of hidden layers is two times the number of output nodes. After initiating the training, we monitor the mean square error MSE at the output layer for every instance in order to determine the termination of training. The initial values of epochs taken is 500. The termination of training is done if the MSE is less than a MSE threshold or when the number of epochs exceeds the initial value of epochs.

Determination of the initial value of epochs and MSE threshold is not easy because the MSE changes among the character set. We setup their values so that the actual recognition rate for the training set reaches a stable point with no further increase.

Three separate neural networks were trained for digits, lowercase and uppercase characters. The number of outputs considered for digit was 10 and for both upper and lower case characters was 26 each. Moreover, 70% of each class was used for training and 30% for testing for each network.

IV. RESULTS ANALYSIS AND COMPARISON

For performance testing of the proposed approach, we have used the predictive analytically parameters, which is a table of confusion also known as confusion matrix, having two rows and two columns that reports the number of false positives (FP), false negatives (FN), true positives (TP), and true negatives (TN) present in the experiment. Total size of dataset will be the sum of all as

$$Total = \{TP + TN + FP + FN\}$$

Such parameters allows more detailed analysis than mere proportion of correct accuracy and error ratio. The accuracy rate (AR) and error ratio (ER) is calculated as

$$AR = \frac{TP + TN}{Total}$$

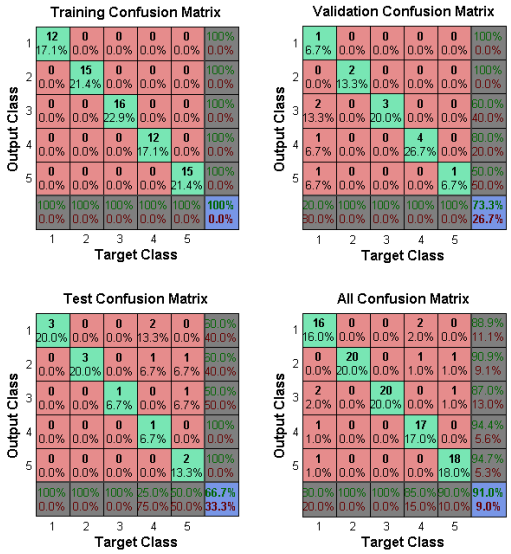


Figure 5. Confusion matrix for numerals.

$$ER = \frac{FP + FN}{Total}$$

For illustration of confusion matrices for digits, we have used five classes of digits which were selected randomly. The confusion matrix in various phases are shown in Figure 5. The performance in different phase of classifier as training, validation and testing is also depicted in Figure 6 for the same. The experimental results for 10 digit class achieved an overall accuracy rate AR and error ratio ER as 93.9% and 6.1% respectively as shown in Table II .

For illustration of confusion matrices for characters, we have used ten classes of characters five from lower case and five for upper case, which were selected randomly. The confusion matrices in various phases are shown in Figure 7. The performance at different stages of classifier like training, validation and testing is also depicted in Figure 8 for the same.

The complement experiments for 26 uppercase characters achieved an accuracy rate AR of 86.19% and error ratio as 13.81% and for 26 lower case characters we achieved an accuracy rate AR of 89.61% and error ratio as 10.39%.)
The complete experimental results with all three datasets is shown in Table II.

A. Results Comparison

It is always a difficult task to compare results for handwritten character recognition with other researchers in the literature. The results vary due to differences in experimental methodology, different experimental settings and difference in the handwriting database used.

In [8], they used the standard transition feature, as proposed by author of [9] and produced results of 70.31%

Table II
RECOGNITION RATE BASED ON NEURAL NETWORK.

Data	Size	True Positive (TP)	False Positive (FP)	True Negative (TN)	False Negative (FN)	Accuracy in %	Error Ratio
Numerals	1000	207	39	732	22	93.9	6.1%
English capital letters	2600	472	307	1769	52	86.19	13.81%
English small letters	2600	543	322	1787	48	89.61	10.39%

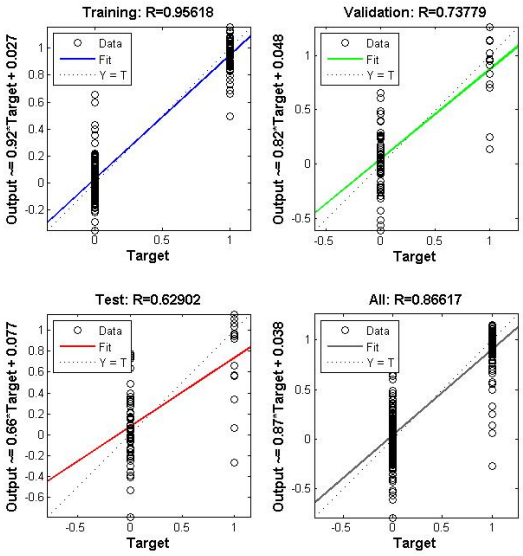


Figure 6. Regression curve for numerals.

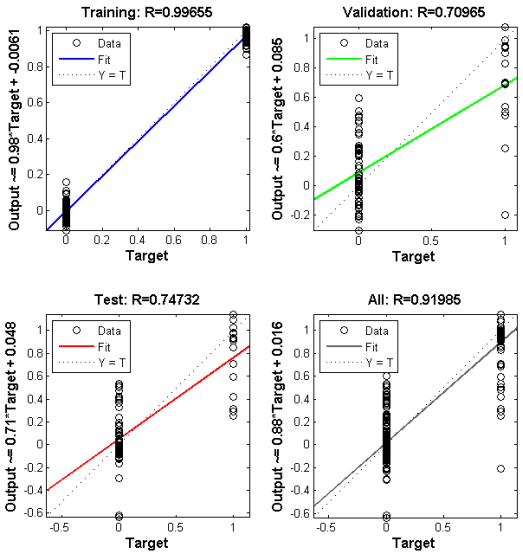


Figure 8. Regression curve of five class of characters.

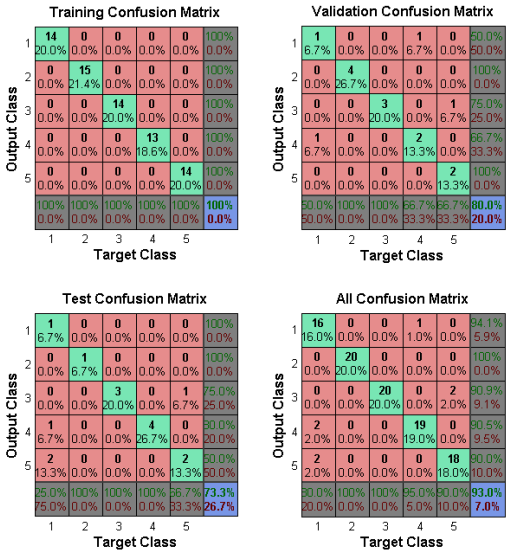


Figure 7. Confusion matrix of five class of characters.

and 79.23% for lowercase and uppercase characters respectively.

A complete state of art and critical analysis of neural network for document image is discussed in [19]. In this literature, reviewers emphasis that all existing preprocessing techniques have some inherent problems. They also provide the comparison of existing feature extraction techniques. All are only suitable for specific predefined condition and for the cursive characters the low accuracy with high computation demand was reviewed.

Additionally, some techniques need training/training data and in most of the cases typical training data is unavailable. Our experiment results are higher in accuracy than those described in above and It also must be noted that our technique requires less processing as the feature vector is a combination of statistical and structural features so it captures both kind of variations. Also a quadratic polynomial is used for encoding the character which is efficient compared to template matching, zoning and direction features as done in various techniques enumerated in [5].

V. CONCLUSIONS

In this paper, we proposed a feature extraction methodology for offline cursive character recognition using both statistical and structural features. This improve the recognition accuracy due to use extensive features and improves the efficiency due to novel curve fitting encoding of character features.

REFERENCES

[1] A. S. Britto, R. Sabourin, F. Bortolozzi, C. Y. Suen,Foreground and Background Information in an HMM-Based Method for Recognition of Isolated Characters and Numeral Strings, 9th International Workshop on Frontiers in Handwriting Recognition (IWFHR-9), 2004, pp. 371-376.

[2] O. D. Trier, A. K. Jain, T.Taxt , Features Extraction Methods for Character Recognition A Survey, Pattern Recognition, 1996, Vol.29, No.4, pp. 641-662.

[3] L. lam, S. W. Lee, C.Y. Suen, Thinning Methodologies-A Comprehensive Survey, IEEE Trans. Pattern Recognition and Machine Intelligence, Vol.14, pp.869-885,1992.

[4] Camastra, F.; Spinetti, M.; Vinciarelli, A., "Offline Cursive Character Challenge: a New Benchmark for Machine Learning and Pattern Recognition Algorithms.",Pattern Recognition, 2006. ICPR 2006. 18th International Conference on , vol.2, no., pp.913,916, 0-0 0 doi: 10.1109/ICPR.2006.895

[5] N. Arica and F. Yarman-Vural, An Overview of Character Recognition Focused on Off-line Handwriting, IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 2001, 31(2), pp. 216 - 233.

[6] Hearn, D. and Baker, P.,"Computer Graphics: C Version", Prentice Hall International, isbn: 9780135786345, 1997.

[7] Luiz S. Oliveira, F. Bortolozzi, C.Y.Suen, Automatic Recognition of Handwritten Numerical Strings: A Recognition and Verification Strategy, IEEE Transactions on Pattern Recognition and Machine Intelligence, 2001, Vol. 24, No. 11, pp. 1448-1456.

[8] M. Blumenstein, B. K. Verma and H. Basli, "A Novel Feature Extraction Technique for the Recognition of Segmented Handwritten Characters", 7 th International Conference on Document Analysis and Recognition (ICDAR 03), Edinburgh, Scotland, 2003, pp. 137-141.

[9] P. D. Gader, M. Mohamed and J-H. Chiang, "Handwritten Word Recognition with Character and Inter-Character Neural Networks",IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics, vol. 27, 1997, pp. 158-164.

[10] K. M. Mohiuddin and J. Mao, A Comprehensive Study of Different Classifiers for Handprinted Character Recognition. Pattern Recognition, Practice IV, 1994, pp. 437- 448.

[11] Vamvakas, G.; Gatos, B.; Petridis, S.; Stamatopoulos, N., "An Efficient Feature Extraction and Dimensionality Reduction Scheme for Isolated Greek Handwritten Character Recognition," Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on , vol.2, no., pp.1073,1077, 23-26 Sept. 2007

[12] A. L. Koerich, Unconstrained Handwritten Character Recognition Using Different Classification Strategies. International Workshop on Artificial Neural Networks in Pattern Recognition (ANNPR), 2003.

[13] K. Woods, K.W. Bowyer, "Generating ROC curves for artificial neural networks",IEEE Transactions on Medical Imaging, 16 (3) (1997), pp. 329337

[14] J. H. Kim, K. K. Kim, C. Y. Suen, Hybrid Schemes Of Homogeneous and Heterogeneous Classifiers for Cursive Word Recognition, 7 th International Workshop on Frontiers in Handwriting Recognition, Amsterdam, 2000, pp 433 - 442.

[15] E. Kavallieratou, N. Fotakis, G Kokkinakis, Handwritten Character Recognition Based on Structural Characteristics, ICPR, 16th International Conference on Pattern Recognition (ICPR'02), 2002, pp. 139-142.

[16] Kuhl, F.P., Giardina, Elliptic Fourier features of a closed contour, Comput. Vis. Graphics Image Process. 18, 1982, pp 236-258.

[17] R. C. Gonzalez, R. E. Woods, Digital Image Processing, Second Edition, Prentice Hall, 2002.

[18] Hu, Visual pattern recognition by moment invariants. In IRE Trans. Inf. Theory 8, 1962, pp. 179-187.

[19] Amjad Rehman, Tanzila Saba, "Neural networks for document image preprocessing: state of the art", Artif Intell Rev DOI 10.1007/s10462-012-9337-z, Springer, Science + Business Media B.V., 2012

[20] Wilkinson, J. Geist, S. Janet, P. Grother, C. Burges, R. Creecy, B. Hammond, J.Hull, N. Larsen, T. Vogl, and C. Wilson. The first census optical character recognition systems conf. NISTIR 4912, The U.S. Bureau of Census and the National Institute of Standards and Technology, Gaithersburg, MD, 1992.

[21] Hull, A database for handwritten text recognition research, IEEE Trans. on Pattern Analysis and Machine Intelligence, 1994, Volume 16, Issue 5, pp. 550 554.

[22] C. Y. Suen, C. Nadal, R. Legault, T. Mai, and L. Lam , Computer recognition of unconstrained handwritten numerals, Proc. of the IEEE, 1992, Volume 7, Issue 80,pp. 1162-1180.

[23] I. Guyon, L. Schomaker, R. Plamondon, M. Liberman, and S. Janet, Unipen project of on-line data exchange and benchmarks, Proc. of the 12th IAPR Int. Conf on Pattern Recognition, Jerusalem, Israel, Oct. 1994, pp. 29-33.

[24] E.Kavallieratou, N.Liolios, E.Koutsogeorgos, N.Fakotakis, G.Kokkinakis, The GRUHD database of Modern Greek Unconstrained Handwriting, In Proc.ICDAR, 2001.

An Annotated Urdu Corpus of Handwritten Text Image and Benchmarking of Corpus

Prakash Choudhary* and Neeta Nain**

* National Institute of Technology Manipur/Computer Science and Engineering, Imphal, India
** Malaviya National Institute of Technology Jaipur/Computer Engineering, Jaipur, India
choudharyprakash87@gmail.com

Abstract - For linguistics related research on a language there is always a need for a large collection of database which includes all features of a language such as grammatical information, style of writing, syntax etc. Corpus provides a platform for investigation on a natural language. As compared to other languages very limited research work is done on Urdu language due to its segmentation dilemma and difficult character shape. Very less number of editable printed text data is available in Urdu language, most of the data is available in graphical or picture format. To increase Natural Language Processing research work on Urdu language there is a need for a large database which contains a range of variance in annotated Urdu handwritten as well as printed text. In our work we purpose a large database of Urdu text including 1000 handwritten text images written by 500 different writers. Each image would be four to six lines of Urdu text having 60 - 80 words per line the estimated number of words would be around .35 million. Selection of words would be done from six different categories so that maximum number of distinct words can be included. Corpus would be annotated for line as well as word segmentation where a word may be an individual character or component. The corpus would be a benchmark for quantitative analysis of Handwritten Text Recognition techniques for Urdu language such as text line extraction, word segmentation and character recognition etc., and for linguistic research in Part of Speech, writer identification, dictionary etc.

Keywords:
Urdu language; Corpus annotation; Urdu corpus; Corpus creation; Urdu handwritten image database

I. INTRODUCTION

In the recent era corpus linguistics has a big role for research and development activities on any particular language. Corpus is a large collection of text data, which is stored in a scientific and systematic manner to represent a hypothesis about a language. This provides a platform for detail investigation on a language. In short form corpus is a big collection of smaller text, where Genre of text may be in printed text, handwritten text or spoken text. According to nature of corpus it is mainly divided into two parts raw corpus and annotated corpus. Raw corpus is just collection of text without any description or labelling of text. Annotated corpus is annotation of text in a systematic and scientific manner according to use of corpus.

Annotation of corpus is done according to the Corpus Encoding Standard (CES). The Layers of Linguistics Annotation is as follows:

- Syntax: - Combination of words to make sentence.
- Semantics: - Relation between words and what they are being used to represent.
- Morphology: - The study of units of meanings in a language.
- Lexicon: - The study of words and phrases used in a language.
- Discourse Analysis: - The study of the flow of information across sentence boundaries.
- Pragmatics: - The study of the context of text affects the meaning of expression.
- Text structures Analysis: - The studies of textual styles are constructed.

Corpus linguistics has good potentials to exhibit most of the features of a natural language with the large collection of empirical databases. Corpus linguistics approach has been used in a number of research areas such as:

- To empirical analysis and study about language based on large collections of text samples.
- Information about all the properties and components used in a language such as grammatical and functional information of letters.
- To build up an intelligent computer system that will be able to make efficient linguistic interaction with human.
- It is a way of collecting literary and non-literary data of a language for language related research and development activities.
- Extra linguistic information relating to time, place, situation, and social cultural backgrounds of linguistic.

According to use of corpus there are different types of corpus design annotations.

- General Corpus: General Corpus (e.g., LOB Corpus) is a simple random collection of texts

belonging to different categories. Selection of text is finite in number, but according to availability of new data it will grow over time.

- Special Corpus: Special Corpus is designed according to certain properties of investigation or demand of researchers. It varies in size and composition according to purpose. Collection of data for this type of corpus is mainly dependent upon the research requirements (e.g., literary Corpus: used for survey on literature collection of data available according to time span).
- Sample Corpus: Sample Corpus is a special type of corpus made with a sample containing sufficient number of text. Selection of text is made with utmost care because once a corpus is developed it will not change later to prevent any kind of imbalance.
- Multilingual Corpus: Corpus in a variety of languages.
- Parallel Corpus: texts are translations of each other (e.g., Hindi to Urdu).

Corpus linguistics use modern computer technology to store big collection of language data, and methods for pre-processing and information retrieval from this database.

As compared to other languages very limited research work has been done in Urdu due its word and character segmentation dilemma and the shape of the character. In our proposed work we design a standard annotated Urdu corpus of handwritten text images, which contains a dataset of 1000 forms of handwritten text image written by 500 writers belonging to distributed geographical region. For ground truth purposed corpus would be labelled in Unicode text file for both line and word according to previously described layers of linguistics annotation.

The paper is organized as: in the next Section 2, we describe literature survey and level of annotation on available corpus. Section 3 is concerned with the motivation and importance of Urdu language and proposed methodology to develop corpus. Section 4 further describes the characteristics of annotation and procedure for data storage and information retrieval. Finally some conclusions are presented in Section 5.

II. LITERATURE SURVEY

Since the 1990s, the corpus methodology has become a revolutionized in all branches of linguistics or language.

[1] Johansson, S., Atwell, E., Garside, R., and Leech, developed Lancaster Oslo/Bergen Corpus (LOB) is the first computer-readable general corpus of texts prepared for linguistic research on modern English. The collection of text in this corpus is in printed (Electronic) form. First version of this corpus was without annotation in which they include raw data as a plain text without labelling of text, but later version of corpus was released in annotated form. The structure of first version corpus released is as follows:

- 500 samples of English language
- 11827 sentences
- 134740 words

[2] Christian, P. Michel, Stefan and Philippe, IRESTE dual handwritten database of French and English. Instead of full length sentences corpus contain isolated words and characters. Corpus was just collection of text without any labelling of words. Corpus was designed for both off-line as well as on-line handwritten recognition. Structure of this multilingual corpus is as follows:

- 28657 words from French language
- 7689 words from English language

U. Marti and H. Bunke., [3, 4] Designed an IAM off-line handwritten database which is first available corpus consisting of full English sentences. The structure of the database is available in both printing and handwritten image format as well as in Unicode format. The segmentation and annotation was done for both line and words according to annotation layers. Due to proper segmentation and annotation the database can be used for English language linguistics research purpose and also for retrieval and information detection. Structure of database is as follows:

- 250 handwritten forms by different writers
- 4751 words
- 4881 lines
- 6625 different word

Lecun et al., MNIST [5] is a database of handwritten digits (Image) in which 60,000 samples of isolated digits handwritten image, is collected from 500 different writers.

M. Waqas, C. Lei, N. Nobile, and C. Y. Suen, (2009) design [6] CENPARMI is a first Urdu handwritten corpus which includes isolated digits and characters of Urdu language. Corpus is available in scanned image format without any annotation. Structure of corpus is as follows:

- 343 handwritten form with different writer
- 44 isolated characters
- 57 isolated words
- Collection of date in different Urdu format

Sarkar, N. Das, S. Basu, M. Kundu, M. Nasipuri and D. Basu, [7] CMATER is a database of unconstrained handwritten Bangla and English mixed script document image. Corpus is available in scanned handwritten image format without any annotation. Structure of the corpus is as follows:

- 150 handwritten form with different writer
- 100 form is pure Bangla language
- 50 forms is mixed of both Bangla and English

Hussain, S., (2008) CRULP [8] corpus is a large collection of Urdu text in Unicode format from Jang news and BBC Urdu services. The corpus has not been made publicly available. In this corpus manually POS tagging is done. The corpus is available in printed not handwritten form. Structure of the corpus is as follows:

- 1,83,08,616 words
- 50635 distinct words
- 1,10,098 annotated words

A. Raza, I. Siddiqi, A. Abidi, and F. Arif, (2012) [9] CENIP-UCCP is the only available corpus of Urdu Handwritten text image with full sentences. The database is labelled with coordinates of each of text for both printed and handwritten text. Manual segmentation is done in this corpus for lines. Data base could be used for handwriting task, but for benchmarking purpose corpus should be annotated for line and words. It was not considered for description of words such as grammatical and segmentation information. Structure of corpus is as follows:

- 396 handwritten forms with 200 different writers
- 2051 lines with 23833 words

Examples of some other widely used databases in the field of handwriting recognition are CEDAR [10], NIST [11], ETL9 (Japan) [12], and PE92 (Korea) [13].

From literature survey it was found that sufficient number of corpus is available for English language in printed and handwritten format. The researcher's community seems anxious to move forward fast in research of South Asian languages and mostly in Urdu language [14], but cannot because a standard corpora is required, which include big collection of linguistics information about a language.

III. MOTIVATION

Urdu is historically associated with the Muslim region of Hindustan. It is the national language of Pakistan, and one of the 22 scheduled languages in the Constitution of India. Also an official language of five Indian states: Uttar Pradesh, Bihar, Andhra Pradesh, Jammu & Kashmir and New Delhi. The population of Hindi-Urdu speakers is the fourth largest of the languages of the world, after Mandarin Chinese, English and Spanish. According to Govt. of India [15] census 2001 around more than 50 million person speaks Urdu in India as shown in Table 1.

Language	Persons who returned the language as their mother tongue			
	1971	1981	1991	2001
Urdu	28,620,895	34,941,435	43,406,932	51,536,111

TABLE 1: URDU SPEAKING POPULATION IN INDIA

Although Persian was the dominant and "official" language of the Mughal Empire, the language of the elite later evolved into a form known as Urdu. History of Mughal Empire (Shahan-e-Mogul) is written in Urdu. Urdu is written in Nastaleeq style using Arabic Script. In recent time the Unicode provides almost complete support for Urdu. Urdu uses the 0600–06FF Unicode range for its characters.

Application of Urdu corpus:

- Corpus is always a strong cognitive and linguistic motivation to envisage how we communicate through language across time and space.
- To build up an intelligent computer system that will be able to make efficient linguistic interaction with humans who know Urdu very well.
- Used to develop systems that contribute for the benefit and advancement of the whole mankind. Such as Machine Translation, Information Extraction, Language Understanding, Speech Understanding, Computer-Aided language teaching, etc.
- To understand empirically the natural languages adorned with all their regular and rare features (e.g., Urdu NLP research).

1. Proposed corpus structure

Our Proposed methodology to design and develop Urdu corpus consists of full Urdu text sentences.

Statistics of the Urdu corpus is as follows:

- Database will consists of about 1000 handwritten text forms. Each form having around 5 to 6 lines and 60 to 80 words.
- To captures the maximum syntactic variations, forms will be filled up by 500 different writers having diverse background and belonging to different geographical locations.
- Number of words in this corpus will be around .35 million.
- In the first phase of corpus development we will collect data in Unicode format as a plain text from various sources.
- Unfortunately there are very few repositories of Urdu text in Unicode format. Most of Urdu source text is present text in graphical form. To cover maximum number of words we will convert graphical image of text to plain text.
- Urdu language has a huge vocabulary because it is used from the time of the Mughal Empire (early 16th century) in India. According to time span each language is constantly evolving by adding some new words instead of existing words. The conversion of the graphical Urdu text to Unicode will make corpus sufficient in quantity term with large vocabulary.
- Database will be completely markup according to hierarchical level of indexing. Markup is done on the basis of multilevel indexing. It will start from handwritten text form, than for segmented lines where line is again segmented into words.
- Proposed corpus would provide facilities to further add more words by same procedure of markup where all annotation information will be entered manually during the insertion of new handwritten text form.

2. Category wise distribution of proposed data collection

To cater to a huge vocabulary the domain of the corpus would be a data collection of six different categories: Art and Literature, History, Science, News, Medicine and Sports. List of category along with some of their subcategory is as follows:

- Art and Literature (Poetry, Shyari, Ghazals)
- History (According to time span) and Architecture
- Science and Technology (Innovations and practices)
- News (News, articles and blogs)
- Medicine (Disease, medicine, procedures)
- Sports (Sports articles, Information about different sports)

The corpus development would start from the collection of raw text and end with the appropriate tagging or labelling of the collected text in the database. Figure 1 shows the complete process in step by step manner to develop the corpus.

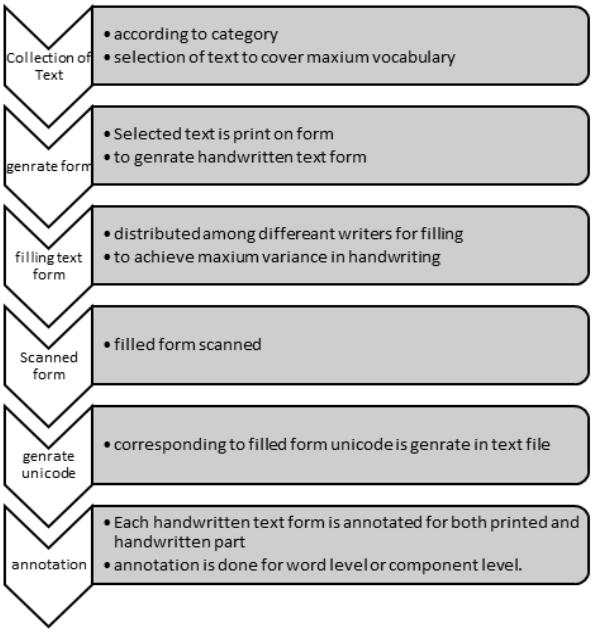


Figure 1. Process of corpus development

IV. OUR CONTRIBUTION

For collection of handwritten Urdu text a form is generated as shown in Figure 2. The design and structure of the form is split into four parts, each part is separated from each other with a horizontal line.

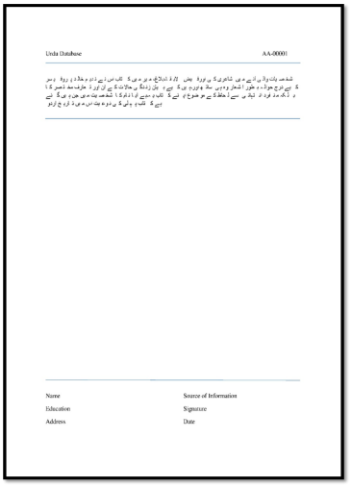


Figure 2. Layout of handwritten text form

Part 1: The first part comprises the title "Urdu Dataset" and a unique identification number of form. For example AA-00001, in which first letter A denotes the category of the form. Each category assigns a particular letter and combination of the first two letters denotes the subcategory (if A denotes category Art and Literature than AA denotes subcategory Poetry and AB denotes Shyari etc.) and 00001 specifies unique identification of form with its category and form number.

Part 2: Second part of the form consists of 4 to 6 lines of printed text which is collected from various sources. Each line having number of words around 60 to 70.

Part 3: Third part of the form is left blank where the writers replicate the printed text in his/her natural handwriting.

Part 4: Fourth part of the form having six attributes: Name, Education, Address, Source of Information, Signature and date of form filling, writers can optionally provide this information.

Sample Corpus Design

The filled form is to be scanned at the resolution of 300 dpi at a grey-level resolution of 8 bits. The images are saved in TIFF-format with LZW compression. Each form is completely scanned, including both the printed and handwritten text. In addition to quantitative tests for image processing and segmentation operations, it can also be used for experiments on machine-printed text and handwritten text separation.

Annotation (labeling) of text in a corpus is prerequisite for any corpus development process, because this task is time consuming and error prone it requires utmost care. Apart from pure text annotation corpus provide some additional linguistics information about data. To create the annotation and markup of handwritten image, the text of a form is copied twice into Unicode, once for the machine-printed text and once for the handwritten text.



Figure 3.Home page of corpus

Figure 3 shows the home page of the proposed corpus. The homepage gives the details of corpus and provide option for access and to update corpus with new data. Handwritten image are distributed according to category, on clicking a category its shows its subcategory and the list of unique-id number of forms stores in this category.



Figure 4. Display of form with annotated information

Along with annotation information the form also shows some additional information as shown in Figure 4. The detail description of handwritten text form contains:

- Unicode corresponding to handwritten text.
- Category and subcategory of form.
- Skew of page.
- Number of handwritten line in the form.
- Each handwritten line is segmented and assigned a unique-id with their Unicode text.
- Name of writer.

A sample display of segmented line and annotated information of line is shown in figure 5. This shows the segmented line separately which is segmented from handwritten text form. The Unicode of the segmented line corresponding to handwritten text line. Each line has a unique id which is extension of form id for example if form id is AA-00001 than it is extended to AA-00001-00 for first line and AA-00001-01 for subsequent line.

For evaluation of handwriting recognition systems, the ground truth data corresponding to the actual transcription of each line of text is stored in Unicode format.

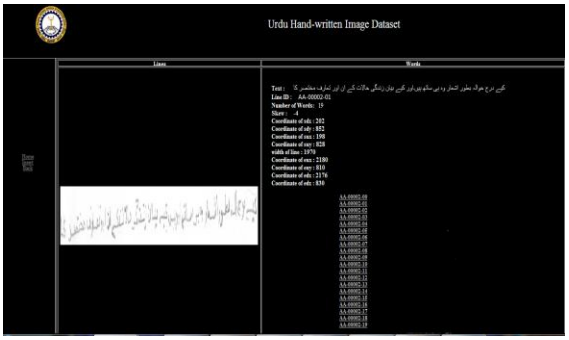


Figure 5. Display of line with its annotated information

Each segmented line of handwritten text form is further mark-up with the detail description which contains:

- Unicode of segmented text line corresponding to handwritten text line.
- Unique identification number of line.
- The x and y coordinates with respect to bounding extent of the line.
- Starting top and bottom pixel.
- Ending top and bottom pixel.
- width and height of the line.
- Skew of the line.
- Number of words in line.

The ground truth data of segmented line with detail description is stored in Unicode format will be useful for benchmarking of handwritten text recognition algorithms.

The segmented line is again segment into words. Each of the segmented word having unique id number:

- Each word has unique id which is extension of line id for example if line id is AA-00001-00 than it is extend to AA-00001-00-00 for first word and AA-00001-00-01 for subsequent word.
- Each word in handwritten text line is segmented and annotated separately.

The segmented word is further mark-up with the information as shown in figure 6 which includes the detailed descriptions:

- Unicode of segmented word or component.
- The x and y coordinates extents with respect to each component.
- width and height of word.
- Skew of word.
- Each word of line is annotated with noun, pronoun, adjective, verb, preposition etc.

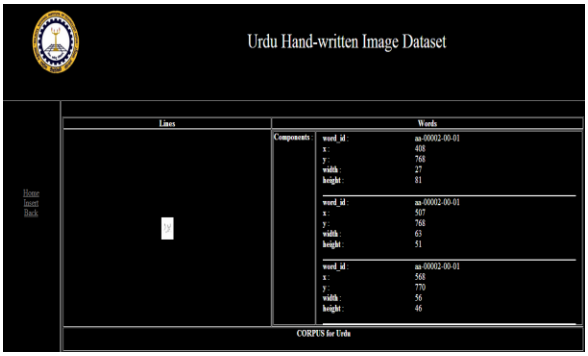


Figure 6. Display of Word with its Annotated information

To retrieve list of words according to alphabetical order we annotated each word's first character with a label according to Urdu alphabet. This will be useful to generate ligature and dictionary searching of Urdu words. Since we are also storing the writer's identity, it could also be used for hand-writer identification detection techniques.

The corpus provides facilities such as benchmarking of handwritten text recognition algorithm. To maintain accuracy in benchmarking, the corpus would be annotated and segmented manually. For benchmarking of handwritten algorithm all the annotated information of line as well as word is stored in Unicode format. These Unicode will be match with algorithm output to check the accuracy.

V. CONCLUSIONS

The proposed corpus includes a very large scale of database with distinct number of words belonging to six different categories to maintain equality and quantity features in the corpus. Annotation of corpus is done for both printed and handwritten format. Database would be useful for linguistics related research on Urdu language. Corpus consists around 1000 Urdu handwritten forms of 500 different writers and each form consists 4 to 6 lines having 60 to 80 words. Information of text will be in both printed and Unicode. Corpus will be completely markup according to hierarchical level of indexing. Markup is done on the basis of multilevel indexing. It will start from handwritten text form, followed with the segmented line where line is again segmented into words. Segmented word may be individual character or unit. Corpus will also provide facilities to further add modify a word. Corpus would be useful for both linguistics related research and benchmarking purpose.

REFERENCES

[1] Johansson, S., Atwell, E., Garside, R., and Leech, "The Tagged LOB Corpus", Norwegian Computing Centre for the Humanities, Bergen, 1986.

[2] Christian , P. Michel, Stefan and Philippe, "The IRESTE On/Off (IRONOFF) dual handwriting database", In Proceedings of 5th international conference on Document Analysis and Recognition, 1999.

[3] U. Marti and H. Bunke, "A full English sentence database for off-line handwriting recognition". In Proc. of the 5th Int. Conf. on Document Analysis and Recognition, pages 705 - 708, 1999.

[4] U. Marti and H. Bunke, "The IAM-database: An English Sentence Database for Off-line Handwriting Recognition". Int. Journal on Document Analysis and R Recognition, Volume 5, pages 39 - 46, 2002.

[5] Lecun et al., "The MNIST Database of Handwritten Digits (Image)", 1999.

[6] M. Waqas, C. Lei, N. Nobile, and C. Y. Suen, "A New Large Urdu Database for Off-Line Handwriting Recognition", In Proceedings of ICIAP, LNCS 5716, pp. 538–546, 2009.

[7] R. Sarkar, N. Das, S. Basu, M. Kundu, M. Nasipuri and D. Basu, "A database of unconstrained handwritten Bangla and English mixed script document image", International Journal on Document Analysis and Recognition (IJDAR), vol 15, Issue 1, pp 71-83, 2012.

[8] Hussain, S. "Resources for Urdu Language Processing", in the Proceedings of the 6th Workshop on Asian Language Resources, IJCNLP'08, IIIT Hyderabad, India, 2008.

[9] A. Raza, I. Siddiqi, A.Abidi, and F. Arif, "An Unconstrained Benchmark Urdu Handwritten Sentence Database with Automatic Line Segmentation, International conference on Frontiers in Handwritten Recognition (ICFHR), Page(s) 491-496, 2012.

[10] J. Hull. A database for handwritten text recognition research. IEEE Trans. on Pattern Analysis and Machine Intelligence, 16(5):550–554, May 1994.

[11] R. Wilkinson, J. Geist, S. Janet, P. Grother, C. Burges, R. Creecy, B. Hammond, J. Hull, N. Larsen, T. Vogl, and C. Wilson. The first census optical character recognition systems conf. #NIST IR 4912, The U.S. Bureau of Census and the National Institute of Standards and Technology, Gaithersburg, MD, 1992.

[12] T. Saito, H. Yamada, and K. Yamamoto. On the data base ETL 9 of hand printed characters in JIS Chinese characters and its analysis. IEICE Transactions, J68-D (4):757–764, 1985.

[13] D. Kim, Y. Hwang, S. Park, E., and S. Bang, "Handwritten Korean character image database PE92", In Proc. of the Second Int. Conf. on Document Analysis and Recognition, Pages 470–473, 1993.

[14] N. Husain, "Language Maintenance of Immigrant community in Britain", International Conference on Languages, Literature and Linguistics, IPEDR vol.26 IACSIT Press, Singapore, 2011.

[15] See http://www.censusindia.gov.in.

Non-Standard Words as Features for Text Categorization

Slobodan Beliga, Sanda Martinčić-Ipšić
Department of Informatics
University of Rijeka
Radmile Matejčić 2, 51000 Rijeka, Croatia
{sbeliga, smarti}@inf.uniri.hr

Abstract - This paper presents categorization of Croatian texts using Non-Standard Words (NSW) as features. Non-Standard Words are: numbers, dates, acronyms, abbreviations, currency, etc. NSWs in Croatian language are determined according to Croatian NSW taxonomy. For the purpose of this research, 390 text documents were collected and formed the SKIPEZ collection with 6 classes: official, literary, informative, popular, educational and scientific. Text categorization experiment was conducted on three different representations of the SKIPEZ collection: in the first representation, the frequencies of NSWs are used as features; in the second representation, the statistic measures of NSWs (variance, coefficient of variation, standard deviation, etc.) are used as features; while the third representation combines the first two feature sets. Naive Bayes, CN2, C4.5, kNN, Classification Trees and Random Forest algorithms were used in text categorization experiments. The best categorization results are achieved using the first feature set (NSW frequencies) with the categorization accuracy of 87%. This suggests that the NSWs should be considered as features in highly inflectional languages, such as Croatian. NSW based features reduce the dimensionality of the feature space without standard lemmatization procedures, and therefore the bag-of-NSWs should be considered for further Croatian texts categorization experiments.

Keywords: text categorization, non-standard words, collection representation, features, accuracy

I. INTRODUCTION

Categorization or classification is tasked to classify a given data instance into a pre-specified set of categories. In the domain of text categorization, the task is called text categorization (TC) – given a set of categories (subjects) and a collection of text documents, the process of finding the correct topics for each document [1]. Typically, texts are represented with vectors of word's frequencies in the bag-of-words model. The problem with this approach is the dimensionality of the features vectors, which is equal to the number of different words in the collection [1, 2].

One of the solutions for the dimensionality problem is **Feature Selection** which can be: heuristic, guided by linguistic and domain knowledge, or statistical [3].

In linguistically guided approach, the dimensionality of the features vector is reduced with stop words (*a*, *an*, *the*, *of*, etc.) omission and lemmatization. Lemmatization changes each inflected word form with its lemma (*walked*, *walks*, *walking* – with lemma *walk*) and therefore reduces the dimensionality of feature vectors, which is of essence in morphologically rich languages such as Croatian. Some feature selection approaches ignore terms that are too frequent or too rare according to empirically chosen thresholds [3]. As data sets become more complex, heuristics are usually not sufficient. Therefore, another solution is to use measures such as Information Gain and Chi-Squere, to define relevance of each feature [4].

Another approach to dimensionality problem is **Dimensionality Reduction by Feature Extraction**. It implies a much smaller set of synthetic features from the original feature set, using synthetic features rather than naturally occurring words. This will avoid the problems inherited from natural language (polysemy, homonymy, etc.). Using word groups, instead of individual words as features, with a high degree of semantic relatedness, proved to be justified in clustering [1]. Systematic approach - Latent Semantic Indexing (LSI) is based on this principle [1].

In this paper we propose opposed approach based on Non-Standard Words (NSWs) as features. Instead of bag-of-words, we construct the bag-of-NSWs, and use them to train different text classifiers. The peculiarity of proposed approach is in using NSWs which would usually be omitted from the text during the text cleaning procedures. Non-standard words are: numbers, dates, time, abbreviations, acronyms, currency, measurement units, etc. Text normalization transforms NSWs in their extended form: the abbreviation */dr./* into */doctor/* or acronym */SMS/* into */Short Message Service/*.

The motivation for this work rises from an idea that discarded characters and numbers are sufficient for discrimination of classification of the text categories, reducing the feature vector dimensionality at the same time [2].

The paper is organized as follows: Section II. presents methodology of text normalization and categorization, Section III. presents the data collection, NSW features extraction procedures and training of classification models. The results are presented in the end of Section III. The paper concludes with discussion and some future work considerations.

This research is partially supported by the grant 13.13.2.2.07 of University of Rijeka.

II. METODOLOGY AND RELATED WORK

A. NSW extraction

Traditionally, in the field of natural language processing (NLP), normalization is the process of transforming the unit into its normal form (i.e. lemma). Normalization is also the first step in the text pre-processing of Text-to-Speech (TTS) systems. It is performed in the normalization module, which is responsible to identify a NSW token and to transform it into its expanded form [5, 6]. Text Normalization for Croatian Speech Synthesis is described in [7]. In this work, the Croatian NSW taxonomy driven rule based approach was used for NSWs extraction [7, 8].

Module for normalization initially detects the NSW and separates it from standard words. Then, detected NSW is transformed into its expanded form, suitable for the TTS system. In this experiment, the process of transforming NSW to expanded form is not required. Instead, we detect each NSW and determine its NSW type as is suggested in taxonomy for classification of non-standard words in Croatian language [8]. In addition, every occurrence of NSW for each NSW type is counted. Our approach identifies NSWs and then normalizes detected NSWs following the taxonomy of Croatian NSWs by combining programmed rules and the lookup dictionary.

B. Text categorization

Text categorization is tasked with assigning one of predefined classes to a document. Usually text categorization is performed using a bag-of-words model borrowed from information retrieval (IR), and most often based on TF-IDF model [1-3]. Documents are represented as feature vectors, which are, according to the bag-of-words model, unordered list of words with the frequency of their occurrence in a document. Inverse document frequency is used to model the rareness of terms in a document collection. Feature vectors are used to train the classifier, usually Bayesian classifier, k-nearest neighbor (kNN), classification tree or support vector machines (SVM), etc. [1, 3, 9]. Some of the techniques for the vector dimensionality or the feature space reduction are: *stop words* and NSW removal [10], lemmatization or morphological normalization [11] - NLP based principles; selecting the most discriminative subset of features (document frequency, information gain, mutual information, odds ratio, etc.) – called feature selection [4, 12] - or deriving new smaller set calculated from original features – called feature extraction [13].

Our approach is presented in Figure 1.: in the first step - we detect all NSWs from the text; in the second step – for detected NSWs we assign its type (number, abbreviation, date, etc.); in the third step – we construct the bag-of-NSWs; in the fourth step – we train the classifiers. Finally, as a result of the text categorization we obtain the texts which are organized as categorized collection of texts – every text is associated with one of the possible categories.

III. NSW BASED TEXT CATEGORIZATION

A. Data (SKIPEZ)

For the purposes of normalization and categorization tasks we prepared the collection of texts in Croatian language – SKIPEZ (Službeno, Književno, Informativno, Popularno, Edukativno, Znanstveno). SKIPEZ contains 390 texts organized in 6 predefined classes: official, literature, informative, popular, educational and scientific. The distribution of text over classes is balanced so each class contains exactly 65 texts. Texts were selected according to their relevance to the class and according to the % of contained NSWs, because we needed a balance between both aspects in order to train a classifier. SKIPEZ statistics is shown in Table I., and additional details about SKIPEZ are in [14].

The content of texts in each subclass was carefully selected, in order to capture as many different and representative text in each class used and contain representative set of NSWs, in order to compile the exhausted list of Croatian NSWs is shown in Table II.

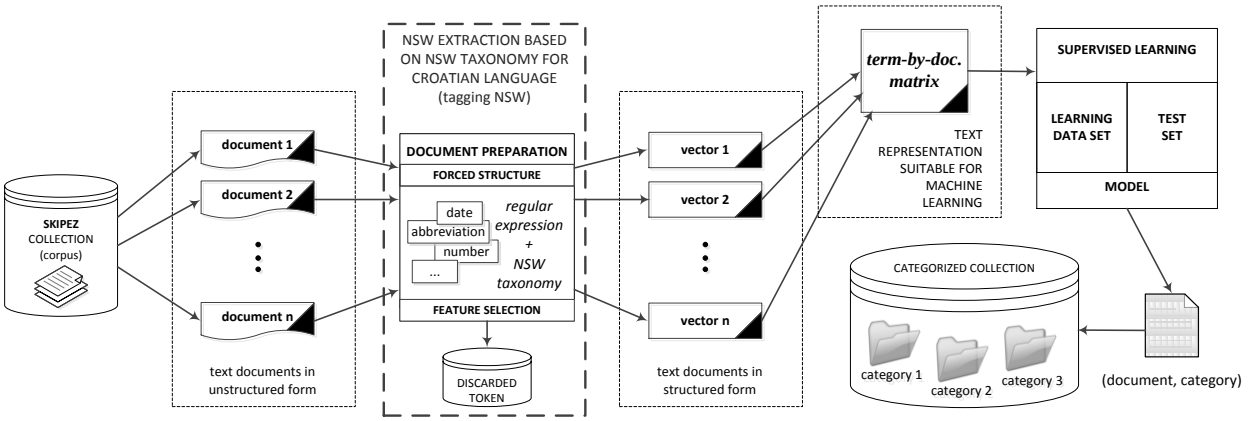


Figure 1. Shematic representation of NSW based text categorization

TABLE I. NUMBER OF TOKENS* AND NON-STANDARD WORDS IN SKIPEZ CORPUS SHOWN PER 6 CATEGORY

CATEGORY	Statistics of SKIPEZ corpus		
	#TOKENS	*#NSWs*	*%NSWs*
official	328.322	81.667	24,87 %
literature	789.555	18.150	2,30 %
informative	86.819	12.641	14,56 %
popular	224.518	18.033	8,03 %
educational	198.661	11.598	5,84 %
scientific	644.571	73.979	11,48 %
OVERALL	**2.272.446**	**216.068**	**9,51 %**

*(e.g. words and punctuation marks)

TABLE II. NUMBER OF TEXT ACCORDING TO THE TEXT TYPE IN EACH SKIPEZ CATEGORY

CATEGORY	text type/origin	#texts	
official	stock reports	16	65
	announc. ads	10	
	legislation gazette	12	
	statistics	8	
	public tenders	9	
	exchange rates	10	
literature	lyrics	2	65
	novel, story play	17	
	publicists	31	
	history, theology	15	
informative	traffic, traveling	19	65
	phone books	6	
	TV guides	12	
	weather forecasts	18	
	commercial adds	10	
popular	forums	8	65
	horoscope	10	
	recipes	14	
	lifestyle, cars, health	12	
	on-line sales catalogs	10	
	blogs	6	
	social network comm.	5	
educational	school books nature	16	65
	school books social	15	
	school books human.	17	
	school books technol.	17	
scientific	encyclopedia	10	65
	atlas	9	
	dictionaries	9	
	master and phd thesis	15	
	books	10	
	journal papers	12	
OVERALL			**390**

B. NSW Extraction Procedure

The NSWs were extracted from the SKIPEZ text collection according to the approach presented in [7, 8]. Here, we only detected NSWs types according to this taxonomy, resulting in feature vectors consisting of NSWs types frequencies for each document in the collection. A feature vector for each text document in collection contains 85 values (calculated from NSW types frequencies):

- STRING part - 15 values: ordinal and nominal roman numbers, simple or compound abbreviations, abbreviations without full stop, chemical elements, measurement units, currency an monetary units, acronyms, inflected and special acronyms, symbols, emoticons and suffixes.
- NUMBER part - 21 values: numeric format date, time period, time, proportion, dimensions, short references, short and long telephone number, interval, positive and negative decimal numbers, decimal interval, positive and negative fractions, exponents, ordinal and nominal numbers, ordinal and nominal intervals, positive and negative nominal numbers, ISDN/UDK numbers.
- COMBINED part - 20 values: 3 different date formats, geographical coordinates, long references, biblical references, enumerations of tables, figures and titles, road designation, TV show label, traffic abbreviations, compound measurement units, vitamins, house numbers, chemical and mathematical formulas, program code, registration number, argo-emoticons, domain specific NSWs, and NSWs of unknown type.
- DERIVED part - 29 values: total of NSWs in the number class, total of NSWs in the string class, total of NSWs in combined class, total of NSWs in text, total of words in text, number of different NSW types, number of empty NSW types, ratio between filled features and subclasses, ratio between filled and empty NSW types, coefficient as number of NSW divided with number of words in text. Moreover, there are also features that summarize the total number of instances of certain NSWs subclasses that are inter-related and can be classified in the same superclass. These are respectively: total of full date, total of incomplete date, total of references, total of numeration, total telephone numbers, total of range, total of decimal numbers, total number of fractions, total number of road designations, total number of electronic addresses, total of ordinal numbers, total of nominal numbers, total of roman numbers, total number of abbreviations, total number of measurement units, total number of acronyms, total number of office designation, total number of formulas and total number of jargon terms.

The first representation of SKIPEZ collection is a feature set composed from NSW type frequencies. For each document the features vector contains 85 values (as listed above).

The second representation of text, based on the NSW frequencies, represents the base from which the values of the features are derived. In this case, the statistics serves as a tool for obtaining metadata from the previously calculated frequency of NSW occurrence. For each element of the population, therefore each vector, we calculate some of its numerical characteristics called statistical features [15, 16]. The main idea when choosing a feature vector is utilizing the basic statistical features, which can describe the dispersion of identified patterns of NSW in a particular text.

The statistical characteristics of each population (each vector) were counted. We observed the vector elements as a sequence of elements and determined the individual values of absolute and relative measures of dispersion: the arithmetic mean, the range of variation, standard deviation, variance, coefficient of variation, and also a measure of flatness and a measure of asymmetry. We then calculated the statistical characteristics exclusively for superclasses (string, number, and string + number) – higher classes from taxonomy (Figure 2.). This second group of calculations observes dispersion patterns within the superclass, i.e. numerically describes the statistical characteristics of the NSW dispersion in three superclasses. We calculated the following characteristics: the arithmetic mean of superclass, interquartile q3, interquartile q1, interquartile Iq, coefficient of quartile deviation and coefficient of variation.

The third representation of texts is formed as a union from the features of the first and second representations, therefore includes all their values. Through the combination of features, we try to take advantage of the knowledge that was previously extracted from all the texts, on the principle "the more, the better".

In experiments with all 3 feature sets we conduct an iterative process of adding the created vectors for each text, we gradually form a bag-of-NSWs model (*term-by-document matrix*) - see Figure 1. Since we used supervised machine learning, during the bag-of-NSWs construction to each feature vector we assign its actual category (see Table II.).

C. Categorization Results

We trained the Naive Bayes, Classification Tree, kNN, CN2, C4.5, and Random Forest using the Orange [17] toolkit. As the measure of classification performance we used accuracy (1) [9].

$$\text{Accuracy} = \frac{Number\ of\ correct\ classifications}{Total\ number\ of\ test\ cases} \qquad (1)$$

where a correct classification means that the learned model predicts the same class as the original class of the test case.

The results of trained classifiers (Naïve Bayes, Classification Tree, kNN, CN2, C4.5 and Random Fortest) are obtained for three different feature representations using 5 fold cross-validation. The text categorization results using the first feature set - NSW frequencies are shown in Table III.; using statistical features in Table IV. and for combination of features in Table V.

TABLE III. RESULTS OF TEXT CATEGORISATION WHERE TEXTS ARE REPRESENTED BY FREQUENCIES OF NSWS

CLASSIFIER	*#tokens*	*Accuracy [%]*
Naïve Bayes	263	67,44
Classification Tree	305	78,21
kNN	296	75,90
CN2	306	78,46
C4.5	309	79,23
Random Forest	**335**	**86,15**
The mean score of all classifiers:		77,56

TABLE IV. RESULTS OF TEXT CATEGORISATION WHERE TEXTS ARE REPRESENTED STATISTICAL FEATURES

CLASSIFIER	*#tokens*	*Accuracy [%]*
Naïve Bayes	176	45,13
Classification Tree	223	57,18
kNN	220	56,41
CN2	212	54,36
C4.5	229	58,72
Random Forest	**260**	**67,95**
The mean score of all classifiers:		56,62

TABLE V. RESULTS OF TEXT CATEGORISATION WHERE TEXTS ARE REPRESENTED BY UNION OF NSWS FREQUENCIES AND STATISTICAL FEATURES

CLASSIFIER	*#tokens*	*Accuracy [%]*
Naïve Bayes	176	62,05
Classification Tree	223	76,92
kNN	220	70,00
CN2	212	74,62
C4.5	229	78,97
Random Forest	**260**	**86,92**
The mean score of all classifiers:		74,91

The experimental results can be analyzed twofold: by accuracy of different classifiers and by accuracy with respect to different types of representations in collection - Figure 3. Random Forest achieves the highest accuracy and Naïve Bayes the lowest in all representation of the collection. The accuracy of all classifiers with statistical features is generally worse than the frequency features

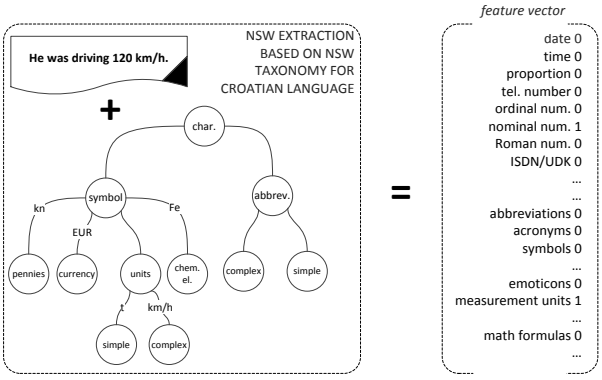


Figure 2. Forming a feature vector in Bag-of-NSWs

(Table IV.). With statistical features, Random Forest achieved a slightly higher accuracy (increase less than 1%). Chakrabarti, Yang and Pedersen in [4] state that numerous techniques of selection of features improve the performance of the classifier only marginally. This was confirmed also with our experiment, although the selection of features used non-standard features of the vector.

Experimental results can be analyzed according to text categories as well. Comparison of text categorization accuracy, only using Random Forest, is presented in Table VI. The highest accuracy was achieved for the class of informative texts, while the worst was achieved in the class of scientific and educational texts. The obtained results of classification leave room for discussion about the possibility of unification or reduction of scientific and educational domains because in this 2 categories is the highest number of misclassified texts. The science and education are closely related and they share the division of the subareas, fields and branches.

The worst scores for educational and scientific domain can be explained as an error that is probably caused by the fact that the educational and scientific texts are equally distributed in the scientific and artistic areas, fields and branches as required by the National Council for Science, and for Republic of Croatia Ministry of Science, Education and Sports, available at [18]. On the basis of this criterion texts are selected for SKIPEZ collection. Classifier failed to recognize the scientific style of writing. Encyclopedic texts, journals, articles and dissertations are apparently very similar to school textbooks, teaching materials and manuals through the NSW perspective.

TABLE VI. COMPARISON OF CATEGORIZATION ACCURACY WITH RESPECT TO 3 DIFFERENT FEATURE REPRESENTATIONS FOR RANDOM FOREST ALGORITHM

CATEGORY	ACCURACY [%]			AVG [%]:
	NSWs freq.	*Statistics*	*NSWs freq. & Stat.*	
official	76,92	66,15	81,54	74,87
literature	98,46	67,69	93,85	86,67
informative	100	83,08	100	**94,36**
Popular	100	66,15	92,31	86,15
educational	72,31	64,62	76,92	**71,28**
scientific	69,23	60,00	76,92	**68,72**
AVG [%]:	86,15	67,95	**86,92**	

IV. DISCUSSION AND CONCLUSION

This paper presents categorization of Croatian texts using Non-Standard Words (NSW) as features. Results suggest that non-standard word forms can be used as features for representation of texts in text categorization. The associated NSW taxonomy is suitable form of background knowledge on the basis of which NSW can be automatically extracted from texts. It has been shown that the NSWs carry enough information about the nature of text, which is suitable for further classification. With this approach, it is possible to significantly reduce the dimensionality of feature vectors (space) and at the same time achieve good text categorization results. Using NSWs feature, the feature vectors have several times lower dimensions then the original ones and this way reduces the problem of sparse data.

Bag-of-NSWs approach is suitable for inflectional languages because extremely reduced dimensionality (only 85 features) which can avoid the procedure of lemmatization of words or morphological normalization.

Obtain results should be compared with standard TF-IDF classification method. Since the literature presents the support vector machine as successful, it would be appropriate to examine whether a classification using SVM provides better results than the Random Forest algorithm. In further work we plan to extend our work to the field of micro texts classification, since SMS and twitter texts contain a lot of NSWs.

REFERENCES

[1] R. Feldman, J. Sanger, The Text Mining Handbook - Advanced Approaches in Analyzing Unstructured Data, New York: Cambridge University Press, 2007.

[2] C. D. Manning, P. Raghavan, H. Schütze, An Introduction to Information Retrieval, Cambridge University Press, 2009.

[3] S. Chakrabarti, Mininh the Web - Discovering knowladge from hypertext data, Morgan Kaufmann Publishers, 2003.

[4] Yang, Y., and Pedersen, J. O. A comparative study on Feature Selection in Text Categorisation. ICML-1997., Morgan Kaufman. San Francisco: 412-420.

[5] Sproat, R., et al.: Normalization of Non-standard Words. Computer Speech & Language. Vol.5. pp. 287-333, 2001.

[6] Sproat, R.: Lightly Supervised Learning of Text Normalization: Russian Number Names, IEEE Workshop on Spoken Language Technology, Berkeley, CA , 2010.

[7] S. Beliga, S. Martinčić-Ipšić, Text Normalization for Croatian Speech Synthesis, Proc. MIPRO junior - Student Papers, 382-387, 2011.

[8] S. Beliga, M. Pobar, S. Martinčić-Ipšić, Normalization of Non-Standard Words in Croatian Texts,« u TSD 2011, Plsen, 2011.

[9] B. Liu, Web Data Mining - Exploring Hyperlinks, Contents, and Usage Data, Berlin: Springer, 2011.

[10] M Radovanović, M. Ivanović, Z. Budimac, Text Categorization and Sorting of Web Search Results. Computing and Informatics, 28(6), pp. 861–893, 2009.

[11] M. Malenica, T. Šmuc, J. Šnajder, B. Dalbelo Bašić, Language morphology offset: Text classification on a Croatian–English parallel corpus, Information Processing & Management, 44(1), 325-339, 2008.

[12] D. Mladenić, Feature subset selection in text-learning, Machine Learning ECML98., 1998.

[13] F. Figueiredo at al. Word co-occurrence features for text classification, Information Systems, 36. 843-858, 2011.

[14] S. Beliga, Klasifikacija tekstova temeljem nestandardnih oblika riječi u hrvatskome jeziku. Diplomski rad. Sveučilište u Rijeci - Odjel za informatiku, 2013.

[15] N. Sarapa, Vjerojatnost i statistika I i II, Zagreb: Školska knjiga, 1993.

[16] I. Šošić, Primjenjena statistika, Zagreb: Školska knjiga, 2006.

[17] Orange - Data Mining, Available: http://orange.biolab.si/. (1.10.2012.).

[18] Nacionalno vijeće za znanost, "Narodne novine, Pravilnik o znanstvenim i umjetničkim područjima, poljima i granama", 22. rujna 2009. Available: http://narodne-novine.nn.hr/clanci/sluzbeni /2009_09_118_2929.html. (2.11. 2012.)

Developing a Text Classifier With Constrained Development and Execution Time

I. Budiselić*, G. Delač* and K. Vladimir*
* University of Zagreb, Faculty of Electrical Engineering and Computing, Zagreb, Croatia
{ivan.budiselic, goran.delac, klemo.vladimir}@fer.hr

Abstract - The aim of this paper is to show that an accurate and efficient text classifier for relatively simple problem domains can be created in only a few hours of development time. The motivating example discussed in the paper is a recent *HackerRank* competition problem that tasked competitors with creating a classifier for questions from the popular question and answer platform *StackExchange*. The paper describes the key components of one solution to this problem, and briefly overviews the naive Bayes classifier that is the basis of the solution. The discussion is focused on feature selection and example representation which were the key challenges to be addressed during the development of this classifier. We also analyze the effect of the number of features on accuracy, training and classification time and the size of the resulting classifier and the representation of the training examples which were all important characteristics for the competition. The described classifier achieved slightly over 89% accuracy on the hidden question set, while the winning submission achieved around 92%.

I. INTRODUCTION

Text classification is a process in which pieces of text, often called *documents*, are assigned to one or more predefined classes or categories. Up until the late 1980s, automatic text classification was based on predefined rules encoding expert knowledge about the problem domain [1]. This approach to text classification requires a high level of expertise, both in understanding the problem domain in which classification is being employed, and in choosing and encoding appropriate rules for classification. However, since then, automatic text classification has shifted towards machine learning techniques where, in some sense, the computer is tasked not only with classification itself, but also with choosing classification rules within a certain model based on a set of already classified examples. This development makes it possible to quickly create useful text classifiers with minimal domain knowledge and prior experience with text classification technology.

This paper is motivated by the September 2013 edition of the *HackerRank* [1] *20/20* online programming competition, where competitors were tasked with creating a classifier for questions from the popular question and answer platform *StackExchange* [2]. We present one classifier developed for this competition.

A competition context presents several specific challenges in designing and developing a text classifier. First, the classifier needs to be developed in a short amount of time, and the classifier described in this paper was developed in around five hours during the course of one day. Second, the evaluation environment of an online competition imposes strict limitations on the training and classification time of the system. The goal of the paper is to describe how these challenges were successfully addressed, and ultimately show that an accurate classifier for this problem can be developed quickly.

The rest of the paper is organized as follows. Section II precisely defines the *StackExchange* question classification problem. Section III describes the key components of the presented solution to the problem. Section IV discusses some key implementation details that were necessary to address the strict efficiency requirements for the classifier, and analyzes the impact of the number of features on several key properties of the classifier. Finally, the paper is concluded in Section V.

II. STACKEXCHANGE QUESTION CLASSIFICATION PROBLEM DEFFINITION

The *StackExchange* platform powers a continually growing number of 112 different question and answer sites spanning various topics, from religion to physics and biology. The objective of the *HackerRank* competition task was to develop a *StackExchange* question classifier. Given the title and the text body of an example question, the classifier should produce the topic in which the question had been asked.

To keep the training and test datasets sufficiently small, the problem was restricted to only these ten topics: *android*, *apple*, *electronics*, *gis*, *mathematica*, *photo*, *scifi*, *security*, *unix* and *wordpress*. Note that there is some overlap between these topics. For example, the *Mac OS X* operating system for Apple computers is based on the Unix operating system. Therefore, some questions could have legitimately been asked on either the *apple* or the *unix StackExchange* site, and even a human expert would not be able to always guess the actual class of the question. A similar overlap exists between any of the *android*, *apple* or *unix* topics and the *security* topic.

[1] http://www.hackerrank.com
[2] http://www.stackexchange.com

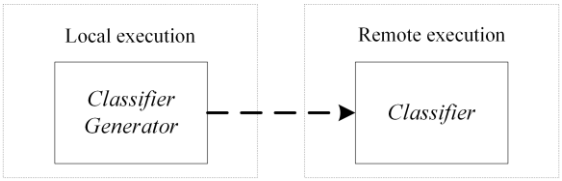


Figure 1. The two programs of the presented solution to the *StackExchange* question classification problem.

Competitors were given a training set of 20209 questions with correct topic labels. The number of questions per topic in the training dataset was between 1369 questions for the *mathematica* topic and 2383 questions for *gis*. This training file was also made available in the evaluation environment.

The problem statement also defined that the hidden question set used for evaluation would contain 22000 questions. This is significant because solutions must label all the questions in the evaluation set within a fixed small amount of time that varies with programming language. For example, C++ solutions are allowed only three seconds of run time. The solution presented in this paper was developed in Python, where solutions are allowed to run for up to 16 seconds. This large difference in allotted processor time is due to the fact that programs in these competitions are typically computationally intensive and the CPython implementation of Python is significantly slower than C++ for such tasks [2][3].

Another significant part of the problem specification is that solutions are submitted in the form of a single file of source code. This solution organization restricts the classifier training process to two options. First, the classifier can be trained offline and the submitted solution can then somehow reconstruct the trained classifier before classifying test examples. Second, the solution can train the classifier during evaluation using the training file available in the evaluation environment. With both options, the time required to either reconstruct or train the classifier must be carefully considered as it is also counted towards the execution time limit. This is an unusual requirement as this process of initializing a classifier is usually done once before the classifier is used for a longer period of time, and only classification time is critically important while longer training times can be tolerated. This issue is discussed in more detail in the next section.

Finally, in addition to the standard Python library, Python solutions to *HackrRank* problems can use the *Natural Language Toolkit* (NLTK)[3] [4][5]. NLTK is a suite of open source Python modules for natural language processing. The solution presented in this paper used the implementation of the naive Bayes classifier found in NLTK.

III. Solution Architecture

The presented solution to the *StackExchange* question classification problem consists of two separate programs shown in Fig. 1. The *Classifier Generator* (called simply the *Generator* in the remainder of the paper) is executed locally and generates the source code of the *Classifier*.

[3] http://nltk.org

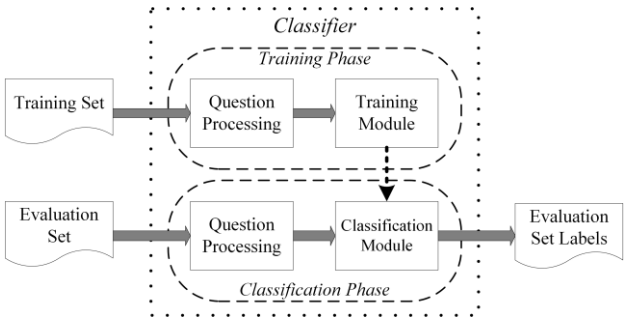


Figure 2. The architecture of the *Classifier*.

The *Classifier* is then submitted to the *HackerRank* server for preliminary evaluation where it is executed on a small question set to make sure that it can run within the evaluation environment.

This organization of the solution allows quick exploration of different options in question processing and representation with frequent submissions for preliminary evaluation. This is essential as it provides feedback on the exact effect of changes in the classifier on the classifier's execution time in the actual evaluation environment.

The architecture and the roles of the *Classifier* and the *Generator* are discussed in detail in the next two subsections. As the *Classifier* is significantly simpler, it is discussed first.

A. Classifier Architecture

The architecture of the *Classifier* program is shown in Fig. 2. Of the two classifier training options mentioned in Section II, the solution presented in this paper is based on the latter option in which the classifier is trained in the evaluation environment. Therefore, the operation of the *Classifier* can be divided into two phases shown with dashed ovals in the figure: the *training phase* and the *classification phase*.

The common process in both phases is *question processing*. Questions are represented in the classifier with a *bag of words* model, which is frequently used in text classification and text processing in general [6]. A bag of words model removes the notion of sequence order from a piece of text, focusing instead only on the word content.

The process of converting the classifier input into this model can conceptually be decomposed into four steps. First, as both training and evaluation questions are given to the classifier in *JavaScript Object Notation* (JSON), they are parsed into an internal format. Second, the question title and body text are merged into a single string that is normalized by replacing punctuation and other special characters with spaces and by converting all letters to lower case. Third, the resulting string is split into *words*, where a word is defined as a continuous sequence of non-whitespace characters.

Finally, in the fourth step, the word list is filtered to only include words that were selected as *features* which is then the final question representation. As the set of features is chosen in the *Generator*, this final step is discussed in greater detail in Subsection III.B.

Question processing is intentionally kept simple to increase efficiency as the classifier must process a large number of questions in a short amount of time. Specifically, neither *stemming* [7] nor *lemmatization* [6] are used. Stemming typically removes affixes from words to create a stem that might not be a word itself, while lemmatization attempts to reduce all of the morphologically different instances of a word to one canonical word. An attempt to use the lemmatizer that is part of NLTK increased the question processing time almost threefold so it was removed from the solution early in the development process.

In the training phase, the processed questions with the given correct topic label are given to the *training module* which generates the *classification module*, as indicated by the dashed arrow in the figure. In the classification phase, the processed questions are given to the classification module which assigns a topic label to each question.

The interface of the *Classifier* is shown with a dotted square in Fig. 2. The *Classifier* takes two input files containing the training question set and the evaluation question set and outputs the topics for the questions in the evaluation set.

The training and classification modules are based on the *naive Bayes classifier* that is briefly described in the remainder of this subsection. Although a more complex classifier model like *Support Vector Machines* (SVM) would potentially yield higher accuracy [8], the training time of such models is typically significantly higher [9][10] and requires more tuning to achieve peak accuracy than the naive Bayes classifier. Both of these characteristics significantly contributed to the choice of naive Bayes, and as it performed well and due to time constraints, other models were not tried.

1) Naive Bayes Classifier

The naive Bayes classifier [11] is based on the idea that training examples can be used to estimate the probability that an example belongs to a certain class. Given an example to classify, the classifier then selects the class that is estimated as the most likely to contain the example.

Examples are encoded with a tuple of values of predefined *features*, where each feature can have some value drawn from a finite set. In the context of text classification, features are words from the text.

It is important to note that *equality* is the only relation defined on the value set of a feature. Specifically, no two values of a feature are more different than any other pair of values. Therefore, the intuitive word frequency value set has the problem that a piece of text with 50 occurrences of a word is modeled as equally different from a piece of text with 49 occurrences of that word and one with no occurrences of that word.

Due to this fact, the solution presented in this paper is based on the bag of words model where features are words from the questions and their values are binary. Specifically, a particular feature has the value one if and only if that word occurs in the question at least once and possibly many times, and otherwise has the value zero.

Let the set of question classes be C, and C_i be the ith class in some fixed order over C. Similarly, let the feature set be F with F_i denoting the ith feature. Examples are then represented with an n-tuple $(v_1, v_2, ..., v_n)$ where v_i is the value of feature F_i and n is the cardinality of the feature set.

When classifying an example, we would like to determine which class attains the maximum conditional probability $P(C_i \mid v_1, v_2, ..., v_n)$, where the condition is a conjunction of events that the example either contains a word or it doesn't. Formally, the given example $(v_1, v_2, ..., v_n)$ would be classified in the class

$$c = \text{argmax}_{Ci}\, P(C_i \mid v_1, v_2, ..., v_n). \quad (1)$$

The naive Bayes classifier introduces a strong assumption that feature values are independent given the class. Using this independence assumption and Bayes' rule, (1) can be rewritten into

$$c = \text{argmax}_{\text{Ci}}\, P(C_i)\prod_j P(v_j \mid C_i). \quad (2)$$

The $|C|$ prior probabilities $P(C_i)$ and the $n|C|$ likelihood parameters $P(v_j \mid C_i)$ can be estimated from the labeled training data. The intuitive maximum likelihood estimator which would assign the ratio of the number of training examples from class C_i that have the feature value v_j and of the total number of examples in class C_i to $P(v_j \mid C_i)$ overfits the training data. Specifically, the fact that a certain feature value never appears in a class in the training data is made universal by assigning the value zero to the corresponding parameter of the classifier. Consequently, no example with that feature value can ever be classified into that particular class.

A simple way to address this limitation is to introduce *smoothing*. The NLTK implementation of the naive Bayes classifier uses *Laplace smoothing* which can be thought of as the maximum likelihood estimator in a training set that is extended with one example for each possible feature value and class pair.

Finally, in practical computer implementations, (2) is replaced with

$$c = \text{argmax}_{\text{Ci}} \left[\log P(C_i) + \sum_j \log P(v_j \mid C_i) \right]. \quad (3)$$

This transformation is done primarily to avoid working with very small real numbers that result from multiplying many potentially small likelihoods which can cause precision issues and even floating point underflow.

B. Generator Architecture

The logical organization of the *Generator* is shown in Fig. 3. The *Generator* takes two question sets as input: the *feature selection set* and the previously discussed *training set*. The output of the *Generator* is the source code of the *Classifier* that can immediately be submitted for evaluation on the competition server.

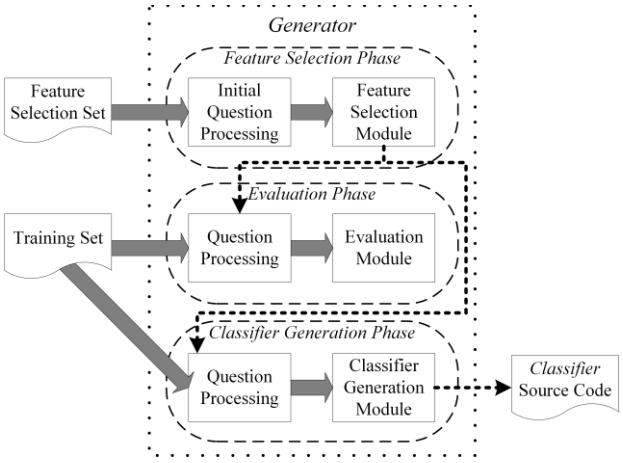


Figure 3. The architecture of the *Generator*.

As shown in the figure, the functionality of the *Generator* can be grouped into three phases which are described in the remainder of this subsection.

1) Feature Selection Phase

In the simplest implementation of a naive Bayes text classifier, all the words appearing in examples could be used as features. However, this choice would lead to two significant limitations. First, by observing (3), it is obvious that the time complexity of example classification scales linearly with both the number of features and the number of classes. Furthermore, the number of parameters that need to be estimated in the classifier training phase similarly increases linearly with the number of features and classes. As the number of classes is typically a parameter of the problem and can't be changed, choosing a smaller number of features is the only way to increase classifier computational efficiency in this model.

Second, many words that appear often in natural language are completely independent of the class to which some text belongs. Typical examples of such words are pronouns and conjunctions. Including such words as features only makes the classification problem harder as they decrease the relative differences between the computed probabilities for each class.

When selecting words as features, we would ideally like to select words that make it possible to distinguish examples from different classes in the general problem domain and not only in the training set. Choosing features that might provide very high accuracy on the training set but fail to generalize to other examples in the problem domain is an instance of the *overfitting problem* [12].

Avoiding overfitting is especially challenging when the training set is small. While there are techniques that successfully deal with this challenge even for small data sets, it is typically better to get more data if possible. As *StackExchange* publishes the whole question set on a quarterly basis, getting more examples is easy.

Therefore, the feature selection question set was generated from the two latest quarterly sets of questions covering the first half of 2013 [4]. From the total downloaded question set, 160 thousand questions were selected at random while preserving the relative frequency of topics to be approximately equal to that of the training set provided in the problem statement. The selected question set was further partitioned into 120 thousand questions for training during feature selection and 40 thousand questions for evaluating the classifier with the selected feature set.

The feature selection phase of the *Generator* consists of *initial question processing* and the actual *feature selection*. Initial question processing in the feature selection phase differs from question processing in the *Classifier* and other phases of the *Generator* in that words can't be filtered as the features have not yet been chosen. Instead, after the question has been split into words, the generator eliminates *stop words* from the question representation. Stop words are ideally words that frequently appear in language and don't help the classification process. In total, 519 words were chosen as stop words using an iterative semiautomated process where the most frequent remaining words were displayed and suggested for removal.

After stop word removal, the remaining least frequent words in the whole set of 120 thousand questions was removed from question representations so that the feature set was reduced to 150 thousand words. With 150 thousand features and 120 thousand training examples, the *Generator* took up to several minutes to train the naive Bayes classifier which is acceptable as this is done locally. This step is essential for feature selection as the resulting *Classifier* needs to use a significantly smaller feature set to meet the efficiency limits of the evaluation environment.

NLTK provides a method *most_useful_features(n)* on a trained naive Bayes classifier which returns the *n* feature values that are estimated to be the most useful for classification given the training set. The usefulness of the feature value v_j is estimated as

$$\frac{\max_{Ci} P(v_j \mid C_i)}{\min_{Ci} P(v_j \mid C_i)}. \qquad (4)$$

In words, the usefulness of the feature value v_j is equated with the ratio of the maximum and minimum likelihood that this feature value appears in any class. If this ratio is large, the feature value v_j is useful for discriminating between examples in the two classes in which the maximum and minimum likelihoods occur.

When features are binary, the usefulness estimate of the two possible values are inverses of each other because the class in which the maximum likelihood is achieved will have the minimum likelihood of the opposite value and vice versa. Therefore, to identify the most useful features, it is sufficient to simply ignore the value of the features returned by the NLTK method. For example, the five most useful features identified by this measure were *squared, code, pre, validity* and *data*.

The chosen set of features is then used in question processing in the remaining phases of the *Generator*. This

[4] http://www.clearbits.net/creators/146-stack-exchange-data-dump

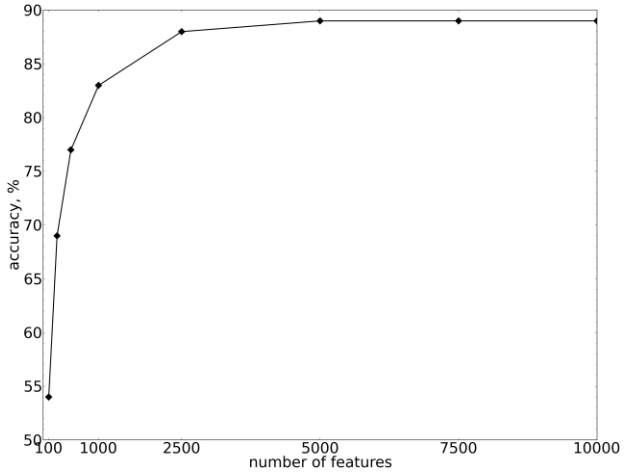


Figure 4. The effect of the number of features on the accuracy of the classifier trained on a set of 16000 questions and evaluated on a set of 4000 questions.

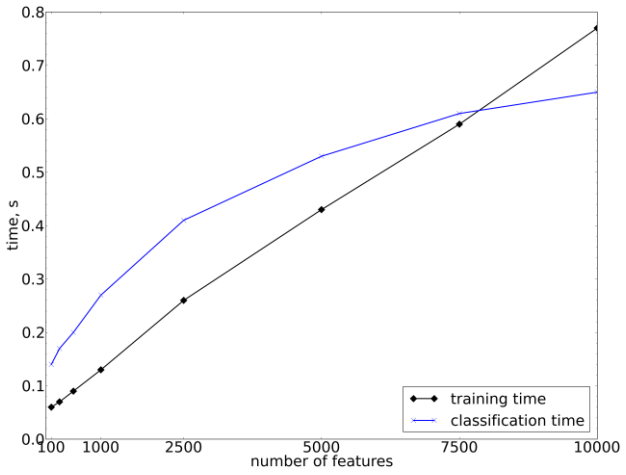


Figure 5. The effect of the number of features on the training time and classification time of the classifier trained on a set of 16000 questions and evaluated on a set of 4000 questions.

transfer of information is shown with dashed arrows entering the question processing modules in Fig. 3.

2) Evaluation Phase

Once a smaller set of features is selected in the feature selection phase, the classifier is retrained and evaluated on the set of 40 thousand evaluation questions. Through about ten iterations of this process, the final number of features was set to 5000. The effect of the number of features on several key properties of the *Classifier* is analyzed in Section IV.

Additionally, although the large feature selection question set significantly reduces the threat of overfitting, it is necessary to check if the selected features actually work well for the question set given to the contestants as part of the problem statement. This is important as questions in several topics are very time dependent and features selected for one time period might not perform well in another time period. To actually estimate the accuracy that the classifier will achieve on the hidden evaluation question set in the evaluation environment, the selected feature set is therefore evaluated using the question set provided in the problem statement.

Specifically, the provided question set is split randomly into 16000 training questions and 4000 evaluation questions. A new classifier is then trained using the selected features and evaluated on the evaluation set.

3) Classifier Generation Phase

The final task of the *Generator* is to actually create the source code of the *Classifier* that can be submitted for evaluation. To enable question processing, the *Classifier* needs the set of features selected in the feature selection phase. The feature set is inserted into the *Classifier* source code using the *pickle* serialization module in Python. Specifically, the data structure representing the feature set is serialized into a string that is compressed and *base64* encoded into the source file. The *Classifier* undoes this process during initialization.

A similar approach can, in theory, be applied to the trained classifier object itself to avoid the training phase during evaluation. However, this was not practical as the serialized and compressed classifier object was large (slightly under 2MB) and its decompression and deserialization took longer than actually training the classifier.

To further optimize the training phase of the *Classifier*, the processed training questions are also serialized into the generated source code. This basically eliminates the question processing step in the training phase described in the previous subsection.

IV. Implementation Details and Evaluation

In this section, we analyze the effect of the number of features on the accuracy and efficiency of the *Classifier*, and discuss some implementation details that were important for achieving the required performance. The analysis was done on the training file provided by *HackerRank* split into a set of 16000 training questions and 4000 evaluation questions. Feature selection was done in the *Generator* using the much larger question set, as described in the previous section.

The effect of the number of features on classifier accuracy is shown in Fig. 4. With a hundred features, classification accuracy is only 54% and grows quickly up to 88% with 2500 features. With 5000 features, the accuracy increases further by 1% and remains constant if more features are used. This graph is a clear indication that 5000 features was an appropriate choice for this problem and classifier model in terms of peak accuracy.

The efficiency of the *Classifier* is significantly impacted by several implementation details. Using the *cProfile* Python profiler on the *Classifier*, we found that almost 30% of run time in the initial implementation was being spent on string processing and manipulation. The simplest way to represent both features and classes in the classifier is to use strings, and this is also the suggested representation in NLTK's tutorials. However, this representation is inefficient both in terms of run time and in terms of the size of the trained classifier. This issue was addressing by mapping all strings to numbers that were then used throughout the *Classifier*. This is a particular

implementation of the general idea of *string interning*. With this change, the execution time of the *Classifier* decreased by around three times and the size of the serialized features and examples decreased up to five times. This ultimately made it possible to increase the number of features from 1000 to 5000 which corresponds to an increase in accuracy of more than 10%.

The change of the training and classification time with different feature set sizes is shown in Fig. 5. The graphs show the mean of five experiments for each data point. Only relative changes should be considered as the actual times on the y-axis were measured for a smaller set of questions than the actual hidden evaluation set, and all experiments were done locally and not on the evaluation server.

As the naive Bayes classifier needs to estimate $n|C|$ parameters during training, the nearly perfect linear dependence of training time with the number of features n is expected. However, the fact that classification time grows sublinearly is surprising as by (3) the naive Bayes classifier needs to add $\Theta(n|C|)$ logarithms which is clearly linearly dependent on the feature count n.

The explanation for this behavior stems from the interplay between the used example representation and the implementation of the naive Bayes classifier in NLTK. As the *Classifier* needs to prepare over 20 thousand training examples, it is impossible to represent every example with all 5000 features. NLTK expects examples to be represented with a feature-value map (a *dict* in Python) which is very inefficient for binary features. Due to this fact, questions were represented with a map that contained only those features with the value 1, i.e. only those feature words that were actually found in the question. However, when classifying examples, the naive Bayes classifier in NLTK simply ignores those features that are not present in the map, and therefore only adds the log-likelihoods of 1-value features. Features added beyond the first 2500 or so are rare words that are therefore not frequently found in example representations and this causes classification time to grow sublinearly.

A similar effect of increasing the number of features can be observed on the sizes of the serialized and compressed feature set (which grows linearly) and processed training question set (which grows sublinearly).

V. Conlusion

In this paper, we've described one solution to the *StackExchange* question classification problem posed in an online competition hosted by *HackerRank*. As the classifier was developed in a competition environment, special attention was given to efficiency, both in the training and the classification phase. The final version of the classifier with 5000 features took 13 seconds to process the evaluation set which fit comfortably within the 16 second time limit.

In terms of accuracy, the classifier took 8th place out of 44 solutions in the competition with 89.46% accuracy on the hidden evaluation set, while the winning solution achieved 92.13%. By using a custom implementation of the naive Bayes classifier it would be possible to significantly reduce the classifier's run time by using a more efficient representation of examples with binary features. This would free up processor time for other improvements, primarily for correctly computing topic assignment probabilities without disregarding 0-value features. Additionally, the method used for feature selection is fairly simplistic, and better feature selection algorithms could be used to potentially increase accuracy or decrease the number of features required for peak accuracy.

A more detailed analysis of accuracy shows that the classifier underperforms on questions in the *apple* (79%), *mathematica* (82%) and *unix* (82%) topics. Searching for features that might increase accuracy for these topics is the most likely avenue for improving the classifier's accuracy with the least amount of additional effort.

Based on the described experience with this classification problem, it can be concluded that it is possible to develop an accurate and efficient text classifier for nontrivial problem domains in only several hours of development time.

References

[1] F. Sebastiani, “Machine learning in automated text categorization,” ACM computing surveys (CSUR), vol. 34(1), pp. 1–47, 2002.

[2] G. Marceau, “The speed, size and dependability of programming languages,” online at http://goo.gl/JkZP, created in 2009, last accessed January, 30th 2014.

[3] B. Fulgham, “The computer language benchmarks game,” online at http://goo.gl/fk8raH, created in 2013., last accessed January, 30th 2014.

[4] E. Loper and S. Bird, “NLTK: the Natural Language Toolkit,” ETMTNLP '02 Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics, vol. 1, pp. 63–70, 2002.

[5] S. Bird, “NLTK: the Natural Language Toolkit,” COLING-ACL '06 Proceedings of the COLING/ACL on Interactive presentation sessions, pp. 69–72, 2006.

[6] W. B. Croft, D. Metzler, and T. Strohman, “Search engines: information retrieval in practice,” Addison-Wesley Reading, 2010.

[7] D. Hull, “The Stemming algorithms: a case study for detailed evaluation,” Journal of the American Society for Information Science, vol. 47(1), pp. 70–84, 1996.

[8] J. Huang, J. Lu, and C. X. Ling, “Comparing naive Bayes, decision trees, and SVM with AUC and accuracy,” Third IEEE International Conference Data Mining, pp. 553–556, 2003.

[9] F. Colas and P. Brazdil, “Comparison of SVM and some older classification algorithms in text classification tasks,” Artificial Intelligence in Theory and Practice, Springer, pp. 169–178, 2006.

[10] F. Colas, P. Paclik, J. N. Kok, P. Brazdil, “Does SVM really scale up to large bag of words feature spaces?,” Advances in Intelligent Data Analysis VII, Lecture Notes in Compute Science, vol. 4723, pp. 296–307, 2007.

[11] T. M. Mitchell, “Machine Learning,” McGraw Hill, 1997, pp. 177–183.

[12] J. Cheng and R. Greiner, “Comparing Bayesian network classifiers,” UAI99 Proceedings of the 15th conference on uncertainty in AI, pp. 101–108, 1999.

Application of the NOK method in sentence modelling

M. Rauker Koch*, M. Pavlić, A. Jakupović*****
*Railway and technical school Moravice, Vrbovsko, Croatia
** Department of Informatics, University of Rijeka
*** Business Department, Polytechnic of Rijeka
marina.rauker@skole.hr, mile.pavlic@ris.hr, alen.jakupovic@veleri.hr

Abstract - Knowledge representation is one of the areas covered by artificial intelligence. One of the methods for graphical representation of text expressed knowledge is the method NOK (Nodes Of Knowledge). NOK method enables transformation of text expressed knowledge into a graphical network of words and group of words. In this paper application of NOK method is presented. This application is based on sentences from an Aesop's Fable in Croatian ("Golden eggs in the chicken") and English ("The Goose with the golden eggs") version. In this way the applicability of this method on two natural human languages is presented, and similarities and differences that are partially conditioned by freedom of translators, and not only by differences in the syntax of the two languages, are observed.

Key words: Node Of Knowledge, Text-based knowledge representation, Network-based knowledge representation.

I. INTRODUCTION

AI studies computational modelling of human intelligence. Intelligent behaviour is conditioned by knowledge – human act based on what they know (or believe) about the world [1], [2]. One of the fields of interest of AI is knowledge representation and reasoning, which is suitable for reaching artificially intelligent behaviour. Many methods of converting knowledge into a form that can be encoded on a computer have been developed and they are applicable to different types of knowledge [3], [4].

Graphical method Nodes Of Knowledge designed for the presentation of sentences written in natural language is applied in this paper [5]. The formalization of this method is shown in [6]. The possibility of its application has been demonstrated on one text in Croatian and English. By analyzing produced graphic presentations answers to the following questions were obtained: Is there a difference in the complexity of making diagrams, which model is more complex, can answers to the same questions be read from both diagrams, regardless of differences in translation, are there any questions applicable on one model only and how can this problem be solved.

II. REVIEW OF RELEVANT PUBLICATION

In the field of artificial intelligence different approaches to knowledge representation formalisms are defined [2], [3], [7]. Conceptual graphs [8], [9], [10], semantic networks [11], frames [12] and predicate logic [13] are traditional formalisms that eventually produced numerous other formalisms and methods for knowledge representation. Ontologies have had a significant impact in the field of knowledge representation in the past twenty years [14], [15], [16]. They emphasize the semantic components of knowledge and language [17]. Furthermore, the field of computer analysis of natural language (NLP) is also involved in the formalization of knowledge expressed in language and text [18], [19], [20], [21], [22], [23], [24], [25], [26], [27].

There are two approaches to the formalisation of text-expressed knowledge: declarative and procedural. Examples of declarative approaches are: BG (Basic Conceptual Graphs) [28], MULTINET (Multi-layered extended semantic networks) [29], HSF (Hierarchical Semantic Form) [30], RDF (Resource Description Framework) [31], and other methods based on the first order predicate calculus (FOPC). The result of the procedural approach is a rule system.

III. RESEARCH MOTIVATION

Text as a knowledge carrier is not suitable for computer-supported knowledge dynamics (changes, paraphrasing, distribution, use, generation of new knowledge, abandonment, etc.). It should be specially formatted - formalised - in order to enable a computer to perform mathematical/logical operations. Nodes of Knowledge (NOK) method is a new method by which sentences written in natural language can be modelled. The purpose of graphical modelling sentences with NOK method is the creation of models (formal representation) of different forms of sentences (declarative, interrogatory, simple, complex, etc.) in order to verify the basic concepts of the method. Designed models and their verification represent the basis for the development of mathematical formalism for knowledge representation and reasoning based on a NOK method - first formalism is shown in [6].

The aim of this paper is to model the same text in two languages, English and Croatian, compare derived models and notice similarities and differences in the model that are conditioned by the freedom of translators.

IV. NODES OF KNOWLEDGE

NOK method belongs to a group of semantic networks [11] where knowledge is represented as a graph [32]. Its

aim is to present textual knowledge as a network of knowledge [29].

NOK is a method for modelling by which textual knowledge is represented graphically in the form of a diagram DNOK (Diagram Nodes Of Knowledge) [33]. Sentences written in human language are converted to graphical model from which it is possible to interpret the meaning because this model kept it. NOK method consists of nodes (node, processing node and linker) and connections between them which contains the role identifier and they together form a network of knowledge.

Characteristics of the NOK method are [5]:

- simplicity – it has only three elements: nodes, links and the role identification in link
- expressiveness – ability to represent different levels of abstraction
- simplicity of reading – it is possible to start reading knowledge from any node

A. *Nodes*

Node represents a word or group of words in a sentence that has some meaning. For example, in the sentence "The toy is very old." nodes are: The toy, is, very, old. Graphical representation of nodes and concepts in the NOK method is given in Table I.

Node name is an attribute of node and it gives it semantic meaning. The name may also consist of several words that have a meaning. To implement the method it is necessary to appoint a node name that clearly identifies it in the network. Homonyms and synonyms can be used [32], [35] but should be linked to their originals on contextual level of knowledge.

We can distinguish between [34], [35], [36]:

- Node – ordinary, static, entity, concept, term
- Context node – node of a higher degree of abstraction (group, sort, class, type)
- Data node – place where data are kept
- Process node – node that connects other nodes (using links) and with them creates a more complex presentation of knowledge. It presents knowledge that cannot be displayed by ordinary node such as connections between multiple nodes. The most common types of words in the names of process nodes are verbs.

B. *Links*

Links connect the nodes in the network of knowledge. They do not have names, but may have role identifications that belong to the nodes [35]. The names of role identifications are questions which connect two adjacent nodes. This link is represented with line, with or without arrows. Arrows are used when connecting nodes on different levels of abstraction [33] or for direct connection of process nodes.

The cardinality of a link is always (1, 1) (1, 1) [35].

TABLE I. SYMBOLS OF NOK METHOD

Simbol	Concept
goose	NODE - part of the knowledge in the reality or minds
1. going	PROCESSING NODE - action, event, occurrence
to	LINKER – complex sentence of more parts
What?	LINK (one-way) and QUESTION (role indentifier) – connecting nodes

V. MODELLING OF SENTENCES

Using the NOK method sentences of natural language can be displayed by using nodes and links between them. From the resulting model, without access to the original text, answers to asked questions can be obtained. The applicability of NOK method will be shown in Croatian and English language on the example of the same Aesop's[1] fable „Zlatna jaja u kokoši“ [37]. The fable is titled "The Goose with the golden egg" in English [38].

For modelling sentences from the fables the symbols shown in Table I were used. Specifying the nodes, processing nodes and linkers, when there was a need for them, and connecting with links which were assigned the role and the meaning of nodes with questions in sentence, both graphic representation were constructed, in Croatian (Figure 1.) and in English (Figure 2.).

Fable „Zlatna jaja u kokoši“ in Croatian has the following 4 sentences:

1. „Neki gospodar imao je među mnogim svojim kokošima i jednu takovu koja je nesla zlatna jaja.
2. On je mislio, da joj u trbuhu mora biti mnogo zlata, pa je, ne razmišljajući mnogo, zakolje.
3. Ali se ljuto prevario, ne našavši u njoj ništa drugo nego li i u ostalim kokošima.
4. Tad se uzalud sam korio što je bio lud, da se polakomio za mnogim blagom, pa tako izgubio i ono što je imao.“

The same fable, "The Goose with the golden egg", in English according to [38] has sentences as follows:

1. „One day a countryman going to the nest of his Goose found there an egg all yellow and glittering.
2. When he took it up it was as heavy as lead and he was going to throw it away, because he thought a trick had been played upon him.

[1] Aesop (620 - 560. BC.), slave and storyteller who lived in Ancient Greece

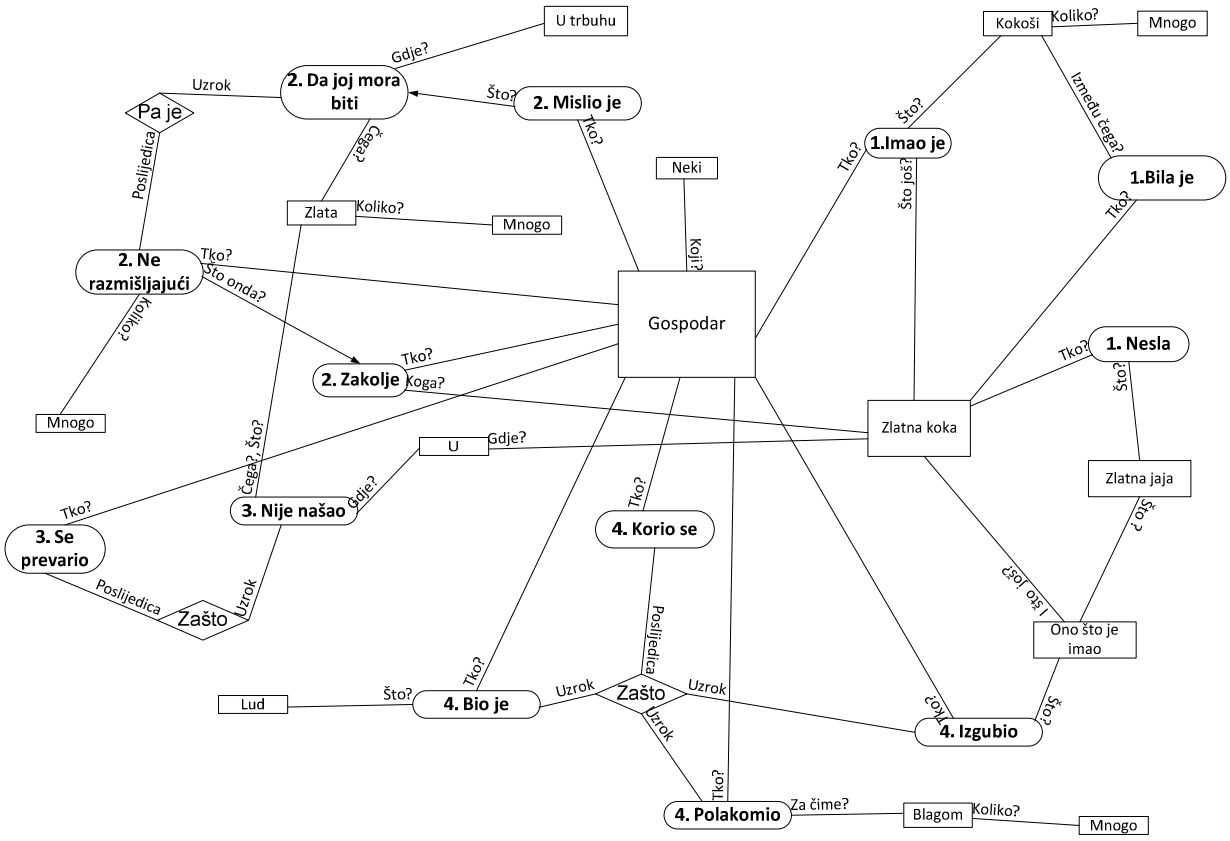


Figure 1. Representation of the fable “Zlatna jaja u kokoši” by NOK method

3. But he took it home on second thoughts, and soon found to his delight that it was an egg of pure gold.
4. Every morning the same thing occurred, and he soon became rich by selling his eggs.
5. As he grew rich he grew greedy; and thinking to get at once all the gold the Goose could give, he killed it and opened it only to find nothing.“

From the number of sentences, or words and groups of words in them, it is obvious that it will be more complex to make diagrams of the English version of the fable as it will consist of more nodes and connections between them.

In complex sentences, and between several separate clauses occurs causal - consequence relationship between certain parts that is displayed by using a linker node "why" which connects the causes and consequences. There is a possibility of multiple causes and / or consequences which determines the complexity of the model.

C. *Interpretation rules*

In the sentence modelling the following rules are applied:

- Each noun becomes a node
- Each verb becomes a processing node
- Each node represents a single term that can consist of a group of words (eg. „zlatna jaja“, „was going to throw“)
- Each node must be linked to at least one node
- Process node is numbered by the number of sentence from which it was taken
- Pronouns are eliminated and the related network is connected with the noun to which eliminated pronouns refer.
- Related actions are connected with a rhombus, and this connection represents cause-and-effect relation.
- Node relationships are identified by questions that determine the semantic connections between nodes.

D. *Comparison*

At first sight it is evident, by comparing the texts of the fable, that there is a difference in the translation which is determined by customs, folk tradition and the environment for which it is intended and, naturally, by the autonomy of translators. The title itself reveals a difference in choosing the animal which lays the golden eggs: it is a hen in the Croatian version and a goose in the English one.

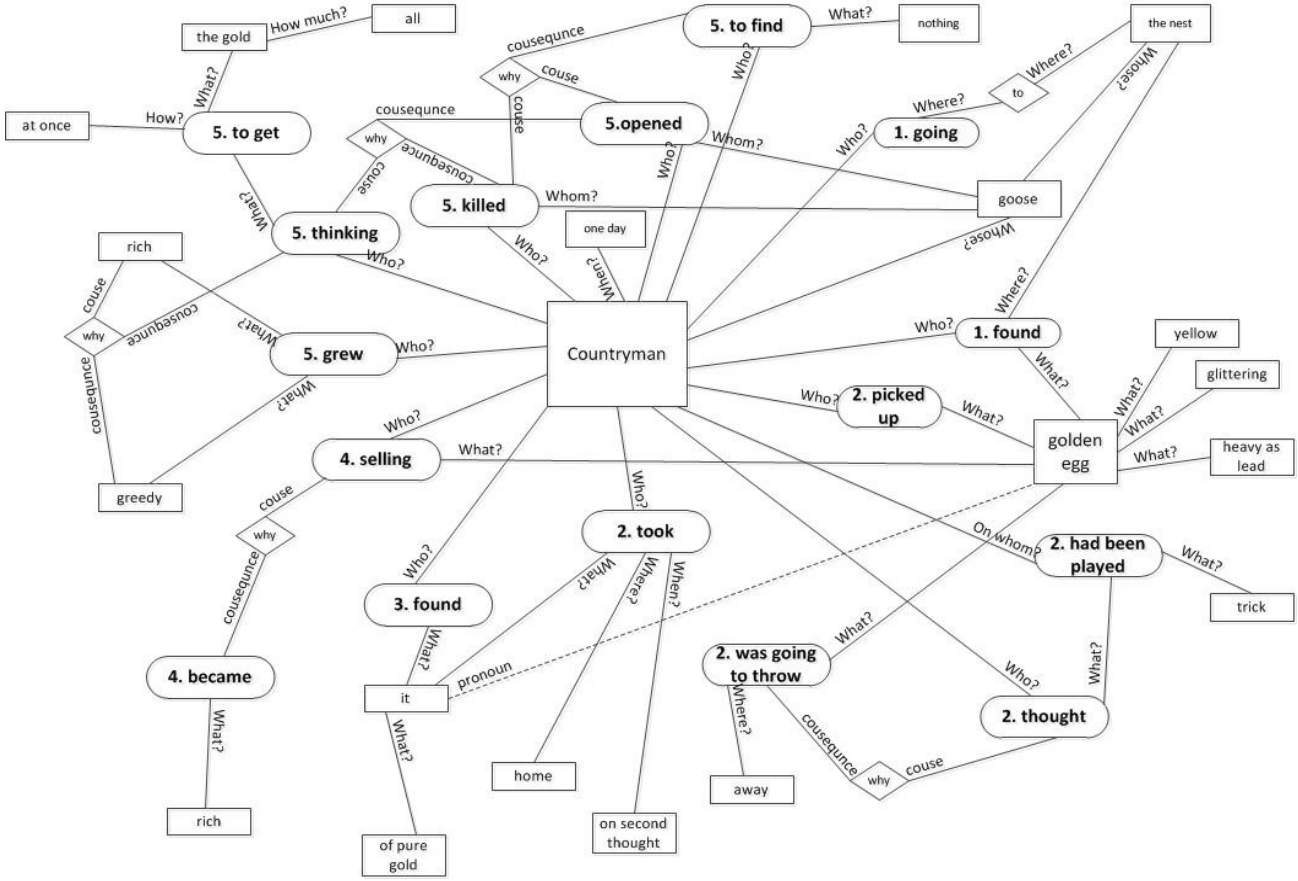


Figure 2. Representation of the fable “The Goose with the golden egg” by NOK method

From this we can conclude that translation is not a mathematically rigorous mapping of terms in one language in the same terms in other language. Also, two different translators will not translate the same text in the same way. Therefore translation does not provide a single best solution. Translators believe that their translation preserves the message and knowledge that the author wanted to convey to the reader.

Despite this, the fable has the same moral and is told in a similar way. This means that the knowledge on the high level of abstraction contained in the fable can be stored in both languages in a different way. This brings us to the conclusion that it is possible to keep text-based representation of the knowledge in a variety of models.

The diagram (Figure 1. and 2.) shows a difference in the presentation. Difference in the complexity of its construction is also evident. English translation of the fable is more complex in terms of the number of nodes and links that are displayed. The English version consists of five, and Croatian version of four sentences. Depending on the number of verbs, ie process nodes, each of these sentences can be divided into simple sentences in which difference becomes even greater. This is evident by the numbering of process nodes on the diagrams (Figure 1. and 2.) that indicate the number of the sentence where the node is.

What is significant for NOK method is getting answers to questions from the both created model. If we ignore the hen/goose difference in translation, we can get similar or the same answers to the same questions from the model, even though the arrangement of nodes in each model is different: for example, who laid the golden eggs?, who was greedy? (Figure 3.), who butchered the hen / goose? etc. On the other hand, there are questions applicable to one model only (for example, English version: who found the golden egg?, who is going to the nest?; Croatian version: who laid the golden egg?, who lost what he had?).

In the graphical representations of fables with NOK method contextual connections between the terms in the text are also recorded. In both versions, there are sentences without explicitly listed content-related terms. For example, in the following part of sentence "he killed it and opened it only to find nothing," one concludes from the context that this refers to the countryman and the goose, and the diagram supports that because the connections between nodes 5. killed, countryman and goose preserved it (Figure 2).

VI. Conclusion

In this paper the same fable in two languages translated by two different authors was presented. For these translations a knowledge network was made, by using the NOK method.

NOK method is applicable to the modelling of a sentence written in a variety of natural human languages.

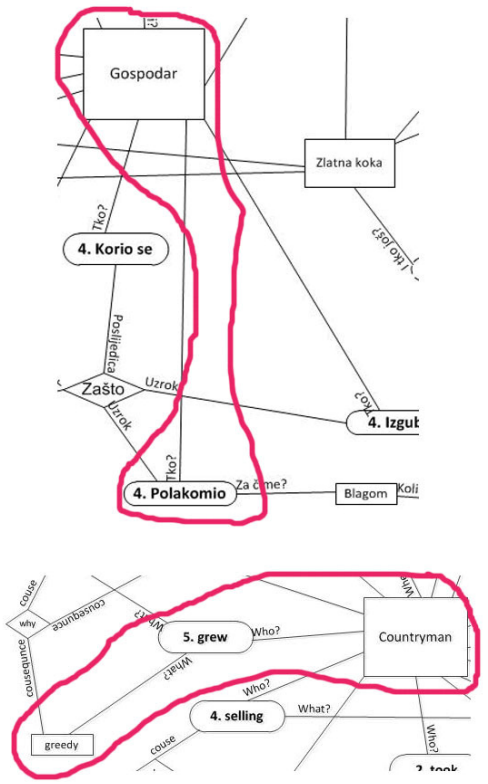


Figure 3. Answer to the question "Who grew greedy?"

Regarding the translator's autonomy, some parts of the model remain similar, but in some parts there are differences in the same story and in the same sentences.

By translating and merging of different models, not just by modelling of translation, we can get the model in different language from which it is possible to get the same meanings.

Choosing texts that are more similarly translated could provide a model with greater overlapping. Furthermore, it is possible to measure the difference in models and translations by a number of used concepts and words. It is possible to explore how NOK method can be used to translate texts and improve the process of translating one language in the DNOK and vice versa. NOK method is a new method for knowledge representation and it yet remains to investigate how it keeps knowledge and what its potential in building a computer system is.

ACKNOWLEDGMENT

The research has been conducted under the project "Extending the information system development methodology with artificial intelligence methods" (reference number 13.13.1.2.01.) supported by University of Rijeka (Croatia).

REFERENCES

[1] Jordan, M. I., & Russell, S., Computational Intelligence. In R. A. Wilson, & F. C. Keil (Eds.), The MIT Encyclopedia of the Cognitive Sciences (lxxiii - xc). Massachusetts: A Bradford book, 2001.

[2] Brachman, R., & Levesque, H., Knowledge Representation and Reasoning. San Francisco: Morgan Kaufmann Publishers, 2004.

[3] F. Harmelen, V. Lifschitz i B. Porter, Handbook of Knowledge Representation, Oxford: Elsevier, 2008.

[4] S. Russell i P. Norvig, Artificial Intelligence: A Modern Approach, Third edition, Prentice Hall, 2010.

[5] M. Ašenbrener Katić, „Predstavljanje znanja: pregled područja", Odjel za informatiku, Sveučilište u Rijeci, kvalifikacijski rad, 2013.

[6] Jakupović, A., Pavlić, M., & Dovedan-Han Z., Formalisation Method for the Text Expressed Knowledge. Expert Systems with Applications. In Press, Accepted Manuscript. from: http://www.sciencedirect.com/science/article/pii/S0957417414001353, 2014., 8.3.2014.

[7] Russell, S. J., & Norvig, P., Artificial Intelligence: A Modern Approach. New Jersey: Pearson Education, 2010.

[8] Sowa, J. F., Conceptual graphs for a data base interface. IBM Journal of Research and Development. 20(4). 336 - 357., 1976.

[9] Sowa, J. F., Semantics of conceptual graphs. Proceedings of the 17th annual meeting on Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics. 39 - 44. 1979.

[10] Sowa, J. F., Knowledge representation: logical, philosophical and computational foundations. Pacific Grove: Brooks/Cole Publishing Co., 2000.

[11] Quillian, R., Semantic memory. Semantic Information Processing. MIT Press. 227 - 270., 1968.

[12] Minsky, M. , Framework for Representing Knowledge. Technical Report. Cambridge. 1974.

[13] McCarthy, J. Programs with common sense. Teddington Conference on the Mechanization of Thought Processes. Teddington: National Physical Laboratory. 75 - 91. 1958.

[14] Gomez-Perez, A., Fernandez-Lopez, M., & Corcho, O., Ontological Engineering. London: Springer., 2004.

[15] Grimm, S., Hitzler, P., & Abecker, A. (2007).Knowledge Representation and Ontologies, Logic, Ontologies and Semantic Web Languages. In

[16] Gruber, T. R., A translation approach to portable ontology specifications. Knowledge Acquisition.5(2). 199 - 220., 1993.

[17] Garcia-Crespo, A., Ruiz-Mezcua, B., Lopez-Cuadrado, J. L., & Gonzalez-Carrasco, I., Semantic model for knowledge representation in e-business. Knowledge-Based Systems. 24 (2). 282 – 296, 2011.

[18] Allen, J., Natural Language, Knowledge Representation and Logical Form. In Bates, M., & Weischedel, R. (1993). Challenges in Natural Language Processing. Cambridge University Press. 1993.

[19] Allen, J., Natural Language Understanding. Boston: Addison-Wesley. 1994.

[20] Allen, J., Swift, M., & Beaumont, W., Deep Semantic Analysis of Text. Proceedings of the 2008 Conference on Semantics in Text Processing. Stroudsburg: Association for Computational Linguistics.343 - 354., 2008.

[21] Chomsky, N., Three models for the description of language. IRE Transactions on Information Theory. Volume 2. 113 – 124, 1956.

[22] Chomsky, N., On certain formal properties of grammars. Information and Control.2(2).137-167. 1959.

[23] Chomsky, N., Studies on Semantics in Generative Grammar. Janua Linguarum Series Minor. Mouton De Gruyter, 1980.

[24] Jurafsky, D., & Martin, J. H., Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. New York: Prentice Hall PTR. 2000.

[25] McQueen, T., Hopgood, A. A., Allen, T. J., & Tepper, J. A., Extracting finite structure from infinite language. Knowledge-Based Systems. 18(4-5). 135-141, 2005.

[26] Shapiro, S. C., & Rapaport, J. W., Knowledge representation for natural language processing, Natural Language Planning Workshop. New York: Northeast Artificial Intelligence Consortium. 56 - 77. 1987.

[27] Wermter, S., Riloff, E., & Scheler, G. Connectionist, statistical and symbolic approaches to learning for natural language processing. New York: Springer-Verlag., 1996.

[28] Chein, M., & Mugnier, M. L., Graph-based Knowledge Representation. London: Springer, 2009.

[29] Helbig, H., Knowledge Representation and the Semantics of Natural Language. London: Springer, 2006.

[30] Stanojević, M., & Vraneš, S., Knowledge representation with SOUL, Expert Systems with Applications: An International Journal, 33(1), 122 - 134., 2007.

[31] Tolle, K. Introduction to RDF, Johann Wolfgang Goethe – Universität Frankfurt am Main. from: http://www.dbis.informatik.uni-frankfurt.de/~tolle/RDF/DBISResources/RDFIntro.html, 2000., 8.3.2013.

[32] M. Pavlić, M. Ašenbrener, A. Jakupović, A. Meštrović, S. Čandrlić i M. Ivašić-Kos, „Inteligentni informacijski sustavi“ u Razvoj poslovnih i informatičkih sustava CASE 25, Rijeka, 2013.

[33] M. Pavlić, Development of a method for knowledge modeling 2013., Rijeka: Odjel za informatiku Sveučilišta u Rijeci, 2013.

[34] M. Pavlić, A. Jakupović, A. Meštrović: “Nodes of knowledge method for knowledge representation”, Informatol. 46, 2013., 3, 206-214

[35] M. Pavlić, A. Meštrović, A. Jakupović, “Graph-Based Formalisms for Knowledge Representation”, Proceedings of the 17th World Multi-Conference on Systemics Cybernetics and Informatics (WMSCI 2013), Vol 2. 2013. 200-204.

[36] A. Jakupović, M. Pavlić i Z. Dovedan Han, „Formalisation Method for the Text Expressed Knowledge“ neobjavljeno, 2013.

[37] Ezop, „Basne“, http://lektire.skole.hr/pdf/ezop/ezop_basne.pdf, 10.12.2013.

[38] Aesop, „The Goose with the golden egg“, http://www.bartleby.com/17/1/57.html, 10.12.2013.

An improved Image Least Significant Bit Replacement Method

Khalid A. Darabkh[1], Iyad F. Jafar[1], Raed T. Al-Zubi[2], and Mohammed Hawa[2]
The University of Jordan
(1) Computer Engineering Department
(2) Electrical Engineering Department
Amman, 11942, Jordan
Emails: {k.darabkeh, iyad.jafar, r.alzubi, hawa}@ju.edu.jo

Abstract— **With the development of internet technologies and communication services, message transmissions over the internet still have to face all kinds of security problems. Hence, how to protect secret messages during transmission becomes a challenging issue for most of the researchers. It is worth mentioning that many applications in computer science and other related fields rely on steganography and watermarking techniques to ensure information safety during communication. In this paper, we propose a new steganographic method to embed the secret data inside a cover image based on least-significant-bit (LSB) replacement method. The embedding process predominantly concentrates on distributing the secret message inside one share of a color image to appear like a 3D geometric shape. The dimensions of the geometric shape are variable pursuant to the size of secret message. Data distribution process makes our method to be of a great interest as of being so difficult for the hackers or intruders to reconstruct the shape from stego-images, thereby the security is improved. Furthermore, we compare the performance of our approach with two other relevant approaches in terms of peak signal-to-noise ratio (PSNR). The contribution of our approach was immensely impressive.**

Index Terms— **Cover image, stego-image, geometric shape, least-significant-bit (LSB), security, imperceptibility**

I. INTRODUCTION

Due to the fast growth of internet applications, digitized data becomes very prevailing bearing in mind that the internet does not use secure links. Hence, data needs to be transmitted may be vulnerable to interception as well. As indicators of its importance, intruders may not only reveal transmitted data to others, but also modify it to give a false idea of an individual or organization or even use it to launch attacks. Owing to the easiness of digital duplication and tampering, data security becomes a challenging matter for most of researchers nowadays. One of great interest solutions is the use of information hiding which encompasses many sub-disciplines, out of which steganography and watermarking [1-7].

There are many watermarking techniques proposed in the spatial domain [8-23]. Among these implemented watermarking techniques, the following are some of interesting and relevant schemes: Pixel-value differencing (PVD) and PVD cascaded with least-significant-bit (LSB). In PVD approach [8], the cover image is simply portioned into non-overlapping two-pixel blocks. The difference between each two consecutive pixels of any block determines how much data can be embedded. In other words, small and large differences denote for working at pixels located in smooth and edge image regions and tolerate small and large changes in their values, respectively. The authors proposed a range table that contains all possible differences in the whole image region in which the selection replies on the sensitivity of human vision to gray value variations. The main goal of doing this is to keep the resulting stego-images unnoticeable. For interested readers, further details can be found in [8]. In [9], both PVD and LSB substitution method have been cascaded for the reason of obtaining more hiding capacity. After embedding using PVD approach, the idea proposed in [9] simply encompasses embedding using LSB method in the first pixel of each pixel pair (block). The distortion, made by LSB method, is then adjusted through modifying the second pixel of that block in a way that the difference between pixels, calculated from PDV approach, is kept.

Taking advantage of LSB scheme, we propose, in this paper, a very efficient watermarking approach. In this approach, the watermark is constructed on the form of a geometric shape that appears like a 3D house in which its dimensions are variable and depend on the size of secret data. The media used is colored images. To the best of our knowledge, this is brand new idea and no prior work similar to ours. However, the house geometric shape is divided into four areas where all are included in the embedding process as will be discussed shortly. For comparison purposes, the red share is used each time to embed the data according to the pixel area bearing into consideration that our approach is flexible to accept incorporating all shares in the embedding process. It is noteworthy to mention that our approach fulfills the security watermarking evaluation objective. In other words, the house geometric shape makes the emission of the attacker harder due to the endless effort needed to find the equations of the house required to find the watermark. The secret bits per pixel (embedding depth) are chosen up to five bits and this is to

conserve the transparency and capacity evaluation objectives. Apart from that, our scheme falls within spatial domain category. Hence, the payload (hiding capacity) is considered to be high compared to those schemes proposed in either transform or quantization domains. Due to changing only one share byte of the pixel at a time and having the size of the house variable pursuant to the size of watermark, the imperceptibility (i.e., the closeness of visual quality between the original and watermarked images) is achieved. As no need, in our watermarking approach, to convert the image into different domains (frequency/quantization) or different spaces (e.g., gray-scale) and there is no heavy computational operations such as multiplication and division required on pixels, the computational time is low. Furthermore, since there is no need for the original image or its features to be present in the exaction phase, our watermarking technique is blind. Additionally, our watermarking technique is a standalone since it is not combined with another transform- or quantization-based watermarking algorithm which makes it easier to improve and modify.

The remainder of the paper is organized as follows. The proposed watermarking approach is elucidated in Section II. Experimental results along with observations and necessary discussions are presented in Section III. The paper is summarized in Section IV.

II. Proposed Approach

The secret data is embedded in color images in the guise of a 3D geometric shape. The embedding occurs in the least significant bits of the red share. The size of this geometric shape is variable according to the secret data size and embedding depth which represents the number of secret bits that need to be inserted and is chosen to be five to maintain the imperceptibility. In order to have a variable shape of the watermark inside a digital image, we should be aware of the following points:

1. Finding the number of pixels needed, which is equal to the ceiling of watermark size (in bits) divided by the embedding depth,
2. Finding the areas entailed to hide the watermark where the overall area indicates the number of needed pixels found in point 1. As seen in Fig. 1, the variable shape is divided into four different areas where each area is calculated with respect to $Area_1$ as the following: $Area_2 = 0.25 * Area_1$, $Area_3 = 0.25 * Area_1$, as well as $Area_4 = 0.125 * Area_1$. It is important to mention that the factors of $Area_1$ (i.e., 0.25, 0.25 and 0.125) are selected according to a specific engineering model that takes into consideration maximizing most of performance objectives, mentioned previously, required for evaluation. Consequently, we can embed in $Area_1$ up to (8/13) times the number of pixels required.

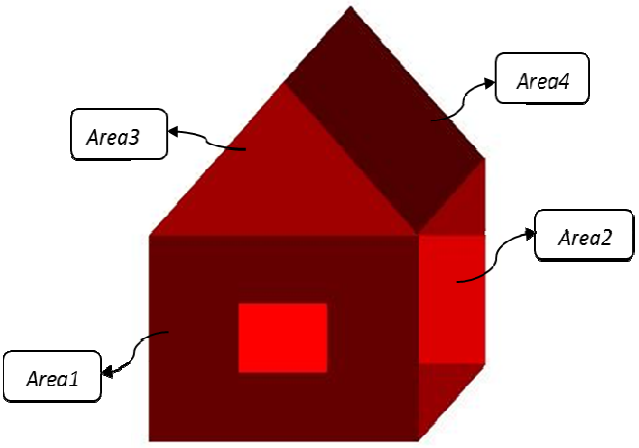


Figure 1. The geometric shape of proposed approach

A. The embedding procedure

On account of being easily applied for gray-scale images and comparing its performance with other pertinent proposed watermarking algorithms, we consider in our work embedding in just one share of the image which is settled on red bearing in mind that involving all images shares may be better for enhancing the security objective. However, Fig. 2 describes the strides entailed to construct the geometric shape that professes to be the watermark. Interestingly, the pixels located in $Area_2$ and $Area_4$ will involve embedding in the semblance of a parallelogram. Similarly, the shapes of triangle and rectangle will be formed if the pixels located in $Area_3$ and small and big windows of $Area_1$, respectively. To improve the security, we start embedding the secret data in the small window and the rest will be resumed in the big window but in the opposite direction. After that, the embedding will recommence in $Area_2$, $Area_3$, and $Area_4$ in sequence. It is good to stress on the point made to improve the imperceptibility that if the secret data is embedded in any pixel of the red share, then the same pixel at the corresponding positions of the other shares will not be affected (immutable). However, after embedding the secret data, the resulting shape will look like a 3D house as shown down in Fig.2. However, the geometric shape is invisible in the watermarked image.

III. Results and Discussions

Our experiments have been conducted using Matlab version 2010R on a PC with an Intel Core i5-2400 CPU working at 3.1 GHz with a 4MB cache and 4 GB RAM. In our experiments, five cover images, which include *Lena*, *Peppers*, *House*, *Couple*, and *Baboon*, were used, each with a size of 512x512. To evaluate the performance of our proposed watermarking scheme to be then compared with what are closely connected in the literature, namely PVD and PVD

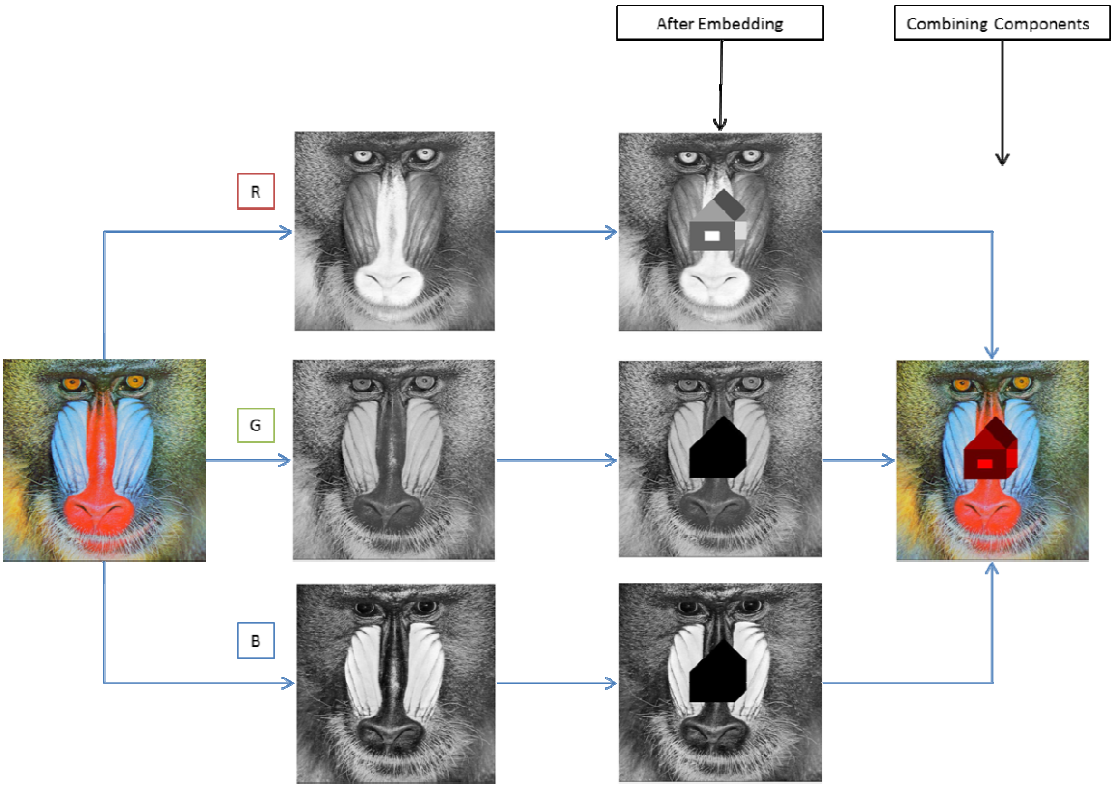


Figure 2. Embedding using the proposed geometric shape

cascaded with LSB, we consider peak signal-to-noise ratio (PSNR) performance metric. It deserves disclosing that for the reason of having a fair comparison, we implement the former approaches on the red share of color images precisely as described. However, PSNR is a very popular function for evaluating the imperceptibility or visual quality of the watermarked result [8, 21].

Fig. 3 shows the results of PSNR performance metric versus different watermark lengths when experimenting (simulating) our proposed approach and two other popular approaches, that is to say, PVD [8] and PVD cascaded with LSB [9]. The host image used is *Couple*. It is worth stating that higher obtained values of PSNR indicate having better results. In other words, as the difference between original and stego-image gets closer and closer, the PSNR gets higher and higher. As seen in this figure, the PSNR values decrease as the watermark size increases. This is to be expected since the number of characters embedded increases. Unlike other proposed approaches, our proposed approach achieves higher values of PSNR for all described watermark lengths. With reference to what are mentioned about PVD and PVD cascaded with LSB in Section I, it is noteworthy to realize and conclude that these approaches may affect the stego-images more than our approach. To show its consistent contribution over other standard test images, Fig. 4 shows the PSNR results for a different cover image (i.e., *Baboon*). It is worth noting that the PSNR values for PVD are higher than those for PVD cascaded with LSB and these completely comply with what are reported in [9].

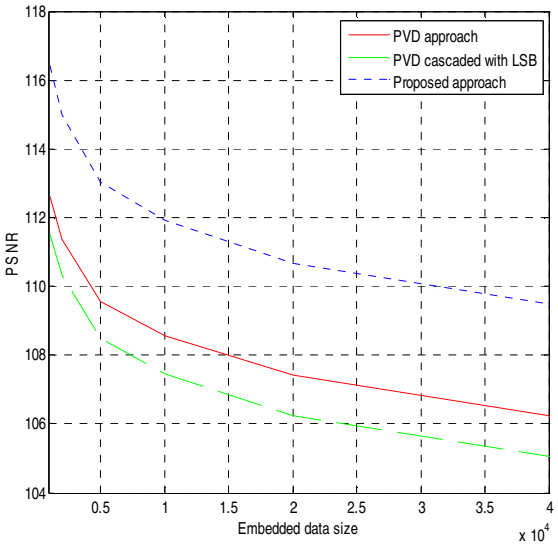


Figure 3. PSNR versus watermark size (in bytes) for different watermarking methods using cover image "*Couple*"

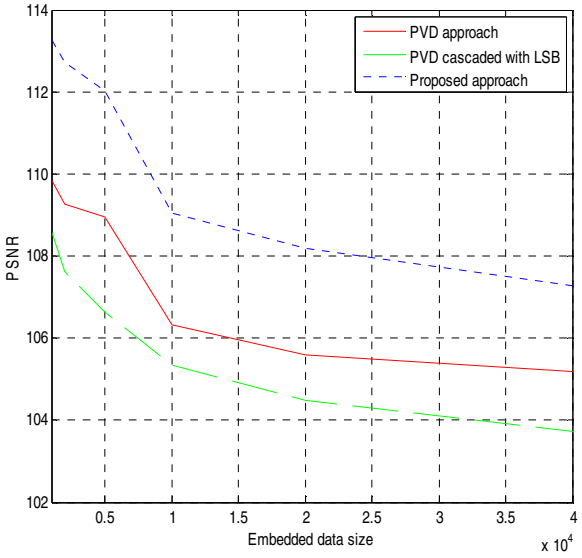


Figure 4. PSNR versus watermark size (in bytes) for different watermarking methods using cover image "*Baboon*"

IV. Conclusion

As the digital technologies over the internet increasingly evolving, the need for digital watermarking to protect the owner's right becomes urgent. In this paper, we propose an efficient watermarking approach based on least significant bits replacement method. Our approach mainly embeds any secret data into a cover image on the form of a geometric shape which looks to observers as a 3D house. This shape is divided into different areas where each has various dimensions and embedding policy. The importance of our approach comes in improving the security and coding time watermarking evaluation objectives. The security is enhanced since too much effort is required to break our algorithm and extract the watermark. Furthermore, its computational complexity causes as little burden on the overall system since no arithmetic operations or translations are required for image pixels. Our proposed approach does not stop to this extent, but rather it improves the imperceptibility and capacity watermarking evaluation requirements. The closeness between cover and stego-images was extremely impressive. Two former algorithms (PVD and PVD cascaded with LSB approaches) were implemented and evaluated to be compared with our approach in terms of PSNR metric. Genuinely, our approach outperforms both of them.

Acknowledgment

The authors would like to thank the students who have participated in developing the system and making this work possible, particularly, Hala Ismail, Dana Hijaz, and Rawan Amayreh.

References

[1] J. Franco-Contreras, G. Coatrieux, F. Cuppens, N. Cuppens-Boulahia, and C. Roux, "Robust Lossless Watermarking of Relational Databases Based on Circular Histogram Modulation," *IEEE Transactions on Information Forensics and Security*, vol.9, no.3, pp.397-410, March 2014.

[2] S. Wang, D. Zheng, J. Zhao, W.J. Tam, and F. Speranza, "Adaptive Watermarking and Tree Structure Based Image Quality Estimation," *IEEE Transactions on Multimedia,* vol.16, no.2, pp.311-325, February 2014.

[3] K. Zebbiche and F. Khelifi, "Efficient wavelet-based perceptual watermark masking for robust fingerprint image watermarking," *IET Image Processing,* vol.8, no.1, pp.23-32, January 2014.

[4] Yafeng Zhou and W.W.Y. Ng, "A study of influence between digital watermarking and steganography," *In Proceedings of the 2013 International Conference on Wavelet Analysis and Pattern Recognition*, Tianjin, pp.49-55, July 2013.

[5] Cameron Whitelam, Nnamdi Osia, and Thirimachos Bourlai, "Securing multimodal biometric data through watermarking and steganography," *In Proceedings of 2013 IEEE International Conference on Technologies for Homeland Security (HST),* Waltham, MA, USA, pp.61-66, November 2013.

[6] B. E. Carvajal-Gamez, F. J. Gallegos-Funes, V. I. Ponomaryov, and R. Cruz-Santiago, "A new steganographic method for RGB color images using estimation of variance field in the wavelet domain," *In Proceedings of 2013 International Kharkov Symposium on Physics and Engineering of Microwaves, Millimeter and Submillimeter Waves (MSMW),* Kharkov, Ukraine, pp.626-628, June 2013.

[7] Hamdy M. Kelash, Osama F. Abdel Wahab, Osama A. Elshakankiry, and Hala S. El-sayed, "Hiding data in video sequences using steganography algorithms," *In Proceedings of 2013 International Conference on ICT Convergence (ICTC),* Jeju, pp.353-358, October 2013.

[8] Da-Chun Wu and Wen-Hsiang Tsai, "A steganographic method for images by pixel-value differencing," Pattern Recognition Letters, vol. 24, no. 9–10, pp. 1613-1626, June 2003.

[9] J. K. Mandal and Debashis Das, "A Novel Invisible Watermarking Based on Cascaded PVD Integrated LSB Technique," *In Proceedings of the 17th International Conference on Engineering of Complex Computer Systems (ICECCS 2012)*, Kochi, India, August 9-11, 2012, book chapter in Eco-friendly Computing and Communication Systems, Series: Communications in Computer and Information Science, vol. 305, Springer, pp. 262-268, 2012.

[10] Mohammed Younes and Aman Jantan, "A New Steganography Approach for Images Encryption Exchange by Using the Least Significant Bit Insertion," *International Journal of Computer Science and Network Security*, vol. 8, no. 6, pp.247-257, June 2008.

[11] Shwu-Huey Yen and Chia-Jen Wang, "SVM Based Watermarking Technique," *Tamkang Journal of Science and Engineering*, vol. 9, no 2, pp. 141-150, 2006.

[12] Jamal A. Hussein, "Spatial Domain Watermarking Scheme for Colored Images Based on Log-average Luminance," *Journal of Computing*, vol. 2, no. 1, January 2010.

[13] Rupinder Kaur, Mandeep Kaur, and Rahul Malhotra, "A New Efficient Approach towards Steganography," *International*

Journal of Computer Science and Information Technologies, vol. 2, no. 2, pp. 673-676, 2011.

[14] Yashar Naderahmadian and Saied Hosseini-Khayat, "Fast and robust watermarking in still images based on QR decomposition," *Multimedia Tools and Applications,* In Press.

[15] W. Bender, D. Gruhl, N. Morimoto, and A. Lu, "Techniques for data hiding," *IBM Systems Journal*, vol. 35, no. 3–4, pp.313–336, 1996.

[16] Y. K. Lee and L. H. Chen, "High capacity image steganographic model," *IEE Proceedings on Vision, Image and Signal Processing*, vol. 147, no.3, pp. 288-294, 2000.

[17] R.-Z. Wang, C.-F. Lin, and J.-C. Lin, "Image hiding by optimal LSB substitution and genetic algorithm," *Pattern Recognition*, vol. 34, no. 3, pp. 671–683, March 2001.

[18] C.-C. Chang, J.-Y. Hsiao, and C.-S. Chan, "Finding optimal least-significant-bit substitution in image hiding by dynamic programming strategy," *Pattern Recognition*, vol. 36, no. 7, pp. 1583-1595, July 2003.

[19] C.-K. Chan and L. M. Cheng, "Hiding data in images by simple LSB substitution," *Pattern Recognition*, vol. 37, no. 3, pp.469-474, March 2004.

[20] W.-N. Lie and L.-C. Chang, "Data hiding in images with adaptive numbers of least significant bits based on the human visual system," *In Proceedings of IEEE International Conference on Image Processing (ICIP 99)*, vol. 1, Kobe, pp. 286–290, October 1999.

[21] Ko-Chin Chang, Chien-Ping Chang, Ping S. Huang, and Te-Ming Tu," A Novel Image Steganographic Method Using tri-way Pixel-Value Differencing," *Journal of Multimedia*, vol. 3, no. 2, June 2008.

[22] H.-C. Wu, N.-I. Wu, C.-S. Tsai, and M.-S. Hwang, "Image steganographic scheme based on pixel-value differencing and LSB replacement methods," *IEE Proceedings on Vision, Image and Signal Processing*, vol. 152, no. 5, pp.611-615, 2005.

[23] S.-L. Li, K.-C. Leung, L.-M. Cheng, and C.-K. Chan, "Data Hiding in Images by Adaptive LSB Substitution Based on the Pixel-Value Differencing," *In Proceedings of the First International Conference on Innovative Computing, Information and Control (ICICIC'06)*, vol. 3, Beijing, pp. 58-61, August 2006.

New Recognition Methods for Human Iris Patterns

Khalid A. Darabkh[1], Raed T. Al-Zubi[2], and Mariam T. Jaludi[1]
(1) The University of Jordan/Computer Engineering Department, Amman, 11942, Jordan
(2) The University of Jordan/ Electrical Engineering Department, Amman, 11942, Jordan
Emails: {k.darabkeh, r.alzubi}@ju.edu.jo, mariam.jaludi@gmail.com

Abstract— **The use of user identification technology has become increasingly in demand in today's society. Biometrics have been capitalized upon for this purpose. The human iris in particular is one of the most unique and intriguing biometrics available to use in the identification of an individual. The process of recognizing a human iris is split into four major steps. These steps are segmentation, normalization, feature extraction, and matching. In this paper, two new methods are introduced to implement feature extraction. Both methods use a sliding-window technique and different mathematical operations on the pixels to produce feature vectors. Experimental results of the methods produced relatively small feature vectors of size 5x120 and 5x130. A small feature vector is valuable as it contributes to the efficiency and speed of the overall recognition system. In addition, a step was included in both methods to minimize the effect of varying light intensity. This reduces the time needed to acquire an image with suitable lighting, which in turn contributes to the speed of the system as well. Analysis of our methods was done by considering various performance metrics such as the False Acceptance Rate (FAR), False Rejection Rate (FRR), and Recognition Rate (RR). Both proposed methods achieved a recognition rate of about 98.3264%.**

Index Terms — **Mean thresholding, Mean-by-median thresholding, Biometrics, Iris recognition rate, Feature vector, FAR, FRR**

I. INTRODUCTION

A biometric is a trait or characteristic of the human body that can be used to uniquely identify a person [1]. Biometrics have been put into practice for a variety of applications. The most common ones used include fingerprints, voice patterns, DNA (deoxyribonucleic acid), or iris patterns [2]. The objective of a biometric system is to be able to represent this unique trait or characteristic into a template. The template is then compared with others in order to identify a match or an imposter [3]. Biometrics systems are further classified into either "Identification Systems" or "Verification Systems". An identification system is one that blindly takes the generated template of a biometric and tries to identify the person it belongs to [4]. A verification system is one that requires an additional input such as a username. The username is linked to a specific template with which the input template is compared to [2, 4].

In security, biometrics are favored over traditional security methods such as keys, PINs (personal identification numbers), and passwords, because for many reasons. Biometrics are robust with time and are always available to a person to use unlike keys or passwords which can be lost or forgotten. Imposters also have an easier time of creating copies of keys or cracking passwords than they would trying to mimic a biometric.

In this paper, two methods for extracting features from the human iris are introduced. The advantages of using the human iris lead its selection over other biometrics. The iris is a thin, colored, circular structure in the eye [5]. The textural details of the iris are unique to every human being and they remain stable throughout a person's life. The function of the iris is to manipulate the diameter and size of the pupil to control the amount of light that passes through to the retina. The iris responds to light by contracting or dilating muscles attached to the aperture at the center of the iris, known as the pupil [5, 6]. Analysis of the iris can be done by acquiring a high resolution image of the iris. DNA, on the other hand, would introduce greater difficulty to analyze and would require more time to do so [7]. Asking a person for a picture is also a more user friendly approach than asking for a saliva swab or blood sample to analyze DNA. Voice patterns have been shown to be less robust with time than other biometrics as a person ages [8].

One of the goals of our feature extraction methods was to create small feature vectors. The reason for this is because small feature vectors contribute to a faster overall system. It also reduces the amount of memory needed to save these vectors in a database. Another problem tackled by our methods was variable light intensity. It is difficult to maintain the same levels of light intensity when acquiring the images for iris recognition. Rather than setting up hardware to solve this problem, we have solved it in our algorithms. This solution contributes to the accuracy of identification.

The remainder of the paper is organized as follows: Section II provides the related work. The proposed methods are elucidated in Section III. Experimental results along with observations and necessary discussions are presented in Section IV. The paper is summarized in Section V.

II. Literature Review

Iris recognition began with Dr. John Daugman, who created the first algorithm to be used to extract features of the human iris [9]. Daugman observed that 2D Gabor filters showed similar frequency and orientation to those of the human visual system [10, 11]. He used them as bandpass filters [12] and modulated them using an elliptical Gaussian function [11] in order to implement his iris recognition algorithm. Other work in iris recognition showed the value of working on individual signal frequency components. Wavelets are used to decompose a continuous-time signal into a set of signals with different resolution levels [13]. Haar wavelets [4, 14] and the zero-crossings of 1-D wavelets are examples of wavelets that have been used in iris recognition. Haar wavelets take the iris image and decompose it into 4 subimages by using 4 different combinations of high and low pass filters [3]. The images are then encoded into binary feature vectors. Zero-crossings of the 1D wavelet are extracted when a signal is decomposed into a set of signals [15]. They are represented using sets of complex numbers to create a feature vector. There has also been work in the time domain to extract features, by working with filters directly on an image and using the resulting values as feature vectors [16]. A method called the difference conversion to complex numbers presented in [17] operates on the pixels in the time domain and presents their values in a form that can be encoded into feature vectors.

III. Proposed Approaches

To create feature vectors, mathematical operations are performed on the input image and the results are used to create the template. A feature extraction method's performance is judged based on how well it can uniquely represent each individual's iris patterns. If every person's iris can be mapped into their own specific feature vector values, this will lead to a highly accurate and fast matching process.

A feature vector needs to be compared to the database of a system in order to identify to whom it belongs to. This database contains the feature vectors of the people allowed access to the system. The type of comparison a system can make is split into two categories: identification and verification. An identification system requires only the input image. It uses this image to compare it with each saved template in the database in order to find a match. A verification system on the other hand does a more specific search. It requires the additional input of a username or something similar in order to specify which part of the database it will search through. For example if a username of "John" is entered along with the iris image, only feature vectors belonging to user "John" are compared with the input image to verify that this person is indeed "John". A good matching algorithm applies both identification and verification, making it more robust and applicable in different situations.

The feature extraction process is done once the image of the iris is normalized. In our methods, the top right part of the iris is used when the iris image is unwrapped using Daugman's Rubber Sheet model for normalization [11]. The differences between pixels were operated on rather than direct operations on the pixels. This is to account for any light intensity changes when the images of the iris were acquired. The proposed methods use "sliding windows." A sliding window means the algorithm is repeatedly applied per pixel and uses the adjacent pixels that fall within the perimeter of this "window." The algorithm is blind to any pixel outside of the window and only sees them once the window is shifted to the next pixel.

A. Method 1: Mean Thresholding

The normalized image section taken is 5x130. A 1x11 sliding window is used on each row, beginning with column 6 up to column 125. The procedure of performing this method includes the following steps in sequence:

1. Center the window at column 6 in the first row.
2. Take the 5 preceding pixels and 5 succeeding pixels.
3. Find the mean value of these pixels and set as the threshold.
4. Test the centered pixel. If the value of the pixel is greater or equal to the threshold, set the pixel as 1, else set it to 0.
5. Repeat the steps until column 125.

The size of the resulting binary feature vector is 5x120.

B. Method 2: Mean-by-Median Thresholding

The section taken from the normalized image is 5x130. A 1x25 sliding window is used from the first column to the last. The following steps summarize how this method precisely works:

1. Center the window at the first pixel in the first row.
2. The last 12 columns from the original normalized iris image are concatenated to the beginning of the section taken.
3. Calculate the mean of the 12 preceding and succeeding pixels.
4. Calculate their median.
5. Calculate the threshold using the equation:

$$Threshold = \sqrt{mean \times median}. \quad (1)$$

6. If the centered pixel is greater than the threshold, set it as 1, else set it as 0.
7. Shift the window to the next pixel and repeat steps 3 through 7 until pixel 130 is reached in the row.
8. Concatenate the first 12 columns from the original normalized iris image to the end of the section taken.
9. Repeat steps 3 through 6 for the pixel.

10. Begin with the next row and repeat steps 2 through 9 until all rows are complete.

The resulting binary feature vector size is 5x130.

Both methods produce binary feature vectors. The Hamming distance matching technique was chosen because it operates best with binary templates. It defines the "dissimilarity" between biometric templates. This means it finds the number of corresponding bits that differ between to templates. Comparison of iris templates was split into two categories:

1. Intra-class comparisons: comparing the input image with templates belonging to a certain person.
2. Inter-class comparisons: comparing the input image with the entire database of users.

During image acquisition, the iris may not be positioned properly and experience rotation. This was accounted for by creating 8 circularly rotated templates of the input image. A total of 9 templates are then compared with the database. This was achieved by circular shifting template columns 4 times to the right and 4 times to the left. With the 9 templates, 9 matching ratios are produced using the Hamming distance formula. The largest matching ratio is the only one taken from among the 9. Once the matching ratio for intra-class and inter-class are calculated, they are used to find the threshold between the two distributions.

IV. Results and Discussions

Two separate codes were constructed with identical segmentation, normalization, and matching techniques, only differing in the feature extraction method. These codes were written using MATLAB® software and the database used for iris images was the CASIA version 1.0 database. Only a subset of 693 images (99 classes) was used from the CASIA database to test our methods. The codes implemented both identification and verification. In the identification mode, the input image was compared with every iris template in the database. The comparison that yielded the highest matching ratio was chosen to be assigned to that person. In verification mode, the pattern that is verified is only matched against a certain user's saved templates or his "class". Matching ratios are generated and the maximum is chosen for that person. To prevent imposters from passing through the system, a minimum level was set for the matching ratio. If the maximum matching ratio found did not exceed this level, it is rejected and no match is found.

Table I compares the recognition rates achieved by our methods with other methods. It shows that the results achieved by our methods are of high accuracy. Our recognition rate for both methods was 98.3264%. The recognition rate is defined as the ratio of the number of input images being correctly recognized to the total number of input images entered into the system.

Our methods produced accurate with a vector of size 600 pixels (5x120) and 650 pixels (5x130). A smaller feature vector contributes to the speed of the system and reduces the amount of memory needed to store them. Table II shows a comparison between the vector sizes of our templates with other methods published. The contribution of our proposed methods is undeniable and very magnificent.

The False Rejection Rate (FRR), or type I error, measures the probability that an individual who has been registered into the system is not identified by the system. FRR can be calculated by:

$$FRR = \frac{\int_T^1 P_{intra}(x)dx}{\int_0^1 P_{intra}(x)\,dx}, \tag{2}$$

where, $P_{intra}(x)$ and T represent the distribution of the intra-class comparisons and threshold between intra- and inter-class distributions, respectively.

The False Acceptance Rate (FAR), or type II error, measures the probability that an individual who may or may not have been registered into the system is identified as another individual. FAR can be calculated by:

$$FAR = \frac{\int_0^T P_{inter}(x)dx}{\int_0^1 P_{inter}(x)\,dx}, \tag{3}$$

where, $P_{inter}(x)$ represents the distribution of the inter-class comparisons.

The intra- and inter- class distributions of method 1 are shown in Fig. 1. Fig. 2 shows the distributions of method 2. The thresholds T were determined experimentally by finding FAR and FRR and choosing the T that minimizes these metrics for each method. Table III shows the thresholds found for our methods.

TABLE I. Recognition Rate Comparison

Method	Recognition Rate
Mean Thresholding	98.3264 %
Mean-by-Median Thresholding	98.3264 %
Proenca (2006)	98.02%
X. Liu (2005)	97.08%
Sung (2004)	94.54%
Camus (2002)	98.0%

TABLE II. Feature Vector Size Comparison

Method	Feature Vector Size
Mean Thresholding	600
Mean-by-Median	650

Thresholding	
Daugman (1994)	2048
Daugman (2004)	2048
Masek (2003)	9600
Addison (2002)	700

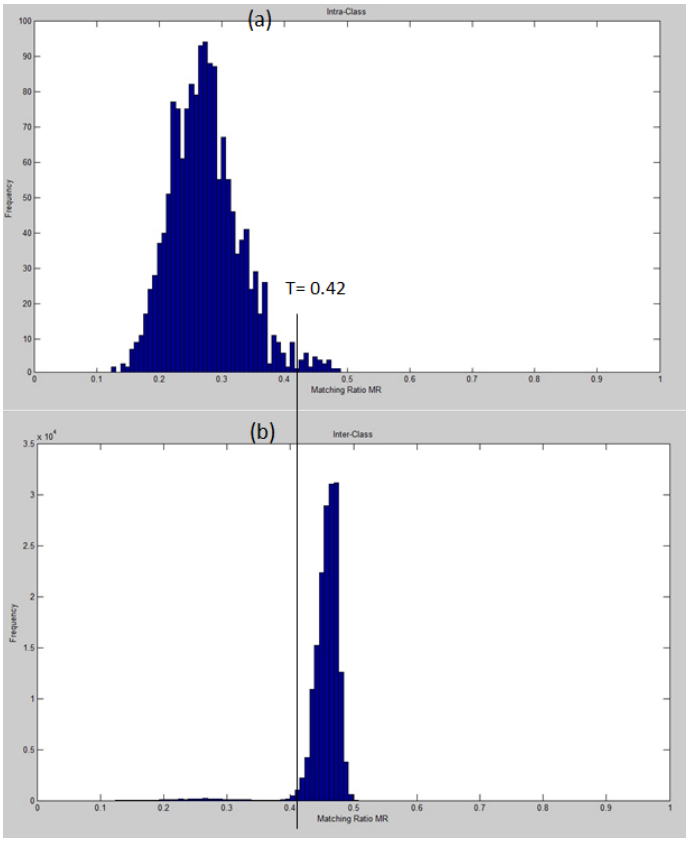


Figure 1: Method 1 distributions: (a) Intra-class distribution (b) Inter-class distribution

TABLE III. THRESHOLDS CHOSEN THAT MINIMIZE FRR AND FAR

Method	**FRR %**	**FAR %**	**Threshold**
Mean Thresholding	0.322	0.039	0.42
Mean-by-Median Thresholding	0.281	0.0389	0.42

V. CONCLUSION

Two different approaches have been introduced to extract features from the human iris. Both have shown to have created small templates. Small templates improve the overall speed and reduce the amount of memory needed to save the templates produced. The effect of light intensity was eliminated during the process of feature extraction so that additional hardware to fix this problem is no longer needed. The methods also achieved high recognition rates. We can conclude from this that our implementation is practical and that an accurate iris recognition system can be created using these methods.

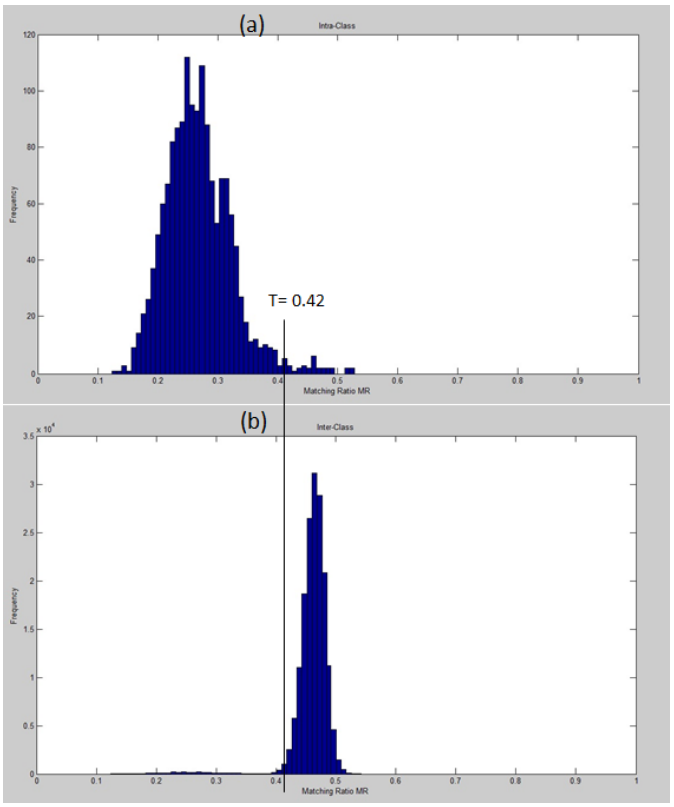


Figure 2: Method 2 distributions: (a) Intra-class distribution (b) Inter-class distribution

ACKNOWLEDGMENTS

The authors would like to thank the student, Hind Al-Kurdi, for her help in developing the proposed algorithms.

REFERENCES

[1] A. K. Jain, A. Ross, and S. Prabhakar, "An introduction to biometric recognition", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no.1, pp. 4-20, January 2004.

[2] A. Ross and A. K. Jain, "Multimodal Biometrics: An Overview," *Proceedings of 12th European Signal Processing Conference*, pp. 1221-1224, Vienna, Austria, September 2004.

[3] A. K. Jain and A. Ross, "Introduction to Biometrics", in Handbook of Biometrics, A. K. Jain, P. J. Flynn and A. Ross (Eds.), pp. 1–22, Springer, 2007. ISBN: 978-0-387-71040-2.

[4] S. Lim, K. Lee, O. Byeon, and T. Kim, "Efficient iris recognition through improvement of feature vector and classifier," *ETRI Journal*, vol. 23, no. 2, pp.61-70, June 2001.

[5] Eye, human, "Encyclopedia Britannica" from Encyclopedia Britannica Ultimate Reference Suite DVD, 2006.

[6] E. Wolff, *Anatomy of the Eye and Orbit*, 7th edition, H. K. Lewis & Co. LTD, 1976.

[7] J. Bienvenue and J. P. Landers, "Rapid DNA Analysis," *Proceedings of 2010 NSF Workshop on Fundamental Research Challenges for Trustworthy Biometrics*, Washington, DC, November 2010.

[8] A. Ross, "Iris Recognition: The Path Forward," *Computer*, vol.43, no.2, pp.30-35, February 2010.

[9] J. Daugman, "Biometric personal identification system based on iris analysis," *U.S. Patent US5291560 A*, March 1994

[10] D. A. Clausi and M. E. Jernigan, "Designing Gabor Filters for Optimal Texture Separability," *Pattern recognition,* vol. 33, no. 11, pp.1835-1849, November 2000.

[11] J. Daugman, "How Iris Recognition Works," *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 14, no. 1, pp. 21-30, January 2004.

[12] A. K. Jain and F. Farrokhnia, "Unsupervised Texture Segmentation Using Gabor Filters," *Proceedings of IEEE International Conference on In Systems, Man and Cybernetics,* pp. 14-19, Los Angeles, CA, November 1990.

[13] P. Addison, *The Illustrated Wavelet Transform Handbook: Introductory Theory and Applications in Science, Engineering, Medicine and Finance*, Taylor & Francis, July 2002.

[14] A. G. Panganiban, N. B. Linsangan, and F. S. Caluyo, "Implementation of Wavelet Transform-Based Algorithm for Iris Recognition System," *International Journal of Information and Electronics Engineering*, vol. 2, no. 3, pp. 328-332, 2012.

[15] W. Boles and B. Boashash, "A human identification technique using images of the iris and wavelet transform," *IEEE Transactions on Signal Processing*, vol. 46, no. 4, pp. 1185-1188, April 1998.

[16] M. M. Memane and S. R. Ganorkar, "RED Algorithm based Iris Recognition," *International Journal of Computer Technology and Applications*, vol. 3, no. 4, pp. 1605-1712, August 2012.

[17] Khalid A. Darabkh, Raed T. Al-Zubi, Mariam T. Jaludi, and Hind Al-Kurdi, "An Efficient Method for Feature Extraction of Human Iris Patterns," *Proceedings of the 2014 IEEE International Multi-Conference on Systems, Signals & Devices, Conference on Communication & Signal Processing,* Castelldefels-Barcelona, Spain, February 2014.

Feature Selection for Object Detection: The Best Group vs. the Group of Best

Luka Fürst* and Aleš Leonardis*†
*University of Ljubljana, Faculty of Computer and Information Science, Tržaška cesta 25, SI-1000 Ljubljana, Slovenia
†Intelligent Robotics Laboratory, Centre for Computational Neuroscience and Cognitive Robotics,
School of Computer Science, University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK
Email: {luka.fuerst,ales.leonardis}@fri.uni-lj.si

***Abstract*—The problem of visual object detection, the goal of which is to predict the locations and sizes of all objects of a given visual category (e.g., cars) in a given set of images, is often based on a possibly large set of local features, only a few of which might actually be useful for the given detection setup. Feature selection is concerned with finding a 'useful' subset of features. In this paper, we compare two approaches to feature selection in a visual object detection setup. One of them selects features based on their individual utility scores alone, regardless of possible interdependence with other features. The other approach employs the AdaBoost framework and hence implicitly deals with interdependence. Using two feature extraction methods and several image datasets, we experimentally confirm the significance of feature interdependence: features that perform well individually do not necessarily perform well as a group.**

I. Introduction

Many object detection methods begin with extracting a potentially large set of local features from a given set of training images. The extracted features may be immediately used as building blocks in the detection model of the given visual category [1]. However, it is often beneficial to reduce the number of features before constructing the detection model. Besides the computational reasons, a properly selected set of features might lead to a better detection performance than the original set, since the selection process may be able to remove the 'useless' or 'harmful' features.

A viable feature selection method has to assess the utility of individual features for the given object detection problem. A feature is 'useful' if it is both distinctive (it often occurs in images depicting objects of the given category and rarely in other images) and predictive (it consistently occurs at similar relative locations within images of objects). For the 'Cars' category, for instance, an example of a 'useful' feature is one that captures a tire, since tires tend to be visually distinctive for cars and to occur at similar relative positions in a majority of car images.

The primary contribution of this paper is the comparison of two methods for selecting a favorable (in terms of detection performance) subset of T features out of the initial set of N features. Both methods use the same measure for the utility of individual features. However, while one approach simply selects the T best features according to the individual utility scores, the other embeds the scoring mechanism into an AdaBoost framework [2], thereby considering interdependence of features [3] in addition to their individual scores. Informally, AdaBoost seeks features that work well as a group rather than just individually. We experimentally show that feature interdependence has to be considered in a feature selection process. As the secondary contribution, this paper compares two feature extraction and representation methods, namely that of Fidler *et al.* [4] and that of Ferrari *et al.* [5]. Both comparisons are evaluated on five real-world image datasets.

In this paper, we use the feature selection approach due to Fürst *et al.* [6]. Viola and Jones [7] proposed a highly successful AdaBoost-based method for selecting features for face detection. Both approaches formulate detection as a classification problem. However, while Viola and Jones use a similar classification approach in both the learning and the test stage, Fuerst *et al.* use classification only to select features in the learning stage. Many other AdaBoost-based object detection approaches have been proposed [8], [9], [10].

The rest of this paper is structured as follows: In Sect. II, we present both feature selection approaches. Section III describes the detection algorithm used for both feature selection and experimental validation. Section IV gives a brief overview of both feature extraction methods. Section V presents and discusses the results of experimental validation. Section VI concludes the paper.

II. Feature Selection

In this section, we present both feature selection approaches. For the sake of clarity, we first present the AdaBoost-based approach, since the approach that searches for the individually best features is actually its simplification.

A. AdaBoost-based Approach

The AdaBoost-based feature selection approach proposed by Fürst *et al.* [6] takes into account both the properties of individual features (distinctiveness and predictiveness) and their interdependence. Originally, AdaBoost [2] is an approach for constructing classifiers in binary classification problems. Given a set of 'weak', threshold-based classifiers, a set of training samples, and a set of weights assigned to individual samples, each iteration of AdaBoost selects the classifier that yields the highest classification accuracy on the sample set with respect to the current values of the weights. After

selecting the current classifier, the weights of the samples misclassified by that classifier are enlarged so that the next iteration will be likely to select a classifier complementary to the one just selected. In this way, AdaBoost will implicitly deal with interdependence in the set of weak classifiers.

We shall now present how to adapt the AdaBoost framework to the problem of selecting features for object detection. Let us assume the existence of an object detection algorithm that accepts an image I and a set of features $\mathcal{G}$ and returns a set of detection hypotheses predicting the locations and sizes of individual objects in the image I, with the additional restriction that I can be represented only by the features from $\mathcal{G}$. Each returned hypothesis H is assumed to be equipped with a positive number $s(H)$ designating its strength. An example of such a detection algorithm is described in Sect. III, but the presented selection approach would work with any algorithm fulfilling the stated requirements.

The AdaBoost-based feature selection process due to Fürst *et al.* requires a set of images that depict various objects of the given visual category on a uniform background (positive training images, $\mathcal{I}_+$) and a set of images depicting anything but objects of the given category (negative training images, $\mathcal{I}_-$). Let $\mathcal{I} = \mathcal{I}_+ \cup \mathcal{I}_-$ denote the entire set of training images, and let $\mathbf{w} = (w_I)_{I\in\mathcal{I}}$ denote a vector of weights assigned to individual images such that $\sum_{I\in\mathcal{I}} w_I = 1$. At the beginning of the AdaBoost algorithm, we have $\mathbf{w} = \mathbf{w}^{[1]}$, where $w_I^{[1]} = 1/(2|\mathcal{I}_+|)$ for all $I \in \mathcal{I}_+$ and $w_I^{[1]} = 1/(2|\mathcal{I}_-|)$ for all $I \in \mathcal{I}_-$. During the algorithm, the weights are iteratively updated.

To apply the classification-oriented AdaBoost framework to our detection-oriented feature selection setup, the problem of detecting objects in the training images is transformed into that of classifying training images themselves, and a threshold-based classifier is constructed for every feature. Let us first define a function $S\colon \mathcal{F}\times\mathcal{I}\to\mathbb{R}$ as

$$S(F,I) = \begin{cases} s(H^*(I \mid F)) & \text{if } I \in \mathcal{I}_+ \text{ and } H^*(I) \text{ is correct;} \\ s(H^*(I \mid F)) & \text{if } I \in \mathcal{I}_-; \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

where $H^*(I \mid F)$ denotes the strongest detection hypothesis on the image I provided that I is represented solely by the feature F. For a feature F and a threshold $\theta \in \mathbb{R}$, let us define a classifier $d_{F,\theta}\colon \mathcal{I}\to\{0,1\}$ that classifies a given training image as positive (output 1) or negative (output 0) according to the following rule:

$$d_{F,\theta}(I) = \begin{cases} 1 & \text{if } S(F,I) \geq \theta; \\ 0 & \text{if } S(F,I) < \theta. \end{cases} \tag{2}$$

If y_I denotes the actual class of the image I ($y_I = 1$ for all $I \in \mathcal{I}_+$ and $y_I = 0$ for all $I \in \mathcal{I}_-$), the classifier's error on the entire training set can be computed as

$$\epsilon(F,\theta;\mathbf{w}) = \sum_{I\in\mathcal{I}} w_I |d_{F,\theta}(I) - y_I|. \tag{3}$$

Let $\theta^*(F;\mathbf{w})$ be a threshold at which the classifier's error is minimal:

$$\theta^*(F;\mathbf{w}) = \arg\min_{\theta} \epsilon(F,\theta;\mathbf{w}) \tag{4}$$

The classifier $d_F^* = d_{F,\theta^*(F;\mathbf{w})}$ attains the smallest classification error and will thus be called the *optimal* classifier. The *utility* of a feature F is defined as

$$u(F;\mathbf{w}) = 1 - \epsilon(F,\theta^*(F;\mathbf{w});\mathbf{w}) \tag{5}$$

For a feature F that is both distinctive and predictive, $S(F,I)$ is large for many positive training images and small for many negative ones. Under the initial weight vector, the optimal classifier for such a feature F attains a low classification error, since a threshold θ^* can be found such that most of the values $\{S(F,I)\}_{I\in\mathcal{I}_+}$ are above θ^* and most of the values $\{S(F,I)\}_{I\in\mathcal{I}_-}$ are below θ^*. The feature F has therefore a high utility. Conversely, if a feature F is non-distinctive, the values of $S(F,I)$ tend to be random, whereas if F is non-predictive, $S(F,I)$ is zero for many positive images. In both cases, even the optimal classifier produces a high classification error, and so F has a low utility.

The AdaBoost-based feature selection method is shown as Algorithm 1. In each iteration, the algorithm selects the feature having the highest utility with respect to the current weight vector and decreases the weights of the training images correctly classified by the optimal classifier for the selected feature. In the next iteration, the misclassified images have higher relative weights, which encourages the selection of a feature complementary to the feature selected in the previous iteration.

B. Static Approach

In the static approach, features are selected by decreasing values of $u(F;\mathbf{w}^{[1]})$. Features are thus selected by individual utility, but their interdependence is not considered.

III. Object Detection

We employed an object detection approach due to Leibe *et al.* [1]. For now, let us assume that all objects are approximately the same size; an extension to arbitrary sizes will be briefly described later. In the training stage, the method of Leibe *et al.* learns the so-called Implicit Shape Model (ISM) from a set of training images depicting different objects of the given visual category on a uniform background. In the test stage, the learned model is used to form detection hypotheses for a given real-world image depicting an unknown number of objects of the given category.

To learn the ISM for a given category, every training image is first represented by the given set of features. For a feature F located at the coordinates $\mathbf{y}$ in an image I, the learning algorithm stores a pair $(F, \mathbf{x}-\mathbf{y})$, where $\mathbf{x}$ denotes the location of the object's centroid in I. Informally, the ISM is thus a collection of the relative locations of individual features with respect to the object centroid. The test stage begins with the representation of a given test image by the feature set. Each feature then 'votes' for all locations of the object centroid based on its ISM entries. Specifically, if a feature F is located

Algorithm 1 The AdaBoost-based feature selection method.

Require:

(a) A set of training images $\mathcal{I}$ with class membership labels y_I (1: positive; 0: negative)

(b) A set of features, $\mathcal{F}$.

(c) The number of iterations, T.

Ensure: A sequence of selected features $\langle F^{[1]}, \ldots, F^{[T]} \rangle$.

Process:

1: $\mathcal{F}' := \mathcal{F}$; // $\mathcal{F}'$: *the set of not-yet selected features*

2: $w_I^{[1]} := 1/(2|\mathcal{I}_+|)$ for all $I \in \mathcal{I}_+$

3: $w_I^{[1]} := 1/(2|\mathcal{I}_-|)$ for all $I \in \mathcal{I}_-$

4: **for** $t := 1$ **to** T **do**

5: Normalize the weights such that their sum becomes 1:

$$w_I^{[t]} := w_I^{[t]} / \sum_{J \in \mathcal{I}} w_J^{[t]} \text{ for all } I \in \mathcal{I}. \quad (6)$$

6: Select the feature with the highest score $u(F; \mathbf{w})$ with respect to the current weights:

$$F^{[t]} = \arg\max_{F \in \mathcal{F}'} u(F; \mathbf{w}). \quad (7)$$

7: Let $\theta^{[t]} = \theta^*(F^{[t]}; \mathbf{w}^{[t]})$, $d^{[t]} = d_{F^{[t]},\theta^{[t]}}$, and $\epsilon^{[t]} = \epsilon(F^{[t]}, \theta^{[t]}; \mathbf{w}^{[t]}) = 1 - u(F^{[t]}; \mathbf{w}^{[t]})$.

8: $\mathcal{F}' := \mathcal{F}' \setminus \{F^{[t]}\}$.

9: Update the weights:

$$w_I^{[t+1]} = \begin{cases} \frac{\epsilon^{[t]}}{1-\epsilon^{[t]}} w_I^{[t]}, & \text{if } d^{[t]}(I) = y_I; \\ w_I^{[t]} & \text{otherwise.} \end{cases} \quad (8)$$

10: **end for**

at the coordinates $\mathbf{q}$ in the test image and if the ISM entries for F are $(F, \mathbf{r}_1), \ldots, (F, \mathbf{r}_k)$, then F votes for the coordinates $\mathbf{q} + \mathbf{r}_1, \ldots, \mathbf{q} + \mathbf{r}_k$. The votes are weighted: the method of Leibe *et al.* considers both the possibility that a given local structure is represented by several features, each with a certain probability, and the possibility that a given feature votes for several object centroid candidates, each with a certain probability.

The output of the voting algorithm is a probability map that, for each pixel of the test image, specifies the probability that the object centroid is located at the coordinates of that pixel. To obtain a small set of robust detection hypotheses, the probability map is smoothed by a Gaussian filter. The locations of the maxima in the smoothed map are taken as individual detection hypotheses. The strength of a hypothesis H, $s(H)$, is defined to be the value of the corresponding maximum.

A single feature may cast votes for many different object centroids, often giving rise to strong secondary hypotheses. Such hypotheses may be suppressed by a greedy MDL-based algorithm [11]. In each iteration, this algorithm selects the currently strongest hypothesis and greedily assigns to it all of its support, i.e., all probabilistic votes that it gathered in the voting process. The competing hypotheses thereby lose part of their support and hence strength. This scheme has been experimentally shown to significantly improve detection performance.

The described training and detection process can also be adapted to multiscale detection. At least two approaches exist: (1) single-scale learning, multiscale detection; (2) multiscale learning, single-scale detection. In the first approach, the training stage is the same as in the single-scale case, but the test-stage voting algorithm is applied to several rescaled versions of the test image. The second approach only works with scale-covariant features (such a feature can be detected at multiple scales, and each of its occurrences provides information about the scale of the corresponding local structure). The learned ISM contains both the offset and the scale of each feature, relative to the position and the size of the object. In the test stage, the features cast votes in a 3-dimensional space. Each feature votes not only for candidate object centroid locations but also for candidate object sizes. The detection algorithm thus produces hypotheses about both the centroids and the sizes of individual objects.

IV. Feature Extraction and Representation

We tested two feature extraction and representation methods: that of Fidler *et al.* [4] and that of Ferrari *et al.* [5]. Features of Fidler *et al.* are arranged into layers according to the 'parts composed of parts' paradigm. The first layer contains only a single feature; this is the kernel of the Gabor filter in the orientation 0°. Each feature of a layer $n > 1$ is a composition of rotated copies of features of layer $n - 1$. In particular, a layer-n feature F_i^n can be represented as $\{(F_{j_1}^{n-1}, \alpha_1, \mu_1, \mathbf{\Sigma}_1), \ldots, (F_{j_s}^{n-1}, \alpha_s, \mu_s, \mathbf{\Sigma}_s)\}$, where for each $k \in \{1, \ldots, s\}$, α_k is the discretized orientation of the k-th component of feature F_i^n (relative to the orientation of the first component), whereas μ_k and $\mathbf{\Sigma}_k$ denote the mean and the covariance matrix of the Gaussian distribution that models the variability of the position of the k-th component (relative to the position of the first component). To represent an image with a set of features, the features are matched against the Gabor-filter representation of the image in a layer-by-layer fashion. The result of the representation is a set of triples $(F_i^n, (x, y), \beta)$, where F_i^n is a layer-n feature, (x, y) is the position of an occurrence of F_i^n in the image, and β is the discretized orientation in which the feature occurs at that position.

Features of Ferrari *et al.* [5] are obtained by a multi-step procedure. First, an edge detector is used to obtain edge elements, i.e., image points at directed intensity transitions. After that, each group of edge elements that form an approximately straight line is taken as a separate 'segment'. In the next step, the extraction procedure builds groups of k segments that touch or almost touch each other. If $k = 2$, for example, such groups include segment pairs in the shape of the letters L and T. Groups of k adjacent segments are then clustered by their visual similarity. Finally, each cluster is taken as a Ferrari *et al.* feature. Features are described in a rotation- and scale-invariant manner. To represent a given image by a given set of features, the image is first processed so as to obtain a set of groups of k adjacent segments. After that, the features are

matched against the groups in the image, giving a set of triples $(\mathbf{r}(G), t(G), \mathcal{R}(G))$, where G is a k-adjacent-segments group, $\mathbf{r}(G)$ and $t(G)$ are its location and size, and $\mathcal{R}(G) = \{(F, r(F,G)), \ldots\}$ is a set of features that can be matched against the group G, equipped with an estimate of the reliability of the match.

V. Experimental Results

Before presenting the results, let us describe how individual experiments were evaluated. The output of the detection algorithm for a given image is a set of detection hypotheses together with their strengths. A hypothesis is a prediction of the bounding box of an object in the image. Let (x_H, y_H, w_H, h_H) denote the image coordinates of the centroid (x_H and y_H) and the width and the height (w_H and h_H) of the bounding box predicted by a hypothesis H, and let (x_O, y_O, w_O, h_O) define the true bounding box of a depicted object. The hypothesis H is proclaimed to be correct if the following three conditions are met [1]:

1) $\sqrt{(2(x_H - x_O)/w_O)^2 + (2(y_H - y_O)/h_O)^2} \leq 0.5$.
2) The predicted bounding box covers at least 50% of the area of the true bounding box.
3) The true bounding box covers at least 50% of the area of the predicted bounding box.

If the above conditions are fulfilled by multiple hypotheses for the *same* object, only one of those hypotheses is considered correct; all others are regarded as false positives.

The detection performance on a given image set is estimated by the following measure:

$$F_{\max} = \max_{\tau} \frac{2 \times \text{recall}(\tau) \times \text{precision}(\tau)}{\text{recall}(\tau) + \text{precision}(\tau)}. \quad (9)$$

The $F_{\max}$ measure ranges from 0 to 1, with 0 being the worst and 1 being the best possible detection performance. The recall and precision at a given threshold τ are computed as follows:

$$\text{recall}(\tau) = \text{TP}(\tau)/\text{OBJ}; \quad (10)$$

$$\text{precision}(\tau) = \text{TP}(\tau)/(\text{TP}(\tau) + \text{FP}(\tau)). \quad (11)$$

Here, OBJ denotes the number of all depictions of objects of the given visual category in the image set, $\text{TP}(\tau)$ ('true positives') is the number of correct hypotheses H with $s(H) \geq \tau$, and $\text{FP}(\tau)$ ('false positives') is the number of incorrect hypotheses H with $s(H) \geq \tau$.

Each combination of feature extraction and feature selection method was tested on five test image datasets, some information about which is provided in Table I. To extract and select features and to learn the detection model for individual categories, we used category-specific training image sets obtained from various sources. All training sets were completely disjoint from the corresponding test sets.

For each combination of feature extraction method, feature selection method, and image dataset, we extracted all features (their number is given in Table II), selected the first T features (where T varied from 1 to 50), and learned and tested the detection model using the T selected features. Fig. 1 shows the $F_{\max}$ values for all experiments.

TABLE I: Test image datasets.

Dataset	Num. of images	Total num. of depicted objects
Apple logos [5][a]	40	44
Bottles [5][a]	48	55
Cars [12][b]	108	139
Horses [5][c]	170	180
Mugs [5][a]	48	66

[a] http://www.vision.ee.ethz.ch/datasets/index.en.html
[b] http://cogcomp.cs.illinois.edu/Data/Car/
[c] http://pascal.inrialpes.fr/data/horses/

TABLE II: The number of extracted features.

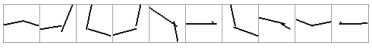
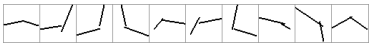
Dataset	Fidler *et al.*	Ferrari *et al.*
Apple logos	592	168
Bottles	896	422
Cars	920	386
Horses	896	398
Mugs	912	314

The AdaBoost-based approach significantly outperforms the static feature selection approach. The reason for such a behavior is the redundancy of the extracted feature set. Many extracted features react to similar local structures and hence attain similar utility scores. Consequently, a feature set selected solely by the utility of individual features is likely to contain many similar features. By contrast, AdaBoost favors features that correctly classify the images misclassified by the already selected features. Tables III and IV list the first 10 selected features for the categories 'Cars' and 'Mugs'. The feature sets selected by the static approach are clearly redundant. As shown in Fig. 1, a much greater number of features selected by the static approach is required to attain a detection performance comparable to AdaBoost-selected features.

For each of the first 10 AdaBoost-selected Fidler *et al.* features for the 'Mugs' category, Table V shows its rank in the decreasing sequence of $u(F; \mathbf{w}^{[1]})$ values. The third selected feature, for instance, appears at a rank as low as 156 out of 912, but it represents the best complement to the first two selected features, at least with respect to the employed training image set.

The features obtained by the method of Fidler *et al.* led to substantially more favorable detection results than those of

TABLE III: The first 10 selections for 'Cars'.

Extraction	Selection	The first 10 selected features
Fidler *et al.*	AdaBoost	
Fidler *et al.*	static	
Ferrari *et al.*	AdaBoost	
Ferrari *et al.*	static	

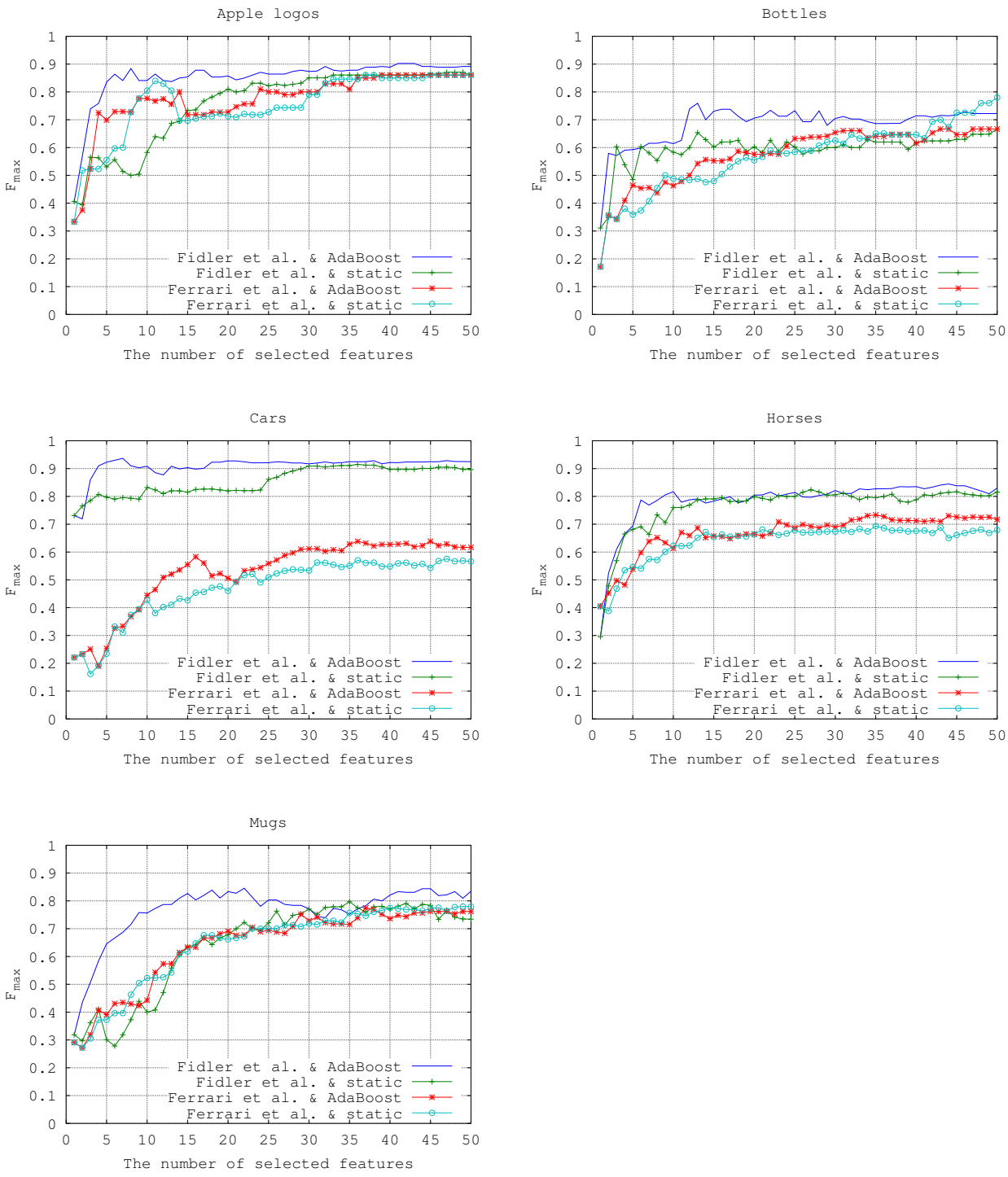


Fig. 1: Detection performance w.r.t. the number of selected features.

TABLE IV: THE FIRST 10 SELECTIONS FOR 'MUGS'.

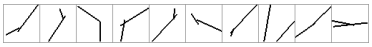
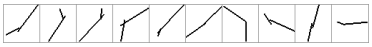
Extraction	Selection	The first 10 selected features
Fidler *et al.*	AdaBoost	
Fidler *et al.*	static	
Ferrari *et al.*	AdaBoost	
Ferrari *et al.*	static	

TABLE V: THE RANKINGS OF THE FIRST 10 ADABOOST-SELECTED FEATURES FOR 'MUGS' IN THE SEQUENCE OF THE FEATURES SELECTED BY THE STATIC APPROACH.

AdaBoost	1	2	3	4	5	6	7	8	9	10
Static	1	7	156	3	28	109	30	4	142	49

Ferrari *et al.*. For instance, the most distinctive and predictive features for the 'Cars' category are those that capture tires, a few examples of which are shown in Table III. While the method of Fidler *et al.* is able to extract such features, this is not the case with the method of Ferrari *et al.*, which builds features from straight segments. These features are therefore unable to capture small circular or semi-circular structures. In addition, although Fidler *et al.* features tend to be fairly indistinctive (with the possible exception of wheel-like features), they occur on equivalent local structures much more consistently than Ferrari *et al.* ones. Fig. 2 shows the first AdaBoost-selected Fidler *et al.* features for the categories 'Cars' and 'Mugs', their respective ISM probability maps (the probability of the object centroid location relative to the location of the feature), and a few occurrences in test images.

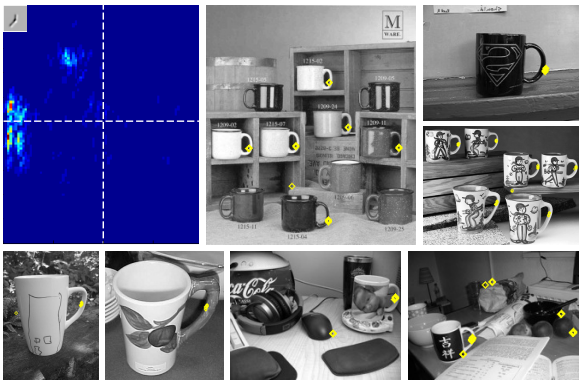
VI. CONCLUSION

We have compared an AdaBoost-based feature selection approach with a simplified, 'static' approach that does not pay regard to feature interdependence. Using two feature extraction methods and five real-world image datasets, we showed that for a fixed target number of selected features, the AdaBoost-based method considerably outperforms the simplified selection method. The main reason is that many extracted features tend to be visually similar and hence attain similar utility scores. The resulting feature sets therefore contain much redundancy. AdaBoost, by contrast, selects features so that they complement each other and thus tends to achieve a comparable detection performance using much fewer features than the static approach.

We assumed that the number of selected features is fixed and known in advance. A natural extension of the presented selection approaches is therefore the automatic determination of the most appropriate number of features with respect to both detection performance and computational burden. Another interesting research direction is feature selection for the multiclass detection setting, where the goal is to find all objects of several categories simultaneously [10].



(a) cars



(b) mugs

Fig. 2: The first AdaBoost-selected Fidler *et al.* feature, its ISM probability map, and a few occurrences in test images.

REFERENCES

[1] B. Leibe, A. Leonardis, and B. Schiele, "Combined object categorization and segmentation with an implicit shape model," in *ECCV'04 Workshop on Statistical Learning in Computer Vision*, 2004, pp. 17–32.

[2] Y. Freund and R. Schapire, "A decision-theoretic generalization of online learning," *Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[3] T. Cover, "The best two independent measurements are not the two best," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 4, no. 1, pp. 116–117, 1974.

[4] S. Fidler and A. Leonardis, "Towards scalable representations of object categories: Learning a hierarchy of parts," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[5] V. Ferrari, L. Fevrier, F. Jurie, and C. Schmid, "Groups of adjacent contour segments for object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 1, pp. 36–51, 2008.

[6] L. Fürst, S. Fidler, and A. Leonardis, "Selecting features for object detection using an adaboost-compatible evaluation function," *Pattern Recognition Letters*, vol. 29, no. 11, pp. 1603–1612, 2008.

[7] P. Viola and M. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[8] H. Grabner, P. Roth, and H. Bischof, "Eigenboosting: Combining discriminative and generative information," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[9] J. Shotton, A. Blake, and R. Cipolla, "Contour-based learning for object detection," in *International Conference on Computer Vision*, 2005, pp. 503–510.

[10] A. Torralba, K. Murphy, and W. Freeman, "Sharing features: efficient boosting procedures for multiclass object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2004, pp. 762–769.

[11] A. Leonardis, H. Bischof, and J. Maver, "Multiple eigenspaces," *Pattern Recognition*, vol. 35, no. 11, pp. 2613–2627, 2002.

[12] S. Agarwal, A. Awan, and D. Roth, "Learning to detect objects in images via a sparse, part-based representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 11, pp. 1475–1490, 2004.

Movie Posters Classification into Genres Based on Low-level Features

Marina Ivasic-Kos, Miran Pobar, Luka Mikec
Department of Informatics
University of Rijeka
Rijeka, Croatia
{ marinai, mpobar}@uniri.hr

Abstract— **A person can quickly grasp the genre (drama, comedy, cartoons, etc.) from a movie poster, regardless of visual clutter and the level of details. Bearing this in mind, it can be assumed that simple properties of a movie poster should play a significant role in automated detection of movie genres. Therefore, low-level features based on colors and edges are extracted from poster images and used for poster classification into genres.**

In this paper, poster classification is modeled as a multi-label classification task, where a single movie may belong to more than one class (genre). To simplify and solve the multi-label problem, two methods for multi-label data transformation are described and evaluated given the classification results obtained by distance ranking, Naïve Bayes and RAKEL.

Experiments are conducted on a set of 1500 posters with 6 movie genres. Results provide insights into the properties of the discussed algorithms and features.

Keywords—multi-label classification, data transformation method, movie poster

I. INTRODUCTION

With just a glance at a movie poster, a person can comprehend a variety of perceptual and semantic information. Capturing essential information about the movie at once, regardless of the visual complexity, can be experienced while driving around town and spotting a billboard or walking around the shopping center and cursorily glancing at an advertisement for a film. A person can grasp the genre (drama, comedy, cartoons, etc.) from a poster independently of the short observation time, clutter and the variety of details.

The same phenomenon has already been observed in complex real-world scenes, where persons can promptly recognize the basic-level category of the scene (e.g., a street) [1], its spatial layout (e.g., a street with tall vertical blocks on both sides) [2], as well as other global structural information (e.g., a large volume in perspective).

Assuming that one of the goals of the poster is to convey the information about the film (genre, etc.) to potential moviegoers without them paying a lot of attention, the important information for determining the genre might be contained in global low-level features such as dominant color of the whole poster, texture or color histogram. We thus wanted to develop a method to automatically determine the genres of movies using mostly visual low-level features of their posters. The task is made more complex by the fact that most movies belong to more than one genre, for example „Titanic“ belongs to Action, Disaster, Drama and Romance genres, and „Back to the Future“ belongs to Adventure, Comedy, Sci-Fi and Family genres, according to TMDB [3]

Film genre classification from additional supporting promotional material (trailers) has recently received some attention. In paper [4] low-level visual features extracted from movie trailers have been used to classify 100 movies into 4 genres (drama, action, comedy, horror). In [5] GIST, CENTRIST and W-CENTRIST scene features were applied to a collection of temporally-ordered static key frames to yield a feature representation. These features are used as visual vocabulary for genre classification and tested on 1239 movie trailers.

In [6] the same visual features were used as in [4]. Movies were classified into three genres (action, drama, and thriller) which were selected because of their frequency among movies that were played in Taiwan from 2004 to 2006. Some additional genres were grouped together and presented as those three (e.g. drama included comedy and romance while thriller included horror). All these approaches [4-6] simplify the problem considering only single genre per movie and deal with only single label classification problem.

In our approach, multi-label poster classification is proposed. In Section II two kinds of data transformation methods are explained. A set of low-level features, classification methods used for poster classification into movie genres and evaluation of described data transformation methods are presented in Section III. The paper ends with description of experimental setup, experimental results with discussion and directions for future work.

II. DATA TRANSFORMATION METHODS

The goal of classification is to assign class labels from the set of class labels C to examples from the set of samples E. In many problems, an example belongs to a

single class (single-label classification). In binary classification, an example e is classified in one of two possible classes from C, $C = \{C_1, C_2\}$ as in spam classification where a mail message is declared as either spam or not-spam. In multi-class classification an example is classified into one of more classes {C1, C2 … Ck}, for example a cloud can be classified as nimbus, cumulus, cirrus, etc. However, there are many problems where an example e_j should be classified into more than one class, as in our experiment. This is the case of multi-label classification which can be formally expressed as:

$$\exists\, e_j \in E : \varphi(e_j) = \{C_l, C_m\} \cup Z,\ Z \subseteq \cup_{i=1}^{k} C_i,\ l, m \in 1..k, l \neq m \text{ and } \varphi: E \to C.$$

Thus, in the set of examples E, an example e_j exists that the classifier φ maps into at least two classes C_l and C_m.

If a hierarchical structure exists among some of the classes in C so that one class can be inferred from another, then it is the case of hierarchical multi-label classification. For instance in scene classification, if a scene can be classified as beach, outdoor scene and natural scene, the parent class – outdoor scene will be inferred from the child class – beach.

Where the parent-child relationship does not exist between classes, the problem is known as flat multi-label classification. In case of e.g. classifying songs into moods and genres, one class label cannot be inferred from the other, because there is no parent-child relationship between them. For example, the song Paperback Writer by The Beatles can be classified into mood carefree and genre pop while the song Sabotage from Beastie Boys can be classified into mood brash and genre hip-hop according to [7]. The same problem exists in many different applications like in text classification where an article could belong to crime and sports classes, in medical diagnosis, information retrieval etc.

The multi-label classification problem cannot be directly solved by most classification methods which are designed for single-label classification. One approach to tackle a multi-label classification problem is to transform the data so it fits a series of single-label classification [8]. The other approach contains methods that extend specific learning algorithms in order to handle multi-label data directly.

In this paper, a multi-label classification problem is transformed into single-label classification, in two ways.

In the first, each of the examples that had two or more class labels was transformed into a set of ordered pairs. The first element of each pair is the example and the second one is the class label. Thus, if an example $e_j \in E$ can be classified into $Y_j = \{C_l, C_m, \dots, C_r\}$, $Y_j \subseteq C$ then we replace that example with $|Y_j|$ ordered pairs $(e_j, C_l), (e_j, C_m) \dots (e_j, C_r)$. For instance, the movie „Top Gun“ belongs to Action, Romance and War genres, and would be transformed into the set of ordered pairs containing individual genres: {(Top Gun, Action), (Top Gun, Romance), (Top Gun, War)}.

Now, model that maps first element of the ordered pair (example) to the second element of the pair (class label) can be defined. If the classifier assigns a confidence value to each class label, classes whose confidence scores are greater than a set threshold will be assigned to the example. A related approach is ranking, where classes are ranked according to their confidence scores, and top n are mapped to the example. In the following, this data transformation method is referred to as M1.

In the second case each set of class labels $Y_j = \{C_l, C_m, \dots, C_r\}$ is transformed into a new combined class $C_{l,m,\dots,r}$ that is assigned to the example e_j. In that way we create the new set $C' \supseteq C$, by adding the new combined classes into the set C'. Using this problem transformation method "The Top Gun" example would be transformed into the ordered pair containing one combined genre (Top Gun, Action&Romance&War). This data transformation method is hereinafter referred to as M2.

Both problem transformation methods are algorithm independent, so any classification method appropriate for the single-label classification can be used to map the examples to the new class.

III. EXPERIMENTS

In this section we present our experiments conducted on posters and metadata obtained from the TMDB. Our goal is to automatically determine the genres of movies using visual features of their posters.

A. Data and Preprocessing Step

We have used 1500 posters of movies dated from 1990 onwards, selected so that 6 chosen genres (Drama, Action, Animated, Comedy, War and Horror) had 250 examples each. Since each movie can have multiple genres, additional genre labels (e.g. Mystery, Crime, Family, and Romance) were present in the data. These genres were transformed into the chosen 6 and only two labels per movie were kept. The following rules were used to transform the original genre labels that were present in the data into genre labels that were used for classification (Table I):

TABLE I. TRANSFORMATION RULES

Original genre labels	Genre class label
Action, Disaster, Fantasy, Western	Action
Animation	Animation
Comedy	Comedy
Crime, Drama, Romance, Thriller	Drama
Horror	Horror
War, History	War

We applied this preprocessing step to prevent data scarcity, caused by too many genre labels in C' with too few examples. After the mapping, the resulting number of movies of each genre differs from the initial 250.

From 1500 collected posters, 1000 were used for training and 500 for testing.

After the preprocessing step, the genre labels in method M1 correspond to individual genres, and are from the set

$$C = \text{Drama, Action, Animated, Comedy, War, Horror}\}.$$

In the method M2 the genre labels except labels of individual genres include combined genre labels, and are from the set:

$$C' = C \cup \{\text{Drama\&Action, Drama\&Comedy,..., War\&Horror}\}.$$

B. *Features*

Motivated by the way people glance at billboards and still capture enough information about the film we wanted to see if only simple low-level features are enough for automatic poster classification into movie genres.

Intuitively it can be assumed that the variance of color has a strong correlation with respect to genres, so we used more low-level color features and experimentally tested their discriminate ability in terms of genre classification. Before the low-level features were extracted, each poster was proportionally scaled to 200 pixel width and converted to HSV color space. The relationship between the hues with maximal saturation in HSV color space with their corresponding RGB coordinates is illustrated in Fig. 1.

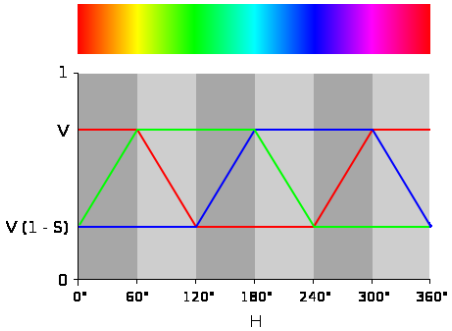


Fig. 1. The relationship between the hue of colors with maximal saturation in HSV color space with their corresponding RGB coordinates [9]

First, the image color histogram was calculated on hue (H) and saturation (S) channels of the whole image. We used 50 bins for the H channel and 60 bins for the S channel. The value (V) channel was not used. The idea was to keep the number of features low enough to emphasize the information that is relevant for classification and adequately train the classifier, especially in approach M2, where the number of classes is increased.

To test if color layout has an important role in classification, we computed 6 local H-S histograms on images divided into 3x2 grids, and one H-S histogram computed on the central part of the image. The central part was of the same proportions as the whole image, but 1/3 of the diagonal size. The arrangement of image grid from which the local color histograms were computed is given in Fig. 2.



Fig. 2. The arrangement of image grid from which the local color histograms were computed

Alternatively, we defined a set of features F which includes dominant colors, features based on edges and the number of faces. Dominant colors correspond to the histogram bins with the highest values and are defined according to [10]. We have used different numbers of dominant colors (3, 6, 8, 12, 16, 24 and 36) and have experimentally determined that 12 dominant colors yield the best classification results in our task.

The edge based feature was computed by first blurring the value (V) channel of each image with the 3 by 3 Gaussian filter, convolving the result with an edge detection mask and counting the number of pixels exceeding the set threshold. Experimentally we selected the value (V) channel of HSV color model and set the threshold to 0.5. We used the Sobel edge detector.

The number of faces in the poster was detected by the OpenCV [11] implementation of Viola-Jones face detector on images blurred with 3 by 3 Gaussian filter. Given that posters are images that are artistically styled, face detection for frontal faces was successful in 66%, while the profiles had significantly worse results, both less than expected for unprocessed images.

C. *Classification Methods*

We used the distance ranking (DR), k-NN and Naïve Bayes classifiers on data adapted with both data transformation methods, and compared the results with the random k-label sets (RAKEL) [12] method which has been shown to be effective in multi-label classification tasks.

For both data transformation methods M1 and M2, we used the same procedure to define representative feature values for the genre labels. For each genre label, we defined a number of representative vectors for each feature except for the number of faces for which the average number was used. The representative values were determined using k-means clustering algorithm. Experimentally, we selected up to 3 representative H-S histograms, 12 dominant colors and 3 centroids of edge based feature values.

For genres that have few movies for one centroids value, less than 3 representative H-S histograms are created. Otherwise genres would be represented with histograms based on too few examples to be relevant. For example, the Horror genre has two representative H-S histograms indicated in the Fig. 3 and Fig. 4, while the Action genre has three representative H-S histograms.

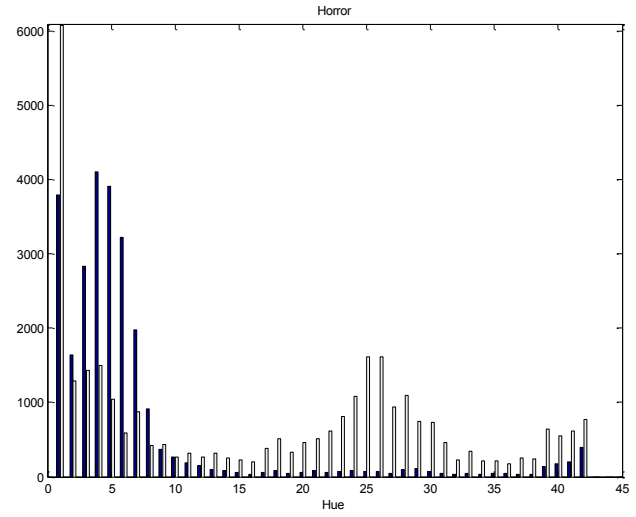


Fig. 3. Representative hue histograms for the Horror genre.

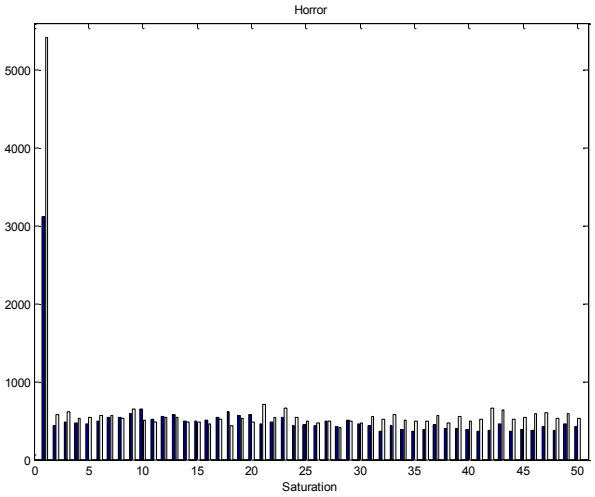


Fig. 4. Representative saturation histograms for the Horror genre.

To determine distance between H-S histograms computed on the poster images, the Chi-Square distance was used (1):

$$d(H_1, H_2) = \sum_I \frac{\left(H_1(I) - H_2(I)\right)^2}{H_1(I)} \qquad (1)$$

where H1, H2 are normalized H-S histograms, I is a histogram bin and d(H1, H2) distance used to express how well both histograms match. For the features from the set F we used the Euclidean distance.

Figures 5 and 6 show the hue and the saturation histograms of Action and Comedy genres. It can be seen for example, that Action posters typically have many more black/gray pixels than Comedy posters, as presented on the saturation histogram. The hue histogram shows the different distribution of colors, e.g. Action genre has more pure red (Hue histogram bin 0) values, while Comedy genre has more orange values (Hue histogram bin 4). Note that hue values range from 0-50 and saturation ranges from 0 to 60.

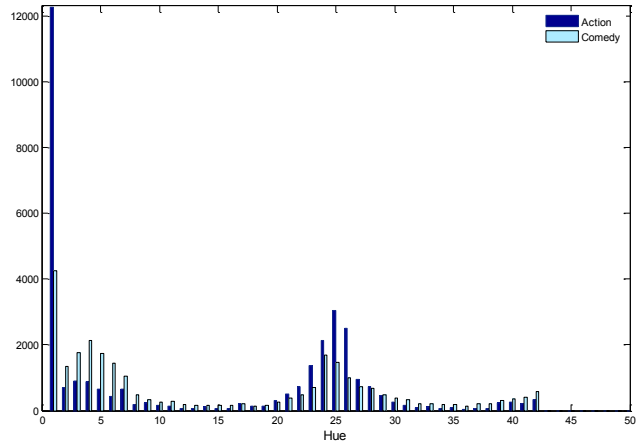


Fig. 5. Action and Comedy hue histograms.

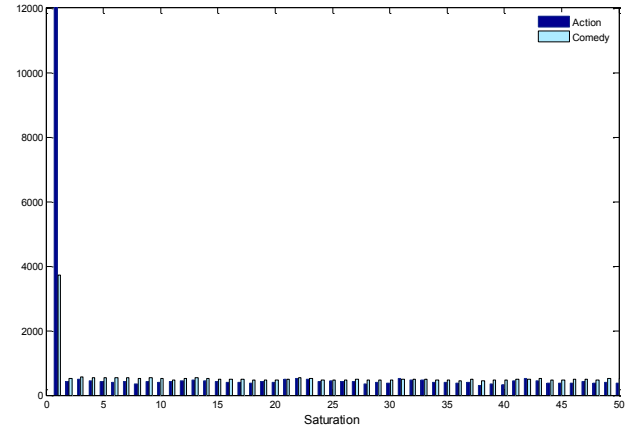


Fig. 6. Action and Comedy saturation histograms.

Moreover, the correlation among dominant colors and movie genres has been experimentally determined. For instance, Animation and Comedy tend to have a large variety of colors (hues), whereas War and Horror films often adopt few dominant colors (hues).

A new unknown movie poster is classified into genres using distance ranking (DR), Naïve Bayes (NB) and RAKEL classification methods.

In the case of DR, the distances of genre centroids and of the feature values extracted from a movie poster are calculated using (1) when histograms are used and Euclidean distance for the features from the set F. The genres are ranked according to the distance of their centroids to the poster features and the top two are selected as classification results.

When using Naïve Bayes, each feature vector x_{e_j} of poster e_j is classified into one of the genre labels according to the maximum posterior probability c_{MAP} (2):

$$c_{MAP} = \text{argmax}_{C_i \in C} P\left(\boldsymbol{x}_{e_j} \middle| C_i\right) P(C_i) \qquad (2)$$

The probability $P(C_i)$ is estimated from relative frequency of label C_i in the training data. To determine the second genre we pick the class with the second highest probability score $P(x_{e_j}|C_i)$.

RAKEL was run using the nearest neighbor (1-NN) classifier based on the Bhattacharyya distance (3),

$$d_B(p,p') = \sum_{i=1}^{N} \sqrt{p(i)p'(i)}, \qquad (3)$$

where p and p' are histograms and N is the number of bins.

The RAKEL subset size was 3 and the number of models was 12.

IV. EXPERIMENTAL RESULTS

All experiments were run using 5 runs of 5-fold cross-validation. The following results were obtained as an average of those runs.

The Table II shows the classification results for the test set of 500 posters using the data transformation approach in which classes correspond to single genres (M1). The presented results are obtained using distance ranking and Naïve Bayes classifier. The results of distance ranking are obtained based on H-S histograms (DR1), local H-S histograms (DR2) or features from the set F (DR3) explained in the Section II. In the Tables II and III, rows marked as "2/2" show the percentage of movies for which both movie genres were correctly recognized and rows marked "1/2" show the percentage of movies for which at least one of the two genres was correct. The best results are obtained using distance ranking based on features from the set F in cases 2/2 and 1/2.Similar results are obtained by Naïve Bayes. Comparing the data, the contribution of color seems to be most important for classification.

TABLE II. CLASSIFICATION RESULTS FOR DATA TRANSFORMATION METHOD M1.

No. correct labels/top ranked labels	DR 1	DR2	DR3	NB
2/2	8.97%	10.10%	**12.1%**	11.2%
1/2	59.72%	54.85%	**66.58%**	62.21%

The results of classification on the same set using the data transformation method with combined genre labels (M2) is presented in the Table III in the similar fashion. In this case, RAKEL method was additionally used.

TABLE III. CLASSIFICATION RESULTS FOR DATA TRANSFORMATION METHOD M2.

No. correct labels/top ranked labels	DR 1	DR 2	DR 3	NB	RAKEL
2/2	**13.61%**	8.90%	10.51%	8.93%	13.24%
1/2	57.69%	58.44%	62.08%	61.24%	**62.54%**

The best results are obtained using distance ranking based on histogram, if all labels need to be correctly classified (2/2). If one correct label should be obtained among the two top ranked (1/2), then RAKEL is the best solution.

Distance ranking based on global H-S histograms used with data transformation method M2 performs best when both genre labels need to be correctly classified. If we allow one incorrect label among the top two ranked, the distance ranking and data transformation method M1 used with the features from the set F (DR3), performed best on test data.

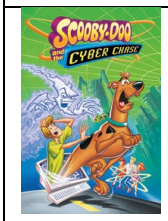
Perhaps surprisingly, results obtained by local H-S histograms are worse in the case of the M2 approach.

The large differences between classification results for cases 2/2 and 1/2 are expected because of the difference in the baseline, where expected random score for 2/2 is 6.7% and for 1/2 is 33%. In both cases, automatic classification performed about twice as good.

The results suggest that both data transformation methods perform similarly regardless of the tested classification method. Consequently, we expect that new features should be extracted from posters to improve the classification.

A few examples of correctly and incorrectly classified posters obtained using DR1 and M2 transformation method are presented in Table IV. Presented movie posters give some insight into the complexity of the problem.

TABLE IV. EXAMPLES OF CORRECTLY AND INCORRECTLY CLASSIFIED POSTERS

Movie poster	No. correct labels/ top ranked labels	Correct labels	Top 3 predicted labels
	2/2	Animation, Comedy	Animation, Comedy, Action
	1/2	Action, War	Action, Horror, Comedy
	0/2	Animation, Comedy	War, Horror, Action

V. Conclusion And Future Work

In this paper, automated detection of movie genres from posters was modeled as a multi-label classification task, where a single movie may belong to more than one genre. The multi-label classification problem was transformed into single-label classification by two kinds of data transformation methods. The first one transforms examples with multiple labels into multiple ordered pairs that are suitable for single label classification. The second extends the set of genre labels by forming combined labels. Both data transformation methods are algorithm independent and were tested with Naive Bayes and distance ranking classifiers and have performed comparably on the test set of 500 movie posters. A well-known multi-label classification algorithm RAKEL was also used with similar results.

The features used in the classification were low-level features based on color and edge combined with the number of detected faces on posters.

The best results were about 14% for two out of two correctly detected labels and 67% for at least one of two correctly detected labels. Since the results were similar using all the tested classification algorithms, we expect that the improvement in classification should primarily come from finding better image features that capture the relevant information about genres in posters. In the future work, we plan to test the dense SURF [13], GIST [14] and other visual low-level features used for scene representation [10] as well as features for text recognition. We also plan to test the classification on a much larger dataset with larger set of genres.

A subjective test will be conducted to determine human ability to detect genres from poster images, and the results will be used for comparison with automatic detection.

References

[1] [1] M. C. Potter, "Short-term conceptual memory for pictures." Journal of Experimental Psychology: Human Learning and Memory, vol. 2, no. 5, p. 509, 1976.

[2] [2] P. G. Schyns and A. Oliva, "From blobs to boundary edges: Evidence for time-and spatial-scale-dependent scene recognition," Psychological Science, vol. 5, no. 4, pp. 195–200, 1994.

[3] (2014, 3) The movie database. [Online]. Available: http://www.themoviedb.org/

[4] Z. Rasheed, Y. Sheikh, and M. Shah, "On the use of computable features for film classification," Circuits and Systems for Video Technology, IEEE Transactions on, vol. 15, no. 1, pp. 52–64, 2005.

[5] H. Zhou, T. Hermans, A. V. Karandikar, and J. M. Rehg, "Movie genre classification via scene categorization," in Proceedings of the international conference on Multimedia. ACM, 2010, pp. 747–750.

[6] H.-Y. Huang,W.-S. Shih, and W.-H. Hsu, "A film classifier based on low-level visual features," in Multimedia Signal Processing, 2007. MMSP 2007. IEEE 9th Workshop on. IEEE, 2007, pp. 465–468.

[7] (2014, 3) Allmusic. [Online]. Available: http://www.allmusic.com/

[8] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown, "Learning multi-label scene classification," Pattern recognition, vol. 37, no. 9, pp. 1757–1771, 2004.

[9] (2014, 3) Wikipedia. [Online]. Available: http://en.wikipedia.org/wiki/Hue

[10] M. Ivašić-Kos, M. Pavlić, P. Poščić, "Descriptors' Specification for Image Representation", Proceedings of International Conference on Image and Video Processing and Computer Vision. Orlando, Florida.

[11] (2014, 3) OpenCV. [Online]. Available: http://opencv.org

[12] G. Tsoumakas and I. Vlahavas, "Random k-label sets: An ensemble method for multi-label classification," in Machine Learning: ECML 2007. Springer, 2007, pp. 406–417.

[13] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in Computer Vision–ECCV 2006. Springer, 2006, pp. 404–417.

[14] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," International journal of computer vision, vol. 42, no. 3, pp. 145–175, 2001.

Detection of Roadside Vegetation Using Features from the Visible Spectrum

I. Harbaš and M. Subašić
Faculty of Electrical Engineering and Computing, Zagreb, Croatia
iva.harbas@fer.hr

Abstract – Detection of vegetation in images is a common procedure in remote sensing and is commonly applied to satellite and aerial images. Recently it has been applied to images recorded from within ground vehicles for autonomous navigation in outdoor environments. In this paper we present a method for roadside vegetation detection intended for traffic safety and infrastructure maintenance. While many published methods for vegetation detection are using Near Infrared images which are particularly suitable for vegetation detection, our method uses image features from the visible spectrum allowing the use of common on-board color cameras. Our feature set consists of color features and texture features. One of our specific goals was to identify a useful texture feature set for the problem of vegetation detection. Based on the feature set, the detection is implemented using a Support Vector Machine algorithm. For training and testing purposes we recorded our own image database consisting of different images containing roadside vegetation in various conditions. We are presenting promising experimental results and a discussion of specific problems experienced or expected in real-world application of the method.

Keywords **- image analysis, image processing, vegetation detection, traffic safety.**

I. INTRODUCTION

Detection and analysis of the state of vegetation from images is considered very important in the field of remote sensing where it is used for detecting green areas on Earth and detecting changes caused by urbanization. Detection of stressed vegetation [2], [10], [5] and other similar applications are aimed at raising environmental awareness and improving our ecological footprint.

Vegetation detection in other areas besides remote sensing, e.g., in robotics, is very young and unexplored. The need for detecting vegetation in this area arose when autonomous vehicles were used for forest exploration. Vegetation is not a static obstacle and all of vegetation-like obstacles don't need to be avoided (i.e. a vehicle can go over tall grass) [9]. Hence, it was necessary to detect vegetation and decide if the vehicle can drive over that obstacle or should it be avoided. Increased use of autonomous vehicles for off-road navigation is the main reason for new research efforts in detection and classification of vegetation [1], [3], [6], [7], [8].

In this paper we present a pixel-based method for roadside vegetation detection from within a moving vehicle intended for traffic safety and infrastructure maintenance (e.g., mowing the grass along the road).

Many published methods for vegetation detection are using Near Infrared (NIR) images where vegetation has distinctive properties. Our method uses image features from the visible spectrum, imitating how people detect vegetation using color and texture features and allowing us the use of a common on-board color camera. The benefit of this approach is that the same camera can be used for other computer vision tasks which require color images recorded from within a moving vehicle. It should be made clear that many publications mentioning successful vegetation detection often indicate some very specific species of vegetation or a very limited dataset. And these approaches are then just used for robots operating in a specific environment or for satellite images but not for roadside vegetation detection using a camera mounted on a vehicle. Due to different goals, those methods cannot be directly compared to our method for detection of vegetation.

A brief overview of some proposed methods for vegetation detection is given in the second section of this paper. Our approach is introduced in the third section where feature selection and classifier training is described. Our methods results will be presented in the fourth section. Conclusions and future work are presented in the last section of this paper.

II. RELATED WORK

People easily detect vegetation on the basis of typical visual characteristics such as color, texture and shape. Often researchers try to mimic the human visual system when developing methods for automatic detection of vegetation. Studies have shown that there is valuable information for vegetation detection in the spectrum invisible to the human eye and therefore, systems for detecting vegetation can be divided into two groups:

- Systems based on the visible spectrum
- Systems based on the invisible spectrum.

Systems based on visible spectrum use color and/or texture features. Using only color as a feature can give good results but objects of the same color as vegetation (e.g., green car) are falsely detected. Such objects can be distinguished from vegetation using additional texture features. A green car does not have a strong texture while the texture of vegetation is unstructured and turbulent.

This research has been partially supported by the European Union from the European Regional Development Fund by the project IPA2007/HR/16IPO/001-040514 VISTA - Computer Vision Innovations for Safe Traffic.

Some researchers use LiDAR (*Light Detection And Ranging*) sensors which give information about the 3D structure of the scene for calculating texture features. Nguyen et al. use a sliding cube across the 3D-point cloud in space to capture the local point statistics features. For every sliding cube they calculate a positive definite covariance matrix and extract the principle components (eigenvalues and eigenvectors) of that matrix. The relative differences of the three eigenvalues indicate a spatial structure. They segment the 3D-point cloud into three classes: surfaces, linear structures and porous volumes (foliage grass, tree canopy) [6], [9]. It can be argued that using only 3D-data can't result in a robust detection of vegetation because no color information is used. Therefore, Nguyen et al. address the 2D-3D fusion approach which combines 2D-3D information for vegetation detection [6]. For 2D features they use mean and standard deviation of brightness and color, and the difference of histograms [6]. These features are used to train a SVM (*Support Vector Machine*). The problem with this approach is mapping the 2D and 3D information because they are obtained by two different sensors. For this in [6] they use a coarse calibration method. This approach is time consuming and it should be used when time is not a criteria [9]. Liu et al., in a similar way, are using a combination of 2D and 3D features. The feature vector consists of the height calculated from the 3D space obtained with LiDAR, and the H and S components obtained by converting RGB (*Red, Green, Blue*) to HSV (*Hue, Saturation, Value*) color space [3].

Except for navigation, detection of vegetation is also used to improve the quality of video, or TV images. In [4] and [11] authors used texture features and color components from the YUV color space as color features. After detection of grass in video sequences they enhance the image by changing the color or brightness of pixels in that segment.

Using features from the visible spectrum has some problems (leaves change color, green cars, different illumination scenes, etc.) and that is why researchers started using features from the invisible spectrum. This idea was present for many years in remote sensing where researchers found that chlorophyll rich vegetation has a high reflectance in NIR part of the spectrum and can clearly be distinguished from the sky and other objects. Using this property of vegetation many different features can be calculated which are called Vegetation Indices. Mostly used is NDVI (*Normalized Difference Vegetation Index*) [8].

Although NDVI has been successfully used in remote sensing the work of Bradley et al. [1] has shown that there is a drastic difference between the view-points of a satellite and an autonomous ground vehicle where there are more problems typical for on-ground recordings such as shadow, shining and underexposure effects. Taking that into consideration and wanting to avoid time consumption needed for calibration and LiDAR scanning Nguyen et al. developed a modified vegetation index MNDVI [9] which they use as a feature for vegetation detection and all they need is NIR intensity and color information. In [8] a combination of vegetation indices with color and texture features is used for vegetation detection. First, by setting a threshold for NDVI and MNDVI they obtain pixels rich in chlorophyll. These pixels are seeds for spreading algorithm that follows. For each seed pixel distance in color and texture between the seed pixel and his neighbors is calculated.

Systems based on invisible spectrum require equipment for recording NIR images and systems based on visible spectrum that use LiDAR for feature extraction need an additional 3D scanner.

We require no additional sensors but a single color camera for data acquisition so the features that will be used for detection are from the visible spectrum only.

III. METHODOLGY

In this section we describe our method for detecting roadside vegetation using only features from the visible spectrum which does not require any additional equipment except a common color camera mounted onboard. Our method for vegetation detection comprises of three steps:

1. extraction of selected features,
2. pixel-based classification and
3. postprocessing.

To the best of our knowledge there is no publicly available database that could be used for our application. That is the reason we made our own database for testing and all results presented in this paper were obtained using this database. The database was recorded using a High Definition Camcorder Canon XF100 from a moving vehicle. The videos have been recorded in daylight at different times of the day to include different lighting conditions. The database includes different traffic scenarios including roadside vegetation in various conditions. Currently our database contains 270 images extracted from the recorded video sequences, each having 1920x1080 resolution. These images were manually segmented into vegetation and non-vegetation regions.

The intended use of our method requires that exact contours of each vegetation region are detected so we perform per-pixel classification into vegetation or non-vegetation classes.

A. Feature extraction

Features that we extract from color images to detect vegetation are color and texture features.

1) Color features

Human perception of color is one of the most important visual elements which help us visually recognize different objects. Color is an important descriptor of vegetation because vegetation doesn't have a specific shape but is usually represented by green, orange or yellow color. Transferring this human ability to a computer algorithm is far from simple. One major problem is the change in intensity and color in different

light conditions. This usually doesn't pose a problem for the human visual system but a change in luminance can cause a significant change in image features.

Some authors consider using different color spaces for color features (e.g., HSV [3], YUV [4], [11]), while others take into account the mean and standard variation values of intensity and color [6]. In [3] they use only H and S components, ignoring the luminance component V so their feature set would be less sensitive to light changes in the scene.

Taking this into account the first step was deciding which color features would be most suitable for our problem. Three color spaces commonly used in published vegetation detection methods were used in our test: RGB, HSV and CieLAB. We didn't test YUV color space because YUV is used for a specific analog encoding of color information in television systems and the only reason it is used in [4] and [11] is because they used it for TV image enhancement. CieLAB was added for testing because it is designed to mimic human perception of color. Additionally, we tested only the H and S components excluding the lightness from HSV and only A and B components excluding the *L* component from CieLAB.

We used five different feature vectors containing different color features to train a SVM classifier with linear kernel. This classifier was used only in this stage of testing to evaluate how appropriate each feature vector is for our problem.

To properly evaluate how each of our color features will generalize to an independent data set we use 10-fold cross-validation to measure the performance and to validate our results. The results are shown in Table I.

From these results we conclude that excluding the lightness component from the feature set does not affect the results drastically. Slightly better performance can be seen when using the lightness component with color information, i.e., better accuracy is achieved using HSV and LAB then HS and AB features respectively. The second conclusion we draw from Table I. is that RGB and LAB color spaces perform similarly and are giving better results compared to HSV. Vegetation color should theoretically be green in HSV color space under most different environment conditions. In reality, this is not always true for scenes containing sky or varying lighting conditions (shades, overexposure, and underexposure). Low intensity of the value V in HSV color space tends to turn the color of image to red, red brown, etc [6]. And that is why using HSV color space has more false positive errors where red objects in scenes are detected as vegetation. On the basis of conclusions previously presented, we decided to use RGB color components as color features.

2) Texture features

TABLE I. COMPARISON OF DIFFERENT COLOR SPACES USED FOR VEGETATION DETECTION

	Feature vector				
	RGB	***LAB***	***HSV***	***HS***	***AB***
Accuracy	92,6873	92,767	87,3196	87,1123	91,1518

Simple color features lack discriminative power required for our problem where non-vegetation objects may have colors similar to vegetation like illustrated in **Fig. 2 b.**. Additionally when dry vegetation, which takes on shades of yellow and red, is added to the training set, the number of falsely detected objects increases. To solve this problem we added texture features to the feature vector.

There are many different methods for calculating texture features in images and choosing one is a difficult task. Knowing that vegetation is diverse and that vegetation regions in images contain more information than homogeneous surfaces, we decided to use entropy as a texture feature. Entropy is a statistical measure of randomness and it can be described as a measure of the amount of disorder in a system. For images it can be expressed as a spread of states (gray levels) which the individual pixel can adopt. If pixels in an image, or in a block of an image, have the same values the entropy is zero and if an image (or a block) contains pixels changing in unexpected ways the entropy is high. We expect that vegetation region in an image have high entropy.

The entropy *H* of an image is defined as [12]:

$$H = -\sum_{k=0}^{M-1} p_k \log_2 p_k \qquad (1)$$

Where *M* is the number of gray levels and p_k is the probability associated with gray level *k*.

B. Feature ranking

Using the entropy of the grayscale image did not yield better performance so calculating entropy in different color spaces was considered. Color spaces taken into consideration were the same ones used for testing color features. There are a lot of possible combinations of color features and entropy in different color spaces that we could use as a feature vector for testing. To find the best one we used feature ranking as a filtering method where all these features are ranked, and based on the calculated rank we choose the best one.

To understand the influence of features on the system or even if the number of features is to large feature ranking is a good method to get baseline results and to assess features individually [13]. We created a feature vector that contains our selected three color features and 9 texture features that are entropy calculated for R, G, B, H, S, V, L, A and B components of the corresponding color space models. Total number of features is 12. Twelve features is not too much to handle for any kind of feature selection, but feature ranking is a good filter method and a good preprocessing step independent of the choice of the predictor [13]. The goal of using any kind of filtering of features is to eliminate possible outliers. Ranking was done using Wilcoxon rank sum test which is a nonparametric test for equality of population medians of two independent samples [14]. A nonparametric test was used because they don't assume normal distributed classes and our samples were tested and showed that they are not normally distributed.

Variable ranking makes use of a scoring function computed from the input variables (calculated features) and output variables (class assignment). For this test the scoring function is the rank sum statistic calculated for every feature in the feature vector. By convention, we assume that a high score is indicative of a valuable variable [13].

This test was repeated 10 times so that statistical significance can be established. For each of the ten different iterations, R, G and B features were always the three top ranking features followed by entropy calculated for saturation (S component of HSV color space). This is the only entropy that was mostly top ranked while others varied drastically from test to test. For this reason we added only this entropy to our feature vector.

C. Training the classifier

SVM is a well known supervised learning algorithm used for classification and in this paper it is used to discriminate vegetation and non-vegetation. An SVM is trained with a set of training examples each marked as belonging to one of two classes, in this case, vegetation and non-vegetation. Positive training examples were selected randomly from hand segmented vegetation regions in images and negative examples were also selected randomly but we made sure that different problematic objects (green, red, and yellow cars, green T-shirt of a pedestrian, yellow stripes on the road, etc.) are present in the negative set so the classifier would be trained on these examples too.

Selecting the training data is a very important step because all further classification is dependent on the trained model. Using only green vegetation for training makes a classifier that detects green vegetation but doesn't detect vegetation that is yellow or red (dry vegetation) (**Fig. 1.**) or different species of vegetation. Adding dry vegetation to the training set makes a classifier that detects dry vegetation but it detects more false objects as vegetation including yellow and red objects (and not only green objects).

After training examples are selected the SVM training algorithm builds a model that assigns new examples into one category or the other [15].

Our database contains 270 images of resolution 1920x1080 which presents 1920*1080*270 pixels available for training and testing the classifier. A smaller sample needs to be selected because training an SVM with a large number of samples can lead to overfitting and a very complex model can be built which can be time consuming during classification.

There are various recipes for calculating the required sample size which require knowledge of the variance or proportion in the population, the maximum desirable error, as well as the acceptable Type I error risk (confidence level).

It is possible to construct a table [16] that suggests the optimal sample size – given a population size, a specific margin of error and a desired confidence interval. Based on the table from [16] we found that the best sample size is 4000 (2000 samples representing vegetation and 2000 samples representing non-vegetation). These 4000 samples are chosen randomly. We have tested the algorithm using a greater number of samples but it only increased the computational time and it didn't improve accuracy.



a) b)

Fig. 1. Detection using: a) only green vegetation for training, b) all vegetation for training

D. Postprocessing

Classification is done for every pixel in an image. The resulting image with two classes marked often has some pixels marked as vegetation inside a non-vegetation region and non-vegetation pixels inside a vegetation region. It is safe to assume that pixels, or groups of pixels surrounded by vegetation also belong to vegetation. Also, solitary pixels or small groups of pixels surrounded by non-vegetation is often non-vegetation but is misclassified because it is too similar to underexposed vegetation (shadows under cars or parts of asphalt) or it has yellow reflectance (dry vegetation is mostly yellow). These gaps and solitary pixels can be removed after classification using a combination of morphological operations on the image.

In the postprocessing stage of our method we use morphological opening. The morphological open operation is an erosion followed by a dilation, using the same structuring element for both operations [17]. After opening further improvement was done by additionally removing solitary groups of pixels of certain size (for which we presume that are misclassified non-vegetation pixels) followed by filling up patches of certain size misclassified as non-vegetation (surrounded by vegetation).

IV. Experiments and Results

To find the optimal parameters for our method we tested several options for every aspect and choose the ones that gave the best accuracy.

Entropy is calculated for every pixel taking into account its neighborhood. The neighborhood size is a changeable variable. We tested four block sizes: 7x7, 9x9, 11x11, and 13x13. Experiments showed that the optimal neighborhood size is 9x9 (the smaller one showed worse accuracy, while the bigger ones did not improve accuracy). For training and classification, beside the linear kernel, we tested quadratic and radial kernels to take advantage of SVMs kernel trick and Cover's theorem. The best classification accuracy was obtained using radial basis SVM. For postprocessing, three variables have to be set: the structuring element, size of patches to remove, and size of patches to fill. We have used a circular structuring element with 5 pixel radius. At this step we presume that vegetation parts of an image are well connected and those parts in the image are large so

we remove the ones that are less than 3000 pixels and filling was done for patches less than 500 pixels big. The small value of 500 was used because there are parts of roadside vegetation that have utility shafts that are small and any bigger size of 500 would misclassify these parts as vegetation (**Fig. 2 b.** and **c.**, **Fig. 3 b.** and **c.**).

Adding texture features improved detection in problematic images. With these features the green objects in **Fig. 2 a.** are much better classified as non-vegetation seen in **Fig. 2 c.** compared to **Fig. 2 b.** where only color features were used. Also, the yellow stripes on the road were falsely detected as vegetation when only color is used as a feature (**Fig. 2 b.**) but adding texture improved detection (**Fig. 2 c.**). In **Fig. 2 c.** misclassified solitary pixels and groups of pixels can be seen which we improve using postprocessing (**Fig. 2.d.**).

For final testing of our method we divided our database into training and testing set. Training set consists of 180 images which is 2/3 of the total number of images in the database, and from this set we are randomly choosing 4000 pixel samples for training. The remaining 90 images were the testing set, and all their pixels were testing samples. Because we select our training samples from the training set randomly we must repeat the selection process to exclude the possibility of choosing "the perfect" training samples and presenting statistically insignificant accuracy. We ran 10 iterations of training and testing, every time randomly selecting the training set and achieved average accuracy of pixel classification is 94.995%. We used high definition images in our training and testing and because of that the classification is somewhat time consuming but at this stage of our research that is not a concern for us as we are not focused on real-time performance.

Some results of vegetation detection using the method presented in this paper are shown in **Fig. 3**. In **Fig. 3 a.** and **b**. more examples of green objects in scenes correctly



a) Original images

b) Detection uing color features only - green and yellow objects detected as vegetation

c) Detection using color and texture features

d) Result of vegetation detection after postprocessing

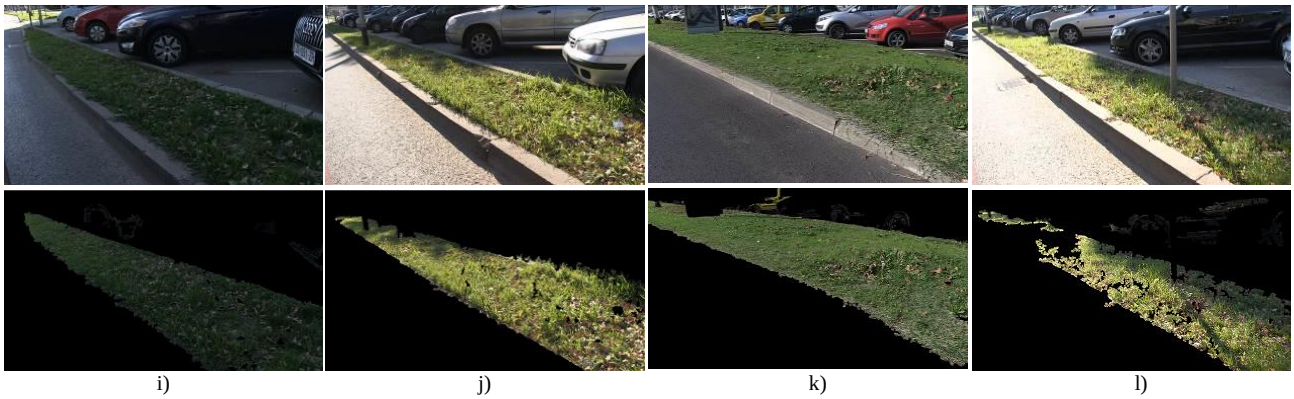
Fig 2. Results of vegetation detection for problem images

detected as non-vegetation can be seen. In **Fig. 3 c. – e.** we see good detection in different traffic scenes, while good performance in detecting vegetation in shade can be seen in **Fig. 3 f. - h.** Also good detection is presented in underexposure and overexposure conditions (**Fig. 3 i.** and **j.** respectively). In **Fig 3 k.** and **j.** we can see poorer detection results. In **Fig 3 k.** parts of the yellow car are still detected as vegetation even though we included these examples in the training set. This color is too similar to dry vegetation examples in the training set and because entropy is high over the edges the classifier decided that this is vegetation. In image **Fig 3 j.** vegetation that is far from the camera is not detected due to a lack of details in that part of the image, combined with a color deviating from green. Parts of the parked dark car are also detected as vegetation due to green reflections on its hull.



a) b) c) d)

e) f) g) h)



i) j) k) l)

Fig. 3. Results of detection using the method presented in this paper (top – original image, bottom – result of detection)

V. CONCLUSION

In this paper we present our method for detecting roadside vegetation using only features from the visible spectrum. This approach allows us to use common multi-purpose color cameras, and we are using both color and texture features. Based on our experiments, RGB color space proved to be the optimal choice for our color features. In the similar manner, we have selected entropy as our texture feature. We require a texture feature because color features alone were insufficient to solve problems with common situations where green non-vegetation objects occur in images. Entropy has mostly solved the problem of false positive green objects, but to further improve detection results we have implemented a post-processing step based on morphological operations. We have compiled our own data set for training and testing, with special attention paid to the training data set. It was compiled from random images and from several especially problematic images. The classifier we have used is SVM with radial kernel. We have obtained promising experimental results, and for our future work we plan to implement complex texture features from the frequency domain to increase accuracy. Real-time execution was not in the focus of presented research so current version of our vegetation detection method is not optimized for speed.

REFERENCES

[1] D.M. Bradley, R. Unnikrishnan, and J. Bagnell. "Vegetation detection for driving in complex environments". In Robotics and Automation, 2007 IEEE International Conference on, pages 503–508, 2007.

[2] E. J M Carranza and M. Hale. "Remote detection of vegetation stress for mineral exploration". In Geoscience and Remote Sensing Symposium, 2001. IGARSS '01. IEEE 2001 International, volume 3, pages 1324–1326 vol.3, 2001.

[3] Liu da xue, Wu Tao, and Dai Bin. "Fusing ladar and color image for detection grass off-road scenario". In Vehicular Electronics and Safety, 2007. ICVES. IEEE International Conference on, pages 1–4, 2007.

[4] Stephen Herman, Johan Janssen, Erwin Bellers, and James Wendorf. "Automatic segmentation-based grass detection for real-time video", December 14 2004. US Patent 6,832,000.

[5] P. Kempeneers, S. De Backer, W. Debruyn, P. Coppin, and P. Scheunders. "Generic wavelet-based hyperspectral classification applied to vegetation stress detection". Geoscience and Remote Sensing, IEEE Transactions on, 43(3):610–614, 2005.

[6] D. V Nguyen, L. Kuhnert, T. Jiang, S. Thamke, and KD Kuhnert. "Vegetation detection for outdoor automobile guidance". In Industrial Technology (ICIT), 2011 IEEE International Conference on, pages 358–364, 2011.

[7] D. V Nguyen, L. Kuhnert, S. Thamke, J. Schlemper, and K. D Kuhnert. "A novel approach for a doublecheck of passable vegetation detection in autonomous ground vehicles". In Intelligent Transportation Systems (ITSC), 2012 15th International IEEE Conference on, pages 230–236, 2012.

[8] D.-V. Nguyen, Lars Kuhnert, and Klaus Dieter Kuhnert. "Spreading algorithm for efficient vegetation detection in cluttered outdoor environments". Robotics and Autonomous Systems, 60(12):1498 – 1507, 2012.

[9] D.-V. Nguyen, Lars Kuhnert, and Klaus Dieter Kuhnert. Structure overview of vegetation detection. "A novel approach for efficient vegetation detection using an active lighting system". Robotics and Autonomous Systems, 60(4):498 – 508, 2012.

[10] Zhuokun Pan, Fang Wang, Lihua Xia, and Xiaoxuan Wang. "Feature extraction for urban vegetation stress identification using hyperspectral remote sensing". In Information Science and Engineering (ICISE), 2010 2nd International Conference on, pages 250–254, 2010.

[11] B. Zafarifar and P.H.N. de With. "Grass field detection for TV picture quality enhancement". In Consumer Electronics, 2008. ICCE 2008. Digest of Technical Papers. International Conference on, pages 1–2, 2008.

[12] Gonzalez, R.C., R.E. Woods, S.L. Eddins, *Digital Image Processing Using MATLAB*, New Jersey, Prentice Hall, 2003, Chapter 11.

[13] Isabelle Guyon and André Elisseeff. 2003. "An introduction to variable and feature selection". *J. Mach. Learn. Res.* 3 (March 2003), 1157-1182.

[14] Japkowicz, N., Shah, M. "Evaluating Learning Algorithms: A Classification Perspective". Cambridge University Press, New York, NY, USA. 2011.

[15] Christopher J. C. Burges. 1998. "A Tutorial on Support Vector Machines for Pattern Recognition". *Data Min. Knowl. Discov.* 2, 2 (June 1998), 121-167. DOI=10.1023/A:1009715923555 http://dx.doi.org/10.1023/A:1009715923555.

[16] Robert V. Krejcie and Daryle W. Morgan. "Determining Sample Size for Research Activities". In Educational and Psychological Measurement, #30, pp. 607-610. 1970.

[17] Dougherty, E. "An introduction to morphological image processing". Tutorial texts in optical engineering. SPIE Optical Engineering Press. 1992.

Predictions of cloud movements and the Sun cover duration

Aleksander Radovan* and Željko Ban**
* University of Applied Sciences Velika Gorica, Velika Gorica, Croatia
** Faculty of Electrical Engineering and Computing, Zagreb, Croatia
aleksander.radovan@vvg.hr
zeljko.ban@fer.hr

Abstract – In order to optimally control the energy flow in the microgrid, consist of several renewable energy sources, including solar photo electric system and storage system, it is essential to know availability of the specific energy source. So, it is necessary to determine the exact moment of the Sun shading the by clouds as well as shadow duration. These parameters could be obtained by cloud movement tracking system. The cloud movement tracking system is based on the detection of cloud edges and determining angular velocity vector which describes the speed and direction of cloud movements. As clouds change their shapes, directions and speeds, the position of every cloud is determined according to his centroid. This paper deals with cloud recognition and movements analysis based on video stream obtained by camera with resolution 1920x1080 pixels and refresh rate of 25 frames per second. The clouds detection is based on the HSV (Hue-Saturation-Value) color model in combination with mathematical morphological operations of erosion and dilation. Video and image processing was performed by application developed in Java programming language using the JavaCV library that offers the functionalities of computer vision.

Keywords – renewable energy sources, cloud movement prediction, Sun cover duration, centroid, HSV

I. INTRODUCTION

Environmental requirements as well as increased demands for energy cause increased share of renewable energy sources in electric power production. On the other hand, increased share of renewable based on solar and wind power increases challenge of maintaining power grid stability, due to their stochastic nature. In order to maintain the desired level of availability it is essential to use different energy storage system in the microgrid and virtual power plants. One of the crucial data in optimizing microgrid power flow is prediction of the availability of the specific power source. The main uncertainty source of the solar irradiance for solar energy sources are clouds and their movement. Predicting the instant of the shading of photo panels, as well as duration of the shading, allows more effective management of the microgrid. As weather forecasts mostly obtain irradiance prediction at the hourly level it is crucial to obtain short term cloud movement prediction [1], [3].

This paper deals with reduction of the cloud movement prediction sampling period to only a few seconds using cloud movement tracking by video stream processing. The camera frame rate was set to 25 frames per second and by processing every 100th or 300th image, it is possible to set the sampling frequency to four seconds (25 x 4 = 100) or to 12 seconds (25 x 12 = 300), etc.

After detecting the cloud edges or contours that accurately determine the position of the clouds on the image being processed, it is necessary to also determine the movement direction and velocity of clouds in the cylindrical coordinate system expressed in radians per second. As the reference point between two cloud positions, the centroid was chosen. Using that information it is possible to estimate the moment when the nearest edge of the cloud will reach the Sun surface and begin to block the Sun. In a similar way it is possible to determine how long the shadow will last.

In the following chapters the principles used to detect the edge of the clouds were presented, based on HSV (Hue-Saturation-Value) color model and morphological operation of erosion and dilation. Every processing step is picture illustrated and the final results of cloud edge detection are shown.

II. HSV COLOR MODEL

In contrast to the commonly used RGB (Red-Green-Blue) color model, which is based on creating different shades of color by adding different intensities of red, green and blue (ranging from 0 to 255), in this paper the HSV color model is used. The HSV color model is based on three parameters: hue, saturation and value. Hue defines the color, saturation defines the amount of color, and value defines the level of brightness in color, or the amount of black added to the color. Instead of the Cartesian coordinate system that is used in the case of the RGB color model, HSV uses a cylindrical coordinate system (Figure 1). [4]

Hue specifies angles from the red color, which is at 0˚, from the yellow color, which is at 60˚, from the green color, which is at 120˚, from the cyan color, which is at 180˚, from the blue color, which is at 240˚ and from the magenta, which is at 300˚. Saturation with value of 1.0 result with the maximum of color amount, and the lower values that 1.0 result in a color that is closer to white color. In case of the saturation value of 0.0, the shade of gray color occurs, and that depends on the level of parameter values of HSV color model. For example, if the color is set to red, saturation and value are set to 1.0, resulting color is pure red. In the case of reducing the saturation red color becomes lighter and eventually

becomes pink, and with additional reduction of the saturation becomes white. [4]

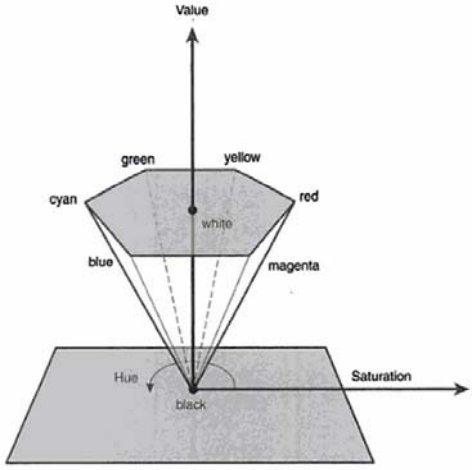


Figure 1. The cylindrical coordinate system of the HSV color model [4]

The conversion of RGB component values to the HSV color model begins by assigning a maximum RGB component value to value parameter of the HSV color model. The value of saturation parameter is determined by calculating the difference between the maximum and minimum values of the RGB components, and this difference must be divided with the highest value among RGB components. The first special case represent RGB values set to 0, which represents the black color, and the HSV parameters saturation and value also have the value 0 (in this case the value of hue parameter has no effect on the resulting color). The second special case represent all RGB values as equal, which represents the shade of gray color, and means that the saturation value is 0. The hue parameter in the HSV color model is determined by one of the methods shown in the following terms, depending on the values of the RGB components. The complete algorithm for determining the HSV values is shown below (1). [4]

$$\max_RGB = \max(R, G, B)$$

$$\min_RGB = \min(R, G, B)$$

$$d = \max_RGB - \min_RGB$$

$$V = m_RGB$$

$$S = \begin{cases} 0, d = 0 \\ \dfrac{d}{\max_RGB}, d \neq 0 \end{cases}$$

$$H = \begin{cases} nedefiniran, S = 0 \\ 60° \times \left(\dfrac{G - B}{d} \bmod 6 \right), R = m_RGB \\ 60° \times \left(\dfrac{B - R}{d} + 2 \right), G = m_RGB \\ 60° \times \left(\dfrac{R - G}{d} + 4 \right), B = m_RGB \end{cases} \quad (1)$$

III. CLOUD EDGE DETECTION

Cloud edge detection consists of several steps: image conversion from RGB to HSV color model, the execution of basic morphological operations of erosion, dilation and detection of the contours that represent the edges of the clouds.



Figure 2. Initial cloud image in RGB color model

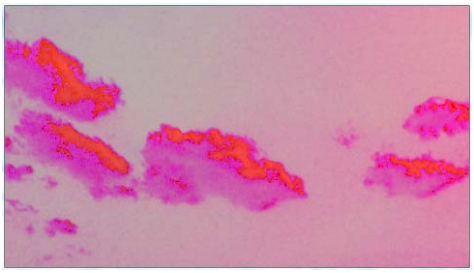


Figure 3. Converted cloud image to BGR and afterwards to HSV color model

A. CONVERSION FROM RGB TO HSV COLOR MODEL

According to the rules for converting the image from RGB to HSV color model described by (1), the transformation results with the same output – the appearance of the initial picture remains unchanged (Figure 2). OpenCV library and thus JavaCV by default use a little different color model called BGR (Blue-Green-Red), which is very similar to RGB, with only difference in the color component order (R and B switch places). Using BGR color model the initial cloud image (Figure 2) is converted to image with predominantly red colors instead of blue colors (Figure 3).

After the conversion to BGR and HSV color model, it is necessary to perform morphological operations of erosion and dilation. Morphological operations use structural elements of the given size and shape based on the values of neighboring pixel perform transformations in the input image. The most basic morphological operations are erosion and dilation. Before performing these operations, the cloud images must be converted to black and white.

B. EROSION

Erosion is used to remove the pixels at the edges of cloud contours within the image. According to [9], erosion is morphological transformation which combines two sets using vector subtraction of set elements. If A and

B are sets in Euclidean *N*-space (E^N), then the erosion of *A* by *B* is set of all elements *x* for which $x+b \in A$ for every $b \in B$. The erosion of *A* by *B* is denoted by $A \ominus B$ and is defined by

$$A \ominus B = \left\{ x \in E^N \,\middle|\, x+b \in A \text{ for every } b \in B \right\} \quad (2)$$

If equation (2) uses the expression for difference of elements *a* and *b* and sets *A* i *B* respectively, the equation (2) becomes:

$$A \ominus B = \left\{ x \in E^N \,\middle|\, \text{for every } b \in B \text{, there exist an } a \in A \text{ such that } x = a - b \right\} \quad (3)$$

According to [9], the utility of the erosion transformation is better appreciated when the erosion of is expressed in a different form. The erosion of an image *A* by structuring element *B* is the set of all elements *x* of E^N, for which *B* translated *x* is contained in *A*:

$$A \ominus B = \left\{ x \in E^N \,\middle|\, (B)_x \subseteq A \right\} \quad (4)$$

Thus the structuring element *B* may be visualized as a probe which slides across the image *A*, testing the spatial nature of *A* at every point. Where *B* translated to *x* can be contained in *A* (by placing the origin of *B* at *x*), then *x* belongs to the erosion $A \ominus B$.

According to [9], erosion of an image *A* by a structuring element *B* is the intersection of all translations of *A* by the points –*b*, where $b \in B$:

$$A \ominus B = \bigcap_{b \in B} (A)_{-b} \quad (5)$$

Erosion rule can be in the case of binary images and sets represented by *X* and *Y* coordinates of pixels shown in Figure 4 interpreted as follows: the value of the output pixel is the minimum value of all neighboring pixels covered by the structuring element. In case of detecting edges of clouds with that operation the more precise contours are produced.

As a structural element which performs erosion operation, the matrix of dimensions [3, 3] is used, wherein for the central pixel (surrounded by the checked adjacent pixels) the pixel with coordinates (1,1) is chosen. The result of executing erosion operation of black and white image (Figure 3) is shown in Figure 4.

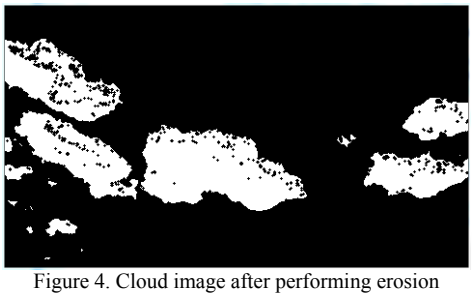


Figure 4. Cloud image after performing erosion operation

In Figure 4. it can be noticed that the edges of the clouds are precisely determined, but on the cloud surfaces black spots are visible – they are undesirable effects of erosion operation. To be able to determine precise contours of each cloud on the image, performing morphological operation of dilation can solve the problem of black spots caused by erosion operation.

C. DILATION

Dilation is used to add pixels to the edges of an object within image. According to [9], dilation is the morphological transformation which combines two sets using vector addition of set elements. If *A* and *B* are sets in Euclidean N-space (E^N) with elements *a* and *b*, respectively, $a = (a_1, \cdots, a_n)$ and $b = (b_1, \cdots, b_n)$ being *N*-tuples of element coordinates, then the dilation of *A* by *B* is the set of all possible vector sums of pairs of elements, one coming from *A* and one coming from *B*. According to [9], assuming that *A* and *B* are subsets of E^N, the dilation of *A* by *B* is denoted by $A \oplus B$ and is defined by

$$A \oplus B = \left\{ c \in E^N \,\middle|\, c = a + b \text{ for some } a \in A \text{ and } b \in B \right\} \quad (5)$$

Furthermore, assuming that *A* is a subset of E^N and $x \in E^N$ then the translation of *A* by *x* is denoted by $(A)_x$ and is defined by [9]:

$$(A)_x = \left\{ c \in E^N \,\middle|\, c = a + x \text{ for some } a \in A \right\} \quad (6)$$

According to [9], the dilation of *A* by *B* can be computed as the union of translations of *A* by the elements of *B*:

$$A \oplus B = \bigcup_{b \in B} (A)_b \quad (7)$$

Dilation rule can be in the case of binary images and sets represented by *X* and *Y* coordinates of pixels shown in Figure 5 interpreted as follows: the value of the output pixel is the maximum value of all neighboring pixels covered by the structuring element.

Dilation operation solves the problem of black spots that appeared on cloud surfaces after erosion operation, thus the image is ready for detecting contours and edges of clouds.

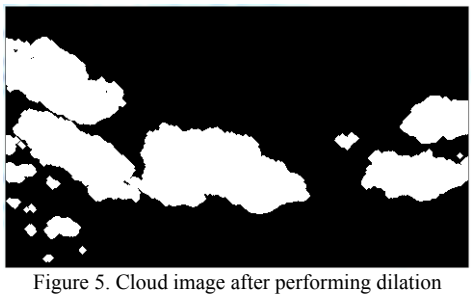


Figure 5. Cloud image after performing dilation operation

Figure 5. shows minimized amount of black points on the surfaces of clouds, while the edges are more rounded than showed in Figure 4. As a structural element that performs

dilation operations, the matrix with dimensions [3, 3] is used, wherein as for the central pixel (surrounded by the checked adjacent pixels) the pixel with coordinates (1,1) is chosen, as in the erosion operation.

D. CLOUD EDGE DETECTION

After performing the dilation operation, in Figure 5 the cloud shapes are clearly visible, which facilitates the detection of edges or contours of the clouds. The contour of the object is represented by a list of points at the edge of the object. [6]



Figure 6. Detected cloud edges

In order to find contours corresponding to the edges of the clouds, the Suzuki contour tracing algorithm [7] was used. Only the outer contours that represent clouds were used. The resulting contours were drawn to Figure 1 and Figure 6 was the final result.

After finding the edges of the cloud it is necessary to determine the direction and velocity of the clouds in order to predict the location of clouds at a certain point in the future, and thus to predict the moment when a cloud reaches the Sun. Regarding how clouds in motion very often change their shape, their edges cannot be used as movement reference points. In this paper, as a reference point for determining the movement of clouds, the centroid of every closed contour (or clouds) was chosen. Also, very small clouds were ignored and the centroid calculation was not performed.

IV. DETERMINING THE ANGULAR VELOCITY VECTOR OF CLOUDS

To determine the direction and velocity of cloud movement, the centroid of every cloud is used to track its position in each image processing cycle. In order to measure the movement, time interval between the measurements must be large enough. Since every 300th image of the video was processed, which means that the interval of cloud position analysis was performed every 12 seconds (300 frames / 25 frames per second = 12s).



Figure 7. Directions of cloud movements

In order to determine velocity and direction of cloud movement, the spherical coordinate system is used. For every cloud the angular velocity vector is determined by analyzing change of the centroid location from the start position $T_0(x_0, y_0)$ to the new position $T_1(x_1, y_1)$.

Figure 7. shows how to determine the direction and movement velocity of clouds. The black point on the cloud surface indicates the new location of centroid after movement, and the red line indicates a change of the centroid position from initial to final position within the analysis cycle (12 seconds). Only clouds with considerable area are taken into consideration.

Angular velocity of each cloud is determined by changes of two components: azimuth and elevation. Azimuth is the angle formed between the North and the observation angle (in range between 0 and 360 °), and elevation is the angle between the horizon and the height of the object being observed (in range between 0 i 90°). Figure 8 shows an example of the object in the sky with azimuth of 156° and elevation of 38°. When determining the azimuth and elevation of objects on the images processed from video, the zenith (the point in the sky above the observer) is placed on the horizontal center of the image (if the image resolution is 1920x1080, the zenith is placed at 960th horizontal pixel so the location of the observer is set on location with coordinates X = 960 and Y = 1080). According to the location of the observer, the angle measured between initial and final centroid points of one measurement interval is calculated and thus determine the first component of angular velocity. Elevation angle is determined by the height of the image (1080 pixels) divided by the maximum elevation angle (90°) and therefore determines the amount of elevation: vertical shift of cloud centroid in pixels multiplied by the quotient of the vertical resolution of the image and the amount in degrees. Taking into account the time interval in which the shift occurred, the second component of angular velocity can be determined.

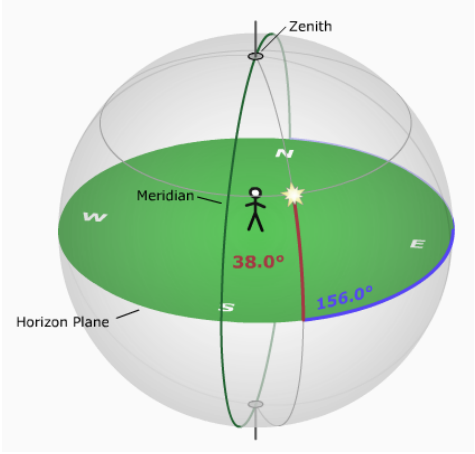


Figure 8. Azimuth and elevation of object on the sky (source: [8])

Figure 9. shows an example of cloud movement, where the centroid as reference point and the observer point is set on coordinated 960x1080. In the analysis of the first image, initial position of cloud centroid was on point

$T_0(x_0, y_0)$, and after 12 seconds (300 frames interval), centroid has moved to point $T_1(x_1, y_1)$ and changed the angle by $\Delta\varphi$. The shift resulted with the change of centroid location, but also the change of azimuth and elevation. According to Figure 10 it is possible to determine the difference between points $T_0(x_0, y_0)$ i $T_1(x_1, y_1)$, consisting of the components Δx i Δy:

$$\Delta x = x_2 - x_1 \tag{8}$$

$$\Delta y = y_2 - y_1 \tag{9}$$

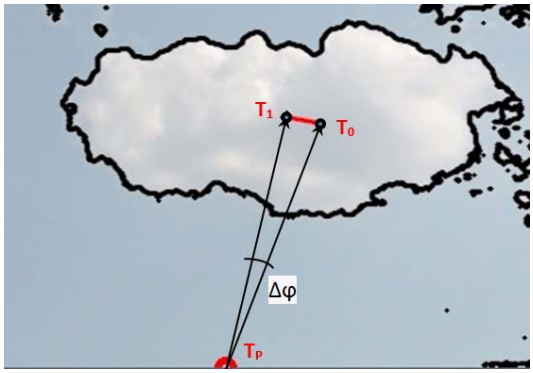


Figure 9. Shift of the cloud (centroid) from the start point T_0 to the final point T_1 from the observer point of view T_P

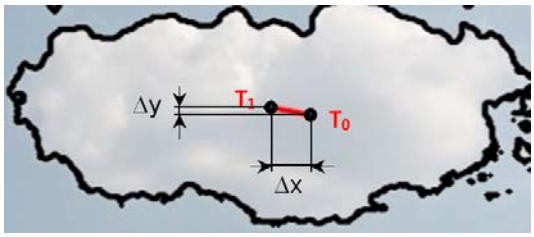


Figure 10. Coordinates change from the start point T_0 to the final point T_1

Assuming that the point $T_P(x_P, y_P)$ is the location of the observer, it is possible to determine the change of the angle $\Delta\varphi$ using simple trigonometric expressions by determining the initial angle at point $T_0(x_0, y_0)$ and afterwards the angle at point $T_1(x_1, y_1)$. The difference between those angles is:

$$\varphi_0 = \arctan\frac{y_P - y_0}{x_P - x_0} \tag{10}$$

Similar to that, it is possible to determine the angle φ_1 of the point $T_1(x_1, y_1)$ and then calculate the change of the angle:

$$\Delta\varphi = \varphi_1 - \varphi_0 \tag{11}$$

To determine the scalar azimuthal angular velocity following expression can be used:

$$\omega_A = \frac{d\varphi}{dt} = \frac{\Delta\varphi}{\Delta t} \tag{12}$$

For calculation of the angle change the expression (12) can be used, and the time interval is defined by the frequency of processing images from video. For example, if every 300th video image is processed, the video quality is set to 25 fps, the time interval is 12 seconds (300 / 25 = 12). After conversion of degrees to radians and use of vector calculus the expression for angular velocity vector is as follows:

$$\vec{\omega}_A = \vec{\omega}_{A1}\omega_A\frac{\pi}{180} \tag{13}$$

where $\vec{\omega}_{A1}$ is a unit azimuth vector with horizontal direction, and ω_A scalar azimuth velocity from the expression (12).

After determining the azimuthal angular velocity, it is necessary to determine the elevation angular velocity, to be able to sum these two vectors and calculate resultant angular velocity. As the elevation is in range between 0 i 90° and vertical resolution of the image is 1080 pixels, the shift of the cloud by 12 pixels (1080 pixels / 90°) corresponds to the shift of one degree of the elevation. According to Figure 10, the vertical shift of the cloud centroid Δy can be determined and according to that shift it is possible to calculate the scalar value of the elevation angular velocity by using following expression:

$$\omega_E = \frac{dy}{dt} = \frac{\Delta y}{\Delta t} \tag{14}$$

where Δt is defined by the frequency of image processing from video. After conversion of degrees to radians and use of vector calculus the expression for angular velocity vector is as follows:

$$\vec{\omega}_E = \vec{\omega}_{E1}\omega_E\frac{\pi}{180} \tag{15}$$

where $\vec{\omega}_{E1}$ is a unit elevation vector with vertical direction, and a ω_E scalar elevation velocity from the expression (15).

The resultant velocity vector can be calculated by the sum of azimuth and elevation angular velocity:

$$\vec{\omega}_R = \vec{\omega}_A + \vec{\omega}_E \tag{16}$$

V. PREDICTION OF CLOUD MOVEMENTS AND SUN COVER DURATION

To accurately predict the duration of Sun cover with clouds, an approximation shapes of ellipse and rectangle were used (Figure 11 shows one cloud whose shape is approximated with ellipse and rectangle). With using these shapes, determining the contours is much simplified and it is easier to describe shapes with mathematical

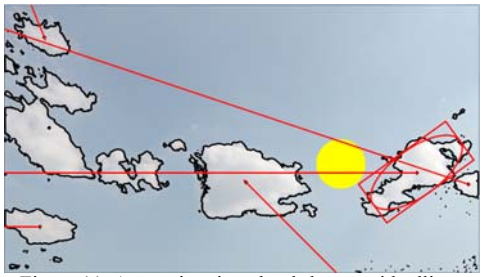


Figure 11. Approximation cloud shapes with ellipse and rectangle

expressions. In addition, Figure 11 shows the movement directions of clouds (marked red), beginning with centroid points to the edges of area covered with camera (image). Simulation the location of the Sun is performed by drawing a yellow filled circle at a given location in the image.

With taking the cloud contour into consideration approximated with ellipse or rectangle and movement direction of the cloud, it is possible to precisely determine the moment of creating the shadow on the Sun and the duration of the shadow. Table 1 shows the results comparison between the approximated cloud shapes and the real cloud contours. Firstly, the starting reference point in time was set. From that point by using approximation of the ellipse and the rectangle, and taking into account the direction of movement of clouds, it is determined the exact moments when the cloud starts to cover the Sun and when the area of cloud finally stops with Sun covering. These results were compared with the real situation and therefore the error rate is shown in Table 1.

TABLE I. RESULTS OF PREDICTING THE MOMENT OF COVERING THE SUN BY APPROXIMATING THE SHAPE OF CLOUDS AND THE REAL CLOUD CONTOURS

Measure ment	**Type of approximation**	**Error in prediction**	
		Start of covering	***End of covering***
1.	Ellipse	16,7%	1,3%
	Rectangle	16,7%	16,7%
2.	Ellipse	115,6%	0,7%
	Rectangle	102,1%	4,39%

VI. EXPERIMENTAL RESULTS

Table 1 shows the results of predicting the moment of cloud covering the Sun using ellipse and rectangle approximation, and determining the error compared to real cloud contours without approximation.

Measurements were performed using various cloud sizes and different movement directions. The most significant impact on error rate had the sudden change of cloud shape and their "connecting" with other clouds. In cases where the data calculation on which the prediction was based is performed more frequently with shorter time intervals, it is possible to significantly reduce the error.

VII. CONCLUSION

The forecast of photo panel shading technique is obtained in this paper in order to decrease uncertainty in prediction of power production which is necessary for effective microgrid control. The paper deals with the photo panel shading caused by clouds. The prediction of shading is based on the clouds movement prediction technique which uses rectangular and ellipsoidal cloud shape approximation and centroid position movement calculation from video stream. Erosion and dilation operations are crucial in recognizing and detecting clouds while the direction and speed of the cloud are calculated according to centroid shift between two measurements.

Developed technique allows precise and frequent short-term shading prediction useful for increasing accuracy of the commercial hourly weather forecast.

VIII. REFERENCES

[1] P.Wood-Bradley, J. Zapata, John Pye: „Cloud tracking with optical flow for short-term solar forecasting", Solar Thermal Group, Australian National University, Canberra, Australia, 2012.

[2] B. Goswami, G. Bhandari: „Automatically Adjusting Cloud Movement Prediction Model from Satellite Infrared Images, India Conference (INDICON), 2011 Annual IEEE

[3] H.Shi: „Cloud Movement Detection for Satelite Images", Signal Processing Proceedings, 1998. ICSP '98, Beijing

[4] J. Gomes, L. Velho, M. Costa Sousa: Computer Graphics: Theory and Practice, CRC Press, ISBN-13: 978-1568815800, 2012., pp. 21-24

[5] Mathworks Documentation Center „Morphology Fundamentals: Dilation and Erosion",

http://www.mathworks.com/help/images/morphology-fundamentals-dilation-and-erosion.html

[6] G. Bradsky, A. Kaehler: "Learning OpenCV", O'Reilly Media, ISBN-13: 978-0596516130, 2008.

[7] S. Suzuki, K. Abe: "Topological Structural Analysis of Digitized Binary Images by Border Following", Computer Vision, Graphics and Image Processing, 1985.

[8] Azimuth/Altitude Demonstrator, http://astro.unl.edu/classaction/animations/coordsmotion/altazimuth.html

[9] R.M. Haralick, S.R. Sternberg, X. Zhuang: Image Analysis Using Mathematical Morphology, IEEE Transactions and Pattern Analysis and Machine Intelligence, vol. PAM1-9, No. 4, July 1987.

[10] R.C. Gonzales, R.E. Woods, Digital Image Processing, Third edition, Prentice Hall, ISBN-13: 978-0-13-168728-8,2008.,pp.627-782

[11] D. Neelima, K. L. Rao: A Moving Object Tracking and Velocity Determination, International Journal of Advanced Engineering Sciences and Technologies, Vol No. 11, Issue No. 1, 096-1

Intelligent Modelling for Predicting Students' Final Grades

Monika Simjanoska, Marjan Gusev and Ana Madevska Bogdanova
University Ss. Cyril and Methodius, Faculty of Computer Science and Engineering,
Skopje, Macedonia
Email: m.simjanoska@gmail.com, {marjan.gushev, ana.madevska.bogdanova}@finki.ukim.mk

***Abstract*—The main objective of this paper is producing an intelligent virtual teacher who will be able to predict the students' final grades at the end of the semester. Our approach is based on continual observation of the student's activities on the particular course during the semester. In order to achieve realistic modelling of the students' devotion to the given lectures and also the degree of how much the student has learned from the given lecture, we take into account both the e-Learning and the e-Assessment results through the semester. In our previous work we did an intelligent students' Profiling to classify the students into a pass, or, fail category. In this paper we go deeper into the problem, achieving more precise modelling according to which we will be able to determine the student's most likely final grade, using multi classification methodology. The advantage of our model is in its ability to take into account all the assessments during the semester, not relying only on the results from the last student's assessment. It can be a good indicator whether the teacher needs to perform additional testing of the student's knowledge in order to derive an overall conclusion on the most appropriate grade.**

***Index Terms*—e-Learning, e-Assessment, Machine Learning**

I. Introduction

E-Learning, or Computer Assisted Learning, refers to the learning by using sophisticated information and communication technologies, whereas, the e-Assessment, or Computer Assisted Assessment, aims to assess the learner's knowledge. Vonderwell et al. [1] make a clear distinction between assessment of learning (assessment for the purposes of grading and reporting with its own established procedures) and assessment for learning (assessment whose purpose is to enable students, through effective feedback, to fully understand their own learning and the goals they are aiming for). A good overview of e-Assessment systems is presented by Gusev and Armenski [2]. In our research the assessment for learning is referred to as e-Learning and the assessment of learning is referred to as e-Assessment.

Rovai [3] in his research on online and traditional learning identifies the major principles of general assessment theory, examines how these principles can be applied to an online environment, and identifies and describes several assessment issues that have special significance in a virtual classroom. Practical suggestions are included to assist both first-time and experienced online instructors develop assessment components for their online courses.

Various conflicting attitudes about distance learning have been discussed by Hannay et al. [4] in order to present the perceptions of both the students and the teachers. Sun et al. [5] developed an integrated model with six dimensions: learners, instructors, courses, technology, design, and environment to investigate the critical factors affecting learners' satisfaction in e-Learning. The results showed that learners' computer anxiety, instructors' attitude toward e-Learning, e-Learning course flexibility, e-Learning course quality, perceived usefulness, perceived ease of use, and diversity in assessments are the critical factors affecting learners' perceived satisfaction.

The research that we conduct in this paper is a result of a long term e-Learning and e-Assessment process during the course on Computer Architecture and Organization (CAO). We believe that besides the benefit of the intelligent prognosis of the students' final grades, the results will be a useful indicator for the teachers of whether the given lectures for the e-Learning process are well formed and represent the real students' knowledge, and also whether the students have adapted to this kind of learning a can fully express their knowledge.

Beside the surveys on the students' satisfaction, universities have been given a lot of advices to turn their interests towards cloud computing, since the cloud offers good resources flexibility and scalability, storage, computational requirements, network access, reduced power consumption and lower costs [6], [7], [8]. As an upgrade to the e-Learning and e-Assessment system that we use in this paper, Ristov et al. [9] propose a SOA architecture of a cloud hosted e-Assessment system which uses scalability and elasticity in order to achieve sustainable performance. Their solution reduces the overall costs since it uses minimum resources utilized only during the e-Assessment.

However, the focus of this research is not to analyse the existing system's performance and cost, but to introduce an intelligence in the e-Learning and e-Assessment processes which will result in producing an intelligent grading agent (IGA) whose decisions will be as similar as possible to the teacher's opinion on the student's final grades. For that purpose we use the results from the online learning system that has been implemented at our university since 2002. Based on 10 years of e-Learning and e-Assessment experience, we have developed a strategy by which we proved that the student has achieved all learning objectives (LOs) by providing correct answers for relevant knowledge items during the online learning process. During the strategy development we confronted issues of guessing the correct answers in a random manner,

which we solved by introducing negative scoring for wrong answers. This is how we expect the students to answers the given questions only when they know the correct answer. During the development of the strategy, we also determined the minimal number of questions the system should ask to get relevant assessment of student's knowledge. After analysing several strategies, we agreed that the least vulnerable strategy that offers satisfactory results is 3R (three correct answers in a row). The probability analysis showed that the probability to randomly choose three consequently correct answers is smaller than the probability to choose any three correct answers from a given set. The overall contribution of this research is that we developed strategy by which the students are asked as minimum questions as 3 and expect 3 correct answers in a row, in order to allow them to continue the online learning process. This strategy somehow confirms the student's knowledge on a particular learning objective.

Therefore, the resulting online learning system is with the following features:

- *Corrective measure* - whenever the student gives a wrong answer the system suggests corrections.
- *Adaptive learning path* - better students can move faster towards the more complex learning objectives.
- *Clear objectives and goals* - the navigation explains what else has left to be learned.

However, this approach does not estimate the final grade of the e-Assessments, nor evaluates the student efficiency, and this is the motivation to conduct a research whose results are reported in this paper.

The work reported in this paper is based on our previous research [10] where we achieved modelling of the students overall work during the semester, taking into account the results from both the e-Learning and the e-Assessment, to determine whether the student will pass or fail the CAO course. We believe that this kind of grading that depends on the students' overall work through the whole course will additionally motivate them to work continuously during the semester and present their real knowledge during the e-Learning process, since it directly affects their final grade.

Our research mainly consists of two parts. At first we verify the existing methodology for classifying the students into a pass or fail class where we use most common Support Vector Machines (SVM) kernels. The second part of the research reported in this paper focuses on building an IGA able to estimate the students' final grade.

Figure 1 presents the interaction between the online learning system and our intelligent agent. As depicted, the online learning system continuously assigns e-Learning exams from the current LOs that have to be learned before the student proceeds to the next learning level. The results from the e-Learning are then stored into the online learning system. Periodically, on teacher's request, the system assigns e-Assessment to the students in order to test their knowledge. All the results are again saved by the online learning system. As soon as the course is finished, the IGA collects all students e-Learning data which then it uses to create individual students' Profiles. The Profiles refer to the students' manner of passing the learning objectives of the particular lecture. On the other hand, the e-Assessment results are used to update those Profiles. After that, the Profiles are used in the ML process, i.e. the SVMs to produce a classifier able to determine the students' final grades. Since the teacher is allowed to monitor the results and each student's knowledge background, the IGA's decision is verified by a human factor. Hereupon, we propose a system that simulates a real teacher and is able to estimate the real achievements of its students.

The rest of the paper is organized as follows. In Section II we present the use of machine learning (ML) methods in education. Our methodology for multi-class classification of the students' knowledge is presented in Section III. In Section IV we present the experiments and the results from the presented approach and we derive a conclusion and present the plans for future work in Section V.

II. Related Work

In this section we give a review of the latest ML methods used for educational problems.

Castro et al. [11] aim to provide an up-to-date snapshot of the current state of research and applications of Data Mining methods in e-learning. They provide a taxonomy of e-learning problems to which Data Mining techniques have been applied, including: students' classification based on their learning performance, detection of irregular learning behaviours, e-learning system navigation and interaction optimization, clustering according to similar e-learning system usage and systems' adaptability to students' requirements and capacities.

Zaiane et al. [12] in their paper suggests the use of web mining techniques to build an agent that could recommend on-line learning activities or shortcuts in a course web site based on learners' access history to improve course material navigation as well as assist the online learning process. These techniques are considered integrated web mining as opposed to off-line web mining used by expert users to discover on-line access patterns.

The paper by Fei et al. [13] describes their work in exploring automatic question classification tests which can be used in e-learning system. Such tests can take the form of multiple-choice tests, as well as fill-in-the-blank and short-answer tests. They proposed a text categorization model using an artificial neural network trained by the backpropagation learning algorithm as the text classifier. Their test results showed that the system achieved the performance of nearly 78%.

Chang et al. [14] state that with the growing demand in e-learning, numerous research indicated that adaptive learning is a critical requirement for promoting the learning performance of students. The first step for achieving adaptive learning environments is to identify students' learning styles. Therefore, in their paper they propose a learning style classification mechanism to classify and then identify students' learning styles. The proposed mechanism improves k-nearest neighbour classification and combines it with genetic algorithms. To demonstrate the viability of the proposed mechanism, the

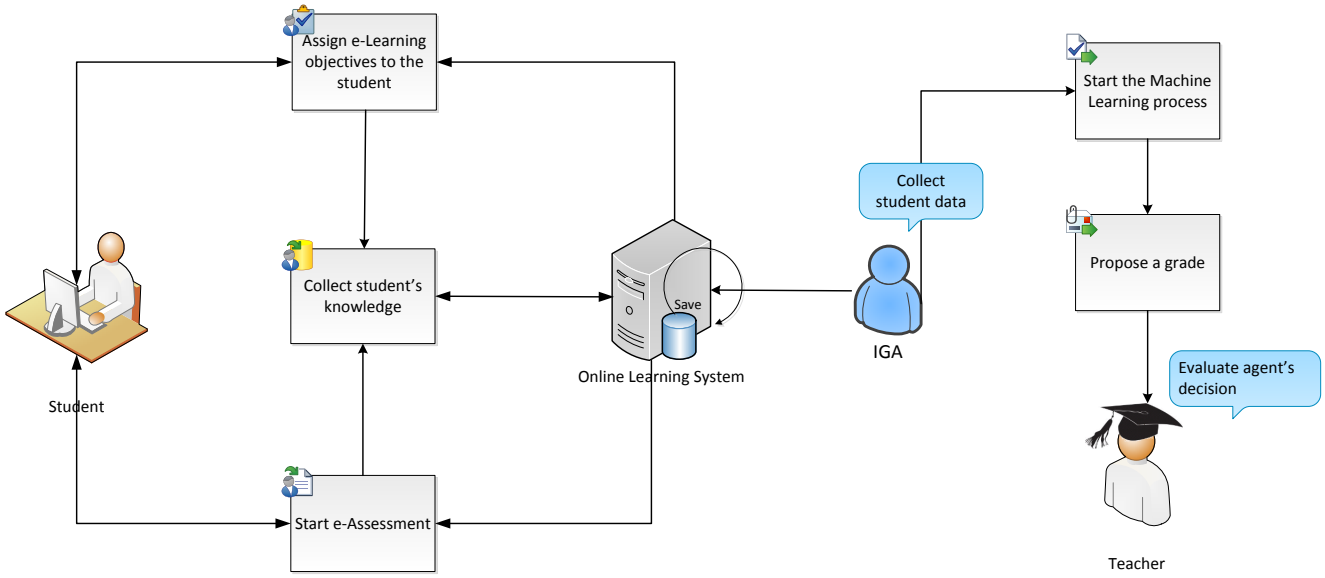


Fig. 1. Interaction between different roles in the machine learning process

proposed mechanism is implemented on an open-learning management system. The experimental results indicate that the proposed classification mechanism can effectively classify and identify students' learning styles.

Applying intelligence in e-Learning services has been the focus of several recent research projects and papers, and with the explosive growth of information sources available on the World Wide Web, the existing e-Learning services requires pedagogical overhauling for keeping pace with the trends in e-Learning. One of several possibilities can be its application to Online Assessment. In this Paper, we present the system architecture for applying intelligent methodologies to online assessment that adapts to the examinee's ability level and the other side we describe that web mining can be applied to Learning Content Management Systems to build knowledge about e-Learning and has potential to help and improve learner's performance.

The design of online assessment tools has been in the focus recently [15], [16], as well as applying intelligence in e-Learning services. For example, Uday et al. [17] present a system architecture for applying intelligent methodologies to online assessment that adapts to the examinee's ability level and the other side we describe that web mining can be applied to Learning Content Management Systems to build knowledge about e-Learning and has potential to help and improve learner's performance.

III. Methods and Methodology

In this section we present the methods and the methodology that we developed for building the students' individual Profiles as well as the classification process itself.

A. Online Learning System's Structure

Before we describe the students' Profiling, we must discuss the structure of the online learning system. The knowledge database for each course can be described in a tree-like manner. If we assume that one course represents the root of the tree, than the course lectures are its branches, and the learning objectives are branches' branches. Each learning objective consists of few questions for which there is one, or, several correct answers. In order to make the 3R strategy applicable for any learning objective, there are at least five questions provided.

B. Student's Profiling

In order to show how the results from the long term e-Assessment of a particular student can affect the previous e-Learning impression, we must provide each student an individual Profile, which can be described as follows.

The first step is extracting their knowledge from the online learning process in a n-dimensional vector, where n is the number of LOs. The students knowledge of a particular LO is represented as a number of trials to pass the given LO and the number of questions answered before the 3R pattern occurred in the passing trial. In order to express the student's efficiency for each LO that has been assigned, we proposed new metric in (1) to measure the student's *Initial Success*, IS. In the beginning, the IS for every student enrolled in the course is a null vector. The dimensionality of the IS vector is the same as the number of all LOs in the knowledge database. Therefore, $IS_i = (LO_1, LO_2, ..., LOj)$, where the highest value of i is the total number of students and j is the total number of LOs in the database. Considering the fact that the students are assigned each week with new questions from the LOs, the IS is calculated as sum of the number of trials (NT) necessary for

the student to pass the given LO, and the product of the scalar 0.01 and the number of questions (NQ) in the last trial. We use the value 0.01 in order to scale the number of questions in range $0.03 \leq NQ < 1$, since the maximum number of answered questions per learning objective is a double digit number. Therefore, the lower the IS score is, the more the student has learned the taught lectures.

$$IS_{ij} = NT_{ij} + (0.01 * NQ_{ij}) \quad (1)$$

Once we built the IS vector, our IGA uses them to conceive preliminary impression on the students' devotion to the course. However, taking into account only the ISs is not enough to derive conclusion of the students' most probable final grade. Therefore, we propose another approach which takes into consideration the students' e-Assessment results before concluding the grade. Thus, the students' success is revised every time the student has taken an e-Assessment. By introducing this approach, we simulate a real surroundings where the teacher evaluates the students' activities during the whole course to determine their grades. Hereupon, we calculate the *Revised Success*, RS, as:

1) *Punishment:*

$$RS_{ij} = IS_{ij} + 1 \quad (2)$$

2) *Reward:*

$$RS_{ij} = IS_{ij} - 1 \quad (3)$$

where i denotes the currently analyzed student and j denotes the particular LO.

The RS fully depends on the student's result from answering the questions from the e-Assessment. If the student's answer on a question from a particular LO is $false$, then the student is punished with adding one more trial on his IS of the particular LO from the e-Learning knowledge background. On the opposite, when the student's answer on a question from a particular LO is $true$, the student is rewarded with reducing the number of trials by 1 in the e-Learning history. That is how we use the e-Assessment results to affect the student's impression during the semester.

C. Machine Learning Analysis

After creating the students' individual Profiles, in this section we show how to use them to train a SVM classifier to perform multi-grade classification. We consider SVM to be an appropriate classification technique since we know the desired outputs during the training process. In order to obtain the best training, we compared the performance of three types of kernels: linear kernel (LK), polynomial kernel (PK) and radial basis function (RBF). In order to avoid overfitting in the process of training the classifier, we used hold-out cross-validation technique which does not permit overlap between training data and test data, yielding a more accurate estimate for the generalization performance of the algorithm [18].

The SVM technique is mainly used for binary classification problems. Since, we want to adapt it to classify the students into six classes, we need to organize the classification across a multiple-machines strategy. Before we discuss the training procedure, we must describe the classes we want our IGA to learn from:

- C_{fail} - contains all the students who have failed the course, i.e., the average score from their e-Assessments is $\leq$ 50%;
- C_6 - contains the students with on average 51 - 60% success;
- C_7 - contains the students with on average 61 - 70% success;
- C_8 - contains the students with on average 71 - 80% success;
- C_9 - contains the students with on average 81 - 90% success;
- C_{10} - contains the students with on average 91 - 100% success;

Considering the classes, we defined two test cases for the training process:

- *Training Case 1*: In the first training case we organize the students Profiles into 2 classes; the first contains the Profiles from C_{fail} and the second contains the Profiles in the remaining $C_{6:10}$ classes. Thus, the first SVM machine will assume that $C_1 = C_{fail}$ and $C_2 = C_{6:10} = C_{pass}$. Hereafter, we need to train the classifier to recognize each grade distinctively. Therefore, we defined five additional SVM classifying machines (CM). The training process is performed in 5 iterations, which means that in each iteration we set the class C_1 to be the class C_i for $i = 6 : 10$, and the class C_2 to be a set from the classes $C_{6:10} \setminus C_i$. The whole process is depicted in Figure 2.
- *Training Case 2*: In the second training case, instead of excluding the Profiles that have failed the course when determining the correct grade, we assume $C_{fail} = C_5$. Thus, the training process is developed as follows. We define six SVM machines and each of them sets the class $C_1 = C_i$ for $i = 5 : 10$, and the class $C_2 = C_{5:10} \setminus C_i$. Figure 3 visually explains the process.

In order to evaluate the classifier's performance we use $Precision$ (4) and $Recall$ (5) measures. The classifier's $Precision$ is a measure of true positives (TP) among all Profiles classified as positive (C_1), which include the true positives and the false positives (FP). The classifier's $Recall$ is a measure of its ability to select the Profiles that are positive from all the Profiles in the positive class.

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

$$Recall = \frac{TP}{TP + FN} \quad (5)$$

The experiments and the results from the analyses are presented in Section IV.

IV. Experiments and Results

In this section we present the results from the methodology described in Section III.

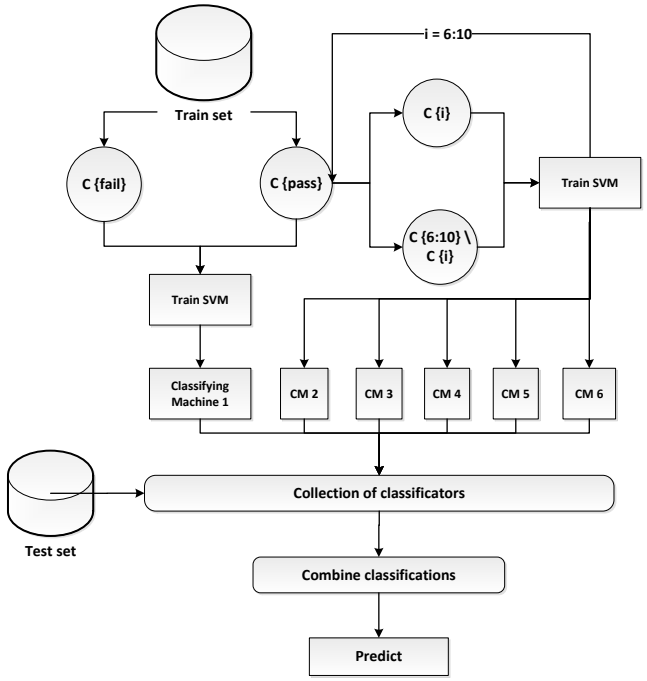


Fig. 2. Training Case 1

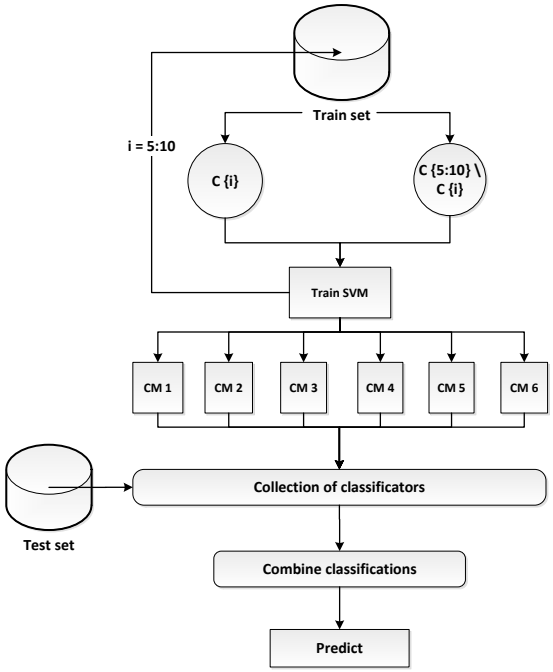


Fig. 3. Training Case 2

In order to obtain reliable results, we analysed the Profiles of 1273 randomly selected students that have attended the CAO course in the past years. The analyses were performed according to the training cases described in Section III-C, by using the *SVMlight* [19] tool. Since the results from the RBF were not promising, we present only the results from the classification with LK and PK.

Considering the fact that in both training cases, the training and the classification procedure is the same for the class of Profiles that have failed the course, $C_{fail} = C_5$, at first we describe the classification results from this analysis. Thus, in order to train the classifier to recognize the Profiles categorized as failed, we created two classes:

- C_5 which contains 629 Profiles that have failed
- $C_{6:10}$ which contains 644 Profiles that have passed

However, the initial results from the classification were not satisfying and we decided to use the bootstrapping method in order to obtain increased number of Profiles that have failed and passed the course. The method resulted in a 1000 Profiles in each class. The results from the binary classification are presented in Table I.

TABLE I
RESULTS FROM THE BINARY CLASSIFICATION

SVM Results	LK		PK	
Class	Precision	Recall	Precision	Recall
C_5 vs $C_{6:10}$	100%	100%	100%	100%

Since there are variable number of Profiles in each of the classes, we also used the bootstrapping method to enlarge the data in the first class so the number of Profiles will be the same as in the second class. Therefore, the number of Profiles before and after applying the bootstrapping method in each Training Case 1 and 2 are presented in Table II.

TABLE II
NUMBER OF PROFILES BEFORE AND AFTER THE BOOTSTRAPPING METHOD

Number of Profiles	Before	Training Case 1	Training Case 2
C_6	247	324	956
C_7	192	379	1011
C_8	101	470	1102
C_9	30	541	1173
C_{10}	1	570	1199

The preparation of the Training Case 1 and 2 according to the methodology defined in Section III-C produced the classification results in tables III and IV.

TABLE III
RESULTS FROM TRAINING CASE 1

SVM Results	LK		PK	
Class	Precision	Recall	Precision	Recall
C_6 vs $C_{6:10} \setminus C_{5,6}$	99.39%	100%	98.18%	100%
C_7 vs $C_{6:10} \setminus C_{5,7}$	99.47%	100%	98.95%	100%
C_8 vs $C_{6:10} \setminus C_{5,8}$	99.16%	100%	98.74%	100%
C_9 vs $C_{6:10} \setminus C_{5,9}$	100%	100%	100%	100%
C_{10} vs $C_{6:10} \setminus C_{5,10}$	100%	100%	98.96%	100%

The analyses from the both training cases showed that the Training Case 2 produced barely improved results, but both of

TABLE IV
RESULTS FROM TRAINING CASE 2

SVM Results	LK		PK	
Class	Precision	Recall	Precision	Recall
C_6 vs $C_{5:10} \setminus C_6$	99.58%	100%	98.96%	100%
C_7 vs $C_{5:10} \setminus C_7$	99.61%	100%	99.02%	100%
C_8 vs $C_{5:10} \setminus C_8$	100%	100%	99.46%	100%
C_9 vs $C_{5:10} \setminus C_9$	100%	100%	99.49%	100%
C_{10} vs $C_{5:10} \setminus C_{10}$	100%	100%	98.36%	100%

them resulted in very accurate and promising classifiers, which confirmed that the methodology we defined in Section III is reliable.

V. CONCLUSION AND FUTURE WORK

In this research we analyse e-Learning and e-Assessment data from the Computer Architecture and Organization course at our faculty. The data is collected from various years between 2002 (when the online learning system was implemented) and 2012.

The purpose of the analysis is creating an intelligent agent which based on the previous students knowledge will be able to determine the students' final grades. Therefore, we proposed a methodology that included modelling students' Profiles based on their activities in the e-Learning and e-Assessment processes during the course. The Profiles are then used in a two types of SVM training processes with multiple kernels, and each of them produced 6 classification machines.

The experiments and the results proved the reliability of the methodology we defined by showing that both training cases are capable of producing very accurate classifiers which is very important for our future research.

We believe that this model is essential and of great advantage for both the students, since they will stay highly motivated to successfully finish their exams until the end of the course, and for the teachers, since it can be a good indicator whether there is a need of additional testing of the student's knowledge in order to derive an overall conclusion on the most appropriate grade.

In our future work we will aim to develop different strategies for determining the final grades which will be based on the students' individual manners for passing the given learning objectives.

ACKNOWLEDGEMENT

This work was partially financed by the CodEval (Evaluating performances of program codes and educational processes using ML methods) project at Faculty of Computer Science and Engineering at the "Ss. Cyril and Methodius University", Skopje, Macedonia.

REFERENCES

[1] S. Vonderwell, X. Liang, and K. Alderman, "Asynchronous discussions and assessment in online learning," *Journal of Research on Technology in Education*, vol. 39, no. 3, p. 309, 2007.

[2] M. Gusev and G. Armenski, "E-Assessment Systems and Online Learning with Adaptive Testing," in *e-Learning Paradigms and Applications - Agent-based Approach*, M. Ivanovic and L. Jain, Eds. Springer Verlag, 2013, pp. 229–249.

[3] A. P. Rovai, "Online and traditional assessments: what is the difference?" *The Internet and Higher Education*, vol. 3, no. 3, pp. 141–151, 2000.

[4] M. Hannay and T. Newvine, "Perceptions of distance learning: A comparison of online and traditional learning," *Journal of Online Learning and Teaching*, vol. 2, no. 1, pp. 1–11, 2006.

[5] P.-C. Sun, R. J. Tsai, G. Finger, Y.-Y. Chen, and D. Yeh, "What drives a successful e-learning? an empirical investigation of the critical factors influencing learner satisfaction," *Computers & Education*, vol. 50, no. 4, pp. 1183–1202, 2008.

[6] A. C. Caminero, A. Robles-Gómez, S. Ros, R. Hernández, R. Pastor, N. Oliva, and M. Castro, "Harnessing clouds for e-learning: New directions followed by uned," in *Global Engineering Education Conference (EDUCON), IEEE*. IEEE, 2011, pp. 412–416.

[7] M. Mircea and A. I. Andreescu, "Using cloud computing in higher education: A strategy to improve agility in the current financial crisis," *Communications of the IBIMA*, vol. 2011, pp. 1–15, 2011.

[8] D. G. Chandra and M. D. Borah, "Cost benefit analysis of cloud computing in education," in *Computing, Communication and Applications (ICCCA), 2012 International Conference on*. IEEE, 2012, pp. 1–6.

[9] S. Ristov, M. Gusev, G. Armenski, K. Bozinoski, and G. Velkoski, "Architecture and organization of e-assessment cloud solution," *Global Engineering Education Conference (EDUCON), IEEE*, pp. 736–743, 2013.

[10] M. Simjanoska, M. Gusev, S. Ristov, and A. M. Bogdanova, "Intelligent student profiling for predicting e-assessment outcomes," in *Global Engineering Education Conference (EDUCON), IEEE*, 2014.

[11] F. Castro, A. Vellido, À. Nebot, and F. Mugica, "Applying data mining techniques to e-learning problems," in *Evolution of teaching and learning paradigms in intelligent environment*. Springer, 2007, pp. 183–221.

[12] O. R. Zaíane, "Building a recommender agent for e-learning systems," in *Computers in Education, 2002. Proceedings. International Conference on*. IEEE, 2002, pp. 55–59.

[13] T. Fei, W. J. Heng, K. C. Toh, and T. Qi, "Question classification for e-learning by artificial neural network," in *Information, Communications and Signal Processing, 2003 and Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint Conference of the Fourth International Conference on*, vol. 3. IEEE, 2003, pp. 1757–1761.

[14] Y.-C. Chang, W.-Y. Kao, C.-P. Chu, and C.-H. Chiu, "A learning style classification mechanism for e-learning," *Computers & Education*, vol. 53, no. 2, pp. 273–285, 2009.

[15] A. Rashad, A. A. Youssif, R. Abdel-Ghafar, and A. E. Labib, "E-assessment tool: A course assessment tool integrated into knowledge assessment," in *Innovative Techniques in Instruction Technology, E-learning, E-assessment, and Education*. Springer, 2008, pp. 7–12.

[16] Z. Harley and E. Harley, "E-learning and e-assessment for a computer programming course," *EDULEARN11 Proceedings*, pp. 2074–2080, 2011.

[17] M. Uday Kumar, J. Mamatha, S. Jain, and D. Jain, "Intelligent online assessment methodology," in *Next Generation Web Services Practices (NWeSP), 2011 7th International Conference on*. IEEE, 2011, pp. 215–220.

[18] P. Refaeilzadeh, L. Tang, and H. Liu, "Cross-validation," in *Encyclopedia of Database Systems*. Springer, 2009, pp. 532–538.

[19] T. Joachims, "Making large scale svm learning practical," Universität Dortmund, Tech. Rep. LS-8 Report 24, 1999.

Determining the impact of demographic features in predicting student success in Croatia

Bruno Trstenjak*, Dženana Đonko**
*Medimurje University of Applied Sciences in Cakovec/Dept. of Computer Engineering, Cakovec, Croatia
**Faculty of Electrical Engineering/Dept. of Computer Science,Sarajevo, Bosnia and Herzegovina
btrstenjak@mev.hr; ddonko@etf.unsa.ba

Predicting the success of students is a topic which has been studied for a long time in different scientific fields. Evaluation of importance of the features used in the prediction and their subsequent selection is an immensely important step in the process of classification and data mining. This paper presents a study on the importance of student demographic features in the process of predicting. The study and performed analyses used the demographic data collected from the Information System for Higher Education (ISVU). For determining the importance of demographic features in the study the following methods have been used: Information Gain (IG), Gain Ratio (GR), Sequential Backward Selection (SBS), Sequential Forward Selection (SFS). The results show the features rank, their importance weight in the prediction and comparison of the results and the use of different methods. Two classification algorithms for evaluating the impact of ranking features to the quality of prediction are used: Naive Bayes i Support Vector Machine (SVM). Final results provide guidelines for the development of a new prediction model.

I. INTRODUCTION

Predicting the success of students has long been a topic which is being studied in various scientific communities. So far they have conducted various studies and analyses related to success in learning, finding the answer to the question how to achieve better results, and the relation regarding student properties and their impact on studying success. Predicting the success of students is closely associated with the concept of data mining and machine learning. In the research various techniques of learning and classification algorithms were used [1],[2]. Demographic data are an integral part of each classification that includes information about people, users, students, etc. For the purposes of the research in this article, we used Information System of Higher Education Institutions (ISVU), the data from Medimurje University of Applied Sciences in Cakovec, Croatia.

Data mining and machine learning are the scientific fields in which the concept of prediction has been explored, various algorithms have been developed in order to achieve better prediction accuracy [3],[4]. When mentioning prediction, it primarily refers to classification of data to determine their membership depending on their features. If the prediction is determined on the basis of input data and achieved results, then we are taking about Supervised Learning [5].

The amount of data, the number of features, data dimensions, vary depending on the prediction purpose. Multi-dimensionality of data and large number of features increase the complexity of algorithms which use in prediction and has an influence on the extension of time for prediction and consumption computer resources. Due to the large number of features that are included in the prediction model and to accelerate the prediction process, some models use algorithms to determine the importance of individual features for the final score prediction. In order to achieve better performance of prediction, it is necessary to pre-process the data. Pre-processing is the process of preparing the data before it is sent to the model of prediction, transforming data suitable for the prediction model. Data pre-processing methods are divided into four activities: data cleaning, data integration, data transformation and data reduction. Data reduction refers to reducing the number of dimensions and features, determining the subset of features [6],[7]. Feature subset selection is the process of identifying and removing irrelevant and redundant information.

Generally, selection algorithms are distinguished between filter algorithms and wrapper algorithms [8],[9]. In evaluation filter algorithm only considers data reduction but does not take into account feature activities. Filters ranked features based on certain statistical criteria. Features with highest ranking values are selected on a subset. In the wrapper approach, feature selection is "wrapped" in a learning algorithm. Wrappers use the learning algorithm itself to evaluate the usefulness of the features. The filter algorithms that are widely used are: Information Gain (IG), Document Frequency (DF), χ^2statistic (CHI), Expected cross entropy (ECE), Weight of evidence for text (WET), Odds ratio (ODD). Selection of algorithm largely depends on the property of multidimensional data and the selected machine learning technique [10]. Wrapper algorithms like Sequential backward selection (SBS), Sequential forward selection (SFS), Plus-L minus-R selection (LRS), Genetic algorithms (GA)[11,12] can be grouped into two categories: greedy algorithms and random / stochastic algorithms [13]. These algorithms typically give better results than filter algorithms, but they are much slower and need to be repeated every time when any changes in the data structure happen, as well as the changes at machine learning algorithm [14].

This article presents the evaluation of selection algorithms on the base of the same data. During the study a prediction over a different set of input features was conducted. Analysis and comparison of results show the achieved accuracy.

II. OBJECTIVES OF RESEARCH

The main objective of this study was to determine the importance and influence of certain demographic features of students in the process of prediction, to rank the features in order of importance and determine the rank threshold where the feature has an impact on the final score of prediction and compare several algorithms selection of features, evaluate the results and determine algorithm that provide the best results.

To predict the student success two techniques of machine learning were used: Naive Bayes and Support Vector Machine (SVM). The prediction was achieved with binary classes: successful and unsuccessful student. The results of prediction were determined by two classes: successful and unsuccessful student.

III. DATA SOURCE

The data used in this study are part of a national database of Information System of Higher Education Institutions (ISVU) [15].The database contains a large number of different data about students organised into several groups: personal demographic information, demographic information about the parents and the type of residence, information on passed exams, data on high school achievement, information about the current student status, social status of student, student grants for food. For a description of one student the system uses more than 60 features. Prediction of student success is focused on determining the success in first year of the study. In addition to demographic data, which are the basis of this research, student high school grades were added to the input data, which is one of the categories in the application to the study.

For training and test data, the data of the students enrolled in the program in the last three academic years 2010- 2012 have been used. As the target data ECTS (European Credit Transfer and Accumulation System) credits which the student achieved at the end of the academic year have been used. For success in the first year of study a student must achieve at least 42 ECTS credits. These students were classified in the "successful students" class, other students in the "unsuccessful students" class.

Cleaning and adjustment of the data was performed before the feature selection and prediction. By removing irrelevant and redundant data (ID number, address, phone number, name of high school, job title ...) a collection of 15 demographic features remained. Each student is described with the following demographic data, shown in Table 1:

TABLE I.
LIST OF STUDENT DEMOGRAPHIC DATA

No	Feature description
1.	Sex - male / female
2.	Place of birth - city post code or the birth place of student
3.	Nationality (Croatian) – Yes / No
4.	Completed secondary school - the name of the secondary school that the student has completed
5.	Vocation - vocation which the student completed at the secondary school
6.	Duration of education - The number of years of education in secondary school
7.	Programme (curriculum) - designation of the curriculum in secondary school
8.	Profession - finished profession in secondary school
9.	Student rights - designation of student rights and subventions
10.	Marital status - married / unmarried
11.	Status of residence – subtenant, lives with his/her parents, his/her own flat/house
12.	Education level of mother - the system defined six levels of education
13.	Mother's profession - the current workplace
14.	Education level of father - the system defined six levels of education
15.	Father's profession - the current workplace

IV. CLASSIFICATION ALGORITHMS

In evaluation of quality of prediction two learning techniques were used: Naive Bayes and Support Vector Machine (SVM).For both of the two learning techniques during forming the model a k-fold Cross-validation method was used, by which the input data are divided into k equal subsets. In k-fold Cross-validation, the training sample is partitioned into k-folds as equally as possible and each fold is left out of the learning process and used as a testing set, k-1 folds used as a training set [16],[17].

A. Support Vector Machine (SVM)

SVM is a machine learning techniques used in the field of data mining. SVM algorithm is based on statistical learning theory developed by Vapnik (1995). The algorithm is defined as a supervised learning algorithm which solves linear and nonlinear binary classification problem. The SVM training algorithm aims to find a hyperplane that separates the multidimensional training data into a discrete predefined number of classes, usual binary class. The operation of the SVM algorithm is based on finding the hyperplane that gives the largest minimum distance to the training examples. Determining the direction and position of the hyperplane where the distance between the training data, both two-class, and hyperplane will be maximised. The hypeplane divides multidimensional training data into two sets: the positive and negative class. Lines that pass the data vector that is closest to the hypeplane are called margins. The training data that lie on the margins, called support vectors, are the only ones that define the hyperplane of maximum margin [18],[19]. SVM ensures maximum distance between the margins throughout the length of the hyperplane.

B. Naive Bayes (NB)

Naive Bayes is a linear classifier based on Bayesian theorem [20]. The algorithm analyses the relations

between features and classes. The algorithm calculates the conditional probability for relations between the values of the features and classes. Features $\{x_1, ...,x_k\}$ represent a vector x, vector x belongs to class c. During training, the algorithm determines the probability of each class. Probability is computed by counting how many times it occurs in the training dataset. This is called the prior probability P(C=c). The algorithm also calculates the probability that the instance x belongs to class c, shown in equation (1)[17]

$$P(C = c \mid X = x) = \frac{P(C = c)\prod_i P(X_i \mid C = c)}{P(X = x)}. \quad (1)$$

In the denominator there is a scalar coefficient, which is calculated as the sum of all the associated probability, shown in equation (2)

$$P(X = x) = \sum_j P(C_j)P(X = x \mid C_j). \quad (2)$$

This algorithm carried out a quality prediction over a small amount of training data, enables the use of multiple output class, but is sensitive to pre-processing of data and prediction over nominal data [21].

V. FEATURE SELECTION

The features selection is performed in order to optimise the operation of the classifier. Prediction is often performed over multi-dimensional data with a large number of features. In order to achieve a simpler and faster classification, some systems reduce the number of input features and determine the features which have a significant role in classification. Features get ranked and determine a features subset which is sent to the prediction model. Selection methods which have been used for the purposes of this article:

A. Info Gain (IG)

Info Gain method calculates the value of the features information taking into account the class to which they belong, thereby determining the entropy. IG uses the following expression for the calculation:

Info Gain (Klasa, Atribut)=H (Klasa) – H(Klasa|Atribut),

where H is the entropy, which is defined by using the following equation (3)

$$Entropy(S) = -\sum_{j=1}^{m} p_j \log_2 p_j, \quad (3)$$

where p is the probability, for which a particular value occurs in the sample space S.

Entropy value ranges from 0 to 1. Value 0 means that all instances of the variables have the same value, value 1 equals the number of instances of each value. Entropy shows how the attribute values are distributed and indicates "purity" of attributes. High entropy indicates the uniform distribution of attributes. Opposite to that, low entropy indicates distribution which is concentrated around the point [22].

B. Gain Ratio (GR)

The Gain Ratio method is a normalized version of Information Gain. Normalization is accomplished by dividing the information gain with the entropy of the attribute with respect to the class. The result of this approach is reduced bias from the information obtained by the algorithm. The formula used by the GR is represented by Eq. (4)[22]

$$GR(Klasa, Atribut) = \frac{(H(Klasa) - H(Klasa \mid Atribut))}{H(Atribut)}. \quad (4)$$

C. Sequential forward selection (SFS)

Sequential forward selection is a simple greedy search algorithm which is trying to determine the optimal set of features. The algorithm uses a special list in which the optimal features are recorded. At the beginning the list is empty. The algorithm adds in the list feature x^+ whose objective function J ($Y_k + x^+$) gives the best result when combined with the feature Y_k which has been selected and is on the list.

Algorithm pseudocode [23]:

1. Start with the empty set, initialise the feature set Y0={Ø}; k=0
2. Select the next best feature x^+=argmax[J(Y_k+x)]; x¢Y_k
3. Update the feature set $Y_{k+1}=Y_k+ x^+$;
4. While k < d
   k=k+1
   Go to Step 2

D. Sequential backward selection (SBS)

Sequential backward selection is an algorithm that is similar to the SFS algorithm, but it works in the opposite direction. Initially, the algorithm in its optimal list has all the features of X from set D. Algorithm removes a single feature that improves (or minimally worsens) an objective function to obtain the best subset of features. The algorithm uses the objective function J (Yx^-) and determines for which attribute function the minimum value is returned. This attribute is permanently removed from the list and increases the value of the target function. J ($Y_k- x^-$) > J (Y_k).

VI. EXPERIMENTAL RESULTS

A. Features ranking

For each student in the database approximately 62 demographic features are located. The original data contained a large number of missing, redundant and irrelevant data. Before the pre-processing of input data and reducing the number of features, feature selection and calculation of the prediction of student success was performed over the "raw" data, all in order to enable later results evaluation, to enable defending individual impact of features in achievement of prediction accuracy.

For the classification of raw data and features ranking we used the Information Gain method and SVM classifier. The study achieved "interesting" results of ranging

features and very poor prediction accuracy. Table 2 shows the results of ranking top 10 features.

TABLE II.
RATING LIST OF FEATURE – RAW DATA

Ranked attributes	
0.94142	1 Student surname
0.86595	2 Date of birth
0.86595	3 Citizenship
0.77155	4 Address
0.71501	5 Average grade
0.71001	6 Mother's name
0.70076	7 Student name
0.41789	9 City
0.41789	8 Place of residence
0.40975	10 Father's name
0.30528	11 Student rights
. . .	

In the classification, input data were divided into 10-fold Cross-validation. The following results were obtained:

Correctly Classified Instances	***56.6038 %***
Incorrectly Classified Instances	***43.3962 %***
Kappa statistic	*0.0225*
Mean absolute error	*0.434*
Root mean squared error	*0.6588*
Relative absolute error	*93.8619 %*
Root relative squared error	*137.1311 %*

Results show a low level of prediction accuracy, with only 56.6% correctly classified instances. The analysis of input data showed a significant number of incomplete data which certainly had an impact on prediction accuracy.

After classification, original data were pre-processed, irrelevant data and redundant features removed. The results of the analysis separated 15 demographic features, features that were used in further process of prediction determination. Ranking using all four methods was conducted on the selected demographic features. Table 3 shows the results of the ranking depending on the chosen selection method.

The results are somewhat different than we expected at the beginning of the studies. All used ranking methods have highlighted "Average grade" feature as the most important, more than any demographic features. This is somehow logical because various demographic data such as situation in the family, type of curriculum, etc. have the impact on the final secondary school grades. Secondary school results are a strong indicator of success / failure of the newly enrolled students.

From the ranked demographic attributes, the analysis highlighted the curriculum that the student completed in secondary school as one of the more important for prediction. According to the rank results, the curriculum that the student learned in high school has significant influences on the success during the study. The second demographic feature that has an impact on the prediction is "Student rights". The feature belongs to a group of motivating features in order to keep the acquired rights that successful students have(cheaper food, transport, students' service, etc.).

TABLE III.
RATING LIST OF SELECTED FEATURE

Rating	
Info Gain (IG)	Ranked feature: 0.74982 1 Average grade 0.29000 2 Programme (curriculum) 0.15111 3 Completed secondary school 0.14609 4 Secondary vocation 0.1079 5 Student rights 0.08909 6 Education level of the father 0.08285 7 Place of birth 0.08262 8 Education level of the mother 0.05238 9 Duration of education 0.04683 10 Mother's profession 0.03954 11 Father's profession 0.03700 12 Status of residence 0.02160 13 Marital status 0.01381 14 Nationality 0.00497 15 Sex
Gain Ratio (GR)	Ranked feature: 0.16713 13 Marital status 0.11926 1 Average grade 0.08177 12 Status of residence 0.07612 2 Programme (curriculum) 0.06896 5 Student rights 0.05782 6 Education level of the father 0.05516 3 Completed secondary school 0.04679 8 Education level of the mother 0.04674 4 Profession 0.04071 9 Duration of education 0.03897 7 Place of birth 0.01885 14 Nationality 0.0139 10 Mother's profession 0.01272 11 Father's profession 0.00561 15 Sex
SFS/ SBS	Ranked feature: 0.206 1 Average grade 0.179 5 Student rights 0.177 2 Programme (curriculum) 0.176 12 Status of residence 0.172 13 Marital status 0.168 6 Education level of the father 0.166 3 Completed secondary school 0.163 9 Duration of education 0.158 8 Education level of the mother 0.154 7 Place of birth 0.150 4 Profession 0.145 14 Nationality 0.138 10 Mother's profession 0.132 11 Father's profession 0.127 15 Sex

Selective methods have also underlined the parents' education as an important element in predicting the student success. According to the character of data, father's education has a significant role compared with mother's education. All methods have determined that the gender of students is ranked the lowest. This information

leads to the conclusion that this attribute should have the least impact on the data classification. This could be due to relatively small number of female population that is enrolled in the study of computing engineering.

B. Features impact

With specifying the individual impact of each attribute, a features selection process can be improved and a more accurate classification provided. After implementing the feature selection, classifications were performed with both techniques, NB and SVM. Table 4 shows the results of classification.

TABLE IV.
CLASSIFICATION RESULTS

Classification summary	**Naive Bayes**	**SVM**
Correctly Classified	61.607%	70.53 %
Incorrectly Classified	38.39 %	29.46%
Kappa statistic	0.2287	0.4027
Mean absolute error	0.4128	0.2946
Root mean squared error	0.5268	0.5428

The results show that the SVM classifier achieved significantly better prediction than NB classifier. SVM also achieved better results with all demographic attributes involved. After measuring, the classification additional analysis of the feature impact on classification accuracywas performed. In testing the prediction, some target features were blocked and after that the accuracy using SVM technique was measured. The obtained results are shown in Table 5.

TABLE V.
FEATURES IMPACTED TO THE RESULT OF CLASSIFICATION

Feature	**Accuracy**	**Impact**
Average grade	66.96	-5%
Programme (curriculum)	64.28	-9%
Completed secondary school	62.50	-11%
Secondary vocation	69.64	-1%
Student rights	55.34	-22%
Education level of father	69.64	-1%
Place of birth	73.21	+4%
Education level of mother	73.21	+4%
Duration of education	65.17	-8%
Mother's profession	71.42	+1%
Father's profession	69.64	-1%
Status of residence	71.42	+1%
Marital status	70.73	0%
Nationality	69.64	-1%
Sex	68.75	-3%

In the table, the column "Impact", shows the impact of each feature on the prediction accuracy expressed as a percentage. A negative value indicates the percentage of accuracy reduction if a specific feature is not taken into consideration during classification. Positive value has the opposite effect on the prediction. In other words, the assumption is that the separation of particular feature from a set of input features would increase the accuracy of prediction.

Such a classification and analysis of the results showed the interesting results. The biggest impact of the features has been calculated for "Student rights" feature. As mentioned in the previous section, this feature indicates the students motive to keep student status and achieve some benefits. Very often some of the students try to maintain their status due to the difficult economic situation or heavy employment. The second ranked feature by impact on prediction is "Completed secondary school" feature.The years of my experience justify this result. This feature is the most significant at the study programmeswhich belong to the area of technical sciences. For the purpose of this research the students come from three types of secondary schools: technical school, high school and school of economics. By comparing the data about their ECTS, it is evident that significantly better results were achieved by the students from technical schools. In Table 5, four features that have shown positive Impact factorwere indicated, which means that they have no significant impact on the accuracy of prediction. To evaluate this assumption a classification was carried out in a way that the indicated features were removed from the input set. After extracting the attribute classification with both techniques better accuracy of prediction, for approximately 5%,was achieved. Table 6 shows the evaluation results:

TABLE VI.
THE CLASSIFICATION WITH REDUCED SET OF FEATURES

Classification technique
Naive Bayes Correctly Classified ***65.1786 %*** Incorrectly Classified 34.8214 % Kappa statistic 0.3004 Mean absolute error 0.3876 Root mean squared error 0.5013 **Classification Improvement ca. 6%**
Support Vector Machine Correctly Classified ***73.2143 %*** Incorrectly Classified 26,7857 % Kappa statistic 0.4582 Mean absolute error 0.2679 Root mean squared error 0.5175 **Classification Improvement ca. 4%**

VII. CONCLUSION

Evaluation of the results definitely has proved that quality pre-processing is a requirement for obtaining good results of classification, in this case the prediction of student success. In the study of selection and ranking features four relatively commonly used methods are used: Information Gain (IG), Gain Ratio (GR), Sequential

Backward Selection (SBS), Sequential Forward Selection (SFS). None of the methods showed identical results of ranking features.It can be concluded that in the selection of attributes and determination of their importance in predicting it is necessary to include additional verification and empirical methods.

Evaluation results of ranking have shown three features that have a significant impact on the success of the study: Average grade, Completed secondary school and Student rights. The analysis of results offers a conclusion that the completed secondary school and the curriculum are very important for success in the first year of the study. Students with poor professional prior knowledge adapt harder to the requirements of the technical study, which in some way justifies the faculties when emphasising that future students should pass some specific courses. In addition to the list of courses, faculties also give great importance to final secondary schoolgrades.

Two other demographic features have a different meaning. Students who come from technical schools, in this case, they can easily adapt on study demands because they have necessary background knowledge which is required for technical studies. In this way some students have reduced the impact of lower grades from the high school. Other demographic feature acts as a motivational factor in education, as previously explained. We conducted a ranking of features with four different methods. Each method has estimated slightly different order of features importance. In estimating the impact of demographic features on the prediction accuracy, analysis was pointed out that feature "Student rights" had the greatest impact. The results of analysis have shown that “Place of birth”, “Education level of mother”, “Mother profession” and “Status of residence” are four features who do not have a significant impact on the classification. After determining the impact of features, four features with the least impact were removed from the input set. After reducing the number features, with new classification was achieved about 5% better results. The obtained results indicate the need for involving additional control in the features selection process. One of possible solutions is to introduce a hybrid model of prediction. Model that with additional algorithms would increase the accuracy of prediction. This assumption gives a scope for further research in the field of classification and prediction of students success.

REFERENCES

[1] C. Romero, S. Ventura, P. G. Espejo, C. Hervás,“ Data mining algorithms to classify students“, in 1st Int. Conf. on Educational Data Mining, 2008, pp. 187-191.

[2] R.R.Kabra, R.S.Bichkar, “Performance prediction of engineering etudents using Decision trees“, in International Journal of Computer Applications, vol. 36, No.11, 2011.

[3] L.David, O.D.Delen, "Advanced data mining techniques", Springer, 2008.

[4] S.Pandya, Dr.Paresh V. Virparia, “Comparing the applications of various algorithms of classification technique of data mining in an Indian“, in IJARCSSE, vol. 3, 2013.

[5] R.Dash, R.L.Paramguru, R.Dash, “Comparative analysis of Supervised and Unsupervised discretization techniques“, in International Journal of Advances in Science and Technology, vol. 2, No. 3, 2011.

[6] T.Karunaratne, H.Boström, U.Norinder, “Pre-Processing structured data for standard machine learning algorithms by supervised graph propositionalization - a Case study with medicinal chemistry datasets“, in Machine Learning and Applications (ICMLA), 2010.

[7] W.Jianping, “Research on data preprocessing in supermarket customers data mining“, Information Engineering and Computer Science, 2010, pp.1-4.

[8] Pablo Bermejo, Luis de la Ossa, José A. Gámez, José M. Puerta, "Fast wrapper feature subset selection in high-dimensional datasets by means of filter re-ranking", Knowledge-Based Systems, vol.25, 2012, pp. 35-44.

[9] S.Beniwal, J.Arora, “Classification and feature selection techniques in data mining“, in International Journal of Engineering Research & Technology, 2012.

[10] Yan Xu, “A comparative study on feature selection in Chinese Spam Filtering“, in Application of Information and Communication Technologies (AICT), 2012, pp. 1-6.

[11] S.F. Pratama, A.K. Muda, “Feature selection methods for writer identification: A Comparative Study“, in International Conference on Computer and Computational Intelligence, 2010.

[12] X. Liu, L.Shang, “A fast wrapper feature subset selection method based On Binary Particle Swarm Optimization“, in IEEE Congress on Evolutionary Computation, 2013, pp. 3347-3353.

[13] I.A.Gheyas, L.Smith, “Feature subset selection in large dimensionality domains“, in Pattern Recognition, 2009.

[14] Hui-Yan Wang, “A comparative study of the feature selection influence on diagnosis In traditional Chinese medicine“, in Proceedings of the Ninth International Conference on Machine Learning and Cybernetics, Qingdao, 2010.

[15] Informacijski sustav visokih učilišta (ISVU), Ministarstvo znanosti i tehnologije Republike Hrvatske, 2003.

[16] Ji-Hyun Kim, “Estimating classification error rate: Repeated Cross-validation, Repeated Hold-out and Bootstrap“, Department of Statistics and Actuarial Science, Soongsil University, 2009.

[17] N. Williams, S.Zander, G.Armitage, "Evaluating machine learning algorithms for automated network application identification", CAIA Technical Report, 2006.

[18] G. Mountrakis, J.Im, C.Ogole, "Support vector machines in remote sensing: A review", in Journal of Photogrammetry and Remote Sensing, 2010.

[19] Pei-Chann Chang, Chiung-Hua Huang, Chi-Yang Tsai, “A hybrid system by the integration of Case-Based-Reasoning with Support Vector Machine for prediction of financial crisis“, in ICIC International, vol. 9, No. 6, 2013.

[20] G. H. John, P. Langley, “Estimating continuous distributions in Bayesian classifiers”, 11th Conference on Uncertainty in Artificial Intelligence, pp. 338-345, Morgan Kaufman, San Mateo, 1995.

[21] S.B.Kotsiantis, "Supervised machine learning: A Review of classification techniques", Informatica 31, 2007, pp. 249-268.

[22] H.Dağ, K.E.Sayın, I.Yenidoğan, S.Albayrak, C.Acar, "Comparison of feature selection Algorithms for medical data", IEEE Innovations in Intelligent Systems and Applications, 2012.

[23] L.Ladha, T.Deepa, "Feature selection methods and algorithms“, in International Journal on Computer Science and Engineering, vol.3, No.5, 2011.

[24] Ricardo Gutierrez-Osuna, "Pattern analysis", CSCE

Intelligent Question – Answering Systems: Review of research

J. Tomljanović*, M. Pavlić ** i M. Ašenbrener Katić***
* Polytechnic of Rijeka
Vukovarska 58, Rijeka, Croatia
** *** Department of Informatics, University of Rijeka,
Radmile Matejčić 2, Rijeka, Croatia
* jasminka.tomljanovic@yahoo.com
** mile.pavlic@ris.hr
*** masenbrener@inf.uniri.hr

Abstract - With the development of ICT the need for automated question-answering systems is becoming increasingly important. Question-answering systems are still under development and experimentation. This paper is an overview of the research area that deals with question-answering systems; it explains the concept of question-answering systems and points out the problems that occur during their development. It also refers to a complex assessment techniques that are necessary when designing such systems. The system described is a real system and so are the test results.

Keywords: question-answering systems, artificial intelligence, technology assessment question-answering systems

I. INTRODUCTION

Along with the development of ICT grows a need for systems that allow the user to ask questions using everyday language and getting fast answers with the smallest quantity of context that is necessary for confirming the answer.

The solution to this problem are question - answering systems (QA systems) (Hirschman et al., 2001). QA systems are a method of retrieving information that answer questions asked in spoken language (Tomljanović et al., 2014).

The success in this field was achieved as part of the Text Retrieval Conference (TREC) (Ittycheriah et al., 2000) which is a series of workshops the mission of which is to retrieve as many information as possible from different fields of research and to backup the research in the field of QA systems.

To answer the question, the system has first to analyze the problem using the context. It finds one or more possible answers offered by the Internet source and offers an answer in an appropriate form. In most cases it is the supporting material that makes the answer understandable to the user or provides some further explanations regarding the answer.

In 1999 TREC started with the evaluation of QA systems that answer the factual questions by consulting the documents contained in TREC collections. A large number of systems successfully combine information retrieval and natural language processing techniques.

One of the groups of methods for knowledge representation is a graphical method for representing knowledge. These methods first transform knowledge into the model, then search knowledge by the model, and then are able to answer the questions. The NOK method (Jakupović et al., 2013), (Pavlić et al., 2013a), (Pavlić et al., 2013b), (Pavlić et al., 2013c) belongs to this group.

This paper will cover the methods which use the existing sources of knowledge registered in different texts.

II. COMPLEXITY ASSESSMENT TECHNIQUES QA SYSTEM

There are two critical research questions in development of QA systems: resources and evaluation. These two questions are closely related; developers need tools to build the system and they need assessment methods (exercises and tests) to evaluate the effectiveness of their system. The introduction of a general evaluation, such as TREC, has created associations interested in the area, and speeded up the research process as well. This is the evaluation with the special emphasis on methods for automatic evaluation.

A. Collection of questions and answers

In order to build a system that answers the questions, scientists need a stream of questions and answers. Ideally, such groups would be often occurring questions, and their answers would be contained in a large collection of documents. Kupiec uses a trivial search for questions as a source of question - answer pairs, as well as the Internet encyclopedia as a collection where to seek the answers (Kupiec, 1993).

A much larger collection is required for efficient use of machine learning and statistical techniques. For example, the syntax of questions and statements is different, and a large collection that explains or narrates usually contains very few questions, so without special settings, the rules based on the collection, will not work for the analysis of questions. The proper analysis of questions requires a large collection of questions and related short answers in order to develop premium features for speech tagging, parsing and questions input. For example, Mann used 2 collections of trivial questions with short answers (Mann, 2001). Auxiliary systems on the Internet include access to the most frequently asked questions as sources of question - answer sets, which are particularly useful in the development of QA systems specialized in certain areas. The number of QA websites is increasing. Good examples are sites that provide tests for foreign languages learning or contain quizzes based on actual events (Roever, 2001; Ushida, 2005).

Additional work in finding and collecting such collections on the Internet would accelerate the progress of the QA systems. The most common source of question - answer sets are multiple choice questions. Such questions are used in standardized tests. MCQs are not as natural as short-answer questions because they are designed to make the evaluation easier. Since the designing of tests is rather expensive, it is difficult to get a collection of materials for research or evaluation.

B. Expansion model

Any collection that contains question - answer sets may be useful. Some types of questions are much easier to be dealt with than the others. Previous researches have focused mainly on simpler questions. At the beginning of TREC assessment, questions were limited to simple factual questions that had the answer in the associated collection of documents. The result showed that the best assessment strategy was a ranked list of proposed correct answers.

The following assessment has increased the complexity of the questions in two dimensions: it allows questions that have no answers and questions with a list of answers (Ittycheriah et al., 2001). This further complexity requires changes in the assessment of the answer and in the construction of the system. To deal with questions that have no answers in the separate collection, systems themselves must measure their reliability regarding the answer.

Questions that require an answer sheet (e.g. a list of countries bordering Croatia) will need to choose an answer from more sentences comprised in one document, or even merge the answer from more documents. Both the former and the latter require the extension of a simple model to simple finding of sentences or parts of the document that give the best answer.

C. Evaluating response problems

The first problem that arises in the assessment is the choice of criteria to be used to assess the answer. *How to evaluate your question answering system every day . . . and still get real work done* (Breck, et al., 2000) calls for six following criteria:

- *relevance:* the answer should answer a given question
- *correctness:* the answer should be factually correct
- *conciseness:* the answer should not contain extraneous or irrelevant information
- *completeness:* the answer should be complete, i.e. partial answer should not get full credit
- *coherence:* the answer should be coherent, so that the questioner can read it easy
- *justification:* the answer should be supplied with sufficient context to allow a reader to determine why this was chosen as an answer to the question

Previous evaluations focused mainly on relevance, although today TREC - QA assessment requires that the answer must be justified by the context and that the size of the answer must be limited (in bytes). This is the first step towards conciseness. Optimization of a single criterion affects the quality of the other. For example, the justification of answers can reduce conciseness. This is exactly why the assessment criteria must depend on the purpose of use, the type of the user who searches for an answer, as well as on the interface.

After the criteria, the assessment procedures occupy the second important position. Experiments carried out during The Eighth TREC Conference introduced human evaluators to read and evaluate each answer. The results obtained showed that the human evaluator was consistent enough to preserve the relative ranking system, which implies that it is actually possible to have only one evaluator.

This has significantly reduced costs, but since there is a need for a human hand, this method does not support the systematic repetition of tests for machine learning. It is always useful to have a human subject dealing with evaluation, and therefore, tests that assess reading comprehension are ideal. In case of using multiple-choice tests, a human evaluator is not required.

D. Automated assessment techniques

Automated grading methods are under development and experimentation. For evaluation of short answers, it is possible to automate the comparison of answers, regardless of whether the answers are provided by the system or by a person, e.g. a student, with answers from the key provided by an expert. Such an assessment is described by Hirschman, Breck, Light,

Burger and Ferro in their book "Automated grading of short answer tests" (Hirschman et al., 2000). Such comparisons, though not as precise as those by human evaluators, nevertheless provide reasonably good results, 93-95 % according to human evaluators, which was good enough for machine learning.

Automated answers and evaluation are of great importance for testing the educational community. Despite the fact that the tests with short answers and essays are better, the multiple choice tests continue to be widely used. Tests where the person himself / herself writes the answers (either short answers or essays) are considered to be too subjective to be used for standardized testing. Recent research in this area (Kukich, 2000) have demonstrated the feasibility of automated methods of assessment, sometimes in combination with only one human evaluator. These results indicate that the automated grading systems are approaching human evaluators. We can already encounter automated assessments and e–learning systems throughout the Internet.

III. EXAMPLES QA SYSTEMS

1. *IBM's statistic system*

Operation of the system that answers the questions is explained using the example of IBM's statistic system that answers the questions (Ittycheriah et al., 2000). The architecture of the system is built around 4 major components, as shown in Figure 1:

- QA Type Classification
- Query expansion/Information Retrieval
- Named Entity Marking
- Answer Selection

The question is input that is classified, as asking for an answer belonging to one of the named entity classes. In addition, the question is presented to the IR engine for query expansion and document retrieval. This engine, given the query, looks at the data base of documents and outputs the best documents or passages annotated with the named entities. The final stage is to select the particular answer, given the answer class and the top scoring passages.

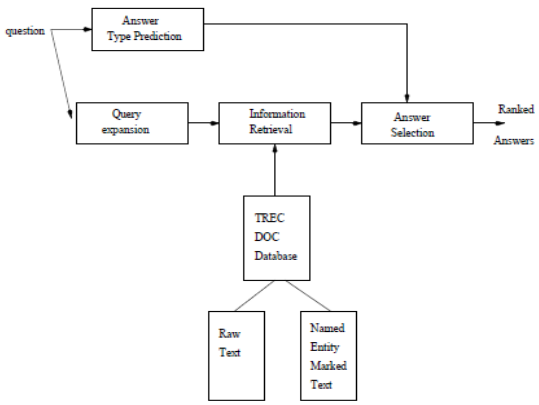


FIGURE 2. QUESTION ANSWERING ARCHITECTURE

A. *QA Type Classification*

The classification refers to the classification of questions and answers. In classifying the type of answer the problem is to label a question with the label of the named entity that the question seeks. These labels are the standard MUC (Chinchor, 1997) categories with the addition of PHRASE (P), which presents all the answers that cannot be put in standard categories. In standard categories we have REASON (R) category, which is tied to *why* questions. Processing of REASON and PHRASE is the same in this IBM system, interpreting it as desiring a clause which has a NOUN PHRASE (NP) imbedded in it.

Ittycheriah, Franz, Zhu, Ratnaparkhi, and Mammone created 1900 questions by presenting a human subject a randomly selected document and having read a portion of the document , a question was phrased, the answer and a document number noted in addition. They also used 1400 questions from a trivia database (Hallmarks, 1999), annotated in a similar way.

Each feature type expands to the property above it. The "Expanded Hierarchy" feature uses WordNet (Miller, 1990) in order to expand words from a question words up to those including noun predicate. The "Mark Question Word" feature identifies the question words and labels them as occurring in the beginning of a question, i.e. bqw, in the middle of a question, i.e. mdw or at end of a question, i.e. eqw.

N- gram language models are used as the method for language modeling. Regarding the length, N-grams can be unigrams (n =1), bigrams (n = 2), trigrams (n = 3)... (Gaspic, 2012).

Table 1 shows the classification of the QA type on the example question *"What year did World War II start?"* First, each word is classified into one of the word classes. WP is *wh* - pronoun such as *who, what, which, how, how many* ... NP refers to a noun phrase, VP verb phrase, NN singular noun, after English noun number, VBD verb in the past tense.

Table 1 shows that *what* is marked as an interrogative pronoun, while *year* is annotated as a noun. Expanding the features of the question pronouns up to those that contain the nominal predicate , i.e. we link *what* to *year*, we can conclude that the *wh* pronoun *what* and a noun predicate *year* require a certain time or a certain time period. Subsequently, we determine

TABLE I. FEATURES USED IN THE ANSWER CLASSIFICATION EXPERIMENTS

Unigrams	What year did World War II start?
Morphed,Part-Of-Speech	what{WP} year{NN} do{VBD} World{NP} War{NP} II{NP} start{NN}
Bigrams	what{wp} what{wp}_year{nn} what{wp}_do{vbd} what{wp}_world{np} ...
Expanded Hierarchy	what{WP} year time_period measure abstraction year{NN} do{VBD} ...
Mark Question Word	what_bqw year time_period measure abstraction year{NN} do{VBD} ...

TABLE II. RETRIEVAL RESULT

	MRR
pass1, TREC	0.4605
pass2, TREC	0.4824
pass2, encyclopedia	0.5031

the position of a question word within a question; in this case it is the beginning of the question, as shown in Table 1.

B. Information Retrieval

The purpose of the IR module is to search the database to select passages of the text, containing information relevant to the query. The IBM IR system uses a two-pass approach. In the first pass, the encyclopedia database is searched. The highest scoring passages are then used to create expanded queries, applied in the second pass. The scoring is based on unigram and bigram features extracted from the text data using tokenization, part-of-speech tagging and a morphological analyzer. (Merialdo, 1990).

In the first pass, Ittycheriah et al. modified the Okapi formula [1] (Robertson, Walker, Jones, Hancock - Beaulieu, Gatford, 1995) in order to score passages extracted from the encyclopedia documents, then converted all the encyclopedia articles into 82 277 overlapping passages, each containing about 100 non-stop words. Based on the first passage ranking, they constructed expanded queries. In the second passage they used extended queries and scored 2 632 807 passages based on the TREC-9 corpus. The passages contained about 200 non-stop words. Table 2 summarizes the Information Retrieval (IR) results on tests (Sparck Jones and Willett, 1997).

The performance is measured by the MRR[2] or Mean Reciprocal Rank. The first line of the Table 2 shows the result of the first pass scoring; the second line shows the result of the second pass scoring, and the third line shows the result corresponding to the system applied by the IBM as an example, with queries expanded using the encyclopedia database.

C. Named Entity Marking

Named entity annotation refers to a markup of the text with the class information. As mentioned above, classes correspond to the standard MUC classes.

Windows of + or - words, morphs, part-of-speech tags and flags raised by pattern grammars (DATE, MONEY, CARD, MEASURE, PERCENT, TIME, DURATION). The window for predicting the tag (0) is shown in Table 3.

Each stream has a fixed vocabulary. N-grams from this vocabulary form the features of the maximum entropy model.

[1] In IR is a ranking function used by search engines to rank matching documents according to their relevance to a given search query.

[2] MRR(Q) = $\frac{1}{|Q|} \cdot \sum_{i=1}^{|Q|} \frac{1}{r(d_{i,rel})}$

TABLE III. FEATURES USED IN THE NAMED ENTITY MODEL FOR PREDICTING TAG(0)

Words	w(-) w(-) | w(0) | w(+) w(+)
Morphs	m(-) m(-) | m(0) | m(+) m(+)
Part-of-Speech	p(-2) p(-1) | p(0) | p(1) p(2)
Grammar Flags	f(-2) f(-1) | f(0) | f(1) f(2)
Previous Tags	t(-2).t(-1) t(-1) |

The training data is arranged to indicate a special category for beginning of each named entity, for example BeginPERSON to find the boundaries of the named entity.

The system explores multiple NE hypotheses in parallel and keeps only those only those with high probability. It proceeds with the algorithm to find the most likely part for the whole sentence.

D. Answer Selection

In this phase we have a question, a class of the answer that the question seeks and a ranked set of passages annotated with the MUC classes. We are searching now the optimal sentence that will contain the answer. The TREC length constraints of 250 byte and 50 byte are then applied on the sentence. The algorithm used in this module is listed here:

1) Each passage is split into sentences.
2) A window is formed around each sentence.
3) The following features are computed:
   a) *Matching Words* (the sum of the number of words that matched identically in the morphed space (+)),
   b) *Aligning the Terms* (the sum of the number of words that are synonyms (+))
   c) *Mis-Match Words* (the sum of the number of question content words that did not match in this answer (-))
   d) *Dispersion* (the number of words in the candidate sentence that occurred between matching question words (-))
   e) *Cluster Words* (the number of words in the candidate sentence that occurred adjacently in both the question and answer candidate (+))
4) The location or absence of desired entities is noted in the score.
5) Each of these distances are weighted, the sentences ranked and the top 5 sentences are then output.

Each of the above distances has its own weight and the corresponding sign shown above to attach to it. The score for an answer is the sum of distances and the top 5 sentences are then output.

In order to select the 250 or 50 byte answer from these sentences, the system identifies the longest mismatched pieces between the answer and the question. It then analyses the answer and the question to find where the center of the match is and using a

subject-verb-object assumptions of the sentence, it takes the question as either desiring the subject or the object, whichever has the least matches with the question.

E. Testing

Testing this IBM example, the authors came to the conclusion that the selection of the answer is still the hugest problem in QA systems. In 24.7 % of cases, their 250B system showed an error when selecting answers, while at 50B system the percentage reaches even 35.6 %.

The method used in the IBM QA system is a complex search that is not trying to get to the semantics of the text, but introduces a number of measures to calculate the overlapping of the content and the answer.

2. DeepQA Project

IBM Research undertook a challenge to build a computer system that could compete at the human champion level in real time on the American TV quiz show, Jeopardy. The Jeopardy Challenge helped to address requirements that led to the design of the DeepQA architecture (Figure 2.) and the implementation of Watson. The goals of IBM Research are to advance computer science by exploring new ways for computer technology to affect science, business, and society. The Jeopardy Challenge requires advancing and incorporating a variety of QA technologies including parsing, question classification, question decomposition, automatic source acquisition and evaluation, entity and relation detection, logical form generation, and knowledge representation and reasoning. Baseline performance is the QA system called Practical Intelligent Question Answering Technology (PIQUANT) (Prager, Chu-Carroll, and Czuba 2004), which had been under development at IBM Research by a four-person team for 6 years prior to taking on theJeopardy Challenge. At the time it was among the top three to five Text Retrieval Conference (TREC) QA systems. PIQUANT was a classic QA pipeline with state-of-the-art techniques aimed largely at the TREC QA evaluation (Voorhees and Dang 2005). PIQUANT performed in the 33 percent accuracy range in TREC evaluations. While the TREC QA evaluation allowed the use of the web, PIQUANT focused on question answering using local resources. A similar baseline experiment was performed in collaboration with Carnegie Mellon University (CMU) using OpenEphyra, an open-source QA framework developed primarily at CMU. The framework is based on the Ephyra system, which was designed for answering TREC questions. In these experiments on TREC 2002 data, OpenEphyra answered 45 percent of the questions correctly using a live web search. Minimal effort was spent in adapting OpenEphyra, but like PIQUANT, its performance on *Jeopardy* clues was below 15 percent accuracy.

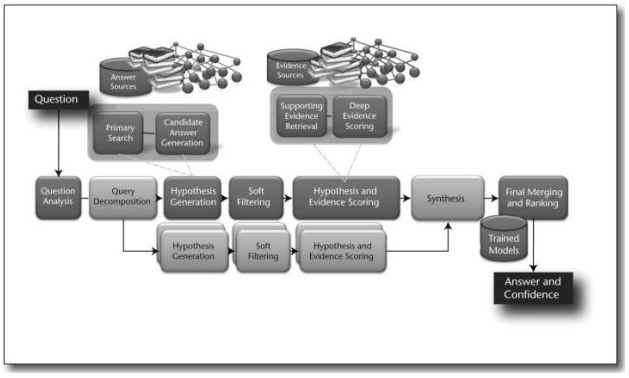


FIGURE 2. DEEPQA ARCHITECTURE

A system called DeepQA, still continuing to develop, is a massively parallel probabilistic evidence-based architecture. DeepQA is an architecture with an accompanying methodology. DeepQA has successfully been applied to both the *Jeopardy* and TREC QA task. It was first adapted to different business applications and additionally to exploratory challenge problems including medicine, enterprise search, and gaming.

The overarching principles in DeepQA are massive parallelism, many experts, pervasive confidence estimation, and integration of shallow and deep knowledge.

Massive parallelism: Exploit massive parallelism in the consideration of multiple interpretations and hypotheses.

Many experts: Facilitate the integration, application, and contextual evaluation of a wide range of loosely coupled probabilistic question and content analytics.

Pervasive confidence estimation: No component commits to an answer; all components produce features and associated confidences, scoring different question and content interpretations. An underlying confidence-processing substrate learns how to stack and combine the scores.

Integrate shallow and deep knowledge: Balance the use of strict semantics and shallow semantics, leveraging many loosely formed ontologies (Ferrucci at al., 2010).

The authors believe that in the future more methods will focus on detecting semantics and propose pretreatment of the texts and their recording in the network of knowledge to be used for asking questions and searching answers. It is planned to do such a research by using the NOK method.

IV. CONCLUSION

The complexity of technical evaluation of QA systems indicates resources and evaluation as two critical issues in their development.

Standardized tests use multiple-choice questions.

The selection of the evaluation criteria presents the problems in assessing the answer. The evaluation criteria are: relevance, correctness, conciseness, completeness, coherence and justification.

Today's QA evaluation of TREC requires the answer to be justified by the context as well as the limitation of the answer (in bytes), which is the first step towards brevity.

Evaluation criteria must depend on the purpose of use, the type of user who searches for an answer and on the interface.

Regarding the short answers, it is possible to automate the comparison of answers. Comparisons that are not accurate, as those by human evaluator, nevertheless provide good results (93-95 %).

Tests where the person himself / herself writes the answers (either short answers or an essay) are considered to be too subjective for standardized testing.

An automated method of evaluation, sometimes in combination with only one evaluator has proven to be worthwhile, which indicates that the automated grading systems are approaching human evaluators.

The IBM, which has played a leading role in the development of information technology, has developed QA systems whose results show an error in 30 % of cases.

They are constantly seeking an algorithm (method, model, procedure) to solve the selection of a short and concise answer.

QA systems are still unreliable, do not always provide good results, require a lot of sources, can't make conclusions, do not solve the context problem, etc.

Further research will go towards the development of QA systems using the NOK.

ACKNOWLEDGMENT

The research has been conducted under the project "Extending the information system development methodology with artificial intelligence methods" (reference number 13.13.1.2.01.) supported by University of Rijeka (Croatia).

REFERENCES

1. Academic Hallmarks (1999). Knowledge master. http://greatauk.com/ (10.02.2014.)
2. Breck, E., Burger, J. D., Ferro, L., Hirschman, L., House, D., Light, M., & Mani, I. (2000). How to evaluate your question answering system every day and still get real work done. *arXiv preprint cs/0004008*.
3. Chinchor, N., & Robinson, P. (1997, September). MUC-7 named entity task definition. In *Proceedings of the 7th Conference on Message Understanding*.
4. Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A. A., ... & Welty, C. (2010). Building Watson: An overview of the DeepQA project. *AI magazine*, *31*(3), 59-79.
5. Gašpić, K., Statističko strojno prevođenje, (2012) http://www.scribd.com/doc/105502391/Statisti%C4%8Dko-strojno-prevo%C4%91enje (10.02.2014.)
6. Hirschman, L., Breck, E., Light, M., Burger, J. D., & Ferro, L. (2000). Automated grading of short-answer tests.
7. Hirschman, L., & Gaizauskas, R. (2001). Natural language question answering: The view from here. *Natural Language Engineering*, *7*(4), 275-300.
8. Ittycheriah, A., Franz, M., Zhu, W. J., Ratnaparkhi, A., & Mammone, R. J. (2000, November). IBM's Statistical Question Answering System. In *TREC*.
9. Ittycheriah, A., Franz, M., & Roukos, S. (2001). IBM's Statistical Question Answering System-TREC-10. In *TREC*.
10. Jakupović, A., Pavlić, M., Dovedan, Z. Han, „Formalisation Method for the Text Expressed Knowledge" neobjavljeno, 2013.
11. Jones, K. S. (Ed.). (1997). *Readings in information retrieval*. Morgan Kaufmann.
12. Kukich, K. (2000). Beyond automated essay scoring. *IEEE intelligent systems*, *15*(5), 22-27.
13. Kupiec, J. (1993, July). MURAX: A robust linguistic approach for question answering using an on-line encyclopedia. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 181-190). ACM.
14. Mann, G. S. (2001, July). A statistical method for short answer extraction. In *Proceedings of the workshop on Open-domain question answering-Volume 12* (pp. 1-8). Association for Computational Linguistics.
15. Merialdo, B. (1994). Tagging English text with a probabilistic model. *Computational linguistics*, *20*(2), 155-171.
16. Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1990). Introduction to wordnet: An on-line lexical database*. *International journal of lexicography*, *3*(4), 235-244.
17. Pavlić, M., Development of a method for knowledge modeling 2013., Rijeka: Odjel za informatiku Sveučilišta u Rijeci, 2013a
18. Pavlić, M., Jakupović, A., Meštrović, A.: "Nodes of knowledge method for knowledge representation", Informatol. 46, 2013b, 3, 206-214
19. Pavlić, M., Meštrović, A., Jakupović, A.,"Graph-Based Formalisms for Knowledge Representation", Proceedings of the 17th World Multi-Conference on Systemics Cybernetics and Informatics (WMSCI 2013), Vol 2. 2013c 200-204.
20. Prager, J. M.; Chu-Carroll, J.; and Czuba, K. 2004. A Multi-Strategy, Multi-Question Approach to Question Answering. In *NewDirections in Question-Answering*, ed. M. Maybury. Menlo Park, CA: AAAI Press.
21. Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M. M., & Gatford, M. (1995). Okapi at TREC-3. *NIST SPECIAL PUBLICATION SP*, 109-109.
22. Roever, C. (2001). Web-based language testing. *Language Learning & Technology*, *5*(2), 84-94.
23. Tomljanović, J., Krsnik, M., Pavlić, M. (2014). Inteligentni sustavi pitanja i odgovora. Zbornik Veleučilišta u Rijeci - Journal of the Polytechnic of Rijeka
24. Ushida, E. (2005). The role of students' attitudes and motivation in second language learning in online language courses. *CALICO journal*, *23*(1), 49-78.
25. Voorhees, E. M., and Dang, H. T. 2005. Overview of the TREC 2005 Question Answering Track. In *Proceedings of the Fourteenth Text Retrieval Conference*. Gaithersburg, MD: National Institute of Standards and Technology.
26. http://www.eureka-centar.hr/upisi/on-line-testiranje/engleski-jezik/ (11.02.2014.)
27. http://www.pnas.org/content/100/15/9096.long (11.02.2014.)
28. http://www.surrey.ac.uk/ELI/ltr.html (11.02.2014.)

Inteligentni sustav upravljanja tunelskom rasvjetom

Zdenko Balaž, Krešimir Meštrović, Petar Jandrlić i Marko Čaržavec
Tehničko veleučilište Zagreb/Elektrotehnički Odjel, Zagreb, Hrvatska
zbalaz@tvz.hr, kmestrovic@tvz.hr, pjandrlic@tvz.hr, mcarzavec@tvz.hr

Sažetak - Moguća racionalizacija svakog projekta rasvjete tunela počiva na njegovim utvrđenim eksploatacijskim parametrima. Obradom parametara rasvjete s pozicije ekspertnih sustava, naročito dugačkih tunela, ukazuje da su moguće značajne uštede kako u potrošnji električne energije tako i u održavanju. Raspoložive baze podataka i baze znanja, te manipulacija znanjem tijekom višegodišnje eksploatacije tunelskih rasvjeta rezultirale su kreacijom inteligentnog sustava za upravljanje tunelskom rasvjetom na način drugačiji od uniformno prihvaćenog i oslonjenog na dosadašnje legislative. Tunelska je rasvjeta kao sustav sačinjena od više međusobno povezanih i zavisnih dijelova rasvjete. Ona u prvom redu pridonosi sigurnosti prometa kroz tunel i stoga je bitno voditi računa o problematici prilagodbe i akceptiranja svih relevantnih parametara luminancije. Inteligentni je sustav zadužen da upravlja rasvjetom, na način da vodi računa o vizualnoj prilagodbi ali uzimajući u obzir prosječnu luminanciju okoline, samog tunela, predportalnih građevina i prometnice, kao i dnevnog svjetla, neba koje je podložno promjenama kako atmosferskih prilika tako i ekliptike. Za tako velik broj utjecajnih parametara razrađene su aktuatorske i perceptorske mogućnosti oslonjene na nove tehnologije opreme, s kojom su studenti Tehničkog veleučilišta u Zagrebu kroz kolegij, Umjetna inteligencija – Espertni sustavi na III. Semestru specijalističkog studija provodili ispitivanja. Dobiveni rezultati i prikupljeni podaci u bazama znanja sintetizirani inteligentnim sustavom su potvrdili opravdanost novog pristupa projektiranju i izvedbi upravljanja tunelske rasvjete.

I. Uvod

Tunel kao objekt autoceste, sa svojom rasvjetom, je bitan tehnički čimbenik u funkciji sigurnosti prometa. Prilaz tunelu, predstavlja dio prometnice autoceste prije samoga ulaza u tunel i za njega se projektira dio tunelske rasvjete nazvan rasvjeta prilazne zone. Nakon ulaza u tunel projektira se dio sustava tunelske rasvjeta nazvan rasvjeta prijelazne zone koja sadrži zonu ulaza i zonu praga, što predstavlja najveću koncentraciju rasvjetnih tijela od ukupno projektiranog sustava rasvjete tunela. Sustav rasvjete tunela sastavljen je od osam podsustava, [1], i tijekom velikog projekta hrvatske izgradnje autocesta zadnjih petnaest godina, zbog nepostojanja domaće legislative u projektima rasvjete tunela, koristila se austrijska i njemačka, [2], [3] i [4].

Problematika upravljanja rasvjetom tunela obrađivala se samo u razvojnim projektima, [5] i [6]. Oni su poslužili kao osnova prikupljanja baza podataka i baza znanja a kroz [7] i kao gradivo za edukaciju. Kroz taj angažman inteligentni sustav, (IS), je poslužio kao izazov i osnova studentima koji su svojim radovima ukazali da ima još prostora za racionalizaciju, [8] i [9], što je namjera prikaza u ovom radu.

Kod vozača prilikom nailaska na tunel, prostorna i vremenska prilagodba uvjetima vožnje, može izazvati osjećaj nelagode i nesigurnosti. Prilikom nailaska, odnosno ulaska i izlaska iz tunela, vrijeme prilagodbe može biti kritičan čimbenik koji utječe na sigurnost prometa.

Promjena sadržaja u vidnom polju vozača, tijekom vizualne prilagodbe pri približavanju ulaza u tunel, ili izlaza, povezana je sa subjektivnim osjećajem percepcije luminiscencije, [1]. Izraz luminiscencija odabran prema [10], ukazuje na svijetljenje, odnosno stvaranje izvora svjetlosti kojeg u prvom redu promatramo od stanja dnevnog svjetla. Iz tih sadržaja bitno je prepoznati i druge parametre prema kojima bi se projektirala rasvjeta tunela jer stvaranje svjetlosti, one dnevne, podložno je klimatskim uvjetima.

Među parametrima moguće je izdvojiti osim svijetle površine neba - svijetli sloj atmosfere između položaja vozača i ulaza u tunel, površinu kolnika dionice ceste ispred vozača, okolicu oko ulaza u tunel, okolicu ceste ispred vozača, otvor i portalne građevine tunela.

II. Novi parametri utjecaja na rasvjetu tunela

A. Ekliptički parametri

Svijetla površina neba, svijetli sloj atmosfere posljedica je položaja Sunca. U svakom trenutku visina i azimut jednostavno opisuju njegovo kretanje. Globalnim sustavima za utvrđivanje pozicije moguće je vrlo precizno pratiti ekliptičke promjene, Tablica 1.

Kretanje sunca po nebu, Sunčev dijagram, prikazuje se u cilindričnim ili polarnim koordinatama. Podaci su dostupni i od izuzetne su važnosti u istraživanjima za fotonaponsko iskorištavanje sunčeve energije, [11]. Rezultati tih istraživanja su korišteni za slaganje algoritama prema kojima funkcioniraju mjerni članova za utvrđivanje trenutnih, ekliptičkih parametara dnevnog svjetla sunca odabranih i u IS. Prema njima su se simulirale matrice potrebnih intenziteta rasvjete ulazne odnosno izlazne zone tunela kako bi bile najbliže trenutnom dominantnom dnevnom svjetlu.

Promjenu dnevnog svjetla prati promjena scenarija prema kojem se odabire referentna veličina za upravljanje intenzitetom rasvjete preko agentskog sustava.

TABLICA I. EKLIPTIKA –RAZLIKA UNUTAR ČETIRI GODINE

Red. broj	Datumi i vremena solsticija i ravnodnevnica		
	Ravnodnevnica	*Ljetni solsticij*	*Zimski solsticij*
2012	20. III.-06:14	21. VI.-00:09	21. IX.-12:11
2013	20. III.-12:02.	21. VI.-06:04	21. IX.-18:11
2014	20. III.- 17:57	21. VI.- 11:51	22. IX.- 00:03
2015	20. III. -23:45	21. VI. -17:38	22. IX. -05:48

B. Klimatski parametri

U svim ekliptičkim stanjima luminiscencija Sunca na nebu kao izvor dnevnog svjetla podložna je klimatskim utjecajima, magla, kiša, oblaci i snijeg. Analizirani su eksploatacijski podaci i sintetizirani s rezultatima dobivenim mjerenjem koja su simulirala određene uvjete. Mjerenja su se dosta vjerno približavala simuliranim rezultatima, što je omogućilo korištenje istih ulaznih parametara za generiranje referentne veličine kao i kod ekliptičkih parametara samo uvođenjem određenih prigušnih faktora, f_{pm}, f_{pk}, f_{po} i f_{ps}.

C. Parametri doba dana

Za analizu dnevnih parametara korišteni su rezultati istraživanja prema [12]. FAA - Savezna je uprava za civilno zrakoplovstvo, koja kao agencija američkog Ministarstva prometa nadležna za reguliranje i nadziranje svih aspekata civilnog zračnog prometa u SAD-u, predlaže promjene legislative. Tako je bio slučaj s dijelom sustava svjetlosne signalizacije u funkciji sigurnosti zračnog prometa za koji je utvrđeno da su najvažniji parametri, koji utječu na rezultate vidljivosti pojedinih svjetlosnih sustava, prijelazi od jutarnjih do večernjih perioda u danu. Prosjek rezultata koji su dobiveni od 42 nezavisna subjekta direktno involvirana u procedure istraživanja, provedena od siječnja do rujna 2004. godine, prikazuje Tablica 1.

TABLICA II. REZULTATI VIDLJIVOSTI SVJETLOSNIH SUSTAVA

Svjetlosni sustav	Parametri, DP-1, 2, 3 i 4			
	Zora /Jutro (m)	*Dnevno svjetlo (m)*	*Sumrak /Večer (m)*	*Mrkla noć (m)*
Svjetlosni znakovi (SIGN)	146	142	122	120
Sustav zaštitnih svjetala (RGL-elevated)	237	95	260	255
Sustav zaštitnih svjetala (RGL-inset)	146	68	243	213
Sustav zaštitnih svjetala „T-con“ (RGL-inset)	76	55	121	127
Oznake bojom (MARK)	63	91	48	48

Tablica se odnosi na tri različita svjetlosna sustava:

- Svjetlosne znakove naredbi, koji su dio obvezne opreme aerodroma prema ICAO i FAA legislativi, [12] i [13]. (engl. Mandatory Instruction Signs). Ugrađuju se sa svake strane manevarske površine. Znakovi su oblika panela visine ispod jednog metra, bijele su boje na crvenoj podlozi.
- Sustav zaštitnih svjetala uzletno-sletne staze koji su podignuti iznad tla, (engl. Elevated Runway Guard Lights). Ugrađuju se sa svake strane manevarske površine. Svaka jedinica ima par svjetala koja naizmjenično svijetle s frekvencijom bljeska nešto kraćeg od jedne sekunde.
- Sustav zaštitnih svjetala uzletno-sletne staze koja su ugrađena u manevarsku površinu, okomito kao zaustavna prečka i sa svake strane, (engl. In-pavement Runway Guard Lights). Jedinice sinkronizirano svijetle s frekvencijom bljeska oko jedne sekunde. Ovaj tip sustava svjetala izveden je u dvije varijante s tim što druga varijanta ima još i par svjetala ispred za „T“ konfiguraciju, (engl. Alert Zone Lighting).

Osim svjetlosnih sustava u tablici je naveden i rezultat za oznake na mjestu čekanja, na manevarskim površinama oslikane bojom, (engl. Runway Hold Position Markings). Ti su podaci korišteni u IS za donošenje odluke o uporabi određenih scenarija upravljanja rasvjetom koji rezultiraju određenom racionalizacijom, baš oslonjeno na analizirana četiri trenutka 24-satnog trajanja dana.

D. Manipulacije parametrima

Manipulacija raspoloživim parametrima ukazivala je na disproporciju rasvjete u pogonu od one koju bi angažirao inteligentni sustav. Za potvrdu dobivenog zaključka pristupilo se novoj analizi baziranoj na fotografskom dokumentiranju stanja. Fotografski snimci rasvjete najčešće ulazne i izlazne zone odabirani su prema kriterijima rezultata američkih istraživanja navedenih u [12]. Snimanja su provedena istom opremom za sve tunele. Iz baza podataka IS, izdvojene su raspoložive fotografije za usporedbu istih dijelova rasvjete u istim uvjetima za dva različita tunela, slika 1. i slika 2.

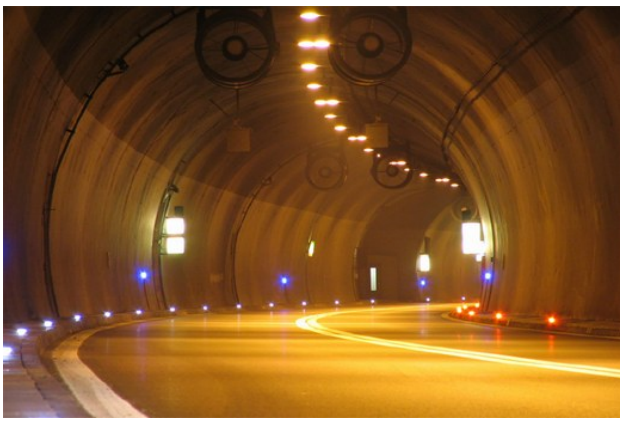


Slika 1. Ulazna zona jednog hrvatskog tunela na autocesti A1 s pripadajućim dijelom sustava rasvjete za parametre DP-3



Slika 2. Ulazna zona jednog francuskog tunela s pripadajućim dijelom sustava rasvjete uz identične parametre DP-3, kao na slici 1.

Za primjer su odabrani i parametri prijelazne zone tunelske rasvjete, koji su isti za DP-1-4, prema Tablici I. u usporedbi hrvatskog i talijanskog tunela, slike 3. i 4.



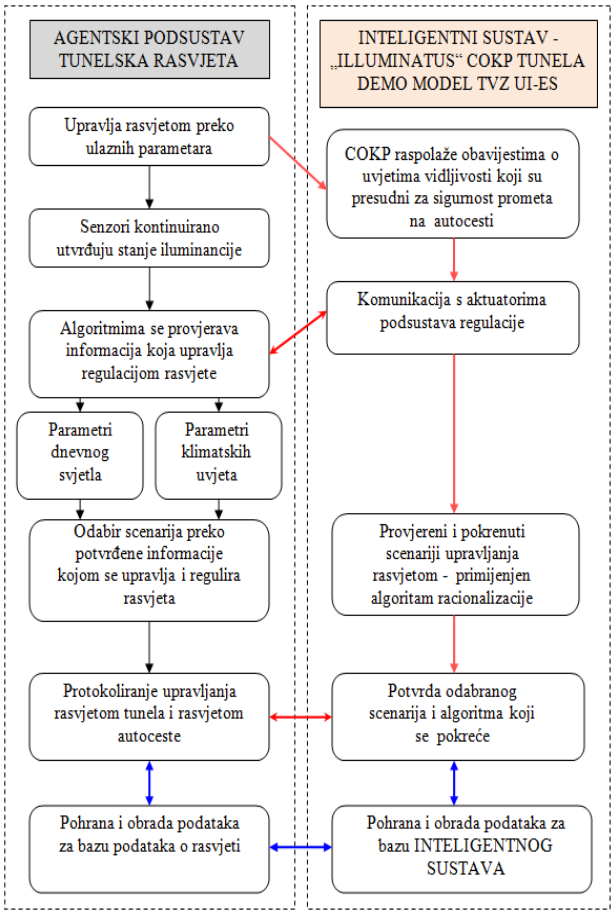
Slika 3. Prijelazna zona tunelske rasvjete hrvatskog tunela na autocesti A1, identična za sve parametre DP-1-4



Slika 4. Prijelazna zona tunelske rasvjete talijanskog tunela, identična za sve parametre DP-1-4

E. Mehanizmi zaključivanja - upravljanje rasvjetom

U razvoju IS su uzete u obzir sve specifičnosti koje unosi odabrani tip rasvjetnih tijela ovisno, o tome kako je tunelska rasvjeta projektirana, evidentne razlike sa slika 3. i 4. To je posebna baza podataka, obogaćena bazom znanja koje je znanje eksperata potvrđeno u eksploataciji. Razradu za provedbu algoritama i scenarija IS za upravljanje rasvjetom prikazuje dijagram, slika 5.



Slika 5. Dijagram provedbe algoritama i scenarija upravljanja rasvjetom tunela

F. Učenje iz inteligentnog sustava

Kreacija baza znanja i baza podataka u IS je posebno interesantna s aspekta planiranja održavanja. Ti podaci i znanja mogu se koristiti odmah, jer djeluju na korekcije ili potvrde odabranih scenarija, čime se direktno može utjecati na racionalizaciju, što je i cilj razvoja inteligentnog sustava. Gledano kroz eksploataciju tunela moguće je vrlo jednostavno sumu troška za održavanje, (ΣT), promatrati kao zbroj troškove preventivnog i interventnog održavanja, ($T_{\pi io}$), i troškova za utrošenu električnu energiju, ($T_{\varepsilon\lambda}$), "(1)".

$$T_{\pi io} + T_{\varepsilon\lambda} = \Sigma T. \qquad (1)$$

U troškovima plaćene električne energije za pogon tunela dominantna je stavka, rasvjeta, [14]. Kao što je eksploatacijskom kreacijom baza znanja i baza podataka IS uočeno, postoji više mogućnosti utjecaja na upravljanje rasvjetom tunela. Mogućnosti pružaju prvenstveno nove tehnologije i aktualna legislativa, ali i empirijska i heuristička znanja eksperata, [15] i [16]. Spoznaja da je ekspert stručnjak odnosno specijalist postavlja ga u poziciju vrsnog poučavatelja koji ne predstavlja nadmoćnu instituciju, instancu za posljednje odluke. Već kao što u korijenu riječi stoji, (lat. experientia docet – iskustvo poučava), skupljanje i posjedovanje iskustva, nije zvanje, ali pohranjeno u IS posreduje u praksi i na taj način IS zamjenjuje eksperta.

G. Rezultati

Odabrani parametri još do sada kako je poznato autorima ovoga rada nisu valorizirani u scenarijima upravljanja na način kako je razrađeno. Rasvjeta u pogonu hrvatskih tunela u velikom broju slučajeva, snimljena i izmjerena na licu mjesta u velikom broju slučajeva bila je i za 50 % veća od one koju bi angažirao inteligentni sustav, a koja se u 80% slučajeva poklapala sa snimljenom rasvjetom nekih europskih tunela, [5] i [6]. Najveća odstupanja u luminiscenciji, (zatečena – generirana od IS), uočena su prilikom obrade dnevnih parametara, DP-1, (svitanje/zora) i DP-3, (predvečer/suton). Ti su se rezultati poklapali sa stanjem koje je bilo posljedica primjene prigušnih faktora, f_{pm}, f_{pk}, f_{po} i f_{ps}. Faktori prigušenja zbog magle, f_{pm}, kao i zbog snijega, f_{ps}, nisu utjecali na luminiscenciju kao što je utjecao faktor prigušenja dnevne svjetlosti zbog naoblake, f_{po}, čiji je efekt najbliži parametru DP-3. Dobiveni su rezultati korišteni za simulaciju novog upravljanja tunelske rasvjete i izračunati su troškovi za potrebnu električnu energiju rasvjete na referentnom tunelu, [17]. Izračunata ušteda električne energije za rasvjetu svih hrvatskih tunela iznosi 8%.

Rezultati provedenih istraživanja razvojem IS, uspoređivani su prema [17], i izdvojen je zaključak koji prezentira drugi pristup troškovne analize izrađene za investitora s 5 različitih scenarija u periodu od 15 godina, što je životni vijek sustava rasvjete. Analize su obuhvatile pet varijanti, od kojih niti jedna ne obrađuje sustav upravljanja, a sve upućuje na zaključak s povećanjem troškova održavanja i troškova za električnu energiju za 2% godišnje, u što je uključen očekivani rast cijena električne energije i inflaciju?

III. ZAKLJUČAK

Inteligentni sustav, (IS), koji se koristi na Tehničkom veleučilištu u Zagrebu, (TVZ), u izbornom kolegiju služi studentima za stjecanje znanja iz područja umjetne inteligencije. Uključivanjem u njegovu izgradnju i razvoj kroz seminarske radove dobivene su konkretne spoznaje i utvrđeni i drugačiji parametri koji upućuju na novi pristup u projektiranju upravljačkih dijelova sustava rasvjete. Karakteristike i efikasnost agentskog podsustava, za upravljanje rasvjetom tunela ovise osim o povezanosti s ekspertnim sustavom za nadzor i upravljanje u centru za održavanje i kontrolu prometa, kojeg predstavlja demo IS za edukaciju na TVZ-u i o novim parametrima. Formiranim bazama znanja i podataka o stanju sustava rasvjete složeni su algoritmi za upravljanje koji su uzeli u obzir nove parametre koji još do sada nisu bili u projektima sustava tunelske rasvjete. Analogijom primjene novih tehnologija i metoda ispitivanja fotonaponskog iskorištavanja sunčeve energije utvrđeni su ekliptički parametri. Istraživačkim projektima u zrakoplovstvu postavljeni su dominantni parametri dnevnog svjetla povezani s prigušnim parametrima zbog klimatskih uvjeta koji su iskorišteni i u demo IS. Prema algoritmima su složeni scenariji kojima se postiže racionalnije korištenje sustava tunelske rasvjete, bez narušavanja sigurnosti prometa, jer su dobiveni rezultati potvrđeni u usporedbama s nekim europskim tunelima. Iako demo IS, studenti na njemu uče konkretne primjere iz prakse i promišljaju na način kojim umjetna inteligencija poučava. Studenti TVZ-a, specijalističkog studija Elektrotehničkog odjela na izbornom kolegiju kroz IS uče i svojim ga znanjem, kako onim kroz seminarski rad, tako i onim iz obrade pretraga, nadograđuju. Raspoloživa znanja iz baza kao i kompetencije koje studenti stječu mogu pomoći u kreiranju planova održavanja.

LITERATURA

[1] E. Širola, „Cestovna rasvjeta, ESING, Energetika marketing, Zagreb, 1997.

[2] RVS - Richtlinien und Vorschriften für das Straßenwesen, Projektierungsrichtslinien: Betriebs und Sicherheitseinrichtungen Tunnelausrüstung, Wien, 20. Mai 2001.

[3] RABT - Richtlinien für die Ausstattung und den Betrieb von Strassentunneln, Forschungsgesellschaft für Strassen und Verkerswesen, Arbeitsgruppe verkersführung und verkehrssicherheit, Ausgabe November 2001.

[4] UN, Vijeće za gospodarstvo i društvena pitanja, povjerenstvo za gospodarstvo za Europu, Komisija za unutarnji promet, Ad hoc multidisciplinarna grupa stručnjaka za sigurnost u tunelima: "Završno izvješće grupe stručnjaka za sigurnost u cestovnim tunelima", Prijevod A.C.7/9, 10. prosinca 2001.

[5] Z. Balaž, "Razvoj inteligentnog sustava za racionalizaciju potrošnje električne energije regulacijom rasvjete", Elaborat Br. 700-006, HAC- Sektor za projektiranje i razvoj, Zagreb, 2009.

[6] Z. Balaž, "Inteligentni sustav regulacije rasvjete Tunela sv. Rok", Seminarski rad, FESB, Poslijediplomski doktorski studij elektrotehnike i informacijske tehnologije, Split, 2009.

[7] Z. Balaž, "Predavanja iz kolegija Umjetna inteligencija – Ekspertni sustavi", TVZ, Elektrotehnički odjel, III. Semestar specijalističkog studija, Zagreb, akademska godina 2013/14.

[8] D. Gemović, „Inteligentni sustav gospodarenja upravljanjem tunelske rasvjete", Seminarski rad iz kolegija Umjetna inteligencija – Ekspertni sustavi", TVZ, Elektrotehnički odjel, III. Semestar specijalističkog studija, Zagreb, akademska godina 2013/14.

[9] P. Jandrlić, M. Čaržavec, „Inteligentni sustav za upravljanje tunelske rasvjete", Seminarski rad iz kolegija Umjetna inteligencija – Ekspertni sustavi", TVZ, Elektrotehnički odjel, III. Semestar specijalističkog studija, Zagreb, akademska godina 2013/14.

[10] B. Klaić, „Novi rječnik stranih riječi",Priredile Ljiljana Jojić i Anuška Nakić, Školska knjiga, Zagreb, 2012.

[11] LJ. Cvitaš, D. Žigman, K. Meštrović, „Utjecaj sjenila na fotostruje optičkih senzora", Polytechnic and Design, Vol.I, No.I, Tehničko veleučilište Zagreb, 2013.

[12] J. W. Patterson, Jr., FAA – Federal Aviation Administration, US Department of Transportation: „Evaluation of Runway Gard Lighr Configurations at North Las Vegas Airport", DOT/FAA/AR-TN06/19, National Technical Information Service (NTIS), Springfield, Virginia 22161, January 2007.

[13] ICAO, International Civil Aviation Organization -Annex 14, Volume I, "Aerodrome design and operations", 4th Edition, Montreal, 2004.

[14] Z. Balaž, "Praksa empirijske racionalizacije utroška električne energije naučena iz ekspertnih sustava - završno predavanje iz kolegija Umjetna inteligencija – Ekspertni sustavi", TVZ, Elektrotehnički odjel, III. Semestar specijalističkog studija, Zagreb, akademska godina 2013/14.

[15] H-G. Gadamer, "Nasljeđe Europe – Granice eksperta", Matica Hrvtska, Biblioteka Parnas, Preveo Kiril Miladinov, Priredio Damir Barbarić, Tisak Targa, Zagreb, 1997.

[16] Z. Balaž, K. Meštrović, "Učenje i poučavanje iz umjetne inteligencije", rad neobjavljen, predan za Polytechnic & Design, Vol. 2, No.2, TVZ, Zagreb, 2014.

[17] Z. Balaž, "Obrada i analiza rezultata studija pripremljenih razvojnih projekata racionalizacije potrošnje električne energije HAC-a", Elaborat Br. 700-001/14, HAC- Sektor za projektiranje, Zagreb, 2014.

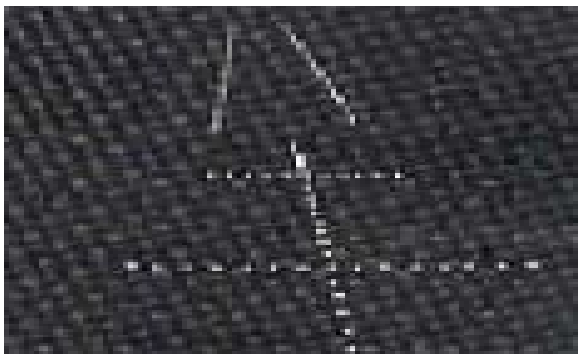
Inteligentni sustav održavanja aerodromskog sustava svjetlosne signalizacije

Zdenko Balaž, Krešimir Meštrović, Goran Bijelić
Tehničko veleučilište Zagreb/Elektrotehnički Odjel, Zagreb, Hrvatska
zbalaz@tvz.hr, kmestrovic@tvz.hr, gbijelić@tvz.hr

Sažetak - Aerodromski sustavi svjetlosne signalizacije su konfigurirani serijskim krugovima s konstantnim izvorom struje. Važan parametar koji ukazuje na funkcionalnost takvog strujnog kruga je otpor izolacije primarnih kabela serijskih strujnih krugova napajanih iz regulatora konstantne struje. Regulator je opremljen modulima koji nadziru i upravljaju rasvjetom. Raspoloživi parametri se nadziru i arhiviraju u računalo službe održavanja. Kreacijom baza znanja i baza podataka te manipulacijom znanjem postiže se efikasnije održavanje. Inteligentni sustav pomaže u procedurama preventivnog i prediktivnog održavanja i služi za učenje. Stečeno znanje ukazuje na moguće racionalizacije i na unapređenja prilikom revitalizacije ili zamjene postojećih sustava. Razvijenim inteligentnim sustavom koji se koristi u programu edukacije za stjecanje licence aerodromskih specijalista na održavanju sustava svjetlosne signalizacije, dobiveni su konkretni podaci o karakteristikama i efikasnosti pojedinih podsustava. Između ostalih to je sustav za nadzor stanja izolacije serijskih strujnih krugova koji pomaže u podizanju kvalitete izolacije. Kontinuiranim praćenjem i stvaranjem baza podataka o stanju izolacije složeni su algoritmi za predviđanje mogućih kvarova a samim tim i za racionalnije održavanje uz visoku pogonsku pouzdanost.

I. UVOD

Svjetlosna signalizacija na zračnim lukama u početku svoje primjene služila je isključivo za vizualno navođenje zrakoplova u procedurama slijetanja i polijetanja, Slika 1.



Slika 1. Svjetlosna signalizacija aerodroma Heathrow 1957. godine

Danas je to složeni sustav s množinom opreme koja osim sustava svjetlosne signalizacije, obuhvaća sustave sigurne opskrbe električnom energijom s osnovnim i sekundarnim izvorima, sustave upravljanja kretanja zrakoplova po manevarskim površinama i sustave za navođenje zrakoplova na parkirne pozicije, Slika 2.



Slika 2. Aerodromski sustavi svjetlosne signalizacije CAT IIIB

U zrakoplovnoj terminologiji to se naziva kategorija zračne luke, (CAT) i predstavlja ukupnu razinu tehničke i tehnološke opremljenosti što uz stroge procedure i osposobljene, (licencirane) djelatnike, garantira pružanje kvalitetne usluga prihvata i otpreme zrakoplova – sigurnosti zračne plovidbe na zračnoj luci. Svi navedni specijalistički tehnički podsustavi povezuju se u ekspertni sustav s bazom podataka, bazom znanja i mehanizmima zaključivanja, koji se nalaze u aerodromskoj službi elektroodržavanja, engl. Airport Electrical Maintenance, (AEM).

U ovom radu bit će prezentirana iskustva stečena kroz realizaciju projekata aplikativne problematike, [1]-[4]. Stupanjem na snagu nove legislative, [5], pokrenuti su programi cjeloživotnog obrazovanja, (AERO TVZ), [6] i opremljen je praktikum za izvođenje vježbi i uvođenje znanstveno-istraživačkog rada, [7]. Uz programe edukacije razvija se inteligentni sustav koji se koristi formiranim bazama znanja i bazama podataka aerodromskih sustava svjetlosne signalizacije. Kreirani algoritmi ukazuju na karakteristike i efikasnosti pojedine opreme i podsustava, što se može vrlo kvalitetno primijeniti u poslovanju za unaprijed planirane zadatke. U edukacijama se koristi pristup kojim su japanski radnici zahvaljujući metodama osposobljavanja kroz kontrolu kvalitete provodili racionalizacije, unapređenja i procjenjivali svoje učinke, kao u [8] i "neobjavljen" [9].

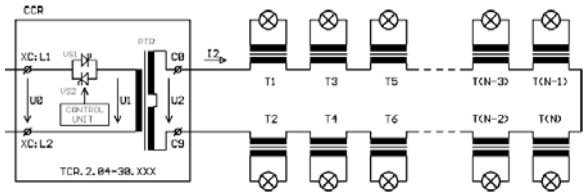
Uporaba informacijsko komunikacijskih tehnologija, (ICT), nameće informatizaciju samog posla. Potencijalno dostupne informacije i parametri mogu se uporabljivati u realnom vremenu te se s njima može manipulirati u smislu potreba odlučivanja, [10]. Inteligentni sustav, (IS), razvijen i instaliran kao demo model sustava svjetlosne signalizacije, uz [6], ima sve realne mogućnosti kao nadzorno upravljački aerodromski sustav. IS ima svoj hardver i softver i omogućava funkcije upravljanja koje se

u stvarnosti na zračnoj luci provode upravljačkom jedinicom na aerodromskom tornju, (TWR). Sve funkcije se nadziru lokalno na samoj opremi kao u AEM-u.

II. AERODROMSKI SUSTAV SVJETLOSNE SIGNALIZACIJE

A. *Koncepcija strujnih krugova aerodromske rasvjete*

Sustav svjetlosne signalizacije, engl. Airfield Lighting System, (ALS), koncipiran je serijskim strujnim krugovima, [11], koji su napajani iz izvora konstantne struje, regulatora, engl. Constant Current Regulator, (CCR). Svako rasvjetno tijelo, svjetiljka ili svjetlosna oznaka, u serijskom strujnom krugu je priključeno preko svog izolacijskog transformatora, (T_1-$T_{(N)}$), Slika 3.

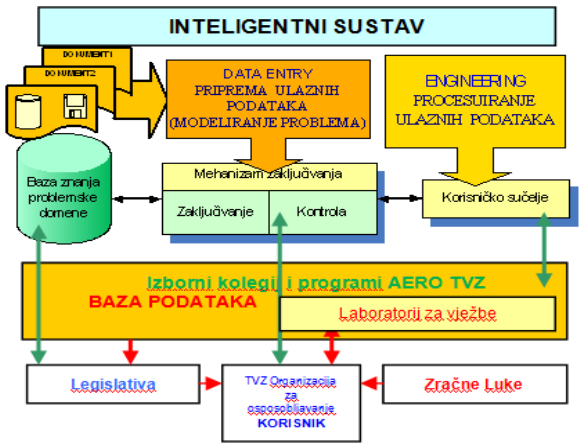


Slika 3. Serijski strujni krug sustava svjetlosne signalizacije

B. *Laboratorijska oprema ALS*

Kompletna oprema za laboratorijski model serijskog strujnog kruga ALS je dobivena zaslugom sponzora[1]. Primarni krug je sastavljen od CCR-a, primarnih kabelskih vodova s primarnim konektorima i izolacijskim strujnim transformatorima. Krug je opremljen senzorima, perceptorima i aktuatorima u modularnoj izvedbe i sve je ugrađeno u ormar CCR. Kao hardver IS sastoji se od modula napajanja, modula upravljačko procesorske jedinice, komunikacijskog modula, modula regulacije CCR-a, modula za nadzor stanja otpora izolacije i modula za kontrolu broja pregorjelih žarulja u serijskom strujnom krugu, sve integrirano operabilnim softverom.

Inteligentni sustav, (IS), je predstavljen računalom s aplikacijskom programskom podrškom kao mehanizmom zaključivanja, bazom znanja i bazom podataka, Slika 4.



Slika 4. Inteligentni sustav laboratorijskog modela aerodromskog sustava svjetlosne signalizacije

[1]-TRANSCON, Electronic Systems,www.transcon.eu. (sponzor)

C. *Inteligentni sustav za učenje i poučavanje*

Baze podataka i znanja, procesuiranje ulaznih podataka i modeliranje problema, kroz IS koriste se za učenje, znanstveno istraživački rad i programe edukacije na Tehničkom veleučilištu Zagreb, (TVZ), i to:

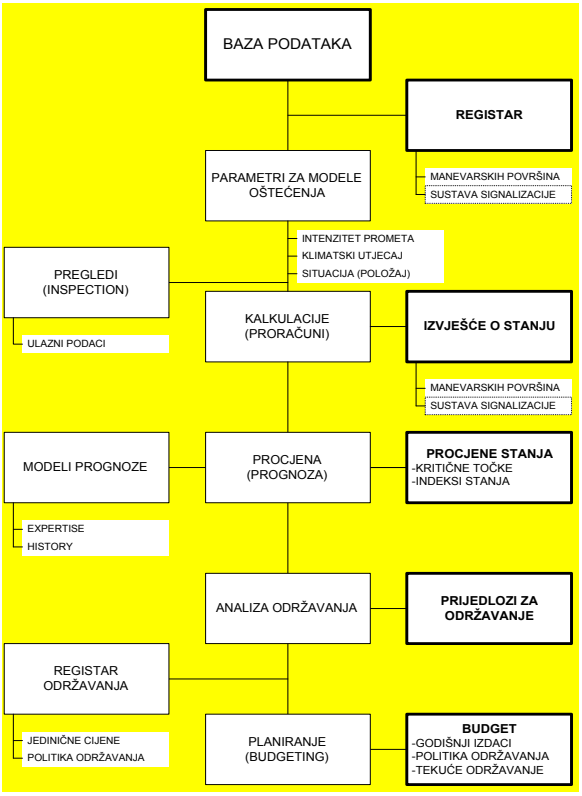
- Kroz teorijska predavanja izbornog kolegija Umjetna inteligencija, (UI) – Ekspertni sustavi, (ES), [7], studenti se uvode u problematiku IS, te kroz interaktivni rad otkrivaju i pronalaze mogućnosti djelovanja. Kroz spoznaje o UI i razvoj ES, studenti osim aplikativne primjene, stječu i usputni pozitivne učinke jer se kroz kolegij težište stavljaju na izučavanje s ciljem boljeg razumijevanja ljudske inteligencije i olakšavanje komunikacije ljudi sa strojem. Klasična UI, usko je vezana s pojmom znanja, njegovog prikupljanja, pohrane i primjene pri rješavanju složenih zadataka. Zbog tih se razloga kroz kolegij potiču razmišljanja o UI naše stvarnosti, odnosno konkretnih praktičnih problema primjera iz prakse, o domenama znanja koje se usvaja kroz kognitivne složenosti i kontekst znanja,"neobjavljen" [9].
- Kroz laboratorijske vježbe izbornog kolegija UI-ES, sintetiziraju se spoznaje do kojih se dolazi pretraživanjima baza znanja, uz pomoć ES. Na taj se način profiliraju stavovi, mišljenja i ekspertiza o znanju i naukovanju i potiče hermeneutičko promišljanje na način da se svjesno dopušta izmak uvjeta njegove ograničenosti kako za poučavatelja tako i za poučavanog,"neobjavljen" [9].
- Kroz specijalističke programe cjeloživotnog obrazovanja na TVZ-u, u skladu s [5], specijalistička znanja su podloga za stručnjake od kojih se očekuje stručnjačko rješavanje problema na razini ekspertize, a za što je neophodno razumijevanje**.** Za potvrdu razumijevanja koriste su raspoložive baze podataka, stvarni eksploatacijski rezultati te uspoređuju s rezultatima dobivenim pretragama. Na taj se način ekspertu "Otkriva mogućnost razumijevanja koje nije nemislivo unaprijed", [12]. Na specijalističkim edukacijama kandidate se upoznaju s djelovanjem eksperta, odnosno specijalista. Promatra ga se s pozicije vrsnog poučavatelja, što je i cilja kojem teže poučavani. U konačno dosegnutom statusu izobrazbe on ne predstavlja nadmoćnu instituciju, instancu za posljednje odluke, jer skupljanje i posjedovanje iskustva, nije zvanje. Ali zvanjem postaje kada posreduje svojim znanjem u stručnoj praksi. Očekuje se, da se njegova uloga odlikuje izrazom expertus, što se osamostalilo iz čitave velike rijeke iskustava stečenog u praksi i potvrđenog na edukaciji. Iz tih su razloga i predviđeni specijalistički programi koji se razlikuju, radili se o stjecanju potvrde o osposobljenosti, njenom obnavljanju ili produljenju. Ekspert se nalazi između znanosti, u kojoj mora posjedovati kompetencije i društvene politike prakse, [13].

- Znanstveno istraživački rad se potiče već izradom seminarskih radova za koje se studenti odlučuju tijekom slušanja kolegija. Predavanja su kao i razrađeni specijalistički programi usmjereni na konkretne probleme koji se rješavaju, pa je tako dostupno u bazi znanja dovoljno gradiva za odabir kvalitetnog završnog rada. Ovaj rad rezultat je rada mentora i studenta koji je započeo seminarskim a bit će finaliziran kroz završni, diplomski rad studenta trećeg koautora.

Iz raspoloživih baza znanja i podataka analiziraju su problemi vezani na konkretne serijske strujne krugove odabranih zračnih luka s učestalim smetnjama koje narušavaju operabilnost sustava svjetlosne signalizacije. Uzrokovani kvarovi na izolaciji primarnih kabelskih veza modeliraju problem i njegovo rješavanje korištenjem svih spoznaja koje se profilirale tijekom svih provedenih edukacijskih programa. Uspoređivanjem rezultata iz prakse sa zračnih luka, uključujući i one s najekstremnijim klimatskim uvjetima sintetiziraju se rješavanja, [14].

III. ODRŽAVANJE AERODROMSKIH SPECIJALISTIČKIH SUSTAVA I OPREME

Održavanje specijalističkih aerodromskih sustava pomoću računala započelo je još sredinom prošlog stoljeća uvođenjem računalne opreme i aplikacijskih programskih podrški oslonjenih najviše na građevinsko održavanje manevarskih površina zračnih luka, [15]. Sustav za gospodarenje kolnika (PMS), koji je prvotno razvila američka vojska doživio je čitav niz prilagodbi, između ostalog i nadogradnju za praćenje stanja sustava svjetlosne signalizacije. Bio je prvenstveno namijenjen za planiranje strategije održavanja, a korišten je i u razvoju edukacijskog modela IS, Slika 5.



Slika 5. Blok dijagram sustava za gospodarenje kolnika

A. *Održavanje sustava svjetlosne signalizacije*

Prema [11], način održavanja sustava svjetlosne signalizacije precizno je propisan. On obuhvaća građevinsko i elektro održavanje. U ovom je radu izložen model održavanja elektro opreme, ALS u AEM. Dvije su veličine kao karakteristike kvalitete ALS-a određene kao referentne. Prva je vezana na kvalitetu luminiscencije i definirana je postotnom vrijednošću intenziteta svakog pojedinačnog rasvjetnog tijela. Druga je vezana na kvalitetu izolacijskih svojstava primarnih kabelskih vodova i definirana je graničnom vrijednošću otpora izolacije kompletnog strujnog kruga. Procedure održavanja striktno su propisane i obuhvaćaju dnevne, tjedne, mjesečne i periodičke preglede. Osim pregleda propisane su i procedure ispitivanja i mjerenja, [11].

B. *Otpor izolacije serijskih strujnih krugova*

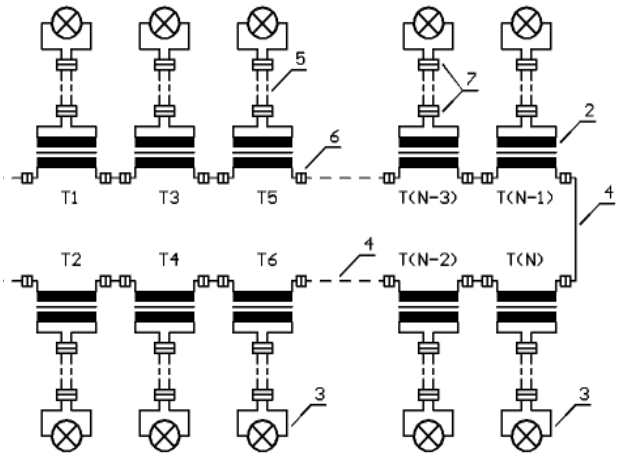
Vrijednosti otpora izolacije serijskih strujnih krugova, (SSK), Tablica I., za aerodromski sustav svjetlosne signalizacije, (primarni krugovi), prema referentnoj dokumentaciji[a], zavisna je o njihovim dužinama. U primarnim krugovima koriste se jednožilni kabeli, različitih tipova izolacije presjeka vodiča od šest kvadratnih milimetara za ukupni napon u krugu od pet kilovolta, kojima su povezani izolacijski transformatori.

TABLICA I. OTPOR IZOLACIJE SERIJSKIH KRUGOVA - ALS

Red.br.	Referentni dokument: MAINTENANCE OF AIRPORT VISUAL AID FACILITIES[a]		
	Dužina serijskog kruga(m)	*(MΩ)* [a]	*(MΩ)*
1	3000 ili manje	50	50
2	3000 - 6000	40	50
3	6000 ili više	30	50

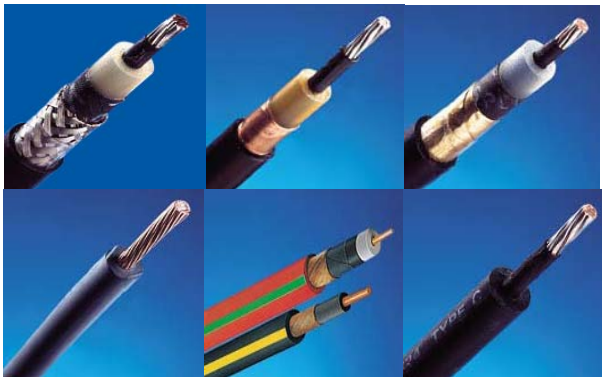
[a] FAA- Federal Aviation Administration AC No: 150/5340-26A, Initiated by: AAS-100

Prema slici 6. u serijskom krugu ALS, segmenti primarnih kabela, (4), povezuju transformatore, (2), primarnim konektorima, (6). Na svakom izolacijskom strujnom transformatoru preko sekundarnih kabela, (5) i konektora, (7), spojena je lampa, (3).



Slika 6. Primarni, serijski strujni krug sustava svjetlosne signalizacije

Izolacijski strujni transformatori u serijskom strujnom krugu bez obzira na pogonsko stanje sekundara, a koje može biti prazni hod, kratki spoj ili radno trošilo – žarulja, ne utječu na stanje otpora izolacije primarnog strujnog kruga. Iz tih razloga utvrđivanje stanje otpora izolacije primarnih strujnih krugova svodi se samo na ispitivanje kruga koji je sastavljen od primarnih kabela, Slika 7., primara izolacijskih strujnih transformatora, Slika 8. i primarnih konektora, Slika 9.



Slika 7. Tipovi primarnih kabela serijskog strujnog kruga sustava svjetlosne signalizacije



Slika 8. Tipovi izolacijskih strujnih transformatora serijskog strujnog kruga sustava svjetlosne signalizacije



Slika 9. Primarni i sekundarni konektori strujnih krugova ALS

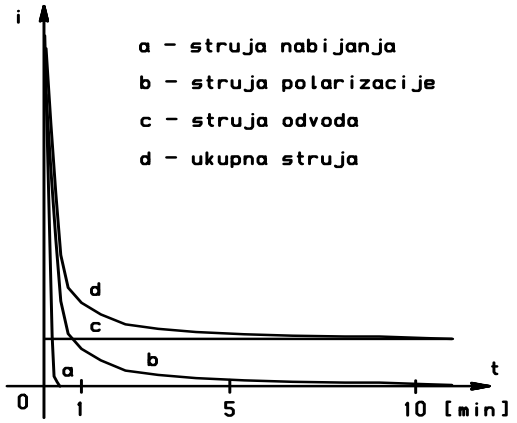
C. Mjerenje i prijenosna mjerila otpora izolacije

Za mjerenje otpora izolacije najčešće se primjenjuje U-I metoda. Koristi se istosmjerni napon, koji priključen na izolaciju u prvom trenutku propusti veliku struju nabijanja koja brzo opadne. Održava se polarizacijska struja koja opada sporije. Tek nakon nekoliko minuta ostane samo struja odvoda kroz izolaciju koja određuje istosmjerni statički otpor izolacije, Slika 10. Zato je bitno da se prilikom mjerenja čeka s očitanjem dok se struja kroz izolaciju približi stacionarnoj vrijednosti i ustali pokazivanje instrumenta kojim se mjeri.

Za provedbu mjerenje bitno je koliki je istosmjerni napon i temperatura jer o tome ovisi vrijednost izmjerenog izolacijskog otpora. Na temelju mjerenja otpora izolacije može se procijeniti stanje izolacije. Naročito, ako se mjerenje otpora izolacije provodi u određenim vremenskim intervalima. Može se na temelju kretanja vrijednosti otpora izolacije dobro zaključivati o stanju izolacije i očekivanim promjenama jer je za takvu ocjenu često važnije praćenje vremenske relativne promjene otpora izolacije nego apsolutni iznos otpora izolacije.

Ove spoznaje oko mjerila i mjerenja otpora izolacije rezultirale su punjenjem baze znanja. Utvrđene su razlike između mjerenje otpora izolacije i kontrole otpora izolacije. U samom početku uvođenja ispitivanja na ALS, za kontrolu izolacijskog otpora u pogonskim uvjetima su se koristili induktorski megaommetar. To su bili istosmjerni generatori s instrumentom s križnim svicima. U novije vrijeme se koriste baterijski uređaji koji daju konstantni istosmjerni mjerni napon uobičajenih vrijednosti 250, 500, 1000 ili 5000V. Preciznost mjerenja u početku nije bila propisana a danas uporabna pogreška mjerila, za mjerno područje ili mjerna područja koja je odredio proizvođač i za nazivne propisane uporabne zahtjeve ne smije biti veća od ± 30%. Za mjerila je potrebno provodi tipska ispitivanja koja se provode na osnovi dokumentacije i/ili rezultata laboratorijskih ispitivanja, [16].

Prema usvojenoj legislativi u Hrvatskoj je na snazi propis koji striktno definira način mjerenja otpora izolacije svih primarnih strujnih krugova sustava rasvjete operativnih površina i prilazne rasvjete. Otpor izolacije kabela nove i postojeće instalacije mora biti najmanje 50 MΩ, prema Tablici I., četvrta kolona nije definirana zavisnost o dužini primarnog kruga. Legislativa navodi, ako se prilikom kontrole mjerenja pokaže da otpor izolacije pada, provjeru otpora izolacije potrebno je obavljati češće s utvrđivanjem uzroka pada. Za ispitni napon se preporuča vrijednost od pet kV. Odstupanje od propisanog minimuma otpora izolacije je dozvoljeno do vrijednosti dobivene izračunom otpora izolacije za svaki strujni krug. Minimalni dozvoljeni otpor izolacije izražen u MΩ može se izračunati na temelju napona ispitivanja i dozvoljene struje odvoda. Proračun otpora izolacije i mjerenje izoliranosti kabela za napajanje električnom energijom i kabelima za signalizaciju i upravljanje provode se u skladu s legislativom, [11].



Slika 10. Struje kroz izolaciju

D. Ugradni moduli za mjerenje otpora izolacije

Spoznaje o postupcima mjerenja kao i procedure ispitivanja otpora izolacije pridonijele su razvoju i primjeni uređaji za kontinuirano praćenje stanja otpora izolacije serijskih strujnih krugova, koji se kao standardna oprema SSK ugrađuje uz CCR opremu ALS jer se pokazalo kako kontinuirani ispitni napon narinut na primarne krugove pridonosi poboljšanju narušenog otpora izolacije zbog vlage, jer se izolacija suši i poboljšava.

Poznato je da prijenosna mjerila otpora izolacije, zbog primjene visokih mjernih napona moraju imati takva konstrukcijska svojstva koja će omogućiti da se njima rukuje sigurno. Također uz svako mjerilo mora biti na raspolaganju odgovarajući visokonaponski kabel koji je za to namijenjen. Ako se k tome doda i nužna obučenost rukovatelja kao i periodično ispitivanje mjerila, prednosti stacionarnih uređaja, kao standardne opreme za kontinuirano praćenje stanja otpora izolacije SSK, i te kako su opravdane, [17].

Analizirani rezultati ispitivanja otpora izolacije nekih karakterističnih SSK, [18], prikupljeni u bazi podataka IS, poslužili su za algoritme održavanja, Tablica II.

TABLICA II. OTPOR IZOLACIJE SERIJSKIH KRUGOVA - ALS

Godina/ /mjesec [a]	REZULTATI MJERENJE NA KARAKTERISTIČNIM KRUGOVIMA		
	Ispitivanje terensko prijenosnim	***Ispitivanje uređaj ugradni -selekcijski***	***Ispitivanje uređaj ugradni - kontinuirano***
	(MΩ)	***(MΩ)***	***(MΩ)*** [b]
1995./III.	5	16	25
1996./VI.	12	19	35
1997./VIII.	15	20	50
1998./XI.	5	17	55
2001./I.	3	18	70

[a] – odabrani vremenski intervali koji uključuju i različite klimatske uvjete kroz godinu

[b] – krug uz ispitivanje otpora izolacije kontinuiranim naponom prolongiran za sanaciju

IV. ZAKLJUČAK

Razvijenim inteligentnim sustavom, (IS), koji se koristi u programima edukacije na TVZ, kako u izbornom kolegiju tako i za stjecanje licence aerodromskih specijalista na održavanju svjetlosne signalizacije, dobiveni su konkretni parametri koji upućuju na karakteristike i efikasnosti pojedinih podsustava. Između ostalih, implementirani modularni sustav za nadzor stanja izolacije serijskih strujnih krugova pridonosi podizanju kvalitete izolacije. Formiranim bazama znanja i podataka o stanju izolacije složeni su algoritmi za predviđanje mogućih kvarova ili za odgađanje sanacije zbog poboljšanja koje je uzrokovano primjenom novih metoda ispitivanja. Samim tim se postiže racionalnije održavanje uz visoku pogonsku pouzdanost.

Kandidati uključeni u specijalističke edukacije i studenti izbornog kolegija kroz IS uče i svojim ga znanjem, kako onim empirijskim tako i onim iz obrade pretraga, nadograđuju. Stečena znanja kao kompetencije educiranim specijalistima pomažu u kreiranju planova održavanja.

ZAHVALA

Za kompletno opremanje laboratorija serijskim strujnim krugom ALS, koji je korišten kao IS i u ovom radu, zahvaljujemo se svojem sponzoru[1], TRANSCON ELECTRONIC SYSTEMS, spol. s r.o., Kvapilova 2133, 738 01 Frýdek-Místek, CZ,, www.transcon.eu

LITERATURA

[1] Z. Balaž, A. Franić, N. Mihelčić, "Tehnička strana opremljenosti Zračne luke Zagreb za uvođenje kategorije CAT IIIB", 17. skup o prometnim sustavima Automatizacija u prometu, Osijek - Budimpešta, 1997.

[2] Z. Balaž, "Primjena računala u procesu održavanja manevarskih površina zračne luke – Programi za građevinsko i elektro održavanje", Stručni skup ISEP 98, Elektrotehnička zveza Slovenije, 1998.

[3] Z. Balaž, "Analiza registriranih poremećaja u radu sustava svjetlosne signalizacije na Zračnoj luci Zagreb - Sicomos", Elaborat br. 10028/01 - Aeroing d.o.o., Zagreb, 2001.

[4] Z. Balaž, "Expertiza specijalističkih sustava u funkciji sigurnosti zračne plovidbe na Zračnoj luci Mostar i prijedlog rješenja za unaprijeđenje", Studija br. 10056/01 - Aeroing d.o.o., Zagreb, 2001.

[5] Narodne novine RH, "Pravilnik o stručnom osposobljavanju zaposlenika i drugih ugovornih djelatnika aerodroma, pružatelja zemaljskih usluga i korisnika usluga aerodroma koji samostalno obavljaju zemaljske usluge", NN br. 7/13, Zagreb, 2013.

[6] Z. Balaž, "Priručnik za osposobljavanje djelatnika koji na zračnim lukama obavljaju specijalistička održavanja", TVZ, Centar cjeloživotnog obrazovanja, Zagreb, 2013.

[7] Z. Balaž, "Predavanja iz kolegija Umjetna inteligencija – Ekspertni sustavi", TVZ, Elektrotehnički odjel, III. Semestar specijalističkog studija, Zagreb, akademska godina 2013/14.

[8] W.E. Deming, "Quality, productivity and competitive position" MA MIT Center for Advanced Engineering Study, 1982.

[9] Z. Balaž, K. Meštrović, "Učenje i poučavanje iz umjetne inteligencije", rad predan za Polytechnic & Design, Vol. 2, No.2, TVZ, Zagreb, 2014.

[10] S. Zuboff, "In the adge oft he smart machine: The future of work and power," New York, Basic Books, 1988.

[11] ICAO -Annex 14, Volume I, "Aerodrome design and operations", 4th Edition, Montreal, 2004.

[12] G. Figal, "Smisao razumijevanja", Matica Hrvtska, Biblioteka Parnas, Prevela, Darija Domić, Priredio Damir Barbarić, Tisak Targa, Zagreb, 1997.

[13] H-G. Gadamer, "Nasljeđe Europe", Matica Hrvtska, Biblioteka Parnas, Preveo Kiril Miladinov, Priredio Damir Barbarić, Tisak Targa, Zagreb, 1997.

[14] Z. Balaž, "Upgrading of Airport at Nafoora Field – Arabian Gulf Oil Company", Benghazi, Lybia, Aeroing d.o.o./SINIT Rovigo, Italy, 2002.

[15] ICAO Journal, "PMS/ATM Systems Implementation", Montreal, 6-24. December, 1993.

[16] Narodne novine RH, „Pravilnik o mjeriteljskim zahtjevima za mjerila otpora izolacije", NN br. 81/02, Zagreb 2002.

[17] G. Bijelić, D. Jakušić, T. Fruk, " Inteligentni sustav za obradu rezultata mjerenja otpora izolacije", Seminarski rad iz kolegija Umjetna inteligencija – Ekspertni sustavi", TVZ, Elektrotehnički odjel, III. Semestar specijalističkog studija, Zagreb, akademska godina 2013/14.

[18] Z. Balaž, "Analiza rezultata ispitivanja stanja otpora izolacije serijskih strujnih krugova Aerodroma Brač, Zračne luke Zagreb i Airport Nafoora", No: F-68/02, Aeroing d.o.o., Zagreb, 2002.

Mobilni uređaji u funkciji racionalizacije upravljanja udaljenim računalima emuliranjem računalnog miša

Dimitrij Vukelić*, Marino Franušić**, Ivan Grakalić**,
* Gradsko kazalište lutaka Rijeka – mitja.vukelic@gmail.com
** Veleučilište u Rijeci, Rijeka – marino.franusic@veleri.hr, ivan.grakalic@veleri.hr

Sažetak - S obzirom na razvijenost mobilnih operacijskih sustava i podršku sklopovlju za povezivanje u bežične mreže, pri razvoju aplikacija za mobilne uređaje može se iskorištavati sve aspekte računalnih mreža. Osim mobilnog prijenosa podataka za spajanje na internet mobilnom mrežom, uz pomoć integriranih mrežnih kartica mobilni uređaji spajaju se i u bežične lokalne mreže. Povezivanjem u lokalnu mrežu ostvaruje se komunikacija s ostalim uređajima spojenim u istu mrežu, a time se otvara cijeli niz mogućnosti korištenja tih uređaja putem mobilnog uređaja. Cilj ovog rada je iskoristiti mobilni uređaj u svrhu upravljanja udaljenim računalom uz pomoć pokreta mobilnog uređaja detektiranog ugrađenim akcelerometrom. U radu se definira koncept upravljanja spajanjem mobilnog uređaja i udaljenog računala u bežičnu lokalnu mrežu. Spajanje se odvija u klijent – server arhitekturi putem protokola UDP. Radom je pokazana mogućnost upravljanja emuliranjem rada računalnog miša i dan je pregled mogućih primjena. Moguće primjene sustava temelje se na sve većoj popularizaciji „pametnih" mobilnih uređaja opremljenih svim potrebnim elementima za rad ovog sustava. Testirana funkcionalnost na više platformi dokaz je valjanosti koncepta i iskoristivosti.

I. UVOD

Sredinom devedesetih godina dvadesetog stoljeća na tržištu su se pojavili mobilni uređaji s vlastitim operacijskim sustavom. Početak razvoja pokrenula je tvrtka IBM s uređajem IBM Simon, mobilnim telefonom i dlanovnikom u jednom. Otvaranjem novog tržišta pokreće se intenzivniji razvoj mobilnih uređaja s operacijskim sustavima te se pojavljuju operacijski sustavi PalmOS, Windows CE za mobilne uređaje, Nokia S40, Symbian i Blackberry OS kao najistaknutiji. No pravi porast razvoja i veći ulazak na tržište pokreće tvrtka Apple plasiranjem svojeg iPhone mobilnog telefona. Od tog trenutka, zbog velikog porasta tržišta i prihvaćenosti u korisnika, pametni mobilni telefoni (engl. smartphone) i njihovi operacijski sustavi postaju vrlo zanimljivi kompanijama

Tehnološki napredak uređaja zahtijeva i vrlo brz razvoj mobilnih operacijskih sustava koji ih pokreću. Mobilni operacijski sustavi, za razliku od klasičnih, prilagođavaju se posebnostima platforme za koju se razvijaju. Tako je potrebno razmatrati potrošnju električne energije zbog ograničenog napajanja, veličinu ekrana, manju količinu raspoložive memorije i procesorske snage. Konstantno poboljšanje tehničkih karakteristika uređaja omogućuje i sve veće mogućnosti mobilnih operacijskih sustava, te se otvara prostor za sve veći broj primjena u područjima za koja oni nisu temeljno namijenjeni te se ranije za njih nisu ni koristili.

Jedan od temelja popularizacije svakako je i velika količina dostupnih aplikacija za mobilne uređaje, te mogućnost razvijanja i prodaje korisnicima. S obzirom na razvijenost mobilnih operacijskih sustava, aplikacije mogu poprimati visoke razine kompleksnosti, te pružaju mogućnosti usporedive sa stolnim računalima. Aplikacije mogu na jednostavan način iskorištavati i posebne mogućnosti koje pružaju mobilni uređaji, kao što su korištenje mobilnog pristupa internetu, spajanje na bežične lokalne mreže, korištenje dodatnih senzora (kamera, GPS (engl. global positioning system), akcelerometar, žiroskop, mikrofon).

Mobilni uređaji mogu se povezati s drugim uređajima koristeći nekoliko metoda komunikacije kao što su IrDA (engl. infrared data association), Bluetooth, SMS (engl. short message service), NFC (engl. near field communication) ili Ethernetom. Povezivanje mobilnih uređaja na računalne mreže pruža sve mogućnosti komunikacije među uređajima, kao da se radi o spajanju s klasičnim računalom. Spajanje na računalnu mrežu neovisno je o mobilnom operacijskom sustavu i uređaju s kojeg se spaja. Pri uključenju u neku lokalnu mrežu otvara se cijeli spektar mogućnosti slanja ili primanja podataka s drugih računala i uređaja spojenih na mrežu. Jedan od mogućih aspekata povezivanja mobilnog uređaja u računalnu mrežu je upravljanje udaljenim računalom putem mobilnog uređaja i/ili upravljanje mobilnog uređaja putem udaljenog računala.

Komunikacija između mobilnog uređaja i udaljenog računala odvija se putem aplikacija koje se razvijaju u nekoj komunikacijskoj arhitekturi, primjerice klijent - server arhitekturi ili komunikacija putem oblaka (engl. cloud). Putem aplikacije mobilni uređaj može primati podatke koje šalje aplikacija na udaljenom računalu, te može slati podatke aplikaciji na udaljenom računalu. Prema primljenim podacima, bilo na mobilnom uređaju ili računalom programiraju se određene akcije koje uređaj mora izvršiti. Tako je moguće upravljati udaljenim računalom putem mobilnog uređaja u svrhu jednostavnijeg i slobodnijeg rada s jednostavnim zadaćama na računalu.

Cilj ovog rada je prikaz povezivosti računala i mobilnog uređaja te kontrola računala. Cilj je racionalizirati upravljanje računalom tako da mobilni uređaj emulira rad računalnog miša gestama, odnosno pomicanjem mobilnog uređaja. Rad je koncipiran tako da su prvo prezentirani slični radovi iz područja upravljanja udaljenog računala s mobilnim uređajem. Zatim je prikazana metodologija izrade ovog projekta te objašnjeni razlozi za korištenje pojedinih elemenata. U posljednjem dijelu dan je prikaz razvijenih aplikacija i opisan je njihov rad. Potom se daje pregled mogućih primjena ovakvog koncepta upravljanja.

II. KRATKI PREGLED ZNAČAJNIH SLIČNIH RADOVA

Raniji radovi na području upravljanja udaljenim uređajima putem mobilnog uređaja koristeći M2M (engl. machine to machine) komunikaciju većinom su se bazirali na upravljanju GSM mrežom (engl. global system for mobile communications) mreže. Koristi se zaseban kontroler koji prima naredbe bilo pozivom (glasovno, automat upravljanje ili DTMF (engl. dual-tone multi-frequency)) ili SMS-om. Ovakve primjene najčešće se nalaze u upravljanju kućnom automatizacijom [1], [2]. Putem GSM-a i SMS-a mobilni uređaj se koristi i za reguliranje potrošnje električne energije [3]. Za upravljanje kućnom automatizacijom koristi se i sustav DTMF [4].

Za povezivanje s mobilnim uređajem koristi se i bežična tehnologija Bluetooth jer je većina danas aktualnih mobilnih uređaja opremljena je modulom Bluetooth. Ovaj način povezivanja najčešće se koristi kad su uređaji u neposrednoj blizini i ostvaruje se komunikacija jedan na jedan. Primjene komunikacije i upravljanja tehnologijom Bluetooth nalaze se u raznim područjima. U [5] bave se prikupljanjem medicinskih podataka i nadzora zdravlja računalom i mobilnim uređajem. U [6] se opisuju implementirani sustav kontrole pristupa u zgradi preko mobilnog uređaja. U [7] prikazuju se upravljanje mobilnim robotom.

S obzirom na činjenicu da je sve veći broj mobilnih uređaja opremljen karticama za pristup bežičnim lokalnim mrežama i činjenicu da većina modernih mobilnih uređaja ima pristup internetu, u tom se smjeru većinom razvija moderna komunikacija između računala i mobilnih uređaja. Radovi [8], [9] i [10] bave se korištenjem mobilnih uređaja spojenih bežičnom lokalnom mrežom na udaljeno računalo za kontrolu kućne automatizacije. U [11] mobilnim uređajem spojenim na bežičnu mrežu spajaju se na računalo koje upravlja robotom, prema određenim radnjama koje se mogu izvršiti povlačenjem prsta po ekranu mobilnog uređaja. Radovi vrlo slični ovom radu u užem smislu načina funkcioniranja i bave se prepoznavanjem pokreta i gesti putem akcelerometra u svrhu upravljanja kućnom automatizacijom [12] ili za emuliranje touchpada korištenog na prijenosnim računalima putem ekrana mobilnog uređaja [13].

Upravljanje mobilnim uređajem većim brojem udaljenih jedinica koje su povezane u ZigBee mrežu te preko modula spojene na bežičnu računalnu mrežu, kojom mogu komunicirati s mobilnim uređajem, bave se u radu koji analizira prikupljanje podataka s udaljenih senzora [14].

Osim znanstvenih i stručnih radova sličnu tematiku su obradili i proizvođači mobilnih aplikacija koje su objavljene i na Google Play dućanu i na Apple Appstoreu. Radi se o aplikacijama Ultimate Mouse Lite [15], Remote Mouse [16] i Mobile Mouse Pro [17].

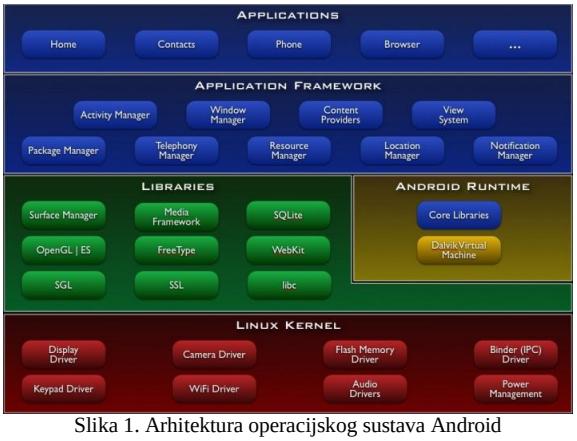
III. KONCEPT ANALITIČKOG DIJELA RADA

Kao temelj analize povezivanja i upravljanja udaljenim računalom putem mobilnog uređaja odabrani su mobilni uređaji na platformi Android. Razlog odabira prvenstveno je mogućnost razvoja i testiranja aplikacija potpuno besplatno, što na drugim mobilnim operacijskim sustavima nije moguće bez značajnijih preinaka. Osim toga, operacijski sustav Android za mobilne uređaje trenutno je najraširenij na tržištu te i dalje ima najveći godišnji rast.

Mobilni operacijski sustav Android baziran je na jezgri (kernel) Linux 2.6 i ima arhitekturu koju je moguće promatrati kao stog koji se sastoji od više razina (grafički prikaz na slici 1):

- Linux jezgra – upravljački programi i međuprocesna komunikacija
- Knjižnice pisane u C/C++ programskom jeziku
- Android Runtime – pokretanje aplikacija ima dva dijela:
    - knjižnice jezgre – core library za Javu
    - Dalvik virtualni stroj – pokreće aplikacije kao zasebne instance
- Aplikacijski sloj – application framework
    - upotreba API-ja
- Aplikacije

S obzirom na to da se razvoj za operacijski sustav Android standardno radi u programskom jeziku Java, i u ovom je radu Java odabrani programski jezik za mobilnu i za aplikaciju na računalu. Zbog otvorenog sustava jednostavno je pisati aplikacije koje koriste hardver uređaja uz pomoć besplatne inačice razvojnog paketa Android – ADK (engl. android development kit) i virtualnog stroja Android - ADT (engl. android development tools) u skladu s Eclipse IDE-om (engl. integrated development enviroment) koji je korišten pri izradi ovog rada.

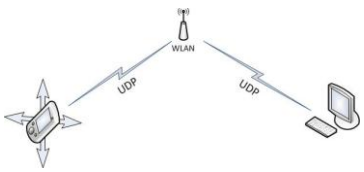


Slika 1. Arhitektura operacijskog sustava Android

Za operacijski sustav na računalu odabran je Microsoft Windows operacijski sustav koji je korišten i za razvoj aplikacija. Ovaj operacijski sustav odabran je prvenstveno zbog činjenice da se na njemu odvija razvoj aplikacija, a potom i zbog činjenice najveće raširenosti i korištenja upravo tog operacijskog sustava. S obzirom na činjenicu da se razvoj vrši u programskom jeziku Java u radu se testira funkcionalnost na više platformi, odnosno na Linux operacijskom sustavu Ubuntu distribucije te na Raspberry Pi, ARM baziranom ugradbenom računalu s Raspbian distribucijom operacijskog sustava Linux.

Za ostvarivanje komunikacije odabrana je klijent – server arhitektura aplikacije. Aplikacija na računalu predstavljat će server na koji se spaja klijentska aplikacija s mobilnog uređaja. Za komunikacijski protokol odabran je protokol UDP (engl. user datagram protocol). Konkretni protokol odabran je zbog svoje brzine i neosjetljivosti procesa na povremeni gubitak paketa. Ovaj posljednji faktor uzet je kao bitan kako bi se aplikacija mogla izvesti bez provjere dospijeća paketa. Razlog tome je pojednostavljivanje aplikacije za potrebe istraživanja i smanjivanje opterećenja na mobilni uređaj u svrhu štednje resursa kao što su memorija, baterija i procesorsko vrijeme. Prikaz arhitekture sustava vidljiv je na slici 2.

Zamišljena metoda upravljanja bazira se na emuliranju pokreta računalnog miša pokretanjem mobilnog uređaja. Dakle, nagibom udesno želi se pomicati pokazivač miša na ekranu udesno, nagibom ulijevo – pomak ulijevo. Slično je i s pomakom gore dolje po ekranu, samo se sad radi o nagibu mobilnog uređaja gore – dolje. Za ovu primjenu potrebno je koristiti očitanja s akcelerometra ugrađenog u mobilni uređaj. Premda se najčešće nalazi troosni akcelerometar, za ovu primjenu dovoljno je promatrati dvije osi (X – horizontalnu i Y – vertikalnu). Emuliranje pritiska tipki računalnog miša zamišljeno je pritiskom na ekran mobilnog uređaja. Zbog definiranja dvosmjerne komunikacije uvodi se i povratna informacija s računala pri prelasku pokazivača miša preko određenih segmenata serverske aplikacije.



Slika 2. Prikaz arhitekture sustava

IV. KRATKI PREGLED ZNAČAJNIH SLIČNIH RADOVA

Server aplikacija razvijana je u alatu Eclipse IDE metodom Model – View – Controller (MVC). Prednost metode MVC je logičko odvajanje modula aplikacije radi olakšanog daljnjeg razvoja, jer se strogo odvajaju programske cjeline.

Aplikaciju se lakše razvija i održava ako se zna da, na primjer, u View klasi postoje isključivo klase za prikaz podataka ili grafike na zaslonu.

Aplikacija Android razvijena u ovom projektu nema Model klasu, već se View klasa dijeli na datoteku XML (engl. extensible markup language) koja definira izgled i kontroler View, koji definira akcije na zaslonu i kontrolu hardvera. Pri razvoju aplikacije za Android povezuju se podaci iz datoteke XML s definiranjem objekata i dodavanjem koda koji se izvršava za svaki objekt, odnosno kontrolu.

Aplikacija na Android platformi zamišljena je i realizirana tako da putem protokola UDP šalje podatke o kutu nagiba mobilnog uređaja s ugrađenog akcelerometra. U samoj aplikaciji na mobilnom uređaju potrebno je omogućiti tipke za simuliranje lijeve i desne tipke miša, te polje za unos IP adrese računala s kojim se želi upravljati odnosno računala na kojem se nalazi serverska aplikacija. Grafičko sučelje aplikacije prikazano je slikom 5 i prikazuje sve glavne elemente od kojih se sastoji aplikacija na mobilnom uređaju.

Otvorenost sustava Android olakšava izradu aplikacija koje koriste specifičan hardver uređaja, kompas, kameru, GPS ili u ovom slučaju akcelerometar. Komunikacija s hardverom, dakle čitanje podataka s njega ili pisanje podataka na njega, svodi se na pozivanje metoda. Metode se pozivaju na objektu koji referencira na hardver, odnosno senzor s kojim se želi komunicirati. Stvaranje instance senzora vrši se putem klase SensorManager koja omogućuje pristup senzorima na mobilnom uređaju. Čitanje podataka s akcelerometra vrši se iz SensorListener objekta onSensorChanged. Ta funkcija se poziva pri promjeni podataka na bilo kojem senzoru mobilnog uređaja te se unutar nje programira čitanje podataka samo u slučaju kada podaci dolaze s akcelerometra, što je moguće jer podatak koji stiže sadrži i informaciju o tipu uređaja s kojeg je stigao.



Slika 3. Grafičko sučelje aplikacije na mobilnom uređaju s elementima

Funkcija prosljeđuje klasu event koja je tipa Sensor te sadrži podatke o tipu senzora i njegovih vrijednosti. Za potrebe ove aplikacije čitaju se dvije osi akcelerometra, X i Y, u svrhu mjerenja kuta nagiba mobilnog uređaja. Pri određivanju kuta uređaja uzima se da akceleracija od ± 1 g referencira kut od ± 90º. Izmjerena akceleracija prosljeđuje se serverskoj aplikaciji koja ne temelju te vrijednosti povećava, odnosno smanjuje vrijednost X i Y pozicije pokazivača miša na ekranu. Držeći mobilni uređaj u ruci korisnik ga naginje u lijevu ili desnu stranu te tako pomiče pokazivač miša na upravljanom računalu u lijevo ili desno. Naginjanjem mobilnog uređaja naprijed ili nazad pomiče se pokazivač miša gore odnosno dolje na ekranu upravljanog računala.

Kontroler klasa u aplikaciji je komunikacijska klasa koja se koristi za UDP povezivanje s udaljenim računalom. Kako bi se ostvarila istovremena dvosmjerna (engl. full duplex) komunikacija koristiti se dva kanala (porta). Jedan kanal koristi se za primanje podataka i jedan za slanje podataka. Metode komunikacijske klase koriste se i u Android i u serverskoj aplikaciji u vrlo malo promijenjenom obliku. Naime, one rade u „cross“ modu, odnosno metoda za slanje podataka iz aplikacije Android koristi kanal koji serverska aplikacija koristi za primanje podataka. Unutar metode za slanje podataka kreira se paket podataka te se šalje na IP adresu koja je definirana kao adresa serverske aplikacije. Paket koji se šalje formira se od podataka s X i Y osi akcelerometra, statusa pritisnutosti lijeve i desne tipke simuliranog miša, te kontrolnog paketa kojim serverska aplikacija provjerava je li primljeni podatak upravo njoj namijenjen.

Metoda za primanje podataka unutar serverske aplikacije raspakirava podatke te provjerava jesu li podaci njoj namijenjeni. Ako jesu, podaci se raspakiravaju i uz pomoć Robot klase upravlja se pokazivačem miša. Operacijski sustav ima prednost u kontroli pokazivačem ispred ulaznih jedinica te se uz pomoć metoda koje se nalaze u Robot klasi može upravljati pokazivačem miša, prosljeđivati pritiske tipki miša ili tipki na tipkovnici.

Serverska aplikacija izrađena je korištenjem alata Swing za izradu grafičkog sučelja. Taj alat koristi se kako bi se ostvarila kompatibilnost s različitim operacijskim sustavima bez potrebe za instaliranjem dodatnih knjižnica. Kako bi se ostvarila i demonstrirala povratna veza, odnosno dvosmjerna komunikacija, serverska aplikacija sadrži nekolicinu tekstualnih objekata preko kojih kad se prijeđe pokazivačem miša mobilni uređaj vibrira. Ovo je realizirano korištenjem metode Listener koja prima podatke kad pokazivač uđe u područje tekstualnog objekta. Pri tome se formira paket koji se šalje na mobilni uređaj slično kao i kad se paket šalje s mobilnog uređaja na serversku aplikaciju, s time da je podatak znatno kraći jer se samo šalje podatak o tome da se aktivira vibracija koja traje 500 milisekundi.

Pokazivač miša na upravljanom računalu pomiče se preko cijelog ekrana i može upravljati u potpunosti operacijskim sustavom na kojem se serverska aplikacija izvodi. Serverska aplikacija izvedena je kao grafička aplikacija za potrebe demonstracije te se za punu učinkovitost planira izvesti kao aplikacija koja će se izvoditi u pozadini (Windows servis ili Linux daemon).

Aplikacija je uspješno testirana na mobilnim uređajima s operacijskim sustavima Android 2.3.6 i 4.0. Povezivanja su testirana s Microsoft Windows 7 (32-bit i 64-bit), Microsoft Windows XP, Ubuntu Linux 12.10 i operacijskim sustavima Raspbian Linux. Testiranja su vršena na više različitih računala te tipova računala, kao što su stolna računala, prijenosna računala i Raspberry Pi ugradbeno računalo. Osim toga, testirano je u raznim konfiguracijama bežičnih mreža bez da su pokazani nedostaci. Trenutnu inačicu aplikacije testiralo je 30 korisnika na svojim mobilnim uređajima u kućnim i poslovnim mrežama.

V. PREGLED MOGUĆIH PRIMJENA

Po razvijenom projektu i dokazanoj valjanosti koncepta otvara se pitanje mogućih primjena ovog načina racionalizacije upravljanja udaljenim računalom. Za punu primjenu ovakvog koncepta upravljanja mobilnoj aplikaciji potrebno je nadodati mogućnost korištenja tipkovnice za prosljeđivanje tipki na udaljeno računalo i otklanjanje potrebe za tipkovnicom na udaljenom računalu.

Trenutni razvoj potrošačke elektronike i razvijenost mikrokontrolera doveli su do razvoja minijaturnih računala koja se mogu spojiti direktno na monitor ili televizor te se koristiti za prikaz multimedijalnih sadržaja ili pretraživanje interneta. Osim toga pokrenut je i razvoj „pametnih“ televizora koji već imaju ugrađeno mikro računalo te se putem samog televizora, spojenog na lokalnu mrežu, može spajati na internet i koristiti aplikacije, slično kao na „pametnim“ mobilnim uređajima, kao što su mobilni telefon ili tablet računalo. Za upravljanje tim uređajima potrebno je spojiti dodatne uređaje kao što su miš ili tipkovnica. Uz pomoć koncepta upravljanja prikazanog u ovom radu moguće je putem vlastitog mobilnog uređaja upravljati tim računalima, pretraživati internet i upravljati multimedijskim sadržajima bez potrebe za dodatnim upravljačkim uređajima. S obzirom na to da su takva računala i televizori pogonjeni operacijskim sustavom baziranom na Linuxu, serverska aplikacija Java na njima može raditi, kao što je i testirano u ovom radu. Razvoj aplikacija za „pametne“ televizore ima sličan proces kao i razvoj aplikacija za „pametne“ mobilne telefone uz razliku korištenja drugog razvojnog paketa (kao što je na primjer SmartTV SDK [18]).

U vrijeme ugrađivanja „pameti“ u veliki broj uređaja, započet je i razvoj „pametnih“ kuća. Kućna automatizacija, kao segment automatizacije u zgradarstvu, podrazumijeva upravljanje rasvjetom, grijanjem i klimatizacijom, kućnim sigurnosnim sustavima, motoriziranim roletama na prozorima i upravljanje bilo kojim uređajem u kući. Sustavi kućne automatizacije, bilo da su pogonjeni nekim namjenskim sustavom ili izvedeni uz pomoć klasičnih ili ugradbenih računala, mogu se spojiti na računalo koje će upravljati cijelim sustavom putem nekog korisničkog sučelja. Premda sustavi kućne automatizacije, naravno, imaju svoje kontrolne ploče i daljinske upravljače, to je uvođenje još jednog dodatnog upravljača za onoga koji posjeduje i koristi takav sustav. Ovakav koncept upravljanja omogućuje korisnicima da svojim mobilnim telefonom upravljaju računalom koje

prosljeđuje naredbe sustavu kućne automatizacije. Ako se isto računalo koristi i za prethodno opisano korištenje spajanja na televizor i za spajanje na sustav kućne automatizacije, tada se putem televizora i mobilnog uređaja pružaju sve mogućnosti upravljanja tim uređajima. Osim predloženog koncepta upravljanja, uz modifikacije programskog koda za čitanje podataka s akcelerometra i interpretaciju podataka na serveru, jednostavno je izvesti i prepoznavanje gesti pomicanja mobilnog uređaja (Kühnel et al., 2011). Pročitani podaci s akcelerometra prosljeđuju se na serversku aplikaciju te se tamo interpretiraju kao skup podataka i pokušava se odrediti i prepoznati uzorak odrađene geste. Prema prepoznatoj gesti serverska aplikacija može izvršavati programirane radnje.

Koncept upravljanja udaljenim računalom putem mobilnog uređaja primjenjiv je i u aspektu računalne sigurnosti. Upravljanje poslužiteljskim računalima najčešće se ograničava tako da na sama računala nisu priključeni ulazne jedinice kao tipkovnica i miš. U prostorijama s poslužiteljskim računalima obično se nalazi jedan komplet ulaznih jedinica koji se spajaju na ono računalo na kojem se radi. No i taj jedan komplet može predstavljati sigurnosni problem. Primjenom koncepta upravljanja mobilnim uređajem otklonila bi se potreba za spajanjem ulaznih jedinica te dodatno smanjila mogućnost korištenja tih računala od strane neautoriziranih osoba. Slično se može primijeniti i na ugradbena računala koja se koriste kao uređaji za prikupljanje i analiziranje podataka s raznih mjernih uređaja s obzirom na to da se i takva računala najčešće ostavljaju bez ulaznih jedinica.

VI. ZAKLJUČAK

Temelj razrade koncepta upravljanja udaljenog računala mobilnim uređajem razrađen u ovom radu temelji se na jednostavnoj arhitekturi i dokazu koncepta mogućnosti upravljanja i primanja povratnih informacija. Cilj samog rada ostvaren je i dokazan testiranjem na većem broju mobilnih uređaja (mobilnih telefona i tableta) u kombinaciji s nekoliko različitih operacijskih sustava kao što su Windows XP i 7, Ubuntu Linux i Raspbian Linux.

Za potpunu funkcionalnost i iskoristivost ovog koncepta upravljanja u mobilnoj aplikaciji bi trebalo dodati mogućnost korištenja tipkovnice. Shodno tome aplikacija na udaljenom računalu trebala bi očekivati pakete s proslijeđenim tipkama s mobilnog uređaja te ih emulirati u operacijskom sustavu na kojem radi. Tako se može ostvariti potpuna funkcionalnost racionalnijeg upravljanja udaljenim računalom i drugim sličnim uređajima putem mobilnog uređaja.

LITERATURA

[1] Yuksekkaya, B. et al. (2006) „A GSM, Internet and Speech Controlled Wireless Interactive Home Automation System“, IEEE Transactions on Consumer Electronics, 52(3), p. 837-843

[2] Alheraish, A. (2004) "Design and implementation of home automation system," IEEE Transactions on Consumer Electronics, 50(4), p. 1087 – 1092

[3] Wahab, M. H. A. et al. (2010) „GSM Based Electrical Control System for Smart Home Application“, Journal of Convergence Information Technology, 5(1), p. 33-39

[4] Hyun, H. et al. (2010) „PC application remote control via mobile phone“, In: Proceedings of the 2010 International Conference on Control Automation and Systems (ICCAS), 27-30 October, Gyeonggi-do, South Korea, p. 2290-2294

[5] Villarreal, V. et al (2011) „Monitoring Architecture to collect measurement data and medical patient control through mobile devices“, In: Proceedings of the 5th International Symposium on Ubiquitous Computing and Ambient Intelligence, 5 – 9 December, Riviera Maya, Mexico

[6] Bauer, L. et al. (2007) „Lessons Learned from the Deployment of a Smartphone-Based Access-Control System“, Institute for Software Research, Carnegie Mellon University

[7] Nasereddin, H. H. O., Abdelkarim, A. A. (2010) „Smartphone Control Robots Through Bluetooth“, International Journal of Research and Reviews in Applied Sciences, 4(4), p. 399-404

[8] Kühnel, C. et al. (2011) „I'm home: Defining and evaluating a gesture set for smart-home control“, International Journal of Human-Computer Studies, 69(11), p. 693-704

[9] Jahn, M. et al. (2010) „The Energy Aware Smart Home“, Future Information Technology (FutureTech), 2010 5th International Conference on, 21-23 May, Busan, South Korea, p.1-8,

[10] Zhao, Y. (2012) „Intelligent Household Application Base on Mobile Phone“, Applied Mechanics and Materials, 155-156, p. 464-469

[11] Song, M. et al (2010) „A Design of Real Time Robot Control System based on Android Smartphone“, The 7th International Conference on Ubiquitous Robots and Ambient Intelligence, 24-27 November, Bexco, Busan, South Korea

[12] Torunski, E. et al. (2011) „Gesture recognition on a mobile device for remote event generation“, Multimedia and Expo (ICME), 2011 IEEE International Conference on, 11-15 July, Barcelona, Spain, p. 1-6

[13] Peers, N. (2012) „Mobile Mouse: Control your PC with smartphone or tablet“, http://betanews.com/2012/02/07/mobile-mouse-control-your-pc-with-smartphone-or-tablet/ (13. 12. 2012)

[14] Lian, K., Hsiao, S., Sung, W. (2013) „Intelligent multi-sensor control system based on innovative technology integration via ZigBee and Wi-Fi networks“, In: Journal of Network and Computer Applications, 36(2), p. 756-767

[15] http://negusoft.com/index.php/ultimate-control/78-ultimate-control/83-ultimate-mouse (21. 3. 2013)

[16] http://www.remotemouse.net/ (22. 3. 2013)

[17] http://www.mobilemouse.com/ (25. 3. 2013)

[18] developer.samsung.com (15. 12. 2012)

Pametna komunikacija na Internetu preko REST protokola

Davor Lozić, dr. sc. Alen Šimec
Tehničko veleučilište u Zagrebu, Vrbik 8, 10000 Zagreb, Hrvatska,
dlozic@tvz.hr, alen@tvz.hr

Abstract - REST protokol je akronim za *Representational state transfer* protokol koji omogućuje komunikaciju na Internetu pomoću svih najpoznatijih programskih jezika kao što su C/C++, C#, Jave, Ruby, PHP i sličnih kao i preko najpoznatijih operacijskih sustava poput Microsoft Windowsa, Mac OSa, Linuxa itd. Zahvaljujući odličnom radu nad velikim brojem klijenata (engl. *scalability*), mogućnosti primjene ovog protokola na mnoge programerske probleme kao što su dohvaćanje podataka na serveru, dodavanje, promjena ili brisanje postojećih podatka te neovisnosti o tehnologiji, razlog je zašto baš ovaj protokol koriste mnoge velike tvrtke kao što su Google (GoogleMaps[1], GoogleSearch[2]), Microsoft (Azure[3], Shrepoint[4]) za komunikaciju između tehnologija na Internetu. Mogućnosti biranja tehnologije odgovora (JSON, XML, HTML itd.), mogućnost slanja filtera podataka direktno serveru i kreiranje odgovora u ovisnosti o poslanim HTTP metodama je ono što čini ovaj protokol naprednijim i boljim od bilo koje druge alternative[5].

I. UVOD

Jedan od najvećih problema Interneta danas je komunikacija među različitim tehnologijama koje se mijenjaju svakodnevno i čiji su standardni masivni i dosta često nedorečeni. Rješenja poput SOAP protokola su često puno teže implementirana i zahtijevaju puno više resursa. [6] Nadalje, rješenja koja su do sada pokazala veliku brzinu su većinom razne binarne standardizacije gdje ne dolazi do parsiranja između teksta i podataka. Takva standardizacija je puno teža za implementaciju od klasičnih rješenja, traženje pogreške i preglednost odgovora (*engl. human readable*) [7]. Trenutno najbolje rješenje je REST protokol koji omogućuje jednostavnu implementaciju, komunikaciju između tehnologija, kao i jedan od najbitnijih aspekata Interneta, a to je sigurnost[8].

II.KORIŠTENJE POSTOJEĆIH PROTOKOLA

Za rad REST protokola nije trebalo izraditi potpuno nove protokole za komunikaciju nego su iskorišteni postojeći. Jedan od iskorištenih protokola je poznati HTTP protokol. Roy Fielding, osoba koja je sudjelovaja u izradi HTTP protokola je napravila i REST protokol.[9]

II.I. HTTP PROTOKOL

HTTP protokol [10] je aplikacijski protokol na čijem se radu zasniva cijeli WWW (*engl. World Wide Web*), točnije na verziji 1.1. Srž rada ovog protokola se zasniva na stvaranju upita i odgovora. Svaki upit i odgovor ima HTTP zaglavlja. Zaglavlje se sastoji od više elemenata među kojima su najpoznatiji: Request URL, Request Method, Status Code, Accept, Host. [11]

Request URL – puna adresa gdje se šalje upit.

Request Method – jedna od HTTP metoda (najčešće GET ili POST).

Status Code – HTTP status kod (najpoznatiji 404 Not Found). [12]

Accept – što server ili klijent prihvaća kao odgovor (json, pdf, sliku itd.).

Host – naziv računala i TCP port na kojem računalo "sluša".

II.II. XML STANDARD

XML (engl. Extensible Markup Language) je jezik za označavanje podataka. Ideja je bila stvoriti jezik koji će biti jednostavno čitljiv i ljudima i računalima.

Jedna od najvećih prednosti XMLa je jednostavnost [13]. Podaci se odvajaju tzv. tagovima ili oznakama (otvorenim i zatvorenim). Oznake mogu imati atribute koji dodatno opisuju podatke.

```xml
<?xml version="1.0" ?>
<Osoba spol="M">
    <Ime>Davor</Ime>
    <Prezime>Lozić</Prezime>
</Osoba>
```

Slika 1 - primjer XML dokumenta

II.III. JSON

JSON (engl. Javascript Object Notation) je otvoreni standard za prikaz podataka kao alternativa XML standardu [14]. Prednost JSONa nad XMLom je što JSON zauzima puno manje prostora tako da je JSON pogodniji za prijenos većih količina podataka, a XML za konfiguracijske datoteke ili kada je potrebna XSLT tehnologija.

```json
{
    "Osoba": {
        "Ime": "Davor",
        "Prezime": "Lozić",
        "Spol": "M"
    }
}
```

Slika 2 - primjer JSON dokumenta

III.RJEŠENJA PRIJE REST PROTOKOLA

III.I. SOAP PROTOKOL

SOAP protokol je baziran na XMLu [15] što ga čini jednostavnim za čitanje (engl. *human readable*). Budući da su poruke koje se šalju između pošiljatelja (klijenta) i primatelja (poslužitelja) zapravo čisti i relativno razumljivi tekst, jednostavno ih je pratiti, ispravljati pogreške i razumjeti što se događa.

Još jedna velika prednost SOAP protokola je mogućnost primjene na različita područja poslovne inteligencije (engl. *extensibility*) [16]. Stvar je implementacije što će određena funkcija raditi, SOAP poruke služe samo za poziv i slanje rezultata obrade.

SOAP je neutralan (engl. *neutrality*) [17]. Jezici kao što su C/C++, C#, Python, PHP itd. imaju biblioteke koje rade s ovim protokolom te nije ograničen niti na operacijske sustave. Ono što je bitno je da se ti jezici drže pravila i standarda te će te poruke moći razumjeti bilo koji operacijski sustavi s bilo kojim programskim jezikom. SOAP protkol koristi najčešće HTTP(S), ali je kompatibilan i s asinkronim protokolima kao što su FTP ili SMTP [18].

```xml
<env:Envelope
  xmlns:env="http://www.w3.org/2003/05/soap-envelope">
  ...
</env:Envelope>
```

Slika 3 - SOAP omotnica

Protokol, kao glavni element, sadrži omotnicu (engl. *envelope*). Unutar omotnice postoji zaglavlje (engl. *header*) koje je opcionalno. Zaglavlje može sadržavati vrijeme, api ključeve (engl. *api keys*), podatke za pristup (engl. *login credentials*) i slično [19].

Unutar protokola mora postojati tijelo omotnice (engl. *body*). Unutar tijela omotnice se najčešće nalazi ime funkcije koje želimo pozvati, zajedno s parametrima funkcije. Opis funkcija, parametara, tipova podataka i slično se nalazi unutar WSDL dokumenta (Web Service Definition Language).

```xml
<env:Body xmlns:m="http://www.primjer.com">
  <m:DohvatiKorisnika
    <m:IdKorisnika>5</m:IdKorisnika>
  </m:DohvatiKorisnika>
</env:Body>
```

Slika 4 - SOAP primjer poziva

Jedan od nedostataka SOAP protokola je relativna kompliciranost, posebno kod implementacije sigurnosti. Specifikacije ne nalažu puno stvari vezane uz sigurnost pa se većinom koriste i ne toliko standardni načini sigurne komunikacije [20].

Drugi nedostatak je što je SOAP relativno velik dokument i za male podatke [20], ali je to više problem XML jezika nego samog SOAP protokola što je prikazano slikom (Slika 5), gdje za običan odgovor s imenom i prezimenom zapravo dobijemo relativno puno teksta.

```xml
<soap:Body xmlns:m="http://www.primjer.com">
  <m:DohvatiKorisnikaOdgovor>
    <m:Korisnik>
        <m:Ime>Davor</m:Ime>
        <m:Prezime>Lozić</m:Prezime>
    </m:Korisnik>
  </m:DohvatiKorisnikaOdgovor>
</soap:Body>
```

Slika 5 - SOAP primjer odgovora

IV. REST PROTOKOL

Trenutno najbolje rješenje za komunikaciju između tehnologija je REST (Representational

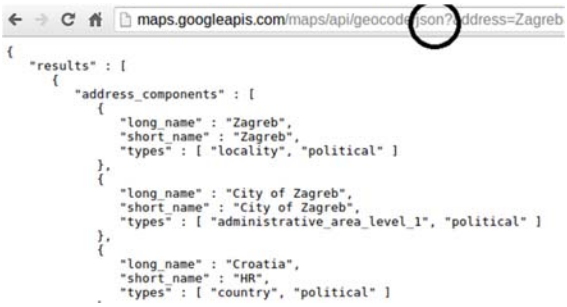
state transfer) protokol. REST je preuzeo dobre stvari iz SOAP protokola kao što su neutralnost tj. mogućnost programskih jezika kao što su C/C++, C#, Java i slični da komuniciraju ili neovisnost o operacijskom sustavu (MS Windows obitelj, Linux itd.). Također je preuzeta i mogućnost primjene protokola na bilo koji aspekt poslovne inteligencije.

IV.I. PREDNOSTI NAD SOAPOM

REST protokol može koristiti više različitih odgovora (XML, JSON itd.). REST servisi daju dobra privremena skladišta (engl. *cache*) pri korištenju GET poziva [21] što pokazuju Slika 6 i Slika 7.



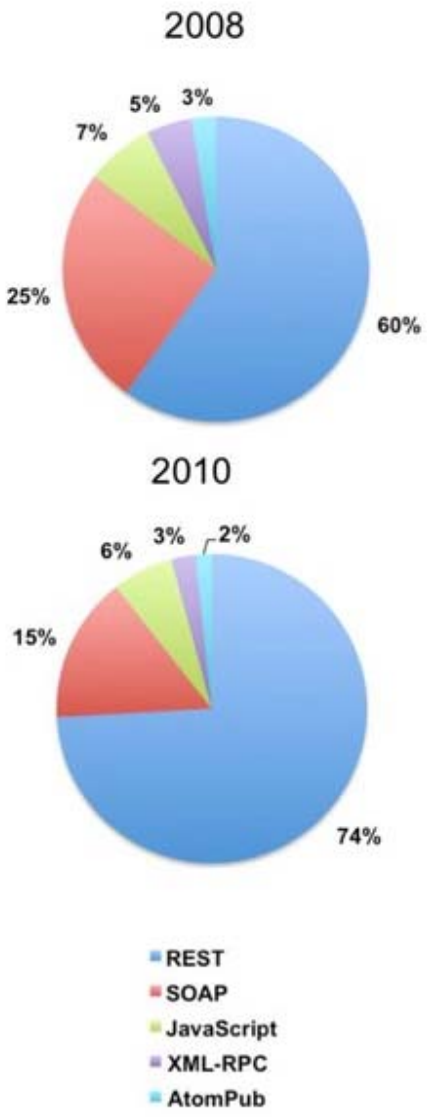
Slika 6 - REST preko XMLa



Slika 7 - REST preko JSONa

IV.II. PODRŠKA REST PROTOKOLA

Jedan od najvećih zagovornika REST protokola je Google. Servisi kao što su Google Translate API [22], Google Earth API [23], Google Maps API [24] itd. Microsoft također koristi REST servise kao što Locations API [25], Traffic API [26] i slični te se tako ti servisi mogu koristiti i u privatnim aplikacijama.



Slika 8 - Sve veći prelazak servisa na REST protokol [28]

IV.III. REST I HTTP

REST protokol je usko povezan s HTTP protokolom, točnije, građen je na njemu. REST, kao i HTTP, može slati GET, POST, PUT i DELETE metode prema serveru. Već se tu može vidjeti povećana sigurnost jer, ako je GET samo dohvaćanje podataka, onda klijent ne može tom REST porukom mijenjati nikakve podatke.

Tablica 1 - REST metode

Metoda	Značenje
GET	Dohvaćanje podataka
POST	Dodavanje podataka
PUT ili PATCH	Promjena podataka
DELETE	Brisanje podataka

REST sintaksa, pravila i ostali dijelovi standarda velikim dijelom ovise o servisu. Najčešći standard je proslijeđivanje parametara kroz URL serveru kao što to koristi Google Maps API [27].

Google Maps url:

http://maps.googleapis.com/maps/api/staticmap?center=Zagreb&zoom=13&size=300x300&sensor=false

Možemo putem URLa proslijediti Google Mapsu koju sliku želimo s mape. Atributi poput center, zoom, size itd. su samo neki koje možemo specificirati. Svaki servis ima svoja pravila i načine korištenja.

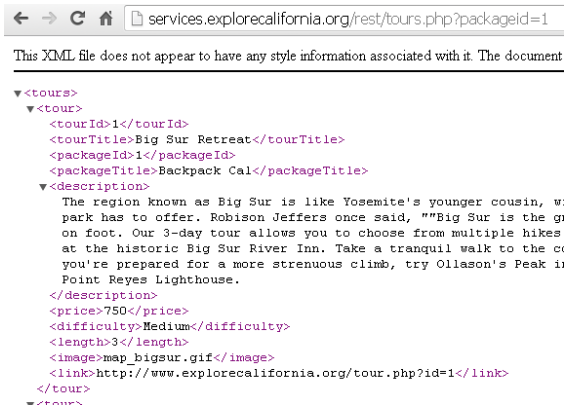


Slika 9 - Google Maps Zagreb

Podatke preko IDa možemo zatražiti također preko URLa.

http://services.explorecalifornia.org/rest/tours.php?packageid=1

Kao rezultat dobijemo XML:

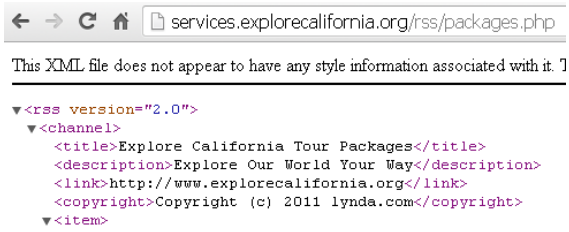


Slika 10 - dohvaćanje preko IDa

Još jedan od formata podataka je RSS.

http://services.explorecalifornia.org/rss/packages.php

Kao rezultat imamo RSS:



Slika 11 - dohvaćanje RSSa

IV.IV. SIGURNOST REST PROTOKOLA

Podaci se mogu slati preko zaštićene HTTP verzije koja ide preko SSL/TLS protokola, tzv. HTTPS protokola. Drugi način je unos podataka za pristup u sam URL:

https://username:password@www.domena.com/webservis

Korisničko ime i lozinka se mogu kodirati pomoću base64 algoritma. Još jedan način je imati neki znakovni niz od više znakova koje je teško pogoditi i poslati ga na server. Server usporedi sa listom znakova koju već ima i ako u bazi postoji taj znakovni niz, dopusti pristup samom servisu.

V.ZAKLJUČAK

Zbog svoje jednostavne implementacije na različite probleme, radom nad HTTP prokolom, interoperabilnosti te primjenjivosti na sve poznate operacijske sustave i programske jezike je razlog zašto je REST protokol danas najpoznatiji način za među-servisnu komunikaciju na Internetu. Lagana primjena GET, POST, PUT i DELETE metoda čini ovaj protokol jednostavnim za implementaciju mnogih rješenja. Sama mogućnost biranja izlaza

(XML, JSON, HTML, RSS) kao i relativno visoka sigurnost je uvelike utjecala na ogromni skok ove tehnologije u komunikaciji na Internetu.

VI.LITERATURA

1.] GoogleMaps API dokumentacija, https://developers.google.com/maps/documentation/webservices/

2.] GoogleSearch REST API, https://developers.google.com/custom-search/json-api/v1/overview

3.] Windows Azure REST API, http://msdn.microsoft.com/en-us/library/windowsazure/hh780717.aspx

4.] Sharepoint 2013 REST servis, http://msdn.microsoft.com/en-us/library/office/fp142380.aspx

5.] Mark Damon Hughes, SOAP or REST for Web Services?, http://stackoverflow.com/questions/76595/soap-or-rest-for-web-services

6.] Tarandeep Singh, 24. 08. 2009., REST vs. SOAP – The Right WebService, http://geeknizer.com/rest-vs-soap-using-http-choosing-the-right-webservice-protocol/

7.] Eric J., Are binary protocols dead?, http://stackoverflow.com/questions/2525188/are-binary-protocols-dead

8.] Beale M, 17. 04. 2013., Secure Your REST API... The Right Way, http://www.stormpath.com/blog/secure-your-rest-api-right-way

9.] Roy T. Fielding, http://roy.gbiv.com/

10.] HTTP specifikacije, http://www.w3.org/Protocols/Specs.html

11.] HTTP zaglavlja, http://www.w3.org/Protocols/rfc2616/rfc2616-sec14.html

12.] HTTP code 404 Not Found, http://www.checkupdown.com/status/E404.html

13.] Why XML, Simon St. Laurent, http://www.simonstl.com/articles/whyxml.htm

14.] JSON, The Fat-Free Alternative to XML, http://www.json.org/xml.html

15.] SOAP specifikacije, http://www.w3.org/TR/soap

16.] SOAP extensibility, http://msdn.microsoft.com/en-us/library/ms995800.aspx#understandsoap_topic4

17.] Web services overview, http://www.visualdataflex.com/features.asp?pageid=797

18.] Kyle Brown, SOAP over FTP, how? http://www.coderanch.com/t/219571/Web-Services/java/SOAP-FTP

19.] SOAP zaglavlje, http://www.w3schools.com/webservices/ws_soap_header.asp

20.] Web Services, REST vs. SOAP, https://blogs.oracle.com/milan/entry/web_services_rest_vs_soap

21.] Swati Dhingra, REST vs. SOAP: How to choose the best Web service, http://searchsoa.techtarget.com/tip/REST-vs-SOAP-How-to-choose-the-best-Web-service

22.] Google Translate REST API, https://developers.google.com/translate/v2/using_rest

23.] Google Earth REST API, https://developers.google.com/earth/documentation/index

24.] Google Maps API, https://developers.google.com/maps/documentation/webservices

25.] Locations API, http://msdn.microsoft.com/en-us/library/ff701715.aspx

26.] Traffic API, http://msdn.microsoft.com/en-us/library/hh441725.aspx

27] Google Maps API, https://developers.google.com/maps/

28] REST in peace, SOAP, http://farm5.static.flickr.com/4092/5083554497_bdf0f485af_o.png

ISS

International Conference on

INFORMATION SYSTEMS SECURITY

Steering Committee

Chairs: Stjepan Groš, University of Zagreb, Croatia
Tonimir Kišasondi, University of Zagreb, Varaždin, Croatia
Željko Hutinski, University of Zagreb, Varaždin, Croatia

PAPERS

Password security – no change in 35 years?

Viktor Taneski*, Marjan Heričko* and Boštjan Brumen*
* University of Maribor, Faculty of electrical engineering and computer science/Institute of Informatics, Maribor, Slovenia
viktor.taneski@uni-mb.si, marjan.heričko@uni-mb.si, bostjan.brumen@uni-mb.si

Abstract - BACKGROUND: Textual passwords were first identified as a weak point in information system's security by Morris and Thompson in 1979. They found that 86% of the passwords were weak: being too short, containing lowercase letters only, digits only or a combination of the two, being easily found in dictionaries. OBJECTIVE: Despite the importance of passwords as the first line of defense in most information systems, little attention has been given to the characteristics of their actual use. Thus, the objective of this paper is to identify any problems that may arise in creating and using textual passwords. METHOD: A systematic literature review of studies in the area of password use and password security. Our research is restricted to articles in journals and conference papers written in English and published between 1979 and 2014. The search is conducted through IEEEXplore, ScienceDirect, Springer Link and ACM Digital Library. RESULTS: The computer community has not made a very much-needed shift in password management for more than 35 years. Users and their passwords are still considered the main weakness in any password system, because users often choose easily guessable passwords: words, names, birthdates, etc., because they are easy to remember. CONCLUSION: Password policies and password checkers can help users create strong and easy-to-remember passwords. This work will serve as a starting point for our further research in this area where we want to determine whether these password policies are useful to the users, and whether the users can easily apply them.

I. INTRODUCTION

Typically, a user is authenticated based on one of the three underlying principles: "what you know" (e.g. textual or graphical passwords), "what you are" (e.g. retina, iris, voice and fingerprint scans) and "what you have" (e.g. smart cards or other tokens). The passwords that fall under the category of "what you know" can be divided into textual passwords and graphical passwords [24][29]–[31][45]. In this study, we focus particularly on the textual passwords and their security simply because they are still the most common method for authentication.

More than 30 years ago, Morris and Thompson [3] identified textual passwords as a weak point in information system's security. They found out that majority of users' passwords (87% of them) were short, contained only lowercase letters or digits or were easily found in dictionaries. Because the use of passwords has always been, and continues to be, one of the most common control mechanisms for authenticating users of computerized information systems, it was expected that Morris and Thompson would have raised the awareness of system users about the consequences of how they choose their passwords. However, despite the widespread use of passwords and their importance as the first line of defense in most information systems, little attention has been given to the characteristics of their actual use.

The objective of this paper is to perform a systematic literature review of studies in the area of textual passwords and textual passwords security. We conduct our review following the guidelines proposed in [1]. The rationale for conducting the review is threefold. The first reason is to identify any problems that may arise in creating and using textual passwords. Secondly, to appraise the current situation of passwords with respect to the passwords strength, password management and password memorability. Thirdly, to find out what is the relationship between users and the textual passwords they use (are the users still "the weakest link?").

The reminder of this paper is organized as follows: the research method is described in Section II, while Section III represent the review results and the answers to our research questions. Section IV concludes the paper and presents plans for future work.

II. RESEARCH METHOD

A. Research Questions

We use the PICOC (Population, Intervention, Comparison, Outcomes, Context) [2] structure for formulating the research questions. Table 1 represents the selected criteria for the PICOC structure. Considering the three parts in Table 1, we formulate the following research questions:

- **RQ1**: What are the major problems with creating and managing textual passwords?
- **RQ2**: What is the current situation of textual passwords with respect to password strength, password management and password memorability?
- **RQ3**: What is the relationship between users and textual passwords? Are the users still "the weakest link"?

Our population represents the textual passwords, while the intervention are the different techniques and technologies used for creating and manipulating textual passwords. We do not compare textual passwords with other types of passwords, since our subject of interest are strictly textual passwords. The outcomes refer to the

TABLE I. THE PICOC STRUCTURE

Population	Textual passwords
Intervention	Techniques and technologies for creating and managing textual passwords
Comparison	/
Outcome	Problems when creating and managing textual passwords
Context	/

problems that may arise in creating and using (or manipulating) textual passwords.

B. Search Strategy

Our literature search strategy is organized in two phases*: initial search* and *reference search*.

The initial search for relevant studies is performed over electronic data sources – digital libraries. When selecting the digital libraries, we took into account the recommendations from the literature [1] and also our knowledge and practical experience and the fact that we do not have access to all the digital databases. This initial search is conducted through the digital databases IEEEXplore, ScienceDirect, SpringerLink and ACM Digital Library.

We composed a search string by deriving major search terms from our research questions and by using Boolean OR to link the major search terms:

("password security" OR "password strength" OR "password memorability" OR "password cracking" OR "password management")

The search of the electronic databases is restricted to articles in journals, conference papers and book chapters written in English and published since 1979, i.e. this is the year when the first article in the area of password security was published.

The reference search is conducted by reviewing the reference lists of primary studies and the review articles addressing the password security [3], [4] and [5].

C. Selection Criteria

The search is performed by using the search string and the search result is a set of documents in which the search string appears partially or entirely. However, in some cases, the search string outputs a study whose topic is in a different field of study. Because of such cases, we also performed a subsequent semantic checking by reading the titles and abstracts of these documents and selecting the documents that are consistent with our research area. This selection is carried out in an objective manner, using inclusion and exclusion criteria. In order to select the relevant documents, we define the following inclusion and exclusion criteria:

Inclusion – a study focuses on textual passwords only ("what you know"):

- the study focuses on password security or presents method(s) for password creation
- the study deals with issues or problems with password use, password management or password memorability

Exclusion – a study is not peer-reviewed, or the study deals with computer security in general, cryptography, graphical passwords or any other type of user authentication (biometrics, tokes or smart cards etc.).

Because the titles and abstracts of the documents are not always extensive, we do not always have a clear indication if the document meets the specified criteria. If this is the case, we take a further step and read the whole document to determine whether the data meet the inclusion and exclusion criteria.

III. RESULTS

Table 2 lists the results from the initial search. The search into 4 electronic databases found **58** relevant primary studies. The third column in the table represents the number of studies that are duplicates, i.e. studies that have already occurred in another digital database. We can see from the table that the ACM Digital Library contributed the largest number of documents (42) and also the largest number of duplicates (20). The column "Total relevant studies" shows the total number of documents from each source, but in this case documents that have already appeared in the a previous source are not included. The review of the references of the primary studies found additional 9 documents that were not found by the initial search of the electronic databases, thereby increasing the number of total relevant documents to **67.**

Next, we present the results of our review and the answers to the research questions.

RQ1*: What are the major problems with creating and managing textual passwords?*

We observe that 17 ([3]-[18][68]) out of 67 studies, address the issue of password security, password use or re-use, and are suitable for answering the first research question.

As already mentioned earlier, Morris and Thompson [3] were the first authors that addressed the issue of password security in 1979. They observed problems regarding to the accessibility and availability of the password file in the UNIX system. They also conducted experiments in order to determine typical users' habits about the choice of passwords and noticed that users of the system chose passwords that are **short**, **simple** and from a **restricted character set** (e.g. alphanumeric password with all lower-case letters). Ten years later authors in [4] concluded that, in order to maximize the difficulty of password cracking and to prevent the fast, simple attacks, systems should implement passwords policies that will require the passwords to contain certain amount of entropy.

On the other hand, higher entropy makes a password more difficult for a user to memorize [13]. Further research has shown that despite password policies' recommendations and the efforts by information system professionals to educate users about secure password policies, users still tend to choose passwords that are

TABLE II. SUMMARY OF FOUND AND SELECTED STUDIES

Electronic database	Found studies	Relevant studies	Duplicates found elsewhere	Total relevant studies
IEEEXplore	89	15	0	**15**
ScienceDirect	76	6	0	**6**
SpringerLink	847	15	0	**15**
ACM Digital Library	1203	42	20	**22**
Total	**2215**	**78**	**20**	**58**

simple and easy to remember [5]. This particular study [5] showed that user-created passwords are still based on user's personal data or a combination of meaningful details. Adams and Sasse [7] conducted a study of password related user behaviors, including password construction, frequency of use, password recall and work practices. They concluded that their participants lacked security motivation and understanding of password policies, and tended to circumvent password restrictions for the sake of convenience.

Because of the rapid growth of the popularity of Internet technology and e-commerce and the increased number of online services requiring password based authentication, users have to maintain many different accounts [15]. Different passwords should be used for every account so that no single password cracking will lead to compromising the other accounts [10], thus users have to remember **multiple passwords**.

In order to provide better password security i.e. to ensure that a stolen password will become unusable more or less quickly, some password security mechanisms and password policies include forced password change (**password aging**). This results with the user having many different accounts to maintain and many different passwords to remember that are changing frequently because of password security mechanisms. Password aging is a very serious problem, because of the lack of warning users receive before their password is about to expire. If the system insist on users changing their password during the login procedure, denies them the time to think about choosing a good one [6]. Forcing password change too frequently may make the user to quickly **forget** the password [5][13], or to **write it down** [5][8][9][14], which is another "bad habit" noted in the literature regarding to password security.

One of the most common vulnerability, which is also most frequently cited in the literature, associated with password security is the **password re-use** (using the same password, or a very similar one, for more than one secure item) [9][10][11][12][14][15][16][17][18]. The results of studies [9][17] show that duplicating passwords is a very common practice. In such an environment, where one password is used for authentication for several different accounts (an average user has about 25 different password accounts [12]), password re-use can cause serious damage, if the password is successfully cracked, for a single (not-so-important) account. Information may be revealed that will aid the hackers in infiltrating other systems, potentially leading to the fall of many other systems (including the ones that are far more secure that the first one) [10]. Users must lower their overall password quality, in order to cope with the password overload [68]. According to [11], duplicating passwords is more likely to become more problematic in the future as password reuse rates are increasing because people are accumulating more accounts but are not creating more passwords.

***RQ2**: What is the current situation of textual passwords with respect to password strength, password management and password memorability?*

From all collected documents, we can observe that in the period of 35 years different methods and approaches for creating and easily memorizing and managing textual passwords were developed ([19]-[58][64]-[67][69] 45 out of 67 studies).

The earliest approach that deals with the problem of creating passwords that are memorable and difficult to guess, is the use of **cognitive passwords**. Cognitive passwords represent a dialog between a user and a system. The dialog is a actually a rotating set of randomly chosen questions out of a set of predefined questions, previously selected by the user [20]. Cognitive passwords are more readily recalled than conventional passwords and are also more difficult to guess [20][21][23]. The only possible issues are the issue of time needed for authentication of user access and that remembering many cognitive passwords may be harder for the user [20][21].

Similar to cognitive passwords are **associative passwords**. This alternative is based on a single-word cue and a one-word response, instead of whole questions and answers [21]. Just like with cognitive passwords, the only possible problem is that a user would likely not remember many associative passwords [21][54].

Passphrases and **mnemonic-based passwords** are another alternative for conventional passwords. A passphrase is basically an extended password, which consists of a meaningful sequence of words. It is designed to form a compromise between ease of memorability and difficulty of guessing it [21]. It turned out that passphrases are no harder to remember and are more resistant to brute-force attacks than conventional passwords [35][21]. But [35] also found out that passphrase users initially experienced a significantly higher number of login failures due to typographical errors, which can be one downside of passphrases. Mnemonic-based passwords present an alternative for passphrases and are also one of the earliest approaches [19]. They are basically a short passphrase where a character (often the first letter) is used to represent each word of the passphrase (e.g. "I love to ski at Seven Springs!" transforms into "Ilts@7S!") [32]. They are difficult to guess, but are also almost equally safe as the random passwords, and equally easy to remember as the naively selected passwords [27][43]. The only downside of mnemonic-based passwords is that users often select phrases from music lyrics, movies, literature, or television shows, which are often available on the internet. This opens the possibility that a dictionary of mnemonic passwords can be built, so authors in [32] conclude that mnemonic passwords should not be treated

as a solution for the problems mentioned earlier in this section.

We mentioned **password policies** earlier as a mechanism for encouraging users to create and use strong and easy to crack passwords. But the literature shows that users do not comply with such recommendations and policies [28][33][47][51][53]. This usually happens because password policies are too inflexible to match users' capabilities, and the tasks and contexts in which they operate [47][66]. So, other studies conclude that password policies should be adjusted for different focus groups and should be more aware of the increasing burden users have managing passwords [39][57][67]. If we start to enforce stricter password policies without a thorough explanation and user training, this has almost no effect [57][64]. This can create a conflict between users' perception and the enforced practice – we can only ran into some of the already mentioned problems regarding to password security [47][52].

Some authors even propose that administrators should pro-actively attempt to crack passwords in their systems, in order to estimate the risk of password-guessing attacks, such as [69]. Password breaking mechanisms may also be used for data recovery purposes, as suggested in [69].

A proactive **password checker** is one possible solution to prevent users from creating easily guessed passwords and to also encourage them to adhere to password policies. Password checkers are proposed in some of the studies we analyze, including [22][25][26] [44][65]. But password security is not a problem that can be solved only by technical means. Human factors are also very important (which is basically the problem with password policies) [25]. A further research is needed in order to find out how to create password policies that are acceptable to users and that can create secure passwords.

A few studies propose a technique of **"persuading"** users to choose better passwords, rather than to instruct them or impose new password schemes, because they believe that improved instructions and new password schemes may not be the solution [36][40].

Passfaces and **graphical passwords** are not originally our subject of interest but, since they are potential substitute for textual passwords, we decided to include only those studies where the security of graphical passwords is discussed. Studies [24][29]–[31][45] show that passfaces and graphical passwords actually provide a better security than textual passwords. The only downside with graphical password is that the password registration and log-in process take very long and the authentication procedure is different than that using textual passwords. This requires redesign of the existing software.

It seems that users are more aware about password security and the importance of creating strong and hard to guess passwords [48]. The results indicate that, according to the amount of characters, passwords became little more secure over time (a study in Greece shows that the average password length is slightly less than 7 characters) [49][56]. Despite that, users still tend to use less secure passwords (which are composed only of lowercase letters, uppercase letters or numbers), to write their passwords down or to share them with their friends and they also rarely change their passwords [34][38][46]. Password requirements may have something to do with this, since [42] found out that the majority of the websites that have passwords requirements, do not keep pace with the progress in the area of password security and rarely change od update their passwords requirements. In the following years we can expect that the security of any system that is based on passwords will be equivalent to the availability of the material for passwords disclosure and not how random and strong passwords are [58]. Therefore, stronger passwords may not be necessary (especially for smaller institutions with hundreds rather than millions of users), as long as the security protocols are well designed (e.g. three unsuccessful logins freezes the account for a time) [37]. Users can also be encouraged to design strong passwords, using elements associated with a given service, together with a personal factor [55]. The growth of web-based services will demand memorizing even more passwords in the future, so some other usable alternatives to textual passwords may have to be found [41][56]. Due to the widespread use of the World Wide Web and the increased number of Web accounts that a user has to maintain, we still encounter with the majority of the already discussed problems [34][50].

***RQ3**: What is the relationship between users and their textual passwords? Are the users still "the weakest link"?*

The rest of the studies (5 out of 67) address the issue of the user as the weakest link in password security [59][60][61][62][63]. From the studies that we have already analyzed, we can observe that the user behavior is a very common issue in the area of password security and many security departments treat users as a security risk that needs to be controlled. Authors in [59] and [60] argue that the usability of the security mechanism has rarely been investigated. Users usually are not aware about the security threats and the importance of security and password mechanism are currently badly matched to users' capabilities and their tasks. These are the challenges arising from the lack of communication between users and organizations (or their security services), which leads to the development of useless security mechanisms [59][60]. Therefore it is maybe necessary for password mechanisms to apply a user-centric approach for designing "usable security" (i.e. human factors should be given priority over technological factors) [62][63]. On the other hand, authors in [61] claim that users are the enemies of the system (willingly or unwillingly) and security policies should be applied appropriately according to the user's type. They argue that ignorant users have to be educated about security mechanisms, and non-copliant users have to be persuaded to follow the security best practices.

According to this, we can not conclude if the users are still the enemy of a system and we aggree that a further research in this area is needed.

IV. CONLUSION

The results of our review show that the computer community has not made a very much-needed shift in password management for more than 35 years. It seems nothing has changed since Morris and Thompson

addressed the issue of password security and published their results in 1979. For example, an average user has 8.5 different Web accounts, 6.5 passwords (each of which is shared across 3.9 different websites), resulting in users to very often forget their passwords or write them down [12][34][50]. Besides that, users still tend to create passwords based on their personal characteristics. Users and their passwords are still considered the weakest link – the main weakness in any password system is that users often choose easily guessable passwords: words, names, birthdates, etc., because they are easy to remember. One way to decrease password guessability is to restrict the passwords accepted from the user by using a system (e.g., a password checker) that filters out easily guessed passwords. Password checkers require that passwords have certain characteristics (usually defined in a password policy) before they are accepted by the system when the user enters a password. Password policies and password checkers can help users create strong and easy-to-remember passwords. However, despite password policy advice, users still tend towards creating weak and easy-to-guess passwords. This work will serve as a starting point for our further research in this area where we want to determine whether these password policies are useful to the users, and whether the users can easily apply them or the policies cause them problems when creating and using passwords.

REFERENCES

[1] B. Kitchenham and S. Charters, "Guidelines fro performing Systematic Literature Reviews in Software Engineering," 2007.

[2] M. Petticrew and H. Roberts, *Systematic Reviews in the Social Sciences - A Practical Guide*, 1 edition. Wiley-Blackwell, 2006, p. 352.

[3] R. Morris and K. Thompson, "Password Security: A Case History," *Commun. ACM*, vol. 22, no. 11, pp. 594–597, Nov. 1979.

[4] D. Feldmeier and P. Karn, "UNIX Password Security - Ten Years Later," in in *Advances in Cryptology — CRYPTO' 89 Proceedings SE - 6*, vol. 435, G. Brassard, Ed. Springer New York, 1990, pp. 44–63.

[5] M. Zviran and W. J. Haga, "Password Security: An Empirical Study," *J. Manage. Inf. Syst.*, vol. 15, no. 4, pp. 161–185, 1999.

[6] M. Bishop, "Password management," *Compcon Spring '91. Digest of Papers*. pp. 167–169, 1991.

[7] A. Adams, M. A. Sasse, and P. Lunt, "Making Passwords Secure and Usable," in *Proceedings of HCI on People and Computers XII*, 1997, pp. 1–19.

[8] E. F. Gehringer, "Choosing passwords: security and human factors," *Technology and Society, 2002. (ISTAS'02). 2002 International Symposium on*. pp. 369–373, 2002.

[9] A. S. Brown, E. Bracken, S. Zoccoli, and K. Douglas, "Generating and remembering passwords," *Applied Cognitive Psychology*, vol. 18, no. 6, pp. 641–651, Sep. 2004.

[10] B. Ives, K. R. Walsh, and H. Schneider, "The Domino Effect of Password Reuse," *Commun. ACM*, vol. 47, no. 4, pp. 75–78, Apr. 2004.

[11] S. Gaw and E. W. Felten, "Password Management Strategies for Online Accounts," in *Proceedings of the Second Symposium on Usable Privacy and Security*, 2006, pp. 44–55.

[12] D. Florencio and C. Herley, "A Large-scale Study of Web Password Habits," in *Proceedings of the 16th International Conference on World Wide Web*, 2007, pp. 657–666.

[13] P. Cisar and S. M. Cisar, "Password – a Form of Authentication." 2007.

[14] P. Hoonakker, N. Bornoe, and P. Carayon, "Password Authentication from a Human Factors Perspective : Results of a Survey among End-Users," pp. 459–463, 2009.

[15] G. Notoatmodjo and C. Thomborson, "Passwords and Perceptions," in *Proceedings of the Seventh Australasian Conference on Information Security - Volume 98*, 2009, pp. 71–78.

[16] P. Tarwireyi, S. Flowerday, and A. Bayaga, "Information security competence test with regards to password management," *Information Security South Africa (ISSA), 2011*. pp. 1–7, 2011.

[17] S. Egelman, A. Sotirakopoulos, I. Muslukhov, K. Beznosov, and C. Herley, "Does My Password Go Up to Eleven?: The Impact of Password Meters on Password Selection," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2013, pp. 2379–2388.

[18] M. Jakobsson and M. Dhiman, "The Benefits of Understanding Passwords," in in *Mobile Authentication SE - 2*, Springer New York, 2013, pp. 5–24.

[19] B. F. Barton and M. S. Barton, "User-friendly Password Methods for Computer-mediated Information Systems," *Comput. Secur.*, vol. 3, no. 3, pp. 186–195, 1984.

[20] M. Zviran and W. J. Haga, "Cognitive passwords: The key to easy access control," *Computers & Security*, vol. 9, no. 8, pp. 723–736, 1990.

[21] M. Zviran and W. Haga J., "A Comparison of Password Techniques for Multilevel Authentication Mechanisms," vol. 36, no. 3, 1993.

[22] M. Bishop and D. V Klein, "Improving system security via proactive password checking," *Computers & Security*, vol. 14, no. 3, pp. 233–249, 1995.

[23] J. Bunnell, J. Podd, R. Henderson, R. Napier, and J. Kennedy-Moffat, "Cognitive, associative and conventional passwords: Recall and guessing rates," *Computers & Security*, vol. 16, no. 7, pp. 629–641, 1997.

[24] S. Brostoff and Ma. Sasse, "Are Passfaces More Usable Than Passwords? A Field Trial Investigation," in in *People and Computers XIV — Usability or Else! SE - 27*, S. McDonald, Y. Waern, and G. Cockton, Eds. Springer London, 2000, pp. 405–424.

[25] J. J. Yan, "A Note on Proactive Password Checking," in *Proceedings of the 2001 Workshop on New Security Paradigms*, 2001, pp. 127–135.

[26] R. W. Proctor, M.-C. Lien, K.-P. L. Vu, E. E. Schultz, and G. Salvendy, "Improving computer security for authentication of users: influence of proactive password restrictions.," *Behavior research methods, instruments, & computers : a journal of the Psychonomic Society, Inc*, vol. 34, no. 2, pp. 163–9, May 2002.

[27] J. Yan, A. Blackwell, R. Anderson, and A. Grant, "Password memorability and security: empirical results," *Security & Privacy, IEEE*, vol. 2, no. 5. pp. 25–31, 2004.

[28] W. C. Summers and E. Bosworth, "Password Policy: The Good, the Bad, and the Ugly," in *Proceedings of the Winter International Synposium on Information and Communication Technologies*, 2004, pp. 1–6.

[29] D. Davis, F. Monrose, and M. K. Reiter, "On User Choice in Graphical Password Schemes," in *Proceedings of the 13th Conference on USENIX Security Symposium - Volume 13*, 2004, p. 11.

[30] F. Monrose and M. K. Reiter, "Graphical Passwords," pp. 161–180, 2005.

[31] X. Suo, Y. Zhu, and G. S. Owen, "Graphical passwords: a survey," *Computer Security Applications Conference, 21st Annual*. p. 10 pp.–472, 2005.

[32] C. Kuo, S. Romanosky, and L. F. Cranor, "Human Selection of Mnemonic Phrase-based Passwords," in *Proceedings of the Second Symposium on Usable Privacy and Security*, 2006, pp. 67–78.

[33] C. P. Garrison, "Encouraging Good Passwords," in *Proceedings of the 3rd Annual Conference on Information Security Curriculum Development*, 2006, pp. 109–112.

[34] S. Riley, “Password Security : What Users Know and What They Actually Do,” vol. 8, no. 1, 2006.

[35] M. Keith, B. Shao, and P. J. Steinbart, “The usability of passphrases for authentication: An empirical field study,” *International Journal of Human-Computer Studies*, vol. 65, no. 1, pp. 17–28, 2007.

[36] A. Forget, S. Chiasson, and R. Biddle, “Helping Users Create Better Passwords: Is This the Right Approach?,” in *Proceedings of the 3rd Symposium on Usable Privacy and Security*, 2007, pp. 151–152.

[37] D. Florêncio, C. Herley, and B. Coskun, “Do Strong Web Passwords Accomplish Anything?,” in *Proceedings of the 2Nd USENIX Workshop on Hot Topics in Security*, 2007, pp. 10:1–10:6.

[38] D. Hart, “Attitudes and Practices of Students Towards Password Security,” *J. Comput. Sci. Coll.*, vol. 23, no. 5, pp. 169–174, 2008.

[39] S. Farrell, “Password Policy Purgatory,” *Internet Computing, IEEE*, vol. 12, no. 5. pp. 84–87, 2008.

[40] A.-M. Horcher and G. P. Tejay, “Building a better password: The role of cognitive load in information security training,” *Intelligence and Security Informatics, 2009. ISI '09. IEEE International Conference on*. pp. 113–118, 2009.

[41] C. Herley, P. C. Oorschot, and A. Patrick, “Passwords: If We’re So Smart, Why Are We Still Using Them?,” in in *Financial Cryptography and Data Security SE - 14*, vol. 5628, R. Dingledine and P. Golle, Eds. Springer Berlin Heidelberg, 2009, pp. 230–237.

[42] B. T. Kuhn and C. Garrison, “A Survey of Passwords from 2007 to 2009,” in *2009 Information Security Curriculum Development Conference*, 2009, pp. 91–94.

[43] D. Nelson and K.-P. Vu, “Effects of a Mnemonic Technique on Subsequent Recall of Assigned and Self-generated Passwords,” in in *Human Interface and the Management of Information. Designing Information Environments SE - 78*, vol. 5617, M. Smith and G. Salvendy, Eds. Springer Berlin Heidelberg, 2009, pp. 693–701.

[44] M. Dell’Amico, P. Michiardi, and Y. Roudier, “Password Strength: An Empirical Analysis,” *INFOCOM, 2010 Proceedings IEEE*. pp. 1–9, 2010.

[45] W. Hu, X. Wu, and G. Wei, “The Security Analysis of Graphical Passwords,” *Communications and Intelligence Information Security (ICCIIS), 2010 International Conference on*. pp. 200–203, 2010.

[46] L. Tam, M. Glassman, and M. Vandenwauver, “The Psychology of Password Management: A Tradeoff Between Security and Convenience,” *Behav. Inf. Technol.*, vol. 29, no. 3, pp. 233–244, 2010.

[47] P. G. Inglesant and M. A. Sasse, “The True Cost of Unusable Password Policies: Password Use in the Wild,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2010, pp. 383–392.

[48] A. Moallem, “Did You Forget Your Password ?,” pp. 29–39, 2011.

[49] A. G. Voyiatzis, C. A. Fidas, D. N. Serpanos, and N. M. Avouris, “An Empirical Study on the Web Password Strength in Greece,” *Informatics (PCI), 2011 15th Panhellenic Conference on*. pp. 212–216, 2011.

[50] E. Hayashi and J. Hong, “A Diary Study of Password Usage in Daily Life,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2011, pp. 2627–2630.

[51] K. Schaffer, “Are Password Requirements too Difficult?,” *Computer*, vol. 44, no. 12. pp. 90–92, 2011.

[52] S. Komanduri, R. Shay, P. G. Kelley, M. L. Mazurek, L. Bauer, N. Christin, L. F. Cranor, and S. Egelman, “Of Passwords and People: Measuring the Effect of Password-composition Policies,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2011, pp. 2595–2604.

[53] P. G. Kelley, S. Komanduri, M. L. Mazurek, R. Shay, T. Vidas, L. Bauer, N. Christin, L. F. Cranor, and J. Lopez, “Guess Again (and Again and Again): Measuring Password Strength by Simulating Password-Cracking Algorithms,” *Security and Privacy (SP), 2012 IEEE Symposium on*. pp. 523–537, 2012.

[54] K. Helkala, N. Svendsen, P. Thorsheim, and A. Wiehe, “Cracking Associative Passwords,” in in *Secure IT Systems SE - 11*, vol. 7617, A. Jøsang and B. Carlsson, Eds. Springer Berlin Heidelberg, 2012, pp. 153–168.

[55] K. Helkala and N. Svendsen, “The Security and Memorability of Passwords Generated by Using an Association Element and a Personal Factor,” in in *Information Security Technology for Applications SE - 9*, vol. 7161, P. Laud, Ed. Springer Berlin Heidelberg, 2012, pp. 114–130.

[56] E. Zezschwitz, A. Luca, and H. Hussmann, “Survival of the Shortest: A Retrospective Analysis of Influencing Factors on Password Composition,” in in *Human-Computer Interaction – INTERACT 2013 SE - 28*, vol. 8119, P. Kotzé, G. Marsden, G. Lindgaard, J. Wesson, and M. Winckler, Eds. Springer Berlin Heidelberg, 2013, pp. 460–467.

[57] B. Lorenz, K. Kikkas, and A. Klooster, “‘The Four Most-Used Passwords Are Love, Sex, Secret, and God’: Password Security and Training in Different User Groups,” in in *Human Aspects of Information Security, Privacy, and Trust SE - 29*, vol. 8030, L. Marinos and I. Askoxylakis, Eds. Springer Berlin Heidelberg, 2013, pp. 276–283.

[58] L. Clair, L. Johansen, W. Enck, M. Pirretti, P. Traynor, P. McDaniel, and T. Jaeger, “Password Exhaustion: Predicting the End of Password Usefulness,” in in *Information Systems Security SE - 3*, vol. 4332, A. Bagchi and V. Atluri, Eds. Springer Berlin Heidelberg, 2006, pp. 37–55.

[59] A. Adams and M. A. Sasse, “Users Are Not the Enemy,” *Commun. ACM*, vol. 42, no. 12, pp. 40–46, Dec. 1999.

[60] M. A. Sasse, S. Brostoff, and D. Weirich, “Transforming the ‘Weakest Link’ — a Human/Computer Interaction Approach to Usable and Effective Security,” *BT Technology Journal*, vol. 19, no. 3, pp. 122–131, 2001.

[61] S. Vidyaraman, M. Chandrasekaran, and S. Upadhyaya, “Position: The User is the Enemy,” in *Proceedings of the 2007 Workshop on New Security Paradigms*, 2008, pp. 75–80.

[62] C. A. Fidas, A. G. Voyiatzis, and N. M. Avouris, “When Security Meets Usability: A User-Centric Approach on a Crossroads Priority Problem,” *Informatics (PCI), 2010 14th Panhellenic Conference on*. pp. 112–117, 2010.

[63] M. Adeka, S. Shepherd, and R. Abd-Alhameed, “Resolving the password security purgatory in the contexts of technology, security and human factors,” *Computer Applications Technology (ICCAT), 2013 International Conference on*. pp. 1–7, 2013.

[64] D. Florêncio and C. Herley, “Where Do Security Policies Come from?,” in *Proceedings of the Sixth Symposium on Usable Privacy and Security*, 2010, pp. 10:1–10:14.

[65] K.-P. L. Vu, R. W. Proctor, A. Bhargav-Spantzel, B.-L. (Belin) Tai, J. Cook, and E. Eugene Schultz, “Improving Password Security and Memorability to Protect Personal and Organizational Information,” *Int. J. Hum.-Comput. Stud.*, vol. 65, no. 8, pp. 744–757, 2007.

[66] B. Grawemeyer and H. Johnson, “Using and Managing Multiple Passwords: A Week to a View,” *Interact. Comput.*, vol. 23, no. 3, pp. 256–267, 2011.

[67] G. B. Duggan, H. Johnson, and B. Grawemeyer, “Rational Security: Modelling Everyday Password Use,” *Int. J. Hum.-Comput. Stud.*, vol. 70, no. 6, pp. 415–431, 2012.

[68] S. Preibusch and J. Bonneau, “The Password Game: Negative Externalities from Weak Password Practices,” in Decision and Game Theory for Security SE - 13, vol. 6442, T. Alpcan, L. Buttyán, and J. Baras, Eds. Springer Berlin Heidelberg, 2010, pp. 192–207.

[69] M. Weir, S. Aggarwal, B. de Medeiros, and B. Glodek, “Password Cracking Using Probabilistic Context-Free Grammars,” Security and Privacy, 2009 30th IEEE Symposium on. pp. 391–405, 2009.

Brute Force Analysis of PsychoPass-generated Passwords

B. Brumen* and A. Černezel*
* University of Maribor, Faculty of Electrical Engineering and Computer Science, Smetanova 17, Si-2000 Maribor, Slovenia
bostjan.brumen@uni-mb.si

Abstract - Background: Recently, Cipresso et al proposed a novel method for generating textual passwords. The PsychoPass password can be created, memorized and recalled by thinking of an action sequence instead of a string of characters.

Objective: The objective of this paper is to analyze how and when the PsychoPass passwords are resilient to brute force attacks and how they compare to randomly generated ones.

Methods: The PsychoPass method was studied analytically. The number of possible different passwords that can be created using PsychoPass method was calculated for different possible settings.

Results: The PsychoPass generates passwords that are resilient to brute force attacks with password lengths comparable to those generated randomly. A PsychoPass-generated password with 18 characters is as strong as a completely randomly generated password of length 11 characters with mixed upper and lowercase letters, numbers and symbols, but can easily be remembered.

Conclusion: The PsychoPass method is an alternative to the existing passwords and is resilient to brute force attack. There is no known dictionary to enable the dictionary attack. The PsychoPass passwords look like random ones, eg. »aEy|dX%7Tu]6hJ«, but a user can easily remember it based on the action sequence instead of remembering its characters.

I. INTRODUCTION

The ability to authenticate the identity of a user via passwords continues to be basic and the most relied-upon security mechanism in information systems. Weak passwords have led to very serious breaches, exposing millions of users and/or causing billions in damages. The history of password-related problems pre-dates the seminal paper written by Morris and Thompson in 1979 [1] – it goes way back to mid-1960s and to the CTSS operating system exposing all the passwords as a daily welcome message [2]. A comprehensive list of breaches since 2005 can be found at [3].

A typical survey evaluating the generation and use of passwords revealed that users have several password uses and the average password has more than one application. Two thirds of passwords are designed around one's personal characteristics, with most of the remainder relating to relatives, friends or lovers. Proper names and birthdays are the primary information used in constructing passwords, accounting for about half of all password constructions. Almost all respondents reuse passwords, and about two thirds of password uses are duplications. Passwords have been forgotten by a third of respondents, and over half keep a written record of them [4].

It seems that nothing has been learnt and changed in the course of almost 50 years – users and their passwords remain the weakest link [5-13], although the basic and the most relied-upon security mechanism in information systems continues to be the ability to authenticate the identity of a user. The passwords used to be [14-16] and still are the main method of authentication [17-19], although research continues on more sophisticated methods of authentication, see e.g. [20-29]. Some novel solutions are even improvements of concepts well known to the automobile industry (i.e. car keys) [30].

II. BACKGROUND ON PASSWORD STRENGTH – WHY ARE GOOD PASSWORDS SO HARD TO REMEMBER?

The general rule is that passwords strength is proportional to their length and the size of the symbols space – i.e. the pool from where the password characters can be used from. This is true only if the process of the character selection is truly random [31]!

The number of possible combinations can be expressed by $s=b^{le}$, where s is the total number of combinations, b is the total number of symbols in a domain and le is the length of the password in number of characters. Typically, the pool size is taken to be 75 (upper and lower case characters, digits, symbols). If n characters are used for a password, there is a total of $s=b^1+b^2+b^3+\ldots+b^n=b^{n+1}-1$. Today, the reported speeds (v) for restoring a plaintext password from a given (unsalted) hash value are in the range of $v=10^9$ combinations/second [32-35], up to 10^{12} combinations/second [31]. Due to the Moore's Law [36] still in effect, the computing power doubles roughly every 18 months. In 10 years, we can expect the speeds to be 100-times of today's (only taking into the account the Moore's Law and no other improvements in the field, e.g. quantum computing [37] or improvements in hash computations [38]). Table I lists the number of all possible combinations for passwords of length from 7 up to 11 characters for the size of 75 character in a pool (letters, numbers and special symbols) and the time required to check all possible combinations at today's speeds and the time it will take in 10 years from now due to speed improvements.

TABLE I. WORST CASE TIME TO CRACK RANDOM PASSWORDS NOW AND IN 10 YEARS FOR DIFFERENT PASSWORD LENGTHS (BASE=75 CHARACTERS; CHARACTERS RANDOMLY DRAWN)

Pwd length	*Number of combinations*	*Time to crack now*	*Time to crack in 10 years from now*
7	1,00113E+15	11 days, 14 hours	2 hours, 46 minutes
8	7,50847E+16	2 years, 139 days	8 days, 16 hours
9	5,63135E+18	178 years, 207 days	1 year, 286 days
10	4,22351E+20	13392 years	133 years, 338 days
11	3,16764E+22	1.004.450 years	10.044 years

The table shows that what is safe today will not be safe tomorrow [31], this especially holds true for off-line files which will remain unchanged for years, containing interesting medical data about an important person. The breach of privacy of data in such a file has a timelessness property [39] – no matter when in the lifetime of data the breach happens, it poses a problem.

A user, when creating a password, needs to observe one the following principle: (1) a password is considered secure if the cost of breaking it exceeds the value of the protected information, or (2) the time required to break the password exceeds the lifetime of the protected data [31]. Today, we cannot rely on the costs because the ratio of bits processed per unit of currency is continuously sinking making the password cracking machine affordable [32, 34, 35]. One can only increase the time required. Thus let us assume a typical 60-year time span of usefulness of stored information [40-44]. In other words, a safe password today (in 2014) would be made of at least nine characters from upper and lowercase letters, numbers and symbols, but it would not be safe in 10 years from now. Thus, 10-character passwords are required today to be safe tomorrow as well, confirming findings by Egelman et al [18]. Today, a typical password has (only) 8 characters [45-47].

When we come to the 10 (or more) characters to remember, they constitute a much larger corpus than is the capacity of human memory, the well-known 7±2 principle [48]. Human memory in addition is temporally limited (short-term) when it comes to memorizing sequences [49] and has a limited capacity with a fixed final error rate [50-52]. For this reason good passwords that are consisting of an abundant number of randomly selected characters are doomed: the users will either forget them [53] or write them down, or both [54-56]. What are the alternatives?

III. RELATED WORK

Typically, a user is authenticated based on one of the three underlying principles (or combinations thereof): "what you know", "what you are" and "what you have" [31, 57]. "What you know" is based on secrecy known only to authorized user, "what you are" is based on user's physical characteristics (e.g. retina image, fingerprint) and is better known by the term biometrics, whereas "what you have" is based on smart cards or other tokens.

The passwords that are generated based on "what you know" principle can be divided into textual passwords and graphical passwords [58, 59]. In our contribution we focus on textual passwords as graphical ones require a different user interface for entering passwords [59] and are thus not directly comparable to our study. In addition they require a redesign of user interface for entering a password [59].

There are several alternatives to (nearly) random long textual password.

First, a password can be based on personal data or characteristics (e.g. birthdate, names, pets, addresses, etc.). They are easy to be cracked [1, 60] and their use is strongly discouraged [61].

Second, cognitive password authentication schemes require user to answer a randomly selected set of personal questions which only an authorized user can answer correctly. They have a high recall rate, but are susceptible to guesses by family and friends [20]. To prevent reuse, each system would require a unique set of questions [62].

Third, pass-sentences and pass-phrases are passwords composed of long, grammatically correct phrases [63]. They are memorable and software-cracking resistant, but their length makes them useless for repeated use [62].

Fourth, randomly generated but pronounceable passwords made up of concatenated pronounceable syllables are resistant to standard dictionary or brute-force attacks [64], but the algorithm is vulnerable to a special dictionary attack [65].

Fifth, mnemonic passwords are those where a user chooses a memorable phrase and uses a character – often the first letter – to represent each word in the phrase, or vice versa, for a given (random) passphrase a system generates a grammatically correct phrase that the user can remember. They were endorsed in the past [16], but were found vulnerable and will even be more vulnerable in the future [20].

Orthogonal to the works on different textual password generating techniques are contributions that deal with password metrics, principally meters that show users how strong their password might be [18, 47, 66]. It was shown empirically [47] and mathematically [67] that Shannon entropy value is not useful when determining the strength of a password creation policy, and other policies need to be used. Common advices on minimum password length and character set requirements provide against online attacks [47]. Yet, by observing these requirements, users tend to forget passwords and/or write them down, usually in an insecure location [60]. Writing down a password is not a bad practice itself, as pointed out by Bruce Schneier: "*…if only users wrote [a password] down on a small piece of paper, and keep it with their other valuable small pieces of paper: in their wallet*" [68].

Orthogonal to our work are also contributions dealing with users' compliance to different password creation policies [5, 6, 8, 69]. General observation is that users tend to choose bad passwords due to inexistent policies or try to comply to them with minimal effort [9, 45, 47, 70]. All users need is a sound piece of advice and an explanation why passwords need to be strong and how to achieve it [8, 71].

A. *PsychoPass Method*

Cipresso et al [72] have proposed a novel method for generating textual passwords: the method was improved by Brumen et al [40]. The proposed improved method is, in our modest opinion, a novelty in producing strong passwords that was not seen for years, maybe ever since the invention of graphical passwords in 1990s [73, 74]. "*The idea of PsychoPass is that a password can be created, memorized and recalled by just thinking of an action sequence instead of a word or string of characters*" [72]. With PsychoPass method, a user creates a password based on visual location of keys, not the key values themselves. Figure 1 depicts a visual circle and a square (actually a rhombus) on the keyboard using blue numbered dots. The keys that draw the circle are A-W-E-D-X-Y, and the keys that draw the square are 5-6-7-T-U-G-H-J.



Figure 1. A circle and a square on keyboard producing a strong password

However, the improved PsychoPass method requires the use of SHIFT and ALT-GR keys and that the keys that are not always adjacent to each other in the sequence.

Suppose that key #1 is pressed without SHIFT or ALT-GR, key #2 is pressed in combination with shift key, and so on, as given in Table II (please note that SI-Slovenian keyboard layout is used).

The password representing the circle and the square in Figure 1 would read »aEy|dX%7Tu]6hJ«, as presented in Figure 2 and Table II.

TABLE II. PASSWORD RESULT OF THE VISUAL REPRESENTATION FROM FIGURE 1 USING PSYCHOPASS

Seq. #:	1	2	3	4	5	6	7	8	9	10	11	12	13	14
SH/AGR[a] :		s		a		s	s		s		a			s
Result :	**a**	**E**	**y**	**|**	**d**	**X**	**%**	**7**	**T**	**u**	**]**	**6**	**h**	**j**

a. S=shift used, A=ALT-GR used

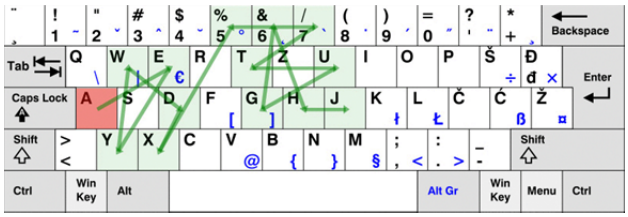


Figure 2. A representation of password »aEy|dX%7Tu]6hJ«

Interestingly, the very same shapes, if shifted one key to the right on a keyboard, would yield another password, namely »sRxefC&8Zih7jK«, as shown in Table III and Figure 3.

TABLE III. PASSWORD RESULT OF THE VISUAL REPRESENTATION FROM FIGURE 1 USING PSYCHOPASS, SHIFTED ONE TO THE RIGHT

Seq. #:	1	2	3	4	5	6	7	8	9	10	11	12	13	14
SH/AGR[a]:		s		a		s	s		s		a			s
Result :	**s**	**R**	**x**	**e**	**f**	**C**	**&**	**8**	**Z**	**i**	**h**	**7**	**j**	**K**

a. S=shift used, A=ALT-GR used

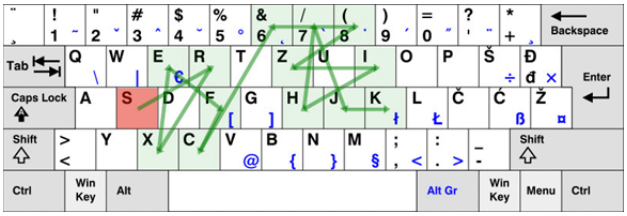


Figure 3. A representation of password »sRxefC&8Zih7jK«

The representations of passwords »aEy|dX%7Tu]6hJ« and »sRxefC&8Zih7jK« are shown in Figure 2 and Figure 3, respectively.

The user thus memorizes a password based on its visual representation (action sequence) and additionally when to press SHIFT or ALT-GR.

It may seem that the password produced (e.g. »aEy|dX%7Tu]6hJ«) is totally random (with 75^{14}= 178.179.480.135.440.826.416.015.625 =1,78e26 different combinations, and the brute force attack would take some 5,6e9 years. However, the Psychopass generated passwords are not random, and the question is, how such passwords are resilient to brute force attacks.

B. *The contribution*

The aim of the present work is to present the findings of the brute force analysis of the improved PsychoPass method.

IV. BRUTE FORCE ANALYSIS OF PSYCHOPASS METHOD

The PsychoPass method was studied analytically. The number of possible different passwords that can be created using PsychoPass method was calculated.

Let us represent a keyboard as an undirected graph, G = (V, E), comprising a set V of vertices together with a set E of edges. Each key capable of producing a character represents a vertex of the graph. Each vertex is weighted; weight 2 denotes keys that can produce 2 or less characters (typically in combination with SHIFT) and weight 3 denotes keys that can produce 3 or less characters (additionally in combination with ALT-GR). Each vertex is connected to the other ones based on the key adjacency on the keyboard of length up to 2, e.g. the "D" key has the following 1-adjacent keys: "SERFCX", and 2-adjacent keys "VYAW345". Additionally, key "D" is adjacent to itself since a password can contain a subset like "dd" or "DD".

Let us represent the graph G by matrix A of dimension *n* x *n*:

$$A = [a_{i,j}] \qquad (1)$$

where

$a_{i,j} =$

$\begin{cases} 0; \text{ if keys } i \text{ and } j \text{ are not adjacent} \\ w_j;\ w_j \text{ is the weight of key } j \text{ if keys } i \text{ and } j \text{ are adjacent or } i = j. \end{cases}$

Since we are interested in all possible combination of passwords of length l, we calculate the $(l\text{-}1)^{\text{th}}$ power of matrix A:

$$B = A^{l-1} \tag{2}$$

To calculate the total number of all possible combinations, we first multiply the ith row with the weight of ith key.

$$C = [c_{i,j}]; c_{i,j} = b_{i,j} * w_i \tag{3}$$

The total number of all possible combinations (s) is the sum of all elements in the resulting matrix:

$$s = \sum_{i=1}^{n} \sum_{j}^{n} c_{i,j} \tag{4}$$

The numbers of all possible combinations for specific l and for Slovenian-SI keyboard layout were calculated and are given in Table 4. Additionally, the table 4 lists the time needed for a brute-force attack with today's computing power and in 10 years from now.

TABLE IV. NUMBER OF POSSIBLE PASSWORD COMBINATIONS FOR GIVEN *L*, AND BRUTE FORCE ATTACK TIMES (NOW / 10 YEARS)

Password length	*Number of combinations*	*Time to crack now*	*Time to crack in 10 years from now*
1	123	0 sec	0 sec
2	1763	0 sec	0 sec
…	…	…	…
10	5,09386E+12	1 hour 24 minutes	51 seconds
11	7,79683E+13	21 hours 39 minutes	13 minutes
12	1,19364E+15	13 days 19 hours	3 hours 18 minutes
13	1,82766E+16	211 days 12 hours	2 days 2 hours
14	2,79886E+17	8 years 319 days	32 days 9 hours
15	4,28669E+18	135 years 339 days	1 years 131 days
16	6,56618E+19	2.082 years	20 years 299 days
17	1,00589E+21	31.896 years	318 years 352 days
18	1,54109E+22	488.675 years	4.886 years

V. RESULTS

From Table IV it can be concluded that passwords generated by PsychoPass method need to be at least 15 characters long to be resilient to today's brute force attacks and 17 to resist those in 10 years from now.

The comparison of strength of PsychoPass passwords and randomly generated is shown in Table V.

TABLE V. COMPARISON OF PSYCHOPASS AND RANDOM PASSWORDS RESILLIENCE TO BRUTE FORCE ATTACKS

Psychopass password length	*Random password length*	*Number of combinations*[a]	*Time to crack now*[a]	*Time to crack in 10 years from now*[a]
12	7	E+15	~10s days	~ hours
13	8	E+16	~100s days	~days
14		E+17	~ years	~10s days
15	9	E+18	~100s years	~years
16		E+19	~1.000s years	~10s years
17	10	E+21	~10.000s years	~100s years
18	11	E+22	~100.000s years	~1.000s years

a. Approximation only

The results show that for comparable password strength the PsychoPass passwords need to be from 5 to 7 characters longer than randomly generated ones.

VI. CONLUSION

In this paper we presented an analysis of brute-force resilience of passwords generated with a novel PsychoPass method that is an alternative to the existing passwords schemes. The idea of PsychoPass is that a password can be created, memorized and recalled by just thinking of an action sequence instead of a word or string of characters.

The results show that psychopass-generated passwords are resilient to today's brute force attacks if their length exceeds 14 characters and are considered safe at lengths exceeding 16 characters when taking into the account the expected increase in the computing power in the next 10 years.

The required length of psychopass-generated passwords exceeds the randomly generated ones for a comparable resilience; however, they are much easier to remember. The length of a randomly generated password constitutes a much larger corpus than the capacity of an average human memory.

Additionally, there is no known dictionary to enable the dictionary attack. The PsychoPass passwords look like random ones, eg. »aEy|dX%7Tu]6hJ«, so the attacker cannot easily prove that the particular password was created using the PsychoPass method instead of being randomly generated.

An additional advantage is that the same action sequence can produce completely different passwords so the popular password reuse is supported in a safe way.

A disadvantage of the proposed method is that it is bound to a specific keyboard layout and the usage in mobile devices is somehow limited.

ACKNOWLEDGMENT

A. Černezel's work was partially supported by the Slovenian Research Agency under grant number 1000-11-310138.

REFERENCES

[1] Morris R and Thompson K. Password security: A case history. Communications of the ACM 1979; 22(11): 594-597.

[2] Corbato FJ. On building systems that will fail. Communications of the ACM 1991; 34(9): 72-81. DOI: 10.1145/1283920.1283947.

[3] Privacy Rights Clearinghouse. Chronology of Data Breaches: Security Breaches 2005 - Present. 2012; Available from: http://www.privacyrights.org/data-breach. (Archived by WebCite® at http://www.webcitation.org/6AfYySWJK).

[4] Brown AS, Bracken E, Zoccoli S and Douglas K. Generating and remembering passwords. Applied Cognitive Psychology 2004; 18(6): 641-651. DOI: 10.1002/acp.1014.

[5] Adams A and Sasse MA. Users are not the enemy. Communications of the ACM 1999; 42(12): 40-46.

[6] Adams A, Sasse MA and Lunt P. Making passwords secure and usable. People and Computers XII. Heidelberg: Springer; 1997. p. 1-19. ISBN: 3540761721

[7] Notoatmodjo G. Exploring the 'Weakest Link': A Study of Personal Password Security. The University of Auckland, New Zealand; 2007.

[8] Sasse MA, Brostoff S and Weirich D. Transforming the 'weakest link'—a human/computer interaction approach to usable and effective security. BT technology journal 2001; 19(3): 122-131.

[9] Tam L, Glassman M and Vandenwauver M. The psychology of password management: a tradeoff between security and convenience. Behaviour & Information Technology 2009; 29(3): 233-244. DOI: 10.1080/01449290903121386.

[10] Juric MB, Rozman I, Brumen B, Colnaric M and Hericko M. Comparison of performance of Web services, WS-Security, RMI, and RMI-SSL. Journal of Systems and Software 2006; 79(5): 689-700.

[11] Juric MB, Rozman I, Welzer T, Hericko M, Brumen B and Podgorelec V. Patterns for enterprise application integration. Frontier in artificial intelligence and applications, Knowledge-Based Software Engineering 2002; 80: 88-101.

[12] Juric MB, Sasa A, Brumen B and Rozman I. WSDL and UDDI extensions for version support in web services. Journal of Systems and Software 2009; 82(8): 1326-1343.

[13] Juric MB, Welzer T, Rozman I, Hericko M, Brumen B, Domajnko T and Zivkovic A. Performance assessment framework for distributed object architectures. Advances in Databases and Information Systems. Lecture notes in computer science 1999; 1691: 349-366.

[14] Loch KD, Carr HH and Warkentin ME. Threats to information systems: today's reality, yesterday's understanding. MIS Quarterly 1992; 16(2): 173-186.

[15] Tzong-Chen W and Hung-Sung S. Authenticating passwords over an insecure channel. Computers & Security 1996; 15(5): 431-439.

[16] Zviran M and Haga WJ. Cognitive passwords: the key to easy access control. Computers & Security 1990; 9(8): 723-736.

[17] Lee C-C, Liu C-H and Hwang M-S. Guessing Attacks on Strong-Password Authentication Protocol. International Journal of Network Security 2013; 15(1): 64-67.

[18] Egelman S, Sotirakopoulos A, Muslukhov I, Beznosov K and Herley C. Does my password go up to eleven?: the impact of password meters on password selection. Proceedings of the 2013 SIGCHI Conference on Human Factors in Computing Systems; 2013. April 27 - May 2, Paris, France: ACM.

[19] Creese S, Hodges D, Jamison-Powell S and Whitty M. Relationships between Password Choices, Perceptions of Risk and Security Expertise. Human Aspects of Information Security, Privacy, and Trust. Heidelberg: Springer; 2013. p. 80-89. ISBN: 3642393446

[20] Kuo C, Romanosky S and Cranor LF. Human selection of mnemonic phrase-based passwords. Proceedings of the second symposium on Usable privacy and security, July 12-14; 2006. Carnegie Mellon University, Pittsburgh, PA, USA: ACM.

[21] Jiang P, Wen Q, Li W, Jin Z and Zhang H. An Anonymous User Authentication with Key Agreement Scheme without Pairings for Multiserver Architecture Using SCPKs. The Scientific World Journal 2013; 2013(Article ID 419592). DOI: 10.1155/2013/419592.

[22] Pang L, He L, Pei Q and Wang Y. Secure and efficient mutual authentication protocol for RFID conforming to the EPC C-1 G-2 standard. in 2013 IEEE Wireless Communications and Networking Conference (WCNC). Shanghai, China: IEEE; 2013. p. 1870-1875.

[23] Holbl M, Welzer T and Brumen B. An improved two-party identity-based authenticated key agreement protocol using pairings. Journal of Computer and System Sciences 2012; 78(1): 142-150. DOI: 10.1016/j.jcss.2011.01.002.

[24] Holbl M, Welzer T and Brumen B. Two proposed identity-based three-party authenticated key agreement protocols from pairings. Computers & Security 2010; 29(2): 244-252. DOI: 10.1016/j.cose.2009.08.006.

[25] Holbl M, Welzer T and Brumen B. Improvement of the Peyravian-Jeffries's user authentication protocol and password change protocol. Computer Communications 2008; 31(10): 1945-1951. DOI: 10.1016/j.comcom.2007.12.029.

[26] Hölbl M, Welzer T and Brumen B. Comparative study of tripartite identity-based authenticated key agreement protocols. Informatica (Ljubljana) 2009; 33(3): 347-355.

[27] Hölbl M, Welzer T and Brumen B. Attacks and improvement of an efficient remote mutual authentication and key agreement scheme. Cryptologia 2010; 34(1): 52-59.

[28] Živkovič A, Heričko M, Brumen B, Beloglavec S and Rozman I. The impact of details in the class diagram on software size estimation. Informatica (Lithuan.) 2005; 16(2): 295-312.

[29] Zorman M, Pohorec S and Brumen B. Opening the Knowledge Tombs - Web Based Text Mining as Approach for Re-evaluation of Machine Learning Rules. Lecture Notes in Computer Science (Advances in Databases and Information Systems) 2010; 6295: 533-542.

[30] Grosse E and Upadhyay M. Authentication at Scale. Security & Privacy, IEEE 2013; 11(1): 15-22. DOI: 10.1109/MSP.2012.162.

[31] Stallings W. Cryptography and Network Security, Principles and Practices. 4th ed. Prentice Hall; 2006. ISBN: 0131873164

[32] Hunt T. Our password hashing has no clothes. 2012; Available from: http://www.troyhunt.com/2012/06/our-password-hashing-has-no-clothes.html. Accessed: 2012-09-20. (Archived by WebCite® at http://www.webcitation.org/6AobEryR5).

[33] InsidePro. Password Recovery Software. 2012; Available from: http://www.insidepro.com/. Accessed: 2012-09-20. (Archived by WebCite® at http://www.webcitation.org/6AoaUZsx3).

[34] Kingsley-Hughes A. Cheap GPUs are rendering strong passwords useless. 2011; Available from: http://www.zdnet.com/blog/hardware/cheap-gpus-are-rendering-strong-passwords-useless/13125. Accessed: 2012-09-20. (Archived by WebCite® at http://www.webcitation.org/6Aob1QCot).

[35] Zonenberg A. Distributed hash cracker: A cross-platform gpu-accelerated password recovery system. Rensselaer Polytechnic Institute, Troy, NY 2009: 27.

[36] Moore GE. Cramming more components onto integrated circuits Electronics 1965; 38(8): 114-117.

[37] Ladd TD, Jelezko F, Laflamme R, Nakamura Y, Monroe C and O'Brien JL. Quantum computers. Nature 2010; 464(7285): 45-53. PMID: 20203602. DOI:10.1038/nature08812.

[38] Feldmeier DC and Karn PR. Unix password security-ten years later. Advances in Cryptology—CRYPTO'89 Proceedings; 1990. Springer.

[39] Brumen B, Heričko M, Sevčnikar A, Završnik J and Hölbl M. Outsourcing Medical Data Analyses: Can Technology Overcome Legal, Privacy, and Confidentiality Issues? J Med Internet Res 2013; 15(12): e283. DOI: 10.2196/jmir.2471.

[40] Brumen B, Heričko M, Rozman I and Hölbl M. Security Analysis and Improvements to the PsychoPass Method. J Med Internet Res 2013; 15(8): e161. DOI: 10.2196/jmir.2366; PMID: 23942458.

[41] Brumen B, Golob I, Welzer T, Rozman I, Družovec M and Jaakkola H. Data protection for outsourced data mining. Informatica (Ljubljana) 2002; 26(2): 205-210.

[42] Brumen B, Golob I, Welzer T, Rozman I, Družovec M and Jaakkola H. An Algorithm for Protecting Knowledge Discovery Data. Informatica (Lithuan.) 2003; 14(3): 227-288.

[43] Brumen B, Welzer T, Družovec M, Golob I and Jaakkola H. Protecting medical data for analyses. Fifteenth IEEE Symposium on Computer-Based Medical Systems; 2002. Maribor, Slovenia; 4 June 2002 through 7 June 2002: IEEE Computer Society Press.

[44] Brumen B, Welzer T, Družovec M, Golob I, Jaakkola H, Rozman I and Kubalík J. Protecting medical data for decision-making analyses. Journal of Medical Systems 2005; 29(1): 65-80.

[45] Dell'Amico M, Michiardi P and Roudier Y. Password strength: an empirical analysis. In Proceedings of 2010 IEEE INFOCOM, 14-19 March 2010, San Diego, Ca, USA; 2010. IEEE.
[46] StClair L, Johansen L, Enck W, Pirretti M, Traynor P, McDaniel P and Jaeger T. Password exhaustion: Predicting the end of password usefulness. Information Systems Security. Heidelberg: Springer; 2006. p. 37-55. ISBN: 3540689621
[47] Brumen B, Rozman I, Heričko M, Černezel A and Hölbl M. Best-fit learning curve model for the C4.5 algorithm. Informatica (Lithuan.) 2014; 25(?): (to be published).
[48] Miller GA. The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychological review 1956; 63(2).
[49] Johnson GJ. A distinctiveness model of serial learning. Psychological review 1991; 98(2): 204-217.
[50] Brumen B, Golob I, Jaakkola H, Welzer T and Rozman I. Early assessment of classification performance. In Hogan J, Montague P, Purvis M and Steketee C (Eds): ACSW Frontiers '04 Proceedings of the second workshop on Australasian information security, Data Mining and Web Intelligence, and Software Internationalisation - Volume 32; 2004. Australian Computer Society, Inc.
[51] Brumen B, Hölbl M, Harej Pulko K, Welzer T, Heričko M, Jurič MB and Jaakkola H. Learning Process Termination Criteria. Informatica (Lithuan.) 2012; 23(4): 521-536.
[52] Brumen B, Jurič MB, Welzer T, Rozman I, Jaakkola H and Papadopoulos A. Assessment of classification models with small amounts of data. Informatica (Lithuan.) 2007; 18(3): 343-362.
[53] Florencio D and Herley C. A large-scale study of web password habits. Proceedings of the 16th international conference on World Wide Web; 2007. ACM.
[54] Yan J, Blackwell A, Anderson R and Grant A. Password Memorability and Security: Empirical Results. IEEE SECURITY & PRIVACY September/October 2004; 2(5):25-31.
[55] Zviran M and Haga WJ. A comparison of password techniques for multilevel authentication mechanisms. The Computer Journal 1993; 36(3): 227-237.
[56] Yan J, Blackwell A, Anderson R and Grant A. The memorability and security of passwords: some empirical results. Technical report # UCAM-CL-TR-500. Cambridge, UK: University Of Cambridge; 2000.
[57] Pfleeger CP and Pfleeger SL. Security in computing. 3rd ed.Upper Saddle River, NY, USA: Prentice Hall PTR; 2003. ISBN: 0130355488
[58] Davis D, Monrose F and Reiter MK. On User Choice in Graphical Password Schemes. Proceeding SSYM'04 Proceedings of the 13th conference on USENIX Security Symposium - Volume 13; Berkeley, CA, USA: USENIX Association; 2004. p. 11-11.
[59] Suo X, Zhu Y and Owen GS. Graphical passwords: A survey. In Proceedings of 21st Annual Computer Security Applications Conference. Tucson, AZ, USA: IEEE; 2005. pp. 463-472.
[60] Zviran M and Haga WJ. Password security: an empirical study. Journal of Management Information Systems 1999; 15: 161-186.
[61] FIPS. PUB 112 Password Usage. National Institute of Standards and Technology; 1985.
[62] Brostoff AS. Improving Password System Effectiveness. PhD Thesis, Department of Computer Science, University College London. London, UK: University of London; 2004.
[63] Spector Y and Ginzberg J. Pass-sentence—a new approach to computer code. Computers & Security 1994; 13(2): 145-160.
[64] Gasser M. A Random Word Generator for Pronounceable Passwords, Technical Report MTR-3006. ESD-TR-75-97, AD-A017676, MITRE Corp., Bedford, Mass; 1975.
[65] Ganesan R, Davies C and Atlantic B. A new attack on random pronounceable password generators. Proceedings of the 17th NIST-NCSC National Computer Security Conference; 1994. Baltimore, MD, USA.
[66] Bishop M and Klein DV. Improving system security via proactive password checking. Computers & Security 1995; 14(3): 233-249.
[67] Verheul ER. Selecting secure passwords. Topics in Cryptology–CT-RSA 2007. Heidelberg: Springer; 2006. p. 49-66. ISBN: 3540693270
[68] Schneier B. Write Down Your Password. Schneier on Security 2005; Available from: http://www.schneier.com/blog/archives/2005/06/write_down_your.html. Accessed: 2013-08-29. (Archived by WebCite® at http://www.webcitation.org/6JEOK0Vdl).
[69] Weirich D and Sasse MA. Persuasive password security. IN Proceedings of Conference CHI'01 Human Factors in Computing Systems. Seattle, WA, USA — March 31 - April 05; 2001. ACM.
[70] Gehringer EF. Choosing passwords: security and human factors. (ISTAS'02). In Proceedings of 2002 International Symposium on Technology and Society; 2002. IEEE.
[71] Horcher A-M and Tejay GP. Building a better password: The role of cognitive load in information security training. In Proceedings of IEEE International Conference on Intelligence and Security Informatics, 2009. ISI'09; 2009. IEEE.
[72] Cipresso P, Gaggioli A, Serino S, Cipresso S and Riva G. How to Create Memorizable and Strong Passwords. Journal of Medical Internet Research 2012; 14(1): e10. PMID: 22233980
[73] Blonder GE. Graphical password, Patent #5,559,961. USA: Lucent Technologies Inc. Murray Hill, NJ, USA; 1996.
[74] Pickover C and Riekert R. Menu Icon with Hidden Geometrical Password IBM Technical Disclosure Bulletin 1990; 32(10B): 463-464.

Security Assessment of Eucalyptus' Web Management Interface

Aleksandar Donevski, Sasko Ristov, Marjan Gusev
University Ss Cyril and Methodius, FCSE,
Rugjer Boshkovic 16,
Skopje, Macedonia
Email: aleksandar.donevski@outlook.com, {marjan.gushev,sashko.ristov}@finki.ukim.mk

***Abstract*—Three approaches exist for a company to migrate its services in the cloud. The first is select the most appropriate commercial cloud provider, such as Microsoft, Amazon, Google or SalesForce, and to rent their resources and deploy its services. These cloud service providers The second approach is to build a private cloud with some of the open source cloud frameworks like Eucalyptus, OpenStack, OpenNebula or CloudStack. Finally, the third option is to build a hybrid cloud, i.e., to split the services and to migrate the confidential and private services in its own private cloud, which could be open source, while the public and non confidential services and data to migrate in some public commercial cloud. Commercial cloud providers encash for their resources as a service, but offer and guarantee secured resources and huge availability of minimum 99.9% through a SLAs (Service Level Agreements). On the other hand, building an open source private cloud is the lowest starting investment, which does not guarantee neither security nor availability. Therefore, it is important for a company to assess the security of the open source cloud frameworks, especially their web management interface since it is the front end application, through which a user can get an unauthorized access to the company's data or applications.**

Eucalyptus is one of the most common open source cloud frameworks for building IaaS (Infrastructure as a Service) private or hybrid clouds, which also has its own web management interface to manage the cloud resources. In this paper, we assess the security of the cloud framework web management interface of the newest version of the Eucalyptus cloud. The results of the security assessment analysis show that the cloud web management interface is vulnerable. We assess the security vulnerabilities and propose measures how to secure them.

***Index Terms*—Cloud Computing, Dashboard, Open Source, Vulnerability, Web ServicesCloud Computing, Dashboard, Open Source, Vulnerability, Web Services!**

I. Introduction

Migrating the services from on-premise to a public cloud reduces the company's cost, but leverages the security risks. If the security is more important than elasticity for a given company, then the solution is to build a private cloud. This can be based either on a commercial cloud solution, such as IBM or VMware, or on one of the most common open source cloud softwares, such as OpenStack [1], Eucalyptus [2], OpenNebula [3], CloudStack [4] etc. However, the security of cloud platform is not guaranteed. For example, Dabrowski and Mills [5] demonstrated that open-source clouds can be attacked to degrade the availability of virtual machines.

Open source cloud is more and more used as a framework for building a private or public cloud. Eucalyptus is one of the major open source cloud software platforms designed to be a counterpart to commercial Amazon's EC2 cloud [6], [7].

All open source cloud frameworks have different purposes. For example, Eucalyptus fits well to build public cloud services (IaaS) with homogeneous pool of hypervisors, while OpenNebula [3] fits well for building private/hybrid cloud with heterogeneous virtualization platforms [6]. Voras et al. [8] devise a set of criteria to evaluate and compare the most common open source IaaS cloud solutions. Mahjoub et al. [9] compare the open source technologies to help customers to choose the best cloud offer of open source technologies.

In this paper, we assess the vulnerabilities of the Eucalyptus web graphical interface for administration, i.e., its dashboards. The rest of the paper is organized as follows. Section II presents the related work in the area of open source cloud security. In Section III, we briefly describe the Eucalyptus cloud architecture, its components and services, while in Section IV we present the assessment environment and the tool that is used for the assessment. The security assessment analysis of Eucalyptus administration service vulnerabilities are presented and discussed in Section V. Finally, Section VI concludes the work and presents future work.

II. Related Work

Virtualization and multitenancy generate additional security challenges and vulnerabilities in the cloud compared to the on-premise solutions. Grobauer et al. [10] define indicators if a particular vulnerability is cloud specific to distinguish general security issues from cloud-specific security issues.

Several papers propose methodologies and evaluate the security of cloud solutions and service layers. Chen and Yoon [11] propose a framework of checklist of IT auditing to secure the cloud. However, third party auditor should not bring new vulnerabilities towards data privacy [12]. Ristov et al. [13] evaluate ISO 27001 [14] control objective importance for on-premise and the cloud service layers IaaS, PaaS (Platform as a Service) and SaaS (Software as a Service).

The security of open source cloud frameworks is also assessed. Gusev et al. [15] analyzed the security vulnerabilities of Eucalyptus cloud from inside and outside and determine that the Eucalyptus server nodes are more vulnerable from inside, rather than from outside of the cloud, while the virtual

machine instances have the same security vulnerability risks both from inside and outside.

Ristov et al. [16] assessed the security vulnerabilities of OpenStack using one network configuration. They proved that OpenStack cloud multi-tenant environment raises new security vulnerability risks from inside the cloud, both for the tenants and the OpenStack cloud provider. Donevski et al. [17] evaluated that segregating the networks in OpenStack cloud rises new security risks, but only for CSP from inside, which need to be patched by the cloud software vendor. The vulnerabilities remain the same for the tenants.

The security of the commercial clouds is also analyzed. Tajadod et al. [18] compared the security of Microsoft's Azure platform and Amazon's Elastic Compute (EC2) cloud architectures and determined that both CSPs have significant security issues, with Microsoft offering a better level of data security than Amazon does.

In this paper we are focused on the security of Eucalyptus' dashboard on the application layer.

III. Eucalyptus Cloud Architecture

Open source clouds have similar architecture [19]. Although the architecture designs consist of several different parts, they are based on two main parts: *Cloud controller* and *Cloud node*. The former controls the system and network, schedules the execution of instances and offers a web interface for management of the whole cloud, while the latter hosts the instances of VMIs. In this section, we briefly describe the Eucalyptus cloud architecture, its components and services.

Eucalyptus is an open source cloud software that allows the users to build public, private and hybrid clouds. The Eucalyptus cloud software platform is developed to support highly scalable, modular and distributed platforms. It offers the customers a possibility to manage VMIs, volumes and snapshots. Nurmi et al. [20] outline the basic principles and important operational aspects of the Eucalyptus cloud solution.

Figure 1 presents the architecture of the Eucalyptus cloud. It consists of six components (five essential and one optional component), separated in three logical tiers:

- *Cloud Tier* - which consists of the following two components:
  - *Cloud Controller (CLC)* - acts as front point of the cloud which offers the web administrative interface for managing the whole cloud and EC2-compatible SOAP and Query interfaces. Also, the CLC makes high-level resource-scheduling decisions, makes requests to all of the available Cluster Controllers (CCs) and is responsible for exposing and managing the underlying virtualized resources (servers, network and storage). This component handles the authentication, accounting, reporting and quota management and communicates with all other cloud components except the Node Controllers (NCs);
  - *Walrus* - offers a persistent storage to all of the installed VMIs in the cloud. This component can contain images of VMIs, volume snapshots, application and user data, and can be used as a Storage-as-a-Service solution;
- *Cluster Tier* - which consists of the following two mandatory and one optional components:
  - *CC* - schedules a certain VMIs execution on specific node and manages the networks. This component communicates with the CLC, SCs and NCs;
  - *Storage Controller (SC)* - offers an option for managing the EBS (Elastic Block Store) volumes and snapshots of VMIs in the cloud. This component functions as Amazon's EBS [21]. It communicates with the CC and NCs;
  - *VMware Broker* - is an optional Eucalyptus component and is out of the scope of this paper. It is activated only in versions of Eucalyptus with VMware support which enables Eucalyptus to deploy VMIs on VMware infrastructure elements;
- *Node Tier* - which consists of the following component:
  - *NC* - controls activities of VMIs and manages the virtual network endpoint. This component communicates only with the CC and interacts with the host OS and the installed hypervisor..

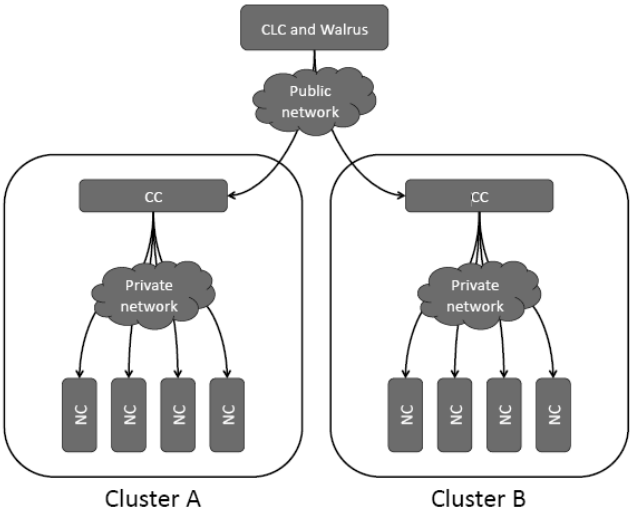


Fig. 1. The Eucalyptus Network [20]

Each component is a stand-alone web service and the communication between the cloud components are based on the SOAP protocol with WS-security extension to add end-to-end security.

The goal of this paper is the security assessment of the web management interface, which is depicted in Figure 2. It allows the administrators to monitor the activated instances, create key pairs, manage the images, etc.

IV. Assessment Environment and Tool

This section presents the assessment environment and tool used in the assessment, that is, the practical deployment

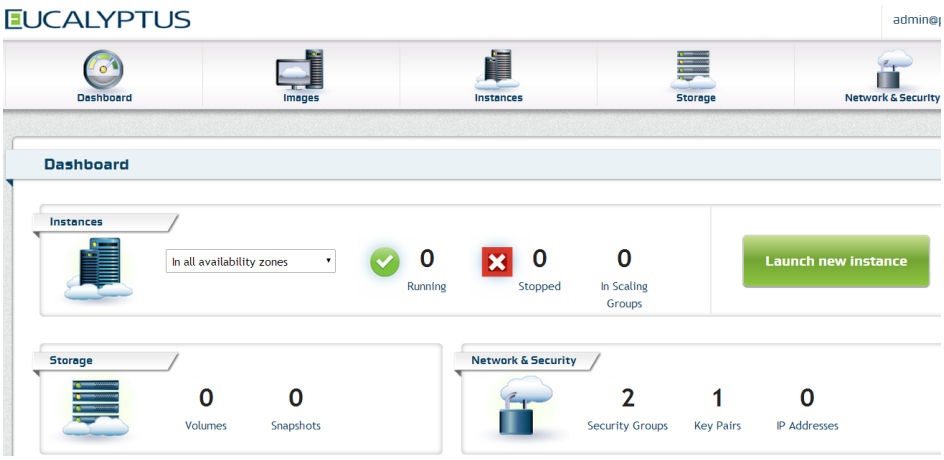


Fig. 2. The Eucalyptus web management interface

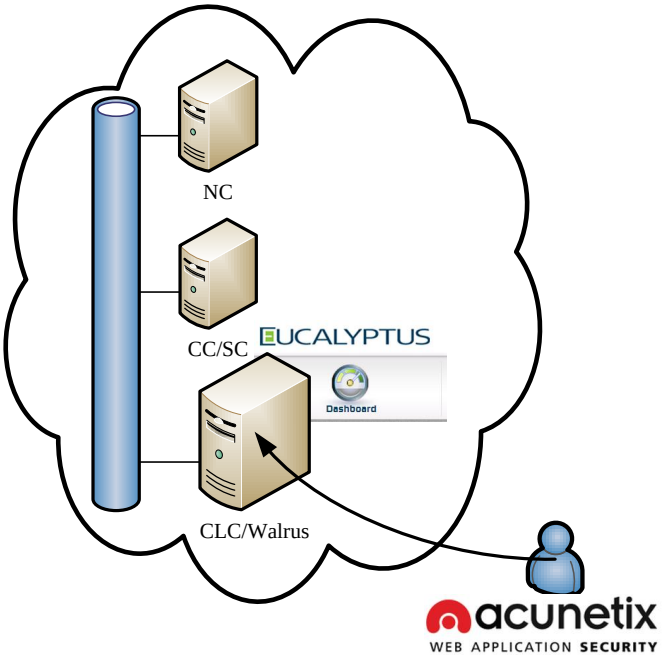


Fig. 3. The assessment environment

of Eucalyptus cloud. The research is based on assessing the security vulnerabilities of web management interface of the Eucalyptus' cloud framework, i.e., the Eucalyptus' User Console.

A. Assessment Environment

The Eucalyptus cloud deployed on three physical servers, as depicted in Figure 3. Each server is installed with Ubuntu Server 12.04 operating system. The web management interface is installed on the the cloud controller.

B. Assessment Tool

The security of cloud framework is assessed using the Acunetix Web Vulnerability Scanner [22] version 8.0. Acunetix scanner is pioneer, worldwide leader and one of the most advanced penetration tools for testing the web application security. Each vulnerability is rated as high-, medium- and low-severity type vulnerability. High-severity vulnerabilities are categorized as the most dangerous, which put a site at maximum risk for hacking and data theft [23]. Medium-severity vulnerabilities are caused by server misconfiguration and site-coding flaws, which facilitate server disruption and intrusion [23]. Low-severity vulnerabilities are derived from lack of encryption of data traffic, or directory path disclosures [23].

Acunetix checks the vulnerability realizing multiple pre-defined checks for the following vulnerabilities: Application Error Message Checks, File Inclusion, Blind SQL injection, Full Path Disclosure, CGI Testing, LDAP Injection, Code Execution, PHP Code Injection, Cookie Manipulation, Remote XSL inclusion, CRLF Injection, SQL Injection, Cross Frame Scripting (XFS), Script Source Code Disclosure, Cross-Site Scripting (XSS), URL Redirection checks, CSRF Injection, XPath Injection, Directory Traversal, and Directory File Checks. It is also a port scanner and enumerates the discovered open services for well-known vulnerabilities.

V. The Results of the Security Assessment

In this section, we present the results of the security assessments made on the Eucalyptus cloud services.

The security assessment of Eucalyptus detected 4 high-severity, 4 medium-severity and 4 low-severity type vulnerabilities.

XSS vulnerability is detected on 4 pages which allows an attacker to send malicious code (Javascript, VBScript, ActiveX HTML or Flash) to another user. The Eucalyptus developers must redevelop the vulnerable pages to filter metacharacters from user input.

Table I summarizes the medium-severity type vulnerabilities. 3 Eucalyptus directories display the list of files, but this is a false positive vulnerability. An attacker can inject plain text into the protocol stream which could allow him to issue

TABLE I
DETECTED MEDIUM-SEVERITY EUCALYPTUS' WEB MANAGEMENT INTERFACE VULNERABILITIES

Vulnerability Type	Assessed Vulnerabilities
Directory Listing	3
TLS1/SSLv3 Renegotiation Vulnerability	1
Total	**4**

TABLE II
DETECTED LOW-SEVERITY EUCALYPTUS' WEB MANAGEMENT INTERFACE VULNERABILITIES

Vulnerability Type	Assessed Vulnerabilities
OPTIONS method is enabled	1
Login page password-guessing attack	1
Password field submitted using GET method	1
Sensitive page could be cached	1
Total	**4**

HTTP requests. The Eucalyptus developer should patch this vulnerability in future version.

Low-severity type vulnerabilities are summarized in Table II.

The OPTIONS method is also false positive since it depends on the web server configuration, and not on the developed software. The Eucalyptus is also unprotected by password-guessing attack. The password field is submitted through GET instead of POST method. Finally, caching should be prevented. The three latter vulnerabilities must be patched by the Eucalyptus' developers.

VI. CONCLUSION AND FUTURE WORK

We assessed the security of Eucalyptus' web management interface using web vulnerability scanner. The experiments addressed the security vulnerabilities of web site administration. The assessment detected some false positive vulnerabilities which depend on web server configuration rather than cloud softwares.

The results of the assessments show that the Eucalyptus' cloud software is vulnerable. That is, we detected 4 high-, 4 medium- and 4 low-severity vulnerabilities. Although some of the detected Eucalyptus' web management interface vulnerabilities are false positive, all 4 high-, 1 medium- and 3 low-severity must be redeveloped by the Eucalyptus' developers.

In this paper we assess the vulnerability of Eucalyptus' cloud web management interfaces. Our future work will be focused on analyzing the other internal cloud services of Eucalyptus, as well as other open source clouds in order to help the clients to evaluate the security of the most common open source cloud frameworks, and select the most appropriate cloud framework to build their private or hybrid IaaS cloud.

REFERENCES

[1] OpenStack. (2012) Openstack cloud software. [retrieved: March, 2013]. [Online]. Available: http://openstack.org

[2] Eucalyptus. (2013) Eucalyptus cloud. [retrieved: March, 2013]. [Online]. Available: http://www.eucalyptus.com/

[3] OpenNebula. (2012) Opennebula cloud software. [retrieved: March, 2013]. [Online]. Available: http://Opennebula.org

[4] CloudStack. (2012) Cloudstack opens source cloud computing. [retrieved: March, 2013]. [Online]. Available: http://cloudstack.org

[5] C. Dabrowski and K. Mills, "Vm leakage and orphan control in open-source clouds," in *Proceedings of the 2011 IEEE Third International Conference on Cloud Computing Technology and Science*, ser. CLOUDCOM '11. IEEE Computer Society, 2011, pp. 554–559.

[6] T. D. Cordeiro, D. B. Damalio, N. C. V. N. Pereira, P. T. Endo, A. V. de Almeida Palhares, G. E. Goncalves, D. F. H. Sadok, J. Kelner, B. Melander, V. Souza, and J.-E. Mangs, "Open source cloud computing platforms," in *Proceedings of the 2010 Ninth International Conference on Grid and Cloud Computing*, ser. GCC '10. IEEE Computer Society, 2010, pp. 366–371.

[7] P. Sempolinski and D. Thain, "A comparison and critique of Eucalyptus, openNebula and Nimbus," in *Cloud Computing Technology and Science (CloudCom), 2010 IEEE Second International Conference on*, 2010, pp. 417–426.

[8] I. Voras, B. Mihaljevic, and M. Orlic, "Criteria for evaluation of open source cloud computing solutions," in *Information Technology Interfaces (ITI), Proceedings of the ITI 2011 33rd International Conference on*, june 2011, pp. 137 –142.

[9] M. Mahjoub, A. Mdhaffar, R. B. Halima, and M. Jmaiel, "A comparative study of the current cloud computing technologies and offers," in *Proceedings of the 2011 First International Symposium on Network Cloud Computing and Applications*, ser. NCCA '11. IEEE Computer Society, 2011, pp. 131–134.

[10] B. Grobauer, T. Walloschek, and E. Stocker, "Understanding cloud computing vulnerabilities," *Security Privacy, IEEE*, vol. 9, no. 2, pp. 50 –57, march-april 2011.

[11] Z. Chen and J. Yoon, "It auditing to assure a secure cloud computing," in *Proceedings of the 2010 6th World Congress on Services*, ser. SERVICES '10. IEEE Computer Society, 2010, pp. 253–259.

[12] C. Wang, S. Chow, Q. Wang, K. Ren, and W. Lou, "Privacy-preserving public auditing for secure cloud storage," *Computers, IEEE Transactions on*, vol. 62, no. 2, pp. 362 –375, feb. 2013.

[13] S. Ristov, M. Gusev, and M. Kostoska, "Cloud computing security in business information systems," *International Journal of Network Security & Its Applications (IJNSA)*, vol. 4, no. 2, pp. 75–93, 2012.

[14] ISO/IEC, "ISO/IEC 27001:2005, Information Security Management Systems - Requirements," 2012. [Online]. Available: http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=42103

[15] M. Gusev, S. Ristov, and A. Donevski, "Security vulnerabilities from inside and outside the eucalyptus cloud," in *Proceedings of the Sixth Balkan Conference in Informatics*, ser. BCI '13. ACM, 2013.

[16] S. Ristov, M. Gusev, and A. Donevski, "Openstack cloud security vulnerabilities from inside and outside," in *Proceedings of The 4th International Conference on Cloud Computing, GRIDs, and Virtualization (CLOUD COMPUTING 2013)*, Valencia, Spain, May 2013, pp. 101–107.

[17] A. Donevski, S. Ristov, and M. Gusev, "Security assessment of virtual machines in open source clouds," in *MIPRO, 2013 Proceedings of the 35th International Convention, IEEE Conference Publications*, Opatija, Croatia, 2013, pp. 1375–1380.

[18] G. Tajadod, L. Batten, and K. Govinda, "Microsoft and amazon: A comparison of approaches to cloud security," in *Cloud Computing Technology and Science (CloudCom), 2012 IEEE 4th International Conference on*, Dec. 2012, pp. 539 –544.

[19] C.-H. Ng, M. Ma, T.-Y. Wong, P. Lee, and J. Lui, "Live deduplication storage of virtual machine images in an open-source cloud," in *Proceedings of the 12th ACM/IFIP/USENIX international conference on Middleware*, ser. Middleware'11. Berlin, Heidelberg: Springer-Verlag, 2011, pp. 81–100.

[20] D. Nurmi, R. Wolski, C. Grzegorczyk, G. Obertelli, S. Soman, L. Youseff, and D. Zagorodnov, "The eucalyptus open-source cloud-computing system," in *Proceedings of the 2009 9th IEEE/ACM International Symposium on Cluster Computing and the Grid*, ser. CCGRID '09. IEEE Computer Society, 2009, pp. 124–131.

[21] Amazon. (2012) Amazon elastic block store. [retrieved: March, 2013]. [Online]. Available: http://aws.amazon.com/ebs/

[22] Acunetix Ltd., "Acunetix web vulnerability scanner," 2012, [retrieved: March, 2013]. [Online]. Available: http://www.acunetix.com/

[23] ——, "Acunetix web vulnerability scanner manual," Feb. 2013, [retrieved: April, 2013]. [Online]. Available: http://www.acunetix.com/vulnerability-scanner/wvsmanual.pdf

CROFlux –Passive DNS Method for Detecting Fast-Flux Domains

Toni Gržinić*, Darko Perhoč*, Marko Marić*, Filip Vlašić* and Tibor Kulcsar*
* Croatian Academic and Research Network – National CERT
{tgrzinic, dperhoc, mmaric, fvlasic, tkulcsar}@carnet.hr

Abstract - In this paper we present our approach to fast flux detection called CROFlux that relies on the passive DNS replication method. The presented model can significantly reduce the number of false positive detections, and can detect other suspicious domains that are used for fast flux. This algorithm is used and implemented in Advanced Cyber Defense Centre – a European project co-funded by the European Commission.

I. INTRODUCTION

Nowadays information systems are threatened by various types of attacks. If we look at the most popular threats in information security like distributed denial-of-service, malware distribution and spam, we can notice that they can be conducted using botnets. Botnets are networks of infected computers that are usually managed from a central point also called command and control center (C&C). These networks use various techniques for hiding their presence or amplifying the damage of their attacks. One commonly used technique is fast flux which abuses DNS (Domain Name System). DNS records of flux domains are frequently changed; in such a way that authoritative DNS returns different bots in various time intervals. Flux domains are usually used for hiding botnet command and controls servers, hosting malware delivery sites or for hosting phishing pages. Fast flux facilitates load balancing and proxy redirection that make malicious servers more resistant against detection or takedown attempts.

There are two different types of fast-flux: single-flux and double-flux. Single fast-flux is a technique where multiple bots, within the botnet, are registered and deregistered in a DNS A record for a single fully qualified domain. In combination with round robin DNS algorithm that has very short TTL values (e.g. 3 minutes), this technique produces a constantly changing list of destination IP addresses for the same domain. Resolved bots act as proxies for malware or phishing delivery sites, these sites are popularly called "mothership" nodes.

Double-flux technique, unlike the single flux, provides an additional layer of redundancy. Specifically, in double flux, both the DNS A record and authoritative name server (NS) record are changed. The used authoritative name server administers a fast-flux DNS zone, with all domains and subdomains in it.

A query that has been sent by client's stub DNS resolver is received by DNS recursor (caching DNS server) which in turn is recursively sent downwards from the root server to the last authoritative name server. The last authoritative name server is a bot used for a double flux scheme. Periodically, other bots take a role of authoritative name servers, thus the corresponding NS records are changed.

After receiving the query the authoritative name server forwards it to the mothership node requesting required information (e.g. domain's A record). The mothership node sends a response to a flux name server which forwards the response back to the DNS recursor. IP addresses in DNS responses belong to bots. Motherships are second layer C&C servers in a fast-flux botnet scheme, these nodes typically host both DNS and HTTP services to accomplish their single and double-flux scheme role.

II. RELATED WORK

Weimer [1] proposed the passive DNS replication (pDNSR) method whose purpose is to extract resource records from response packets of authoritative DNS servers. Using the paired query and response, pDNSR builds a replica of DNS zones that are available from the data. To preserve privacy pDNSR uses only queries between DNS servers. A similar approach was used by Zdrnja and his colleagues [2]; they deployed a custom solution for collecting domains. Analyzed data contained: misspelled domains, fast flux domains and spam related domains.

Holz et al [3] proposed a Flux score to differentiate malicious domains from benign ones. Flux score uses features derived from active DNS queries. These features are: number of unique A records in DNS lookups, number of name servers in lookups and number of unique autonomous numbers (ASN) from type A records. Their system sends several active queries to collect relevant data about domains. After the first query system waits for TTL to expire, and then it sends another query. Using popular spam domains they obtained an optimal hyper plane that separates flux domains from benign ones.

Fast flux domains are classified using Flux score as following:

$$1.32 \cdot \text{numIP} + 18.54 \cdot \text{numASN} + 0.0 \cdot \text{numNS} > 142.38 \quad (1)$$

Where the parameters are:

- *numIPs* – number of distinct IP addresses
- *numASN* – number of unique autonomous system numbers of IP addresses
- *numNS* – number of used name servers for the given domain (ignored because of the zero coefficient in the Equation 1)

Holz et al also realized that Fast-Flux domains have similar characteristics like content delivery networks (CDN) or round robin DNS's that are used to increase website availability. Thus, CDNs are usually classified from the aforementioned algorithm as false positive flux domains. A handful implementation of the algorithm can be found on GitHub [9].

Fluxy [4] uses a similar approach; their authors implemented an adapted equation to compute the Fast-Flux score. In addition to previously mentioned Holz et al method, Fluxy uses reverse DNS queries to detect dynamic IP addresses probably belonging to infected end-users.

EXSPOSURE [5] is a system based on a passive DNS replication. Used data was provided from Security Information Exchange (SIE), which contained response data from authoritative servers from North America and Europe. From this vast amount of data, approximately 4.8 million domains collected through 2.5 months, they extracted time based features of collected domains and analyzed them with a change point detection algorithm CUMSUM. CUMSUM is used to detect short lived domains and domains with repeating patterns. Their classification model was built using a C4.5 decision tree classifier and was trained with features of popular malicious domains. Beside the time related features, they used other groups of features based on: DNS answers, TTL values and domain name characteristics. An extensive list of features can be found in the paper.

Fluxbuster [6] is another system which uses passive DNS replication. Fluxbuster prunes the collected domains based on conservative criteria and outputs candidate domains; pruning procedure is based on data returned in responses (TTL values and diversity of servers' networks). After that, candidate domains are clustered using a hierarchical clustering algorithm. Classification of these domains is based on features collected by the passive analysis. They introduced some novel features like growth ratio of IP addresses and networks. Given clusters are classified using a decision tree classifier. The used training set was obtained using a semi-manual process, where authors combined known malware domains and manual corrections in labeling.

III. CROFLUX

A. System Architecture

The developed system CroFlux uses a Farsight's SIE framework [1]for DNS data collection. The framework is similar to the solution proposed by Weimar [1] which reconstructs a partial zone replica preserving users' privacy.

The SIE sensor, as shown in Fig. 1, captures DNS messages from the connected authoritative DNS server.

Processing stage returns queried domains and IP addresses from a particular time interval. We can look at those results as aggregated responses to specific domains. SIE outputs the processed DNS messages in NMSG format. NMSG is a wire format optimized for storing and transmitting binaries over UDP [10]. NMSG stores the following information: domain name, DNS query type, name server that sent the query, authoritative name server that answered the query, IP addresses pointing to domain, record class and record type.

B. Fast Flux Detection

One limitation of designing a fast flux detection algorithm is the passive nature of DNS data collecting. So the classification process needs to rely on data gathered by completely unpredictable timing of DNS queries sent by various users. This posed a problem, since we need as many as possible resolved IP addresses of fast flux domains which enhance the detection result. We did not use any active DNS data checking like described in [3] and [4], because we wanted to avoid potential exposure to botnet operators. This limitation may become negligible if passive DNS replication method is installed on a big network with many users.

Another problem that may occur is detection of benign domains as flux (false positives). False negatives, i.e. fail to detect fast flux domain, are of a less pertinence. As the result of fast-flux detection algorithm is considered to be a part of public services and actions, false positive results may produce certain inconvenience, distortion of reputation or legal issues. Thus false positive elimination presents a problem of current fast flux detection solutions [8].

CROFlux algorithm runs through three phases. Namely, as shown on Figure 1:

(a) Prefiltering/pruning of collected domains
(b) Candidate domains clustering
(c) Detection of fast flux clusters

a) Prefiltering fast flux candidates

SIE outputs NMSG files that contain captured domains which are later filtered. This filtering stage is similar to the pre-filtering stage proposed by Perdisci [6]. On Figure 1 this step is called domain list pruning procedure.

[1] Farsight SIE framework is available at https://archive.farsightsecurity.com/

Selected candidate domains must meet the following requirements:

- Domain's time to live (TTL) should be below 3 hours (TTL<=3600 s)
- Minimum number of resolved IP addresses for the following domain |RS|>= 3, or if RS requirement is not met take only domains with a small TTL<=30 seconds. The number of the resolved IP addresses is aggregated from various DNS queries and depends on cache timeout setup on the SIE sensor.
- Diversity of networks for a domain should be div(RS)>0.333. To be a part of the same network, resolved IP addresses must have the same /16 prefix. Diversity is calculated as entropy of /16 networks of resolved IPs for a given domain.

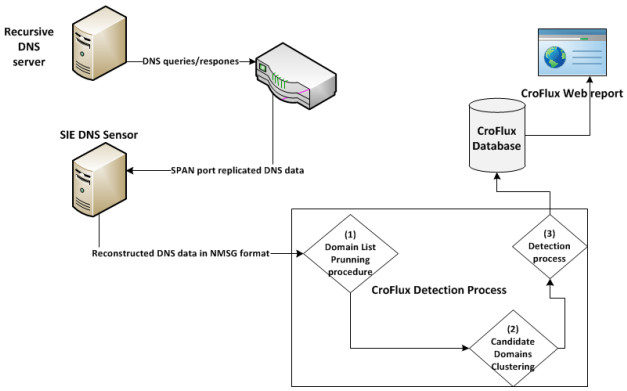


Figure 1. – Overview of CroFlux architecture

After the pruning procedure we have a dictionary with candidate flux domains, where domain names are keys and discovered IP addresses are values. Periodically, measured in time of few minutes, all dictionary records are added to a global dataset. Candidate domains from the global dataset are evaluated from: number of distinct IP addresses, number of distinct AS origins of those addresses and number of estimated dynamic IP addresses of those domains.

Behind flux domains are infected computers or bots which are spread across multiple ISP networks over the world, mostly linked with a broadband connection. The following method is used to mark domain as fast flux candidate [4]:

$$\text{Fluxy Score} = 1.32 * \text{numIPs} + 40.0 * \text{numSPLink} + 20.0 * \text{numASNs} \quad (2)$$

Where the parameters are:

- *numIPs* – number of distinct IP addresses
- *numSPLink* – number of dynamic IP addresses
- *numASNs* – number of distinct autonomous systems

We used a conservative threshold score of 450 for marking domain as a fast flux candidate. For comparison, Fluxy authors and Holz and colleagues used a lower threshold score 142.38 [3] [4].

b) Clustering

Domain clustering is a process in which we group domains that are operated in the same network. As we may expect, hosting providers, content delivery networks (CDN) and botnets that use fast flux are those types of networks. In this step we try to target more closely fast flux domains. Number of common IP addresses is used as clustering criteria. This number of overlapping IP addresses is empirically tuned. Every week clusters are expanded or new ones are created, based on newly collected DNS data.

c) Determining fast flux clusters

The final step is to determine fast flux clusters from all candidate clusters. Fast flux domains are often used for malware delivery or hosting phishing domains so we can say that they act as proxies. The way to determine fast flux cluster is to compare candidates' clusters against publicly available and private malware lists collected by internally used tools. Based on the number of malicious domains in cluster we label that cluster as fast flux cluster. Otherwise, cluster is labeled as some other type (CDN, hosting providers etc.). We use a minimum number of malicious domains to classify a candidate cluster as fast flux. This number is tunable and is found empirically. Aforementioned blacklists are available via multiple online services [11] [12] [13] [14] [15] [16] [17]

IV. RESULTS

DNS data has been replicated from August 2013 till January 2014, using a CARNet recursor. CroFlux has collected 427.502 de-duplicated second level domains. From these domains we have near 265 malware domains with flux characteristics. In Figure 2 we can see the distribution of top level domains with flux characteristics. Collected flux domains are primarily from Russia (.ru) and Soviet Union (.su) domain which is still used.

In Figure 3 are presented calculated Flux scores of CroFlux's detected flux domains. Scores are shown on logarithmic axes, and as we can see collected flux domains mostly have higher Flux scores. The bright dashed line represents the Fluxy's threshold for flux domains. Also, we can notice that detected flux domains tend to have higher flux scores.

Fluxy score, similarly to Holz score, generates many false positive results [9], i.e. benign domains labeled wrongly as flux. It is also useful to note that given classification hyperplane weights (calculated in [3] and [4]) are trained on elder flux domains - from circa 2008.

Many collected legitimate sites have a high score. They also have similar characteristics to flux domains like: short TTL, many IP addresses scattered through geographically distant autonomous systems etc. Examples of these sites are: content delivery networks like Akamai, cloud services like Amazon AWS, sites used for time synchronization NTP and video streaming services like Netflix and Hulu.

Similarly, distributed peer to peer networks like Bitcoin use fallback nodes hardcoded in Bitcoin miner applications that use a short TTL (60 seconds) and contain many peer nodes with dynamic IP addresses.

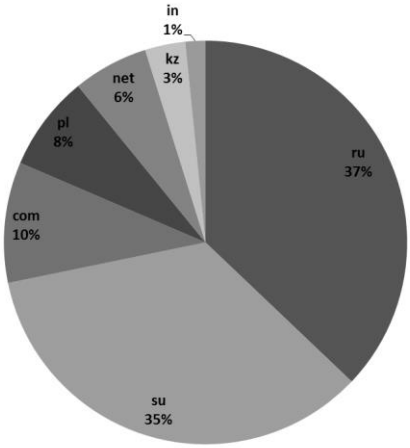


Figure 2 – TLD Distribution of detected flux domains

Using our data we recalculated the Fluxy separating hyperplane. Because of the difference between active querying and passive replication we used the last two inserts (of the specific domain) for calculating feature weights.
On a dataset with 9300 benign domains and 100 flux domains we performed a 10-fold cross validation, like in [3]. We obtained feature weights using the Support Vector Machine method with a linear kernel.
This resulted with an optimal hyperplane for flux domains:

$$-0.2853 \cdot \text{numIPs} + 0.4076 \cdot \text{numASNs} + 0.1837 \cdot \text{numSPLink} > 1.5704 \quad (3)$$

Benign domains were classified with a precision of 99% and flux domains with a precision around 91%.
Detected flux domains contained many double flux domains which rotate name servers and IP addresses in responses.
An example of double flux is the following:

Ns1.nulled-db.com
Ns2.nulled-db.com
…
NsXX.nulled-db.com

We also noted multiple level flux domains, like mentioned in [8]. These domains have multiple levels of name servers, where there is an overlap between the existing IP addresses.
An interesting type of domains are domains with a short TTL (less than 10 seconds), that rotate several IP addresses in every query. Usually, they return a single IP address. These domains tend to evade popular approaches to flux detection.

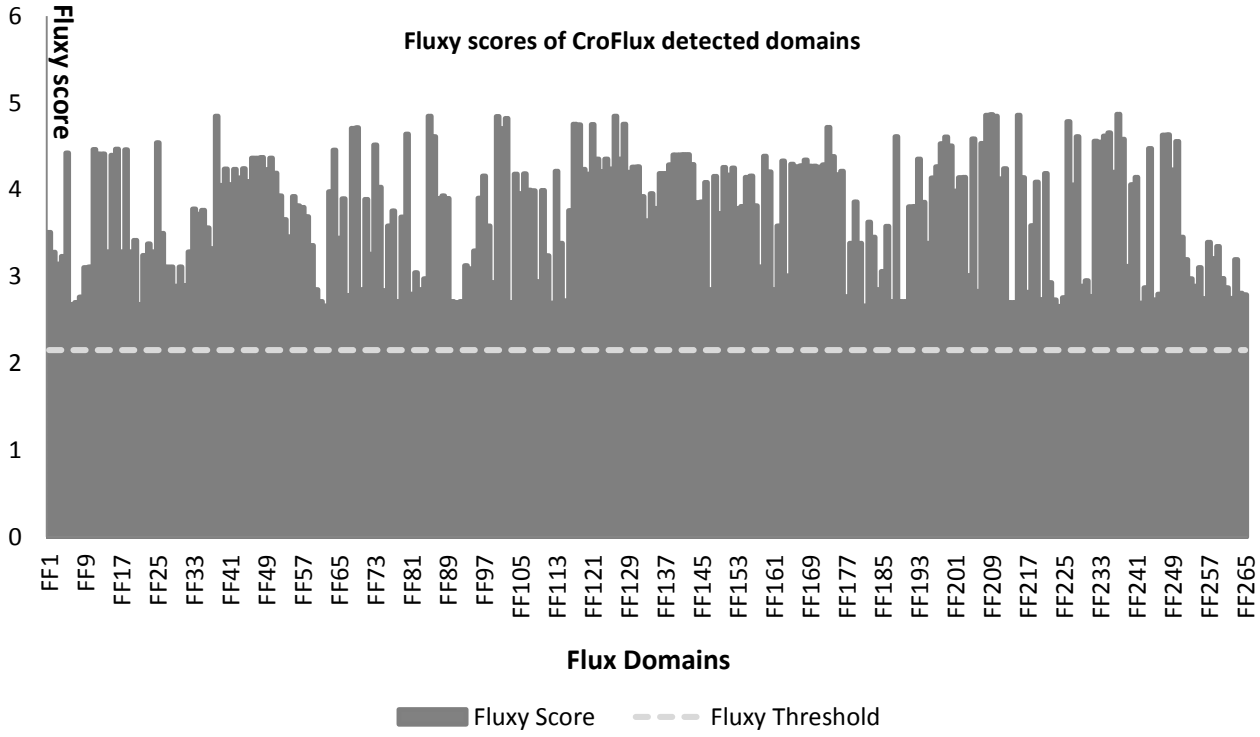


Figure 3 – Fluxy scores of collected Fast Flux domains

V. CONLUSION

In this paper we introduced CROFlux, a method for detecting fast flux domains in the wild. Our approach leverages on publicly available knowledge about malicious delivery sites.

Our system can continuously detect unknown fast flux domains with an advantage that reduces false positives, relaying on detected domains with flux characteristics which are usually used for sharing malware.

With our conservative approach we get a list of real flux domains, and we avoid reporting benign domains with similar characteristics.

We also have similar lists of malicious domains that are collected using other systems developed in house and used in an EU funded project - Advanced Cyber Defence Centre (ACDC). CroFlux has been developed within the ACDC project, and temporally is tested on one DNS recursor hosted in our network. We have plans to expand the number of monitored DNS servers, in order to gain more flux domains. At the moment information about detected flux domains is sent to a centralized ACDC database called central clearing house. Further actions like domain takedown will be done in collaboration with law enforcement agencies and domain registrars with the aim of reducing botnets' damage.

As we can see from the results, botnet operators constantly work on new techniques to avoid detection and takedown of malicious domains. In order to have an updated classification, we updated the classifier using newly collected domains. Edge cases like small TTL domains with one IP address in response cannot be detected using an active querying approach, but passive replication helps on aggregating resolved IP addresses if this type of domains is queried enough times.

ACKNOWLEDGMENT

Writing of this paper is funded by the European Commission. The views and opinions expressed in this paper are those of the authors and do not necessarily reflect the opinions and views of the European Union.

Also, we would like to thank anonymous reviewers whose comments helped us to improve this article.

REFERENCES

[1] Weimer, Florian. Passive DNS replication., FIRST Conference on Computer Security Incident. 2005.

[2] Zdrnja, Bojan, Nevil Brownlee, and Duane Wessels. Passive monitoring of dns anomalies, Detection of Intrusions and Malware, and Vulnerability Assessment. Springer Berlin Heidelberg, 2007. 129-139.

[3] Thorsten Holz, Christian Gorecki, Konrad Rieck, Felix C Freiling Detection and mitigation of fast-flux service networks, Proceedings of the 15th Annual Network and Distributed System Security Symposium (NDSS'08). 2008.

[4] Wegrzynowicz Patrycja, Jantura Jaroslaw, Detection of Fast Flux Botnets, 21th CENTR Technical Workshop,2009, Lisbon, Portugal

[5] Bilge, L., Kirda, E., Kruegel, C., & Balduzzi, M. EXPOSURE: Finding Malicious Domains Using Passive DNS Analysis, NDSS 2011.

[6] Perdisci, Roberto, Igino Corona, and Giorgio Giacinto. Early Detection of Malicious Flux Networks via Large-Scale Passive DNS Traffic Analysis, Dependable and Secure Computing, IEEE Transactions on 9.5 (2012): 714-726.

[7] Farsight SIE, https://archive.farsightsecurity.com/, accessed 29.1.2014.

[8] Wei Xu. Xinran Wang, Huagang Xie, New Trends in FastFlux Networks, Black Hat USA 2013

[9] Github, PFFDetect - https://github.com/z0mbiehunt3r/pffdetect, accessed 6.2.2014.

[10] Robert Edmonds – ISC Passive DNS Architecture, Internet Systems consortium, 2012

[11] Zeus Tracker, https://zeustracker.abuse.ch/. Accessed 2.4.2014.

[12] SpyEye Tracker, https://spyeyetracker.abuse.ch/, Accessed 2.4.2014.

[13] Palevo Tracker, https://palevotracker.abuse.ch/, Accessed 2.4.2014.

[14] SpamHaus, http://www.spamhaus.org/, Accessed 2.4.2014.

[15] Feodo Tracker, https://feodotracker.abuse.ch/, Accessed 2.4.2014.

[16] Malware Domain List, http://www.malwaredomainlist.com/, Accessed 2.4.2014.

[17] Arbor Atlas, http://atlas.arbor.net/, Accessed 2.4.2014.

A survey of static code analysis methods for security vulnerabilities detection

Melina Kulenovic, Dzenana Donko
Faculty of Electrical Engineering, University of Sarajevo
Sarajevo, Bosnia and Herzegovina
E-mails: mk14746@etf.una.ba , ddonko@etf.unsa.ba

Abstract—Software security is becoming highly important for universal acceptance of applications for many kinds of transactions. Automated code analyzers can be utilized to detect security vulnerabilities during the development phase. This paper is aimed to provide a survey on Static code analysis and how it can be used to detect security vulnerabilities. The most recent findings and publications are summarized and presented in this paper. This paper provides an overview of the gains, flows and algorithms of static code analyzers. It can be considered a stepping stone for further research in this domain.

Keywords—static code analysis; security; vulnerability; survey;

I. INTRODUCTION

Millions of users accomplish their daily tasks using web and desktop applications on different electronic devices. The prevalence of software enforces the software industry to think of how to build quality in. Most aspects of software quality are related to the skills and knowledge of the development team [1]. Unfortunately, developers make mistakes that lead to vulnerable and defect software. Exploited security vulnerabilities lead to unreliable software that can become harmful for the user and software provider.

The CWE (Common Weakness Enumerations) initiative provides a unified, measurable set of software weaknesses which can lead to serious security vulnerabilities [2].The top 25 vulnerabilities list contains the most widespread and critical errors that can lead to security vulnerabilities that can be easy exploited by an attacker. Table I presents an overview and description of the most common and harmful attacks. CWE has documented more than 680 software weaknesses that can lead to impacts like modified data, read data, denial of service (DOS) attacks, resource consumption, execution of unauthorized code, gain of privileges, protection bypass or the hiding of activities.

Many security focused methodologies like the Open Software Assurance Maturity Model (Open SAMM) [3],The Building Security In Maturity Model (BSIMM) [4], or Microsoft's Security Development Lifecycle (SDL) [5] are designed to enforce developers to build security in. Still, the research provided by Cenzic showed that 99% of all reviewed applications in 2012 had at least one serious security vulnerability [6].Conveniently, tools that are able to scan millions of lines of source code and detect defects and vulnerabilities in relatively short time are improving rapidly. A variety of such utilities is now available as commercial or free tools that can be used either as a standalone tool or as a plug-in for a development framework. Automated code analysis tools can be utilized during the development phase and as part of the build process. The detection of defects in such an early phase can reduce costs significantly.

The main characteristics of static code analysis tools, their gains and flows are outlined in section II. Section II also provides information on how to measure the contribution of static analysis to vulnerabilities detection. Section III presents the techniques and algorithms static code analysis is based on. This section contains an overview of the most recent work and a summary of related researches. The Conclusion is presented in section IV.

Table I. Top 5 security attacks [2]

Name	*Description*	*Effects*
SQL injection	SQL – Injection is one of the major threats for data rich software. Every non validated textbox that utilizes user input and uses it in a SQL context can be used by an attacker to inject SQL commands.	Attackers use the commands to bypass authentication mechanisms or insert tainted data into the database.
OS command injection	Applications are considered vulnerable to the OS command injection attack if they utilize non validated user input in a system level command what can lead to the invocation of scripts injected by the attacker.	Attackers attempt to execute system level commands in order to achieve the desired goal.
Buffer overflow	Buffer overflows is an anomaly where a program, while writing data to a buffer, overruns the buffer's boundary and overwrites adjacent memory. It can be triggered by non-validated inputs that are designed to execute code.	Buffer overflow may result in erratic program behavior, including memory access errors, incorrect results, a crash, or a breach of system security.
Cross-site scripting	Cross-site scripting is typically found in Web applications and it enables attackers to inject client-side scripts into Web pages viewed by other users.	Attackers use it to bypass access controls, steal user sessions or to redirect users to a different website.
Missing authentication	Missing authentication is a security vulnerability that occurs in software that does not perform any authentication for functionalities that require a provable user identity or consume a significant amount of resources.	Exposing critical functionality. Provides an attacker with the privilege level of that functionality.

II. STATIC CODE ANALYSIS

Every software application has its business context, environment and purpose. Potential risks and threats should be defined in regards to the protected values, and the weaknesses prioritized accordingly. The Common Weakness Risk Analysis Framework (CWRAF) [7] and the Common Weakness Scoring System (CWSS) [8] help the community to focus on its risk-reduction activities. CWRAF supports the automatic selection and prioritization of relevant weaknesses, customized to the specific needs of the organization's business or mission. Such initiatives help developers and consumers introduce more robust and resilient software into their operational environments [9]. Once the business value has been linked with weaknesses it is easier to detect and fix the security vulnerabilities important for the application domain. Since manual code reviews are costly, automated tools seem to be a good solution. Automated tools can perform static or dynamic code analysis.

Static analysis tools detect vulnerabilities by scanning the applications source code and examining all possible execution paths and variable values without executing the application. This makes the process fast, repeatable and not dependent on the fact whether the program can be executed. Static code analysis can be applied during the development phase which can help programmers to find and fix defects and security lapses early. Static code analysis is related to the source code examination which makes it dependent on the programming language. One of the major roadblocks to the wide use of automated code analyzing tools is the fact that such tools tend to report false positives and false negatives [10]. False positives refer to results that report non existing security bugs, similar to false negatives which occur when tools don't report existing bugs. Unfortunately, false positives and false negatives seem inevitable if a reasonable amount of resources should be used for the analysis. Although, there is a theoretical possibility to create a tool that has almost no false negatives, close to zero false positives and that works for many plausible programs. Such tool would find errors by systematically exploring all possible program states with all possible inputs and looking for error states. In theory, this is possible because digital computers are ultimately finite state machines, but the number of states is so vast that such a tool would take an unreasonable amount of time to complete [11]. Since such a tool doesn't exist, the main goal is to achieve a tolerable false positive and false negatives rate and a satisfying execution time. Most static analysis tools are configurable; they allow users to define the set of vulnerabilities to search for. Static analysis tools present a list of warnings that have to be inspected manually, and classified as true weaknesses or false alarm – this process is known as *triaging* [10].

Dynamic analysis tools require an installed and configured application. The vulnerabilities detection is based on test scenarios and the findings are related to the current configured environment. Dynamic analysis tools are susceptible to false positives and false negatives also. However, such tools are agnostics in terms of the used programming language because they are utilized on object code, which makes them preferable in some situations.

The common known shortcomings of static code analysis often led to discussions about the benefit of such tools. The most effective tool is one that strikes the right balance among false positives, false negatives, and performance. The benefit of a static analysis tool can be measured in regards to the weaknesses reported.

Recall is defined as the probability that a tool will find a real defect [11].

$$R = \frac{TP}{TP + FN} \tag{1}$$

Precision is defined as the probability that a result corresponds to a real defect [11].

$$P = \frac{TP}{TP + FP} \tag{2}$$

TP represents the number of true positives found by a tool;*FN* represents the number of false negatives and *FP* the number of false positives reported by a tool.

Once the reports have been triaged and the true positives sorted out it is easy to calculate the precision measuring, but it is difficult to measure the recall value since it relies on the number of false negatives which cannot be determined accurately. Also, a tool can be very effective in finding one kind of vulnerabilities but it doesn't mean that it will have a satisfying recall and precision value for other types of vulnerabilities. There is a general understanding that the real measure of a static-analysis tool's usefulness is whether it can find a reasonable number of important defects in a program without overwhelming the user with useless reports or using an unreasonable amount of computing resources.

A step forward in improving user experience is a well-structured and intuitive UI that groups related defects. This could reduce the number of reports and suppresses repeated defect reports that can be addressed with one common fix. Some tools like Parfait provide an easy way to store and share defects [12]. Such reports can be reviewed by other team members or used for further bug prediction. Some results might be misjudged or discarded because of way they were presented. Distinguishing a true positive from a false positive takes some care, time and experience. Tools have to be intuitive, easy to install and forthcoming in the way they present the results or they won't be used, no matter how precise or fast they are.

III. HOW STATIC CODE ANALYSIS WORKS

Static code analysis is performed on a model of the code to be analyzed. The model usually consists of a set of intermediate representations such as symbol tables, abstract syntax trees, control-flow graphs, and the program call graph. This model is queried using different mechanisms like simple searches of data structures to very sophisticated algorithms based on advanced concepts such as dataflow analysis,

symbolic execution, abstract interpretation, or model checking [11]. Since the analyzing mechanisms aren't applied directly on the program but on a model of the program, false positives and false negatives arise because the model is almost never exact. The model of the program is most often either an over- or under-approximation. An exact model or over-approximation is often referred to as *sound* because it captures all possible aspects of the program's behavior. Such models will most likely find the majority of true positives but they will also tend to report false positives. An under-approximation is a model that fails to capture some relevant aspect of the program, and therefore a variety of false negatives are possible.

There are various techniques and mechanisms developed to detect security vulnerabilities in a program model. Table II presents the most common vulnerabilities and their classification based on the source of the security vulnerability. The following subsections outline in more detail the techniques and mechanisms that static code analysis is based on for the tree identified vulnerability classes.

Table II. Vulnerability classification[13]

Access Control Vulnerabilities	***Information Flow Vulnerabilities***	***API Conformance Vulnerabilities***
Broken access control	Un-validated input	Insecure storage
Broken authentication and session management	Cross-site scripting	
Application denial of service	Injection flaws	

A. Access control

The mechanisms for access control restrict access to security-sensitive resources based on a user's identity and membership in predefined groups [14]. Runtime authorization failures may result when an access control policy doesn't grant sufficient permissions to a user. Likewise, if an access control policy grants users unnecessary permissions, the system may be exposed to security attacks. There are two major approaches used in the Java and .Net Common Language Runtime platforms:

- Stack-based Access Control (SBAC)
- Role-based Access Control (RBAC)

SBAC systems ensure that only programs that satisfy a set of permission requirements gain access to restricted resources. When a program attempts to gain access to restricted resources the runtime system checks whether all callers on the thread's stack satisfy a set of permissions requirements. Therefore SBAC systems prevent a caller which is trying to gain access to restricted resources by invoking more trusted code. In SBAC systems permissions can be granted declaratively in a policy database. The permission requirements for a SBAC security policy depend of the dynamic program behavior. This can be quite complicated since several components are interacting in diverse contexts. The policy configuration can be determined through testing. Permissions are added iteratively until no failures occur. Even when all test complete successfully there is still a chance that some paths of execution were not covered which could lead to runtime authorization failures.

SBAC systems should always stick to the Principle of Least Privilege [15] without violating the functioning of the program. Fortunately SBAC systems provide *privilege asserting* [16]. During stack inspection, the runtime system must enforce that privilege-asserting code holds the necessary permissions, while the callers of the privilege-asserting code do not need to hold the permission. Although privilege-asserting code is necessary to enforce security policies, malicious clients can exploit the privilege-asserting mechanisms to introduce security holes. In order to prevent security vulnerabilities developers should verify that:

- The privilege is necessary
- The privilege is not redundant
- All security-insensitive code that can be moved out of privilege-asserting code is moved
- The privilege asserting code block does not contain more than one security sensitive operation
- The privilege asserting block must respect all integrity and confidentiality requirement concomitant with the security-sensitive operation

Papers [17]-[25]present some new concepts and algorithms related to SBAC implementation.

RBAC is a mechanism to restrict system access to authorized users. RBAC systems provide access control based on permissions and roles. Permissions represent the right to perform restricted operations. Roles represent a set of permissions that can be assigned to a user. Each user that is attempting to invoke an operation must be assigned to a role with that particular permission granted.

RBAC is an access control mechanism based on operations instead of resources, which has its challenges. In order to restrict access to sensitive data all operations accessing such data need to be identified. A resource based approach would be much more intuitive in this scenario. Also, there is no simple way of denying a user access to sensitive data. All operations that have directly or indirectly access to the particular resource must be denied for the role assigned to the user. Mapping the operations to the data they access would provide verification that the intended data based security policy does not conflict with the configured operation-based security policy [26] [27]. Another difficulty with RBAC is related to the invocation chain of an application. The invocation of an application entry point will only be successful when the user has all roles assigned that are necessary to perform all the operations transitively triggered by the application. Important is that a user shouldn't have any unnecessary roles assigned since that would be a violation of the principle of least privilege.

Papers [28]-[35] offer some interesting researches related to the implementation of RBAC mechanisms.

B. Information flow

The two main types of information flow vulnerabilities are integrity and confidentiality. The principle of noninterference dictates that no information flows from higher to lower security levels [36]. Contempt of this principle can lead to violation of the integrity and confidentiality of sensitive data. Vulnerabilities such as those caused by non-validated input and injection flaws constitute integrity violations. SQL-Injections and OS command injections are some of the most common attacks caused by using tainted data (data gathered from an untrusted source) in a trusted environment [37]. The requirement for confidentiality establishes that a value flowing out of a privilege-asserting block should remain confined inside that component unless a check has been performed to verify that the value can be safely released [13].The basic idea behind using static analysis for detecting information flow is to statically check that flow of information between variables in a program is consistent with the security labeling of variables. Each variable is labeled with a certain security level. If a variable x is used to derive, or influence, the value of another variable y, there is potential information flow from x to y. The flow is permissible under a security policy only if the policy allows the security level of x to flow to the security level of y [13].

Studies [38]-[42] present the most efficient approaches and algorithms used for the detection of information flow vulnerabilities.

C. API Conformance

Security vulnerabilities can be related to violations of API specifications. Enterprise software platforms provide a number of APIs for system security services. The misuse of such interfaces can lead to security holes. This kind of misuse can't be detected by compilers but it can be checked with trivial syntactic code scanning or by using some more complex methods. A client that wants to invoke some services from a package needs to follow API conformance rules described in the class documentation. If the client fails to follow these conformance rules, the application will receive a security-related exception.API conformance rules can be expressed as temporal safety properties in a deterministic finite state automaton, or a *typestate specification* [43]. A large number of researches are related to the implementation of such rules [44]-[51], while other papers present a purely analytical approach [52]-[60].

IV. Conclusion

Static code analysis can be utilized to detect defects and security vulnerabilities in source code. It is a fast and repeatable technique for scanning millions of lines of code during the development phase. This is a time and a cost saving way to detect vulnerabilities before an attacker has the chance to exploit them. As shown in the paper, false positives and false negatives are still a major roadblock for the acceptance and widespread use of such tools. The third section outlines that static analysis is based on powerful algorithms that are improving constantly. This indicates that static analysis is one of the most promising mechanisms that can be utilized to achieve secure and reliable software.

References

[1] Etics, "Software Quality Basics", http://etics.web.cern.ch/etics/sq-basics.htm [Online; accessed 08/23/13]

[2] Steve Christey, 2011 CWE/SANS Top 25 Most Dangerous Software Errors. The Mitre Corporation, http://cwe.mitre.org/top25/, 2011. [Online; accessed 3.7.2013.].

[3] Open Software Assurance Maturity Model, "Software Assurance Maturity Model", http://www.opensamm.org [Online; accessed 08/23/13]

[4] The Building Security In Maturity Model, "The Software Security Framework", http://bsimm.com/online [Online; accessed 08/27/13]

[5] Security Development Lifecycle, "What is the Security Development Lifecycle", http://www.microsoft.com/security/sdl/default.aspx [Online; accessed 08/26/13]

[6] Cenzic, "Application Vulnerability Trends Report 2013",http://info.cenzic.com/rs/cenzic/images/Cenzic-Application-Vulnerability-Trends-Report-2013.pdf[Online; accessed 09/01/13]

[7] Common Weakness Risk Analysis Framework, http://cwe.mitre.org/cwraf/[Online; accessed 09/01/13]

[8] Common Weakness Scoring System, http://cwe.mitre.org/cwss/ [Online; accessed 09/01/13]

[9] Martin, R.A.; Christey, S.M., "The Software Industry's "Clean Water Act" Alternative," Security & Privacy, IEEE , vol.10, no.3, pp.24,31, May-June 2012

[10] Walden, J.; Doyle, M., "SAVI: Static-Analysis Vulnerability Indicator," Security & Privacy, IEEE , vol.10, no.3, pp.32,39, May-June 2012

[11] Anderson, P., "Measuring the Value of Static-Analysis Tool Deployments," Security & Privacy, IEEE , vol.10, no.3, pp.40,47, May-June 2012

[12] Cifuentes, C.; Keynes, N.; Lian Li; Hawes, N.; Valdiviezo, M., "Transitioning Parfait into a Development Tool," Security & Privacy, IEEE , vol.10, no.3, pp.16,23, May-June 2012

[13] Pistoia, M.; Chandra, S.; Fink, S.J.; Yahav, E., "A survey of static analysis methods for identifying security vulnerabilities in software systems," IBM Systems Journal , vol.46, no.2, pp.265,288, 2007

[14] J. H. Saltzer and M. D. Schroeder, "The protection of information in computer systems", Proceedings of the IEEE 63, No. 9, 1278–1308 (September 1975).

[15] Pistoia, M.; Banerjee, A.; Naumann, D.A., "Beyond Stack Inspection: A Unified Access-Control and Information-Flow Security Model," Security and Privacy, 2007. SP '07. IEEE Symposium on , vol., no., pp.149,163, 20-23 May 2007

[16] D. S. Wallach, A. W. Appel, and E. W. Felten, "SAFKASI: A Security Mechanism for Language-Based Systems", ACM Transactions on Software Engineering and Methodology (TOSEM) 9, No. 4, 341–378 (2000).

[17] F. Pottier, C. Skalka, and S. F. Smith, "A Systematic Approach to Static Access Control", Proceedings of the 10th European Symposium on Programming Languages and Systems, Springer-Verlag (2001), pp. 30–45.

[18] Jensen, T.; Le Metayer, D.; Thorn, T., "Verification of control flow based security properties," Security and Privacy, 1999. Proceedings of the 1999 IEEE Symposium on , vol., no., pp.89,103, 1999

[19] M. Bartoletti, P. Degano, and G. L. Ferrari, "Static Analysis for Stack Inspection,'' Proceedings of International Workshop on Concurrency and Coordination, Electronic Notes in Theoretical Computer Science 54, Amsterdam, The Netherlands, Elsevier (2001), pp. 69–80.

[20] A. Banerjee and D. A. Naumann, "A Simple Semantics and Static Analysis for Java Security", Stevens Institute of Technology, Hoboken, NJ, USA, Technical Report CS2001-1 (July 2001).

[21] U. Erlingsson and F. B. Schneider, "IRM Enforcement of Java Stack Inspection", Proceedings of the 2000 IEEE Symposium on Security and Privacy, Oakland, CA, USA, IEEE Computer Society (May 2000), pp. 246–255.

[22] Ryder, Barbara G. "Dimensions of precision in reference analysis of object-oriented programming languages." In Compiler Construction, pp. 126-137. Springer Berlin Heidelberg, 2003.

[23] M. Pistoia, "A Unified Mathematical Model for Stack- and Role-Based Authorization Systems", Ph.D. dissertation, Polytechnic University, Brooklyn, NY, USA (May 2005).

[24] M. Pistoia, R. J. Flynn, L. Koved, and V. C. Sreedhar, "Interprocedural Analysis for Privileged Code Placement and Tainted Variable Detection", Proceedings of the 9th European Conference on Object-Oriented Programming, Glasgow, Scotland, UK, Springer-Verlag (July 2005).

[25] David F. Ferraiolo, Janet A. Cugini, D.Richard Kuhn, "Role based access control (RBAC): features and motivations", National Institute of Standards and Technology U.S. Department of Commerce Gaithersburg MD 20899.

[26] PaolinaCentonze, GlebNaumovich, Stephen J. Fink, Marco Pistoia, "Role based access control consistency validation", ISSTA'06, July 17–20, 2006, Portland, Maine, USA.

[27] A. Schaad and J. D. Moffett, "A Lightweight Approach to Specification and Analysis of Role-Based Access Control Extensions," Proceedings of the 7th ACM Symposium on Access Control Models and Technologies, Monterey, CA, USA, ACM Press (2002), pp. 13–22.

[28] E. Damiani, S. De Capitani di Vimercati, S. Paraboschi, and P. Samarati, "A Fine-Grained Access Control System for XML Documents," ACM Transactions on Information Systems Security 5, No. 2, 169–202 (2002).

[29] M. Kudo and S. Hada, "XML Document Security Based on Provisional Authorization," Proceedings of the 7th ACM Conference on Computer and Communications Security,Athens, Greece, ACM Press (November 2000), pp. 87–96.

[30] M. Murata, A. Tozawa, M. Kudo, and S. Hada, "XML Access Control Using Static Analysis," Proceedings of the 10th ACM Conference on Computer and Communications Security, Washington, DC, USA, ACM Press (October 2003), pp. 73–84.

[31] F. Ricca and P. Tonella, "Analysis and Testing of Web Applications," Proceedings of the 23rd International Conference on Software Engineering, Toronto, ON, Canada, IEEE Computer Society (2001), pp. 25–34.

[32] D. Clarke, M. Richmond, and J. Noble, "Saving the World from Bad Beans: Deployment-Time Confinement Checking," Proceedings of the 18th Annual ACM SIGPLAN Conference on Object-Oriented Programing, Systems, Languages, and Applications, Anaheim, CA, USA, ACM Press (2003), pp. 374–387.

[33] M. Pistoia, S. J. Fink, R. J. Flynn, and E. Yahav, "When Role Models Have Flaws: Static Validation of Enterprise Security Policies," Proceedings of the International Conference on Software Engineering (ICSE 2007), Minneapolis, MN, USA (forthcoming, May 2007).

[34] P. Centonze, G. Naumovich, S. J. Fink, and M. Pistoia, "Role-Based Access Control Consistency Validation," Proceedings of the International Symposium on Software Testing and Analysis (ISSTA '06), Portland, Maine, USA (July 2006), pp. 121–132.

[35] A.W. Roscoe, M.H. Goldsmith, "What is intransitive noninterference?"

[36] J. Newsome and D. Song, „ Dynamic taint analysis for automatic detection, analysis, signature genetarion of exploits on commodity software", Proceedings of the 12th Annual Network and Distributed System Security Symposium, San Diego, CA, USA, IEEE Computer Society (February 2005).

[37] D. Volpano, C. Irvine, and G. Smith, "A Sound Type System for Secure Flow Analysis," Journal of Computer Security 4, Nos. 2-3, 167–187 (January 1996).

[38] U. Shankar, K. Talwar, J. S. Foster, and D. Wagner, "Detecting Format String Vulnerabilities with Type Qualifiers," Proceedings of the 10th USENIX Security Symposium, Washington, DC, USA (August 2001), pp. 201–220.

[39] J. Newsome and D. Song, "Dynamic Taint Analysis for Automatic Detection, Analysis, and Signature Generation of Exploits on Commodity Software," Proceedings of the 12th Annual Network and Distributed System Security Symposium, San Diego, CA, USA, IEEE Computer Society (February 2005), http://jimnewsome.net/papers/taintcheck.pdf.

[40] K. Ashcraft and D. Engler, "Using Programmer-Written Compiler Extensions to Catch Security Holes," Proceedings of the 2002 IEEE Symposium on Security and Privacy, Oakland, CA, USA, IEEE Computer Society (May 2002),pp. 143–159.

[41] G. Snelting, T. Robschink, and J. Krinke, "Efficient Path Conditions in Dependence Graphs for Software Safety Analysis," ACM Transactions on Software Engineering and Methodology 15, No. 4, 410–457 (2006).

[42] V. B. Livshits and M. S. Lam, "Finding Security Vulnerabilities in Java Applications with Static Analysis,"Usenix Security Symposium (2005).

[43] J. A. Goguen and J. Meseguer, "Security Policies and Security Models," Proceedings of the 1982 IEEE Symposium on Security and Privacy, Oakland, CA, USA, IEEE Computer Society Press (May 1982), pp. 11–20.

[44] R. E. Strom and S. Yemini, "Typestate: A Programming Language Concept for Enhancing Software Reliability", IEEE Transactions on Software Engineering 12, No. 1, pp. 157–171 (1986).

[45] R. E. Strom and D. M. Yellin, "Extending Typestate Checking Using Conditional Liveness Analysis," IEEE Transactions on Software Engineering 19, No. 5, 478–485 (May 1993).

[46] R. DeLine and M. Fa¨hndrich, "Enforcing High-Level Protocols in Low-Level Software," Proceedings of ACM Conference on Programming Language Design and Implementation (June 2001), pp. 59–69.

[47] R. DeLine and M. Fa¨hndrich, "Adoption and Focus: Practical Linear Types for Imperative Programming," Proceedings of ACM Conference on Programming Language Design and Implementation, Berlin (June 2002), pp. 13–24.

[48] J. S. Foster, T. Terauchi, and A. Aiken, "Flow-Sensitive Type Qualifiers," Proceedings of ACM Conference on Programming Language Design and Implementation, Berlin (June 2002), pp. 1–12.

[49] C. Flanagan, K. R. M. Leino, M. Lillibridge, G. Nelson, J. B. Saxe, and R. Stata, "Extended Static Checking for Java," Proceedings of ACM Conference on Programming Language Design and Implementation, Berlin (June 2002), pp. 234–245.

[50] V. Kuncak, P. Lam, and M. Rinard, "Role Analysis," Proceedings of ACM Symposium on Principles of Programming Languages, Portland, Maine (January 2002), pp. 17–32.

[51] A. Aiken, J. S. Foster, J. Kodumal, and T. Terauchi, "Checking and Inferring Local Non-aliasing," ACM SIGPLAN Notices 38, No. 5, 129–140 (May 2003), in Conference on Programming Language Design and Implementation (PLDI).

[52] J. C. Corbett, M. Dwyer, J. Hatcliff, C. Pasareanu, Robby, S. Laubach, and H. Zheng, "Bandera: Extracting Finite- State Models from Java Source Code," Proceedings of International Conference on Software Engineering (June 2000), pp. 439–448.

[53] T. Ball and S. K. Rajamani, "Automatically Validating Temporal Safety Properties of Interfaces," Proceedings of 8th International SPIN Workshop of Model Checking of Software (SPIN 2001), in Lecture Notes in Computer Science 2057, 103–122 (2001).

[54] K. Ashcraft and D. Engler, "Using Programmer-Written Compiler Extensions to Catch Security Holes," Proceedings of IEEE Symposium on Security and Privacy, Oakland, CA (May 2002).

[55] M. Das, S. Lerner, and M. Seigle, "ESP: Path-Sensitive Program Verification in Polynomial Time," ACM SIGPLAN Notices 37, No. 5, 57–68 (May 2002), in Conference on Programming Language Design and Implementation (PLDI).

[56] N. Dor, S. Adams, M. Das, and Z. Yang, "Software Validation via Scalable Path-Sensitive Value Flow Analysis," Proceedings of the ACM SIGSOFT International Symposium on Software Testing and Analysis (2004), pp. 12–22, http://doi.acm.org/10.1145/1007515.

[57] S. Fink, E. Yahav, G. Ramalingam, N. Dor, and E. Geay, "Effective Typestate Verification in the Presence of Aliasing," Proceedings of the International Symposium on Software Testing and Analysis (2006), pp. 133–144.

[58] J. Field, D. Goyal, G. Ramalingam, and E. Yahav, "Typestate Verification: Abstraction Techniques and Complexity Results," Proceedings of Static Analysis Symposium (SAS'03), in Lecture Notes in Computer Science 2694, 439–462 (June 2003).

[59] G. Ramalingam, A. Warshavsky, J. Field, D. Goyal, and M. Sagiv, "Deriving Specialized Program Analyses for Certifying Component-Client Conformance," Proceedings of ACM Conference on Programming Language Design and Implementation, ACM SIGPLAN Notices 37, No. 5, 83–94, New York: ACM Press (June 17–19 2002).

[60] R. Shaham, E. Yahav, E. Kolodner, and M. Sagiv, "Establishing Local Temporal Heap Safety Properties with Applications to Compile-Time Memory Management," Proceedings of Static Analysis Symposium (2003), pp. 483–503, http://link.springer.de/link/service/series/0558/bibs/2694/26940483.htm.

Random flip-flop and its applications

Mario Stipčević[1]
[1]Ruđer Bošković Institute, Bijenička 54, 10002 Zagreb, Croatia
*corresponding author e-mail: Mario.Stipcevic@irb.hr
Date: February 22, 2013

Abstract – We define a new non-sequential logic element: the random flip-flop. By definition it operates similarly to the conventional flip-flop except that its clock input functions with probability of 1/2 otherwise the flip-flop does nothing. We discuss several applications of that novel device.

I. INTRODUCTION

Practically all contemporary computers are constructed from deterministic logic circuits (gates, flip-flops, etc.) and are therefore equivalent to a Turing machine [1]. However, it has been realized that a more powerful notion of a computer, so called probabilistic Turing machine [2], [3], can execute *randomized algorithms* [4] which offer either greater speed or greater simplicity (or both) than their deterministic counterparts and on top of that they may be more resilient to errors which may occur during the calculation process. The price to be paid for that advantages is that a randomized algorithm have a finite, usually small and tunable probability to yield a wrong result. For example, one of the earliest known randomized algorithms, the Rabin-Miller primality test [5-6], estimates efficiently and with a high probability whether a given large number is prime or not which is useful in cryptography. This algorithm is much faster than any known deterministic algorithm [7]. Examples of use of probabilistic Turing machine include numerical simulations, Monte Carlo calculations, PSI factor research and gambling automata. And last but not least the cryptography, as one of the cornerstones of online trade and internet security, depends on local availability of freshly generated, unguessable random numbers at each communication endpoint.

Probabilistic Turing machine can be thought of as an ordinary Turing machine complemented with a random number generator and/or random decisions. Having in mind importance of random numbers or random decisions for countless applications it is quite astonishing that modern computer processors do not contain a *standardized* randomness source and a set of machine language instructions dealing with random decisions and/or random numbers. While rare attempts to include a random number generator in the microprocessor exist [8-9] they are by no means guaranteed to appear in any future processors, there are no industry standards on how to access them and are slower than a PRNG ran on the same microprocessor. Probably because of these drawbacks, developing software that would use these generators does not seem to be an appealing option to software designers. Instead, established state of the art is to use pseudo-random formulae seeded by user supplied data or with a handful of bits collected in various ways from the computer hardware (hard disc access times, internal clocks/interrupts, temperature sensors, keystrokes, etc.) in order to generate long pseudo-random numbers which are then used as a substitute for true random numbers.

Modern computers are built from deterministic sequential logic (e.g. the NAND gate) and non-sequential logic circuits (e.g. a flip-flop). Here we propose a single, new elementary non-sequential logic building block: *the random flip-flop* which can be used, among many other applications, to build a probabilistic Turing machine. Random flip-flop may be useful in many other applications such as signal processing, cryptography, laboratory equipment, analog computing, etc. In this study we will define, describe a few possible experimental realizations of the random flip-flop and discuss examples of its usage which extend over both digital and analog applications.

II. DEFINITION OF THE RANDOM FLIP-FLOP

Conventional edge-triggered D-type flip-flop (DFF) shown in Fig. 1a) is an important building element of logic circuits, microprocessors, computers etc. It is typically used as a one bit data storage or a frequency divider-by-two. The "edge-triggered" action means that flip-flop performs its action when clock pulse input (CP) changes state from LOW (logic 0) to HIGH (logic 1), that is upon the "rising edge" of the input pulse. Flip-flop does not care about pulse duration or the falling edge at the clock input (CP). In this work we will only refer to edge-triggered flip-flops and therefore this characterization will be dropped henceforth. Although there are several types of flip-flops in use (D, T, RS, JK, etc.) the D-type is universal in the sense that it can most easily mimic all other types of flip-flops. In principle, it can be realized from basic logic gates [10]. D-type flip-flop operates in the following way. With each occurrence of the rising edge at the input CP (clock pulse), the logic level present at the data input D is transferred to the output Q. The transfer takes some propagation time, usually denoted by t_{CPQ}. After the transfer is finished the value of Q is frozen i.e. becomes independent of the subsequent changes at the input D until occurrence the next rising edge at the clock input. This operation is illustrated by an example timing diagram shown in Fig. 1b). By definition, the D-type flip-flop is a *deterministic* device. It means that under the same input conditions it will always give the same output. Of course, conventional logic elements such as flip-flops, AND gates, OR gates etc. are designed specifically to behave in a deterministic fashion in order to ensure stable and predictable operation of logic circuits, computers, etc. Although in practical devices problem known as "metastability" [11] occurs with a small probability, it can be minimized by careful design of the flip-flop, by stacking flip-flops or by avoiding critical operating conditions. This elusive effect is of no interest here.

As a contrast to deterministic flip-flops we here introduce a family of *random flip-flops* (RFF). Each deterministic type of flip-flop has its random counterpart. For example, by definition, the D-type random flip-flop (DRFF) (symbol shown in Fig. 1c)), operates the same as the conventional D-type flip-flop with only difference that upon the clock pulse, state of the data input D is transferred to the output Q with probability of ½, randomly.

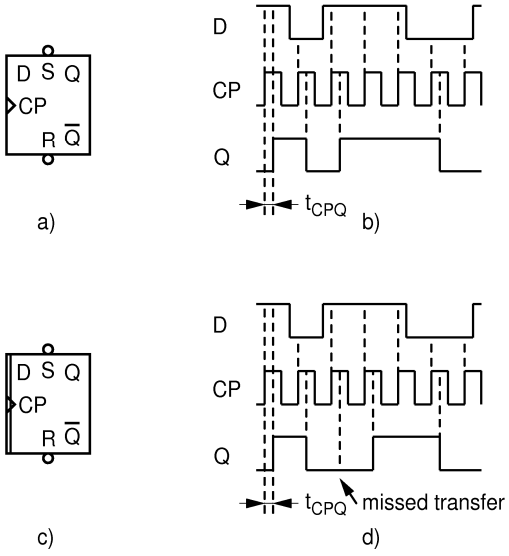


Figure 1. The conventional D-type flip-flop (DFF) symbol (a) and an example of its input-output waveforms (b). The D-type random flip-flop (DRFF) symbol (c) and an example input-output waveforms (d). The t_p is propagation delay between the clock pulse input CP and the outputs Q and $\overline{Q}$.

An example of the input-output diagram of DRFF is shown in Fig. 1d). We have adopted the usual convention that the state of the output changes with positive going edge of the clock pulse. Another type of random flip-flop of particular convenience for this presentation is T-type (TRFF). Deterministic T-type flip-flop (TFF) toggles output state with each clock if input T is held at logical 1 while if T=0 the output is left unchanged. By definition, TRFF is different from TFF only in that if T=1 then each clock pulse reverses the output state with probability ½. In both TFF and TRFF clock input is disabled while T=0. Note that the two definitions (of DRFF and TRFF) are mutually compliant in that whatever a given type of deterministic flip-flop does upon receiving a clock pulse, its random counterpart does randomly with probability of ½, and that in the same manner the definition could be extended to all other types of flip-flops. A RFF optionally features inverted output $\overline{Q}$ which is a binary complement of the output Q. Besides random action of the clock input CP, all random flip-flops optionally feature deterministic Set (S) and Reset (R) inputs or their inverted versions ($\overline{S}, \overline{R}$), which unconditionally *set* (to HIGH) or *reset* (to LOW) the output Q. Although the two random flip-flops can be emulated by each other as shown in Fig. 2, it is sometimes more convenient to realize given circuit by use of a particular type. Interestingly, the same emulation schemae are true for deterministic counterparts (DFF and TFF).

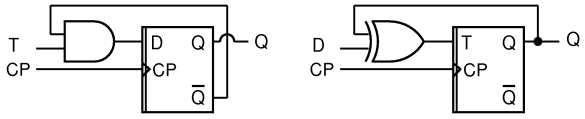


Figure 2. Emulations of DRFF and TRFF by each other. These exact same schematic diagrams hold also for deterministic counterparts.

Of course, one could consider a random flip-flop with the action probability different from 1/2 but the probability of 1/2 is the most convenient for general applications.

Random flip-flops are digital devices with logic inputs and logic outputs that are fully compatible with conventional logic circuits and can be used in mixed designs. Research on full potential of RFFs has just began. In present article we outline some of the most interesting known areas of application.

III. RANDOM NUMBER GENERATOR

Probably the simplest and most obvious application of the random flip-flop is the random bit generator (RBG), a physical device that produces one random bit upon a request. By definition of RBG, any generated bit must have no correlation with the pool of previously generated bits *and* probability of zeros p_0 must be equal to the probability of ones p_1. Figure 3 shows two equivalent random bit generators, one realized by DRFF (left) and the other by TRFF (right). Every rising edge at the CP input will trigger a generation of a fresh new random bit at the Data output synchronously with the Strobe's rising edge. By definition of RFF, the sequence of random bits is truly random and there is no correlation of any sort among bits, but of course in practice it would depend on the quality of the RFF. A random number generator that produces multiple random bits per single clock can be obtained by parallelizing a desired number of bit generators. For example by parallelizing eight RFFs one could generate a random byte in a single clock cycle.

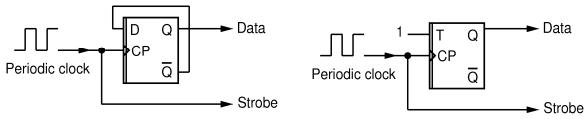


Figure 3. Stream random number generator. Shown are equivalent realizations with DRFF (left) and TRFF (right). Each rising edge of the periodic clock generates a fresh new random bit at the Data output synchronously with the Strobe's rising edge.

Having in mind possible miniaturization of RFFs to a chip level, random number generators of this sort can be particularly useful in cryptographic hardware (smart cards, password tokens, cell phones, computers,...) with a purpose of secure generation of keys, passwords,

challenge-response data, as well as basis data in quantum cryptography [24].

IV. RANDOMNESS PRESERVING FREQUENCY DIVIDERS

Dividing periodic frequency of periodic signals by a fixed integer or fractional number can be performed by various well known techniques [17-19]. For example, a single TFF will divide by two frequency of its clock signal while division by n can be made by a counter-to-n.

Let us now define a random pulse train (RPT) as a sequence of logic pulses in which waiting times between subsequent pulses follow Exponential p.d.f. RPTs are time-wise maximally random and are an important concept with many applications. Frequency of a RPT can also be divided by counters but the Exponential p.d.f. will not be preserved. Randomness preserving frequency division wasn't feasible so far. However, this can be made with random flip-flops.

A random divider by 2, shown in Fig. 4a, omits randomly on average every second clock pulse. If a RPT of average frequency f is fed to the input CP, a random pulse train of frequency two times lower will appear at the output Q. We denote the function of this circuit as $f_{OUT} = f_{IN} / 2_R$ where subscript "R" denotes "random" division action. Stacking two DRFFs as shown in Fig, 4b one achieves division by 4_R etc. A surprising possibility of dividing frequency by 4 with a single DRFF is shown in Fig 4c, but this division is composed of first dividing by 2 randomly and then by 2 deterministically, which we denote as $1/2_R/2_D$ where subscript "D" stands for "deterministic".

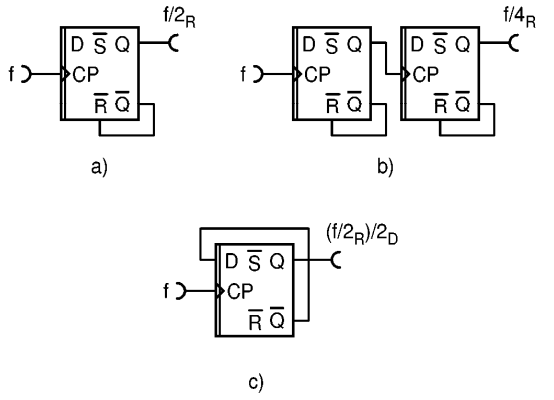


Figure 4. Basic dividers with RFFs: a) random divider by 2; b) random divider by 4; c) divider which performs random division by 2 followed by deterministic division by 2.

Although dividers in Figs. 4b) and 4c) both divide input frequency by 4 their action is not the same because the first divider preserves exponential distribution whereas the second converts exponential distribution into Erlang distribution with shape parameter n=2. Generally, a RPT with mean frequency f deterministically divided by n will result in another random pulse train with mean frequency of f/n but its distribution will be Erlang distribution with shape parameter n:

which for large n tends to the Normal distribution due to the central limit theorem. The Erlang distribution for n = 2 which corresponds to waiting times of an exponential RPT divided by 2_D is shown in Fig. 5. A random division of an Erlang(t; f=1, n=2) pulse train by factors 2_R, 4_R and 16_R (normalized to unity frequency). The conclusion is that deterministic division by n drives the output signal away from randomness and Exponential distribution whereas random division by n drives *any* initial distribution asymptotically closer to the Exponential p.d.f., that is to the highest possible time-wise randomness. In that respect random frequency division and synthesis of RPTs as well as RPT computing are stable, self-healing processes which tend to diminish effects of imperfections in hardware.

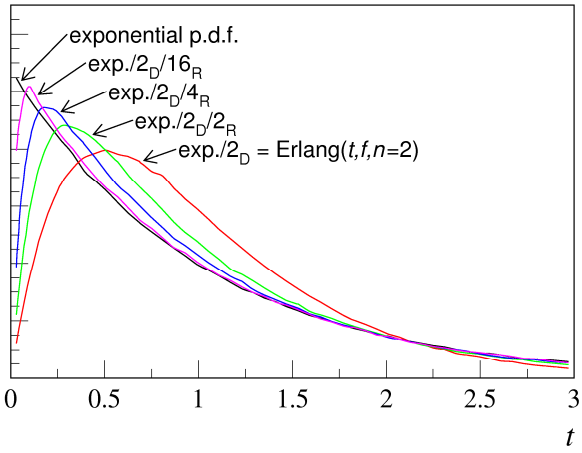


Figure 5. Waiting time distributions of an RPT (exponential) and of an RPT divided by factors: 2_D, 2_D2_R, 2_D4_R and 2_D16_R.

A complex circuits for frequency division by means of multiplication by a number smaller than 1, as well as for synthesis of an arbitrary random frequency, have been developed and presented in our earlier work [20]. Such circuits can be used to realize a paradigm of random pulse computing [21] which allows for complex, fast and reliable calculations even with slow and unreliable hardware – probably a principle used in biological systems.

V. RANDOM COUNTER

Various circuits possible with deterministic flip-flops are also realizable with RFFs but function of the random counterpart can be quite different. Building tools for systematic understanding of properties of logic circuits containing random flip-flops is an interesting open challenge.

For example the $1/2_R/2_D$ frequency divider explained above (Fig. 4c) can be considered as a single bit random counter. Indeed, usual approach to frequency division is by use of counters. To that end let us now explore an equivalent of a deterministic synchronous 4-bit counter but realized with TRFFs rather than with TFFs, as shown

in Fig. 6. This circuit has a rich set of complex properties not all of which have been explored yet.

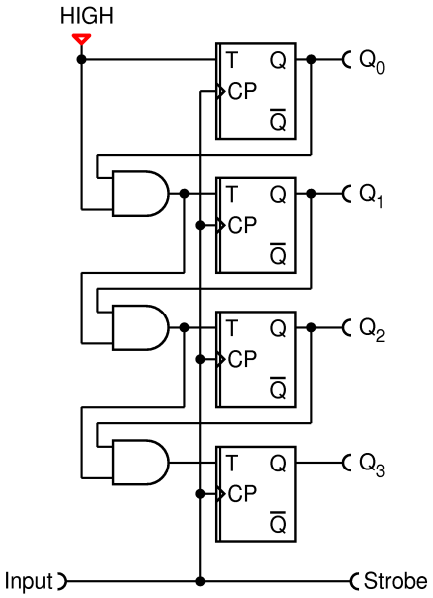


Figure 6. Synchronous 4-bit random counter. With each clock pulse the counter randomly advances up similar to random walker on a line.

The counter has 16 possible states which can be represented by a binary number $\mathbf{Q}=\{Q_3,Q_2,Q_1,Q_0\}$ whose value is in the range from 0 to 15. Unlike its deterministic counterpart which counts from 0 to 15, then folds back to 0 etc., the output value of the random counter behaves like a random walker that wonders around the mean value of 15/2. The 4-bit binary numbers **Q** behave like random draws from the Binomial distribution with n = 15 trials and trial success probability p = ½. It can be used as a generator of such values, but at the moment it is not known how to make a random counter up to an arbitrary integer number n.

When considered individually, each of the outputs Q_i (taken in synchronization with the Strobe) yields a random number sequence with zero bias but a strong autocorrelation characterized by the serial autocorrelation coefficient with a lag k (defined for example in [22]) equal to: $a_k(Q_i) = (1-2^{-i})^k$. The coefficients obey geometric law with respect to the lag k.

If the Input is fed by a random pulse train, each output Q_i gives another random pulse train. The mean frequency of i-th output (Q_i) is will be equal to the input mean frequency divided randomly by 2^{i+1} and then deterministically by 2 (in our notation: $f(Q_i) = f_{in}/(2^{i+1})_R/2_D$) but random pulse trains appearing at outputs Q_0-Q_3 are not statistically independent and cannot be combined for frequency synthesis.

VI. Random Programming Language Instructions and Probabilistic Computing

As noted in the introduction, today's computers are deterministic and cannot *per se* generate any random action which poses serious limitations. Should computer microprocessors be equipped with random logic circuits, such as random flip-flop, they would be able to generate random numbers, make random decisions and execute random instructions, for example random branching instruction. Conventional, deterministic branching instruction typically takes the following form:

```
IF {condition} THEN
 {block1}
ELSE
 {block2}
ENDIF
```

where *condition* is a Boolean-valued function, such as "A>B". If *condition* is fulfilled then program flow deterministically branches to the *block1* otherwise to *block2* after which it continues at the first instruction after ENDIF. Following the logical reversibility of computing [16] this (and in fact any) instruction can be inverted. To facilitate inversion, special universal programming languages have been formulated (e.g. R, PISA, LISP). This led to an automatic program inverter in LISP [23] capable of inverting any program written in LISP. Invertible programming is important because it promises computing with reduced power dissipation - a problem that has virtually stopped advance of clock frequency of computer processors about year 2005. Probably the most devastating practical consequence of inverse programming is that any cryptographic primitive, such as the encryption algorithm AES, could have an inverse that only needs to be "calculated" once and that could be made beforehand. Having such an "inverse" would be a great help in cryptanalysis: a single attack on a ciphertext could in principle yield both plaintext and the key. On the other hand, programming instructions relying on true randomness are non-invertible and therefore hold promise of safer cryptography, such as quantum or stochastic [24]. In those, Alice and Bob iterate via authenticated messages a certain randomized algorithm that converges to a common key with a probability arbitrary close to 1. The algorithms itself are non-invertible and use abundant quantity of random numbers. Here we give two examples of randomized instructions. A random branching instruction may have the following form:

```
RIF {prob} {condition} THEN
 {block1}
ELSE
 {block2}
ENDIF
```

where *prob* could be a real-valued probability of branching to *block1* if *condition* is TRUE, otherwise branching to *block2*. In a real-world implementation parameter *prob* could be compared to a single-precision float number which consists of 24 bit random mantissa and fixed 8 bit exponent and sign part. A 24 random flip-flop register could produce one such random number in a single processor clock – incomparably faster than any pseudo-random number algorithm executed in software, and on top of that truly random and non-invertible. A version of this would be random GOTO instruction that

would branch to any of the labels in the list with equal probability:

RGOTO label1, label2, label3, ...

Randomized algorithms [4] make use not only of random numbers but also of random decisions. In fact, computers *per se* do not compute or use numbers: they only make Boolean operations between binary values which we interpret as numbers (or letters, pictures, etc.). Even simplest numerical operations such as addition or multiplication are "higher level" operations build of very many basic logic operations. It is therefore tempting to believe that the most optimal use of randomness would be at the level of basic logic operations instead.

Another issue is the speed of random number production. In Monte Carlo programs with intensive use of random numbers their explicit production in software by mathematical operations can amount a significant portion of CPU resources. Having a specialized register consisting of random flip-flops, to generate the whole random number in a single CPU cycle, would greatly improve the CPU performance.

In order to fully grasp computational power of a randomized computer, a new type of programming language would be required which could use a standardized and efficient access to random bit registers, execute random decisions and random logic operations.

VII. GAUSSIAN WHITE NOISE GENERATOR

Gaussian white noise is a very useful resource for various techniques such as: real time speech and music synthesis, measuring and testing of audio equipment and for radio communication circuits. Random logic can be used quite elegantly for production of analog white noise with exceptionally flat power density spectrum. Circuit shown in Fig. 7, driven by the periodic oscillator of frequency f_{osc}, generates at its output Gaussian white noise whose frequency spectrum envelope is nearly flat from 0 Hz till some portion of f_{osc}. The envelope of the power density is given by:

$$\frac{dP}{df} = \left(\frac{\sin\left(\pi \frac{f}{f_{\text{clock}}}\right)}{\pi \frac{f}{f_{\text{clock}}}} \right)^2 \qquad (2)$$

The power density continuously drops down falling only -0.1dB at 0.12 f_{clock}, -1dB at 0.26 f_{clock} at and half of the power (-3dB) at 0.44 f_{clock}. If faster drop at the end of the spectrum is desired, one could pass the output through an active analog low-pass filter. In that case, the flatness of the spectrum is also affected by the transfer function of the filter. As is well known in the art, for maximum flatness in the pass region one should use the *n*-th order Butterworth filter, whereas for the maximum steepness in the stop region and minimum ripples in the pass region one should use the Chebyshev type of low-pass filter [25]. Digital filtering offers further possibilities to shape the power density virtually at will.

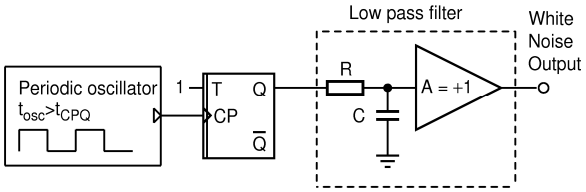


Figure 7. A practical realization of analog Gaussian white noise generator by means of a single random flip-flop.

A detailed mathematical treatment of this type of white noise generator and its spectrum is given in [26] and an improved version in [27] while suitable low-pass filters are explained for example in [25]. If the pseudorandom source was used, the spectrum would consist of equidistantly spaced spikes (excluding 0Hz) rather than being truly flat. The comb-like spikes in the pseudo-random noise come from the finite period of the pseudo-random number generator, whereas the period of a non-deterministic generator is infinite and consequently the spectrum is continuous. An interesting characteristic of the analog noise generator presented here is that the spectrum is extremely flat at the low frequency end and extends all the way to 0 Hz, which is to be contrasted to various semiconductor sources, such as Zener diodes or bipolar transistors whose spectrums suffer from 1/f rise in that region, as well as with PRNG based generators that exhibit spectral ripples and a cut-off near zero. Random flip-flops can also be used to generate other random amplitude analog signals.

VIII. SYMMETRIC BINARY CHANNEL WITH NOISE

In information-theoretic secure cryptographic protocols based on noisy channels [28-29] the crucial resource is a physical device called "symmetric binary channel with noise" (BSCN), which flips the data bit with a predetermined probability p_{err}, known also as *bit error probability*. Such a device effectively adds a precise, predetermined quantity of non-deterministic noise (entropy) to the stream of digital data bits. Circuit shown in Fig. 8 is a practical realization of a BSCN using previously described random frequency division circuit. We assume that the data bits at the Data Input appear in synchronization with the positive going edges of the Strobe. The data bits and the random sequence from the divider's output are synchronized by means of the pair of D-type flip flops and then XOR-ed to obtain a bit flip with probability $p_{err} = \lambda$.

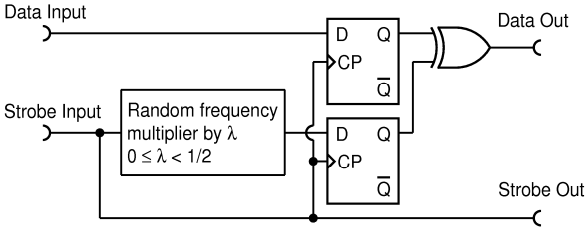


Figure 8. Realization of symmetric binary channel with noise by means of the random divider circuit. Input data bits are either copied to the output with probability (1 - λ) or inverted with the probability λ.

For binary channels with noise considered in information theory bit error probability is usually a small number, and is absolutely limited by $0 \le p_{err} < 0.5$.

IX. DICE THROWER

And finally some fun. Circuit in Fig. 9 after each request generates one random number between 1 and 6 (in binary or BCD notation) thus emulating throws of a perfectly fair dice. What happens is that upon a request signal on the input "Request" the device produces one random number in the range 0-7. If it happens to produce 0 or 7 then the periodic oscillator sends additional requests until the random number happens to fall between 1 and 6. New requests are not possible until the last one is completed which is signaled by "Ready" going HIGH. Note that this particular circuit is permutation symmetric in the three flip-flops meaning that it is irrelevant which binary value is assigned to which output. Moreover, the assignment can be changed at will and at any time - the circuit will still produce sequence of absolutely fair and uncorrelated random throws! This "dice" is unpredictable: collection of all throws made so far contains zero information about the outcome of the next throw. This example illustrates usability of RFFs in hazard games, online gambling and gaming automata.

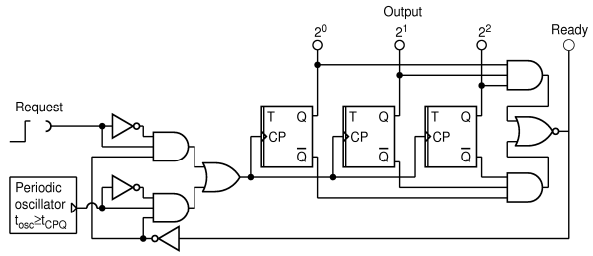


Figure 9. A random logic circuit which generates equiprobably one random binary number in the range of 1 to 6 upon each request, thus mimicking throws of a perfectly fair dice.

In a similar fashion one could create circuits to mimic any lottery game.

X. REALIZATION OF (PSEUDO) RANDOM FLIP-FLOP

Apparently, one can think of a random flip-flop as consisting of an ordinary flip-flop (of the same type) and a physical source of randomness that will be "consulted" at the time when clock pulse arrives in order to make decision of whether the clock pulse will be accepted (action) or rejected (no action). In that respect a RFF is a simplest mechanism to obtain a random bit from a physical source of randomness. Conversely, a random bit generator (RBG) can be used to build a random flip-flop. Figure 10. shows a schematic of a RFF with an internal pseudo-random bit generator (PRBG) or physical true random number generator (TRBG) .

It may seem odd to use a RBG to build RFF which can be used as a RBG (Section 3). However our point is that packaging randomness in form of a RFF rather than RBG gives much more versatility because RFF can be conveniently used for so many digital and analog applications. In that respect a random number generator is just one out of many possible applications of random flip-flop.

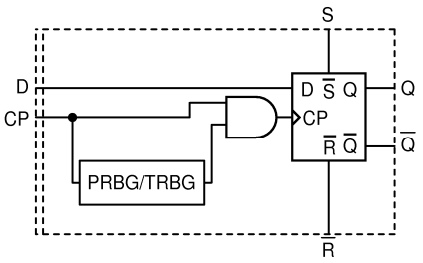


Figure 10. Random D-type flip-flop containing an internal random number generator.

It turns out that for usability of RFF should preferably have a well defined propagation delay which should be the same for all RFFs used in a given circuit. (Just as it is the case for flip-flops found on the same chip.) This is necessary for synchronization of several RFFs which is important for synchronous counters, random frequency dividers, random frequency synthesis etc.

By definition random flip-flop is a non-deterministic device. Nevertheless in some cases pseudorandom behavior can be tolerated or perhaps preferred. To that end a linear feedback shift register (LFSR) [13-14] can be used and it already features the constancy of the delay time. If nothing else, pseudo-random flip-flop (PRFF) is technically interesting because it can be built from standard logic circuits. The leftover problem of initializing a PRFF, especially in case of using many of them in a single design, can be conveniently solved by use of recently found Boolean chaos [12]. Having this, all circuitry required for realization of a PRFF can be programmed in an FPGA.

It is generally straightforward to make PRFF with a constant delay time, a feature which is highly desirable but as we shall see not easy to obtain with sources of true randomness. Additional bonus of using LFSR is that the total propagation time of the RFF is constant and independent of the length of the LFSR so highly complex LFSRs can be used at no cost in speed. The propagation time of the resulting PRFF is very small: just about twice the propagation time of the DFF employed. However one has to bare in mind that pseudo-random flip-flop is in fact deterministic device, built only from deterministic logic elements and therefore its output is predictable. This compromises its usability in cryptographic and some other applications. Therefore generally, a truly random flip-flop would be highly desirable. Quantum physics is known to allow for true randomness [15], [30] but general problem with physical random phenomena is that they either create random events at random times or at known times but with probability smaller than 1 and therefore random events might not be available at the time of request leading to non-constant propagation delay. The concept of RFF is very new and research on its possible realizations by using physical random phenomena is under way.

XI. CONCLUSION

We introduce a new non-sequential logic element, the random flip-flop, and discuss its applications in generating

random numbers, realizing random counters, randomness preserving frequency dividers, probabilistic Turing computers, analog noise generators, binary channels with noise and gambling automata. Random flip-flops feature logic inputs and outputs compatible with ordinary logic circuits and therefore can be easily combined with them. While Boolean logic allows us to predict functioning of deterministic logic circuits, building of tools for systematic understanding of properties of logic circuits that contain random flip-flops is an interesting opened challenge.

REFERENCES

[1] Alan Turing, 1948, "Intelligent Machinery". Reprinted in "Cybernetics: Key Papers" Ed. C.R. Evans and A.D.J. Robertson. Baltimore: University Park Press, 1968. p.31.

[2] K. De Leeuw, E.F. Moore, C.E. Shannon, and N. Shapiro, "Computability by probabilistic machines", In Automata Studies, C. E. Shannon and J. McCarthy, Eds., Princeton University Press, Princeton, NJ, (1955) 183–212.

[3] M. J. Atallah, "Algorithms and Theory of Computation Handbook", p. 20-29, Copyright 1999 by CRC Press LLC

[4] R. Motwani and P. Raghavan, "Randomized Algorithms", ACM Computing Surveys, 28 (1996)33-37.

[5] G. L. Miller, "Riemann's Hypothesis and Tests for Primality", Journal of Computer and System Sciences 13 (1976), 300–317.

[6] M. O. Rabin, "Probabilistic algorithm for testing primality", Journal of Number Theory 12 (1980), no. 1, pp. 128–138.

[7] Wolfram MathWorld, "Primality test", http://mathworld.wolfram.com/PrimalityTest.html

[8] Cryptography Research Inc., "Evaluation Summary: VIA C3 Nehemiah Random Number Generator", URL: http://www.via.com.tw/en/downloads/whitepapers/initiatives/padlock/evaluation_summary_padlock_rng.pdf

[9] G. Taylor, and G. Cox, "Behind Intel's New Random-Number Generator", IEEE Spectrum, 24 Aug 2011, URL: http://spectrum.ieee.org/computing/hardware/behind-intels-new-randomnumber-generator

[10] "SN5474/SN7474 A DUAL D-TYPE POSITIVE-EDGE-TRIGGERED FLIP-FLOPS WITH PRESET AND CLEAR", note SDLS119, Texas Instruments Incorporated, December 1983-revised March 1988

[11] A. Cantoni, J. Walker, T.D. Tomlin, "Characterization of a flip-flop metastability measurement method", IEEE T Circuits-I 54(2007)1032-1040.

[12] R. Zhang, H. L. D. de S.Cavalcante, Z. Gao, D. J. Gauthier, J. E. S. Socolar, M. M. Adams, and D. P. Lathrop, "Boolean chaos", Phys. Rev. E 80 (2009) 045202

[13] P.H.R. Scholefield, Shift Registers Generating Maximum-Length Sequences, Electronic Technology 10 (1960) 389-394.

[14] L. Wang, N. A. Touba, R. P. Brent, H. Xu, and H. Wang, "On Designing Transformed Linear Feedback Shift Registers with Minimum Hardware Cost", Technical Report UT-CERC-12-03 November 8, 2011, http://www.cerc.utexas.edu/reports/UT-CERC-12-03.pdf

[15] M. Stipčević, B. Medved Rogina, "Quantum random number generator based on photonic emission in semiconductors", Rev. Sci. Instrum. 78(2007)045104:1-7

[16] C. H. Bennett, "Logical reversibility of computation", IBM Journal of Research and Development, 17 (1973) 525–532.

[17] V. B. Minuhin, "3/2 frequency divider", patent US4866741, priority date Sep. 12, 1989.

[18] N. Regimbal, F. Badets, D. Bordeaux, J. B. Begueret, "Fractional frequency divider",patent US2012/0001665, priority date Jun 29, 2011.

[19] Best, R. E. (2003), "Phase-locked Loops: Design, Simulation and Applications", McGraw-Hill, ISBN 0-07-141201-8

[20] M. Stipčević, "Quantum random flip-flop", arXiv:1308.5719 [quant-ph]

[21] R.C. Lawlor, "Computer ulitizing random pulse trains", US patent US3612845

[22] D. E. Knuth, The art of computer programming, Vol. 2, Third edition, (Addison-Wesley, Reading, 1997).

[23] R. Glueck, M. Kawabe, "Revisiting an automatic program inverter for Lisp", ACM SIGPLAN Notices 40 (2005) 8-17.

[24] M. Stipcevic, "Quantum random number generators and their applications in cryptography", Proc. SPIE 8375, Advanced Photon Counting Techniques VI, 837504 (May 1, 2012).

[25] K. Lacanette, "A basic introduction o filters: active, passive and switched-capacitor", National Semiconductor application note AN-779, April 1991.

[26] P. Horowitz, W. Hill, "The Art of Electronics", Cambridge University Press; 2 edition (July 28, 1989, p.658, ISBN: 0521370957.

[27] J.C. Linz, "Method of synthetizing low frequency noise", patent US3614399, priority date Aug. 30, 1968.

[28] I. Csisz'ar and J. Koerner, "Broadcast channels with confidential messages", IEEE T. Inform. Theory, Vol. 24, No. 3, pp. 339-348, 1978.

[29] U. M. Maurer, "Secret key agreement by public discussion from common information", IEEE T. Inform. Theory, Vol. 39, No. 3, pp. 733742, 1993.

[30] M. Stipčević, and Ç. K. Koç, "True Random Number Generators", to appear in Open Problems in Mathematical and Computational Sciences, Ed. Ç. K. Koç, Springer 2014.

Comparison of ID-Based Blind Signatures from Pairings for E-Voting Protocols

Pance Ribarski
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, Macedonia
Email: pance.ribarski@finki.ukim.mk

Ljupcho Antovski
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, Macedonia
Email: ljupcho.antovski@finki.ukim.mk

***Abstract*—Elliptic curves are gaining momentum as scientists are continuously proving their security and performance. Pairings over elliptic curves are relatively new in the world of cryptography. Researchers are coming with new cryptographic usage of pairings for over ten years. ID-based cryptography is also gaining popularity because of the certificate-less mode of work. Blind signatures are appropriate schemes when user anonymity is wanted as property. One possible type of blind signature is ID-based blind signature based on bilinear pairings. For easy computation we look into pairing-friendly elliptic curves for implementation of pairings. This paper will review state of the art ID-based blind signature schemes from pairings over elliptic curves which are suitable for building blind signatures as part of e-voting protocols. We give comparative results about the computation cost of arithmetic operations. In our knowledge, this is first paper which gives head-to-head bandwidth comparison of the interactive protocol in the signing algorithm of blind signature schemes. The results are easy to use when choosing appropriate blind signature scheme for e-voting protocols.**

***Keywords*—elliptic curves, cryptography, pairing, blind signature, e-voting**

I. Introduction

For almost thirty years scientist and practitioners are trying to come up with elegant solution which will be a substitute for traditional voting. E-Voting is relatively recent tool in the democratic processes. Many cryptographic protocols have been invented for more secure and efficient e-voting processes. These protocols are heavily burdened with the many properties that they have to fulfil.

As soon as cryptographers come up with new security mechanisms, e-voting protocols are starting to follow path. The base notion of secure protocols is the identity certification. The standard public key cryptosystem (PKC) has one private key (SK) and one public key (PK) for each user. Additionally, every user has some identification (ID) which is linked to his PK by means of certificates. Therefore, a prerequisite step of communication is to verify the certificate of the user. This additional step in communication also means that the system has to keep all certificates which leads to additional storage and computing requirements. In 1984 Shamir [1] introduced new kind of encryption and signature schemes, so-called ID-based schemes, in order to simplify the key management in standard PKC's. ID-based PKC has the property that a user's identity is his public key, and the private key can be derived by a trusted authority called private key generator (PKG).

Bilinear pairings were firstly used to attack the DLP problem and reduce DLP on some elliptic curves (or hyperelliptic curves) to the DLP of representative finite fields (MOV attack [2] and FR attack [3]). However, recently bilinear pairings have been employed in "good" light, in construction of new ID-based cryptographic primitives. In 2000 Joux [4] made one-round three party Diffie-Hellman key agreement protocol. In 2001 Boneh and Franklin [5] made ID-based encryption scheme based on bilinear pairings. Again in 2001, Boneh, Lynn and Shacham [6] made basic signature scheme (BLS signatures) using bilinear pairings which has the shortest length among all signature schemes.

Blind signature is interactive signature scheme between user and signer, devised by Chaum in 1982 [7]. Blind signature S_b is signature on a blinded message M (let us call it M_b). The signer can not deduce any information about M and anyone can verify that the signature S_b is from the signer for the message M. Blind signatures provide anonymity of users for electronic money, or in our case, for e-voting. There are more types of blind signatures, such as fair (involve third party for blindness revocation in case of dispute) or partially blind (user and signer agree on some mutually known piece of information from the message M). Horster et al. [8] divide blind signatures into four categories: hidden, weak, interactive blind and strong blind. The later categorization treats the message as (a) document and (b) signature parameters, and the four categories of blind signatures determine the knowledge of the signer about the relation of (a) and (b).

Blind signatures have found their way into many protocols. E-voting uses blind signatures to obtain votes without violating the anonymity of the voter. In this protocol, the voter usually asks for blind signature on his encrypted vote. Then the voter unveils the signature and sends the signed ballot for counting. This scenario ensures that the administrator which checks the eligibility of the voter doesn't see the plaintext vote. Further more, this means that the e-vote process can have more roles, divided among more parties for higher robustness of the system. Blind signatures have bee used in many e-voting protocols [9], [10], [11], [12]. Because ID-based cryptography has certain advantages, especially in the cheaper setup of a PKI, there are current e-voting protocols which use ID-based blind signatures [13], [14], [15]. Because blind signatures are used in the casting of every vote, it is of great importance that this operation does not become the bottleneck of the protocol. This paper reviews ID-based blind signatures and compares the

arithmetic operations for their execution. Implementors have many algorithms to choose from, depending on the needs of the protocol. This paper is just part of the bigger research in e-voting protocols for low-resource devices, such as mobile phones. If we are to implement decent e-voting protocol intended for big audiences then we have to choose the most optimal algorithms in terms of processing requirements.

This paper reviews the ID-based blind signatures in [16], [17], [18], [19], [20], [21], [22]. These blind signatures were devised in the last ten years of research papers. Here we bring the protocols under one metric and compare them. First we give detailed algorithms of each of the blind signatures scheme. With this step we can easily see the similarities and the inventions that authors found in ID-based blind signatures. The comparison includes required assumptions, security model, computational cost and bandwidth cost. We can understand the computational cost and the bandwidth cost only if we have in mind implementation of these schemes. In this paper we use the Panda project [23] for calculating the cost of each arithmetic operations and the size of the operands passed around in the interactive protocols. This is the framework under which we compare the ID-based blind signatures from bilinear pairings.

The rest of the paper is organized as follows: Section 2 covers the preliminary terms used in the blind signatures. Section 3 explains the reviewed blind signature protocols. In Section 4 we give comparative data about the cost of implementation and running of the protocols. Finally, in Section 5 we conclude the paper.

II. Preliminaries

A. Elliptic curves and pairings

Definition 1. *Let E be elliptic curve defined over a finite field F_p denoted by the equation $E : y^2 = x^3+ax+b$. Then, $E(F_p)$ is denoted as the set of solutions (x, y), $x, y \in F_p$, where x and y satisfy E. There is another element on the curve which doesn't belong to F_p, denoted as point-at-infinity $\mathcal{O}$ and acts as the neutral addition element.*

Definition 2. *Consider the map $e : G_1 \times G_2 \to G_3$ where G_1 and G_2 are additive groups, and G_3 is a multiplicative group. The function e is a bilinear pairing if it satisfies the following properties:*

1) *Bilinearity: for all $P, P^{'} \in G_1$, $Q, Q^{'} \in G_2$ $e(Q+Q^{'}, P) = e(Q, P)e(Q^{'}, P)$ and $e(Q, P+P^{'}) = e(Q, P)e(Q, P^{'})$.*
2) *Non-degeneracy: $e(Q, P) = 1$ if and only if $P = \mathcal{O}$ or $Q = \mathcal{O}$.*
3) *Computable: There exists an algorithm A which efficiently computes $e(P, Q)$ for all $P \in G_1$, $Q \in G_2$.*

For the defined bilinear function e it holds $e(aQ, bP) = e(bQ, aP) = e(abQ, P) = e(Q, abP) = e(Q, P)^{ab}$, for arbitrary $a, b \in Z_r$.

Definition 3. *Random Oracle Model (ROM): Protocols under ROM are proved secure under additional assumptions that there exist publicly available functions chosen at random. Then, concrete implementations are obtained by replacing the random oracles with cryptographic hash functions (like SHA or MD5) [24].*

Definition 4. *ID-Based Blind Signature Scheme (IBSS) consists of four algorithms: Setup, Extract, Issue and Verify. The Setup algorithm is run by trusted party called PKG which generates the public parameters and the master key. The public parameters are published and the master key is kept secret by PKG. Extract algorithm has as input the public parameters and the master key, and generates the private key D_{ID} for given ID. Issue algorithm blindly issues a signature for a message. The Verify algorithm inputs a signature, message, ID and the parameters of the scheme, outputs 1 if the signature is valid for the message and ID, or outputs 0 otherwise.*

B. Security assumptions

Next we explain the security assumptions used by the authors as difficult problems on which they rely in the creation of blind signatures.

- **Decisional Diffie-Hellman (DDH):** given P, aP, bP, cP output *yes* if $c = ab$, or output *no* otherwise.
- **Computational Diffie-Hellman (CDH):** given P, aP, bP output abP.
- **Gap Diffie-Hellman group (GDH):** Group G is GDH group if there is efficient polynomial time algorithm which solves the DDH problem and there is no polynomial time algorithm which solves the CDH problem.
- **Bilinear Diffie-Hellman (BDH):** given P, aP, bP, cP output $e(P, P)^{abc}$.
- **Decisional Bilinear Diffie-Hellman (DBDH):** given $P, aP, bP, cP, r, a, b, c$ output *yes* if $r = e(P, P)^{abc}$, or output *no* otherwise.
- **ROS problem:** Given an oracle random function $F : Z_q^l \to Z_q$, find coefficients $a_{k,i} \in Z_q$ and a solvable system of $l + 1$ distinct equations in the unknowns $c1, c2, ..., c_l$ over Z_q: $a_{k,1}c_1 + ... + a_{k,l}c_l = F(a_{k,1}, ..., a_{k,l})$, for $k = 1, 2, ..., t$.
- **Bilinear Diffie-Hellman Inversion (BDHI):** Given aP, bP, cP compute S, T s.t. $e(S, T) = e(P, P)^{abc}$.

III. Blind signatures

This section overviews the ID-based blind signature schemes in [16], [17], [18], [19], [20], [21], [22], in order of appearance in the scientific literature. Each of this schemes is explained in its own subsection.

A. ZK1 scheme [16]

This scheme is the first ID-based blind signature scheme from bilinear pairings. The IBSS scheme is given as follows:

[Setup]: Let P be a generator of GDH group G_1. PKG chooses a random $s \in Z_q^*$ and set $P_{pub} = sP$. Then two hash functions are defined as $H_1 : \{0, 1\}^* \to Z_q$ and $H_2 : \{0, 1\}^* \to G_1$. The published parameters are $\{G_1, G_2, q, P, P_{pub}, H_1, H_2\}$ and s is the master key.

[Extract]: Given identity ID, PKG calculates public key $Q_{ID} = H_2(ID)$ and returns private key $S_{ID} = sQ_{ID}$.

[Issue]: This algorithm is proposed as interactive protocol between the signer and the user:

- **Commitment:** Signer randomly chooses $r \in Z_q^*$, computes $R = rP$ and sends R to User.
- **Blinding:** User randomly chooses $a, b \in Z_q^*$, computes $c = H_1(m, e(bQ_{ID} + R + aP, P_{pub})) + b$ and sends c to Signer.
- **Signing:** Signer computes $S = cS_{ID} + rP_{pub}$ and sends S to User.
- **Unblinding:** User computes $S^{'} = S + aP_{pub}$ and $c^{'} = c - b$ and outputs $\{m, S^{'}, c^{'}\}$.

[Verify]: Verify that $c^{'} = H_1(m, e(S^{'}, P)e(Q_{ID}, P_{pub})^{-c^{'}})$.

The authors in [16] proved that the scheme is blind and that the security on the generic parallel attack depends on the ROS problem. But, Wagner in [25] claimed that ROS can be broken in subexponential time and in order to be resistant q needs to be at least 1600 bits long. This makes this scheme inefficient in a practical way.

B. ZK2 scheme [17]

The authors in [16] proposed new IBSS scheme which does not depend on the ROS problem. Their scheme is given as follows:

[Setup]: This algorithm is the same as in ZK1 scheme.

[Extract]: This algorithm is the same as in ZK1 scheme.

[Issue]: This algorithm is an interactive protocol between the signer and the user:

- **Commitment:** Signer randomly chooses $r \in Z_q^*$, computes $U = rQ_{ID}$ and sends U to User.
- **Blinding:** User randomly chooses $\alpha, \beta \in Z_q^*$, computes $U^{'}\alpha U + \alpha\beta Q_{ID}$ and $h = \alpha^{-1}H_1(m, U^{'}) + \beta$, and sends h to Signer.
- **Signing:** Signer computes $V = (r + h)S_{ID}$ and sends V to User.
- **Unblinding:** User computes $V^{'} = \alpha V$ and outputs $\{m, U^{'}, V^{'}\}$.

[Verify]: Verify that $e(V^{'}, P) = e(U^{'} + H_1(m, U^{'})Q_{ID}, P_{pub})$.

This scheme has proven the blindness property, but there is no formal proof against one-more signature forgery.

C. HCW scheme [18]

The authors in [18] attack the unforgeability claim under generic parallel attack of ZK2 scheme [17] and present new efficient IBSS scheme secure under the ROM model:

[Setup]: Let P be a generator of GDH group G_1. PKG chooses random $r \in Z_q^*$ and set $P_{pub} = sP$. Then two hash functions are defined as $H_1 : \{0,1\}^* \times G_2 \rightarrow Z_q^*$ and $H_2 : \{0,1\}^* \rightarrow G_1$. The published parameters are $\{G_1, G_2, e, q, P, P_{pub}, H_1, H_2\}$ and s is the master key.

[Extract]: Given identity ID, PKG computes $P_{ID} = H_2(ID)$ and returns private key $S_{ID} = sP_{ID}$.

[Issue]: This algorithm is an interactive protocol between the signer and the user:

- **Commitment:** Signer randomly chooses $r \in Z_q^*$, computes $R^{'} = e(P_{ID}, P_{pub})^r$ and sends $R^{'}$ to User.
- **Blinding:** User randomly chooses $t_1, t_2 \in Z_q^*$, computes $R = R^{'t_1}e(P_1, P)^{t_2}$, $h = H_1(m, R)$, $h^{'} = ht_1$ and sends $h^{'}$ to Signer.
- **Signing:** Signer computes $V^{'} = (rh^{'} + 1)S_{ID}$ and sends $V^{'}$ to User.
- **Unblinding:** User checks if $e(V^{'}, P) = R^{th^{'}}e(P_{ID}, P_{pub})$, computes $V = V^{'} + ht_2P_1$ and outputs $\{m, R, V\}$.

[Verify]: Verify that $e(V, P) = R^{H_1(m,R)}e(P_{ID}, P_{pub})$.

Authors have proven unforgeability of this scheme under ROM, but they don't give formal proof against one-more signature forgery, just like the scheme in [17].

D. GWWF1 scheme [19]

This scheme is the first IBSS that proposes optimal rounds in the interactive protocol. GWWF1 has one-round interactive protocol between the user and signer in the phase of issuing of blind signatures. The scheme is given as:

[Setup]: Let P be a generator of GDH group G_1. PKG chooses random $r \in Z_q^*$ and set $P_{pub} = sP$. Then two hash functions are defined as $H_1, H_2 : \{0,1\}^* \rightarrow G_1$. The published parameters are $\{G_1, G_2, e, q, P, P_{pub}, H_1, H_2\}$ and s is the master key.

[Extract]: Given identity ID, PKG computes $Q_{ID} = H_1(ID)$ and returns private key $D_{ID} = sQ_{ID}$.

[Issue]: This algorithm is an interactive protocol between the signer and the user:

- **Blinding:** User randomly chooses $r_1 \in Z_q^*$; computes $P_m^{'} = r_1H_2(m)$ and sends $P_m^{'}$ to Signer.
- **Signing:** Signer randomly chooses $x_{ID} \in Z_q^*$, computes $A^{'} = x_{ID}P_m^{'}$, $B^{'} = x_{ID}^{-1}D_{ID}$, $C^{'} = x_{ID}P$, and sends $A^{'}, B^{'}$ and $C^{'}$ to User.
- **Unblinding:** User verifies $e(A^{'}, P) = e(P_m^{'}, C^{'})$ and $e(Q_{ID}, P_{pub}) = e(B^{'}, C^{'})$, then chooses random $r_2 \in Z_q^*$, computes $A = r_2r_1^{-1}A^{'}$, $B = r_2^{-1}B^{'}$ and $C = r_2C^{'}$ and publishes $\{m, A, B, C\}$.

[Verify]: Compute $P_m = H_2(m)$ and verify that $e(A, P) = e(P_m, C)$ and that $e(Q_{ID}, Ppub) = e(B, C)$.

Authors have proven the blindness property and security against one-more forgery under parallel chosen message and given ID attacks. These proofs are based on one-more BDHI

assumption that authors devised in their proof (similar to the one-more RSA-inversion in [26] which proves Chaum's blind signature scheme [7]).

E. KKS scheme [20]

The KKS scheme is actually a generalization of modified ElGamal signatures which are extended to blind signatures scheme [27]. Horster [28] showed that using the general signature equation $\tilde{A} \equiv \alpha\tilde{B}+k\tilde{C} \pmod q$ one can obtain many variations of the blind signature scheme. The KKS scheme has 18 variants based on permutations of the general signature equation. Here we present just one variant of the possible permutations:

[Setup]: This algorithm is the same as the Setup algorithm in HCW scheme.

[Extract]: This algorithm is also the same as the Extract algorithm in HCW scheme.

[Issue]: This algorithm is an interactive protocol between the signer and the user:

- **Commitment:** Signer chooses random $k \in Z_q^*$, computes $\tilde{t} = e(P,P)^k$ and sends $\tilde{t}$ to User.
- **Blinding:** User chooses random $a, b \in Z_q^*$, computes $t = \tilde{t}^a e(P,P)^b$, $r = H_1(m,t)$, $\tilde{r} = a^{-1}r$ and sends $\tilde{r}$ to Signer.
- **Signing:** Signer computes $\tilde{U} = \tilde{r}S_{ID} + kP$ and sends U to User.
- **Unblinding:** User verifies that $e(\tilde{U},P)e(Q_{ID},P_{pub})^{-\tilde{r}} = \tilde{r}$, computes $U = a\tilde{U}+bP$ and outputs $\{m,r,U\}$.

[Verify]: Verify that $e(U,P)e(Q_{ID},P_{pub})^{-r} = r^m$.

The authors give proof for the blindness property, but do not give official proof against one-more signature forgery.

F. RARG scheme [21]

The RARG scheme in construction is very similar to the schemes [18] and [20] and is based on Hess ID-based signature scheme [29].

[Setup]: This algorithm is the same as the Setup algorithm in HCW scheme.

[Extract]: This algorithm is also the same as the Extract algorithm in HCW scheme.

[Issue]: This algorithm is an interactive protocol between the signer and the user:

- **Commitment:** Signer randomly chooses $k \in Z_q^*$, computes $R = e(P,P)^k$ and sends R to User.
- **Blinding:** User randomly chooses $a, b \in Z_q^*$, computes $R^{'} = e(bQ_{IDs}+aP, P_{pub})R$, $V = H_1(m,R^{'})+b$ and sends V to Signer.
- **Signing:** Signer computes $S = Vd_{IDs}+kP$ and sends S to User.
- **Unblinding:** User computes $S^{'} = S + aP_{pub}$, $V^{'} = V - b$ and outputs $\{m, S^{'}, V^{'}\}$.

[Verify]: Verify that

$$V^{'} = H_1(m, e(S^{'},P)e(Q_{IDs},P_{pub})^{-V^{'}}).$$

The authors prove the blindness property of the scheme and the unforgeability due to Hess ID-based signature [29].

G. GWWF2 scheme [22]

This scheme is an improved version of the GWWF1 scheme [19]. In [22] the authors give two versions of their scheme in [19], named IBBS-1 and IBBS-2. Both schemes are very similar to the GWWF1 scheme. We would not go into details, but will list the resulting cost of the scheme together with all the schemes reviewed in this paper.

In the next section we present comparable results of the computation and bandwidth cost of the schemes explained in this section.

TABLE II. BANDWIDTH COST OF SCHEMES: G_1 - ELEMENT FROM GROUP 1; G_P - ELEMENT FROM PAIRING GROUP; S - ELEMENT FROM THE BASE FIELD (SCALAR); C - COMMITMENT PHASE; B - BLINDING PHASE; S - SIGNING PHASE; P - SIGNATURE PUBLISH PHASE.

scheme	direction	G_1	G_p	s
ZK1	**C: S to U**	1		
	B: U to S			1
	S: S to U	1		
	P: U to BB	1		1
ZK2	**C: S to U**	1		
	B: U to S			1
	S: S to U			1
	P: U to BB	2		
HCW	**C: S to U**		1	
	B: U to S			1
	S: S to U	1		
	P: U to BB	1	1	
GWWF1	**B: U to S**	1		
	S: S to U	3		
	P: U to BB	3		
KKS	**C: S to U**		1	
	B: U to S			1
	S: S to U	1		
	P: U to BB	1		1
RARG	**C: S to U**		1	
	B: U to S			1
	S: S to U	1		
	P: U to BB	1		1
GWWF1,IBBS-1	**B: U to S**	1		
	S: S to U	3		
	P: U to BB	3		
GWWF1,IBBS-2	**B: U to S**	1		
	S: S to U	3		
	P: U to BB	3		

IV. RESULTS

The aim of this paper is to give comparative review of ID-based blind signature schemes in the scientific literature. We reviewed seven schemes [16], [17], [18], [19], [20], [21], [22] and gave detailed description of their algorithms in the previous section and the security assumptions and proofs used by the authors. In this section we present the results in several forms.

The first comparative table I is representing computation cost - it gives count of operations with elements in the base field and the group elements. For better understanding of this

TABLE I. COMPUTATIONAL COST OF SCHEMES: P - PAIRING; M - SCALAR AND GROUP 1 ELEMENT MULTIPLICATION; A - ADDITION OF TWO GROUP 1 ELEMENTS; H - HASH FUNCTION $H : \{0,1\}^* \rightarrow G_1$; M_S - MULTIPLICATION OF TWO SCALARS; A_S - ADDITION OF TWO SCALARS; I_S - INVERSION OF SCALAR; C_S - COMPARISON OF TWO SCALARS; H_S - HASH FUNCTION $H_s : \{0,1\}^* \times G_2 \rightarrow Z_q^*$; E_P - EXPONENTIATION OF PAIRING ELEMENT; M_P - MULTIPLICATION OF TWO PAIRING ELEMENTS; C_P - COMPARISON OF TWO PAIRING ELEMENTS.

scheme		P	M	A	H	M_s	A_s	I_s	C_s	H_s	E_p	M_p	C_p
ZK1	**S**		3	1									
	U	1	3	3			1			1			
	V	2						1		1	1	1	
ZK2	**S**		2				1						
	U		3	1		2	1	1		1			
	V	2	1	1						1			1
HCW	**S**	1	1				1				1		
	U	3	1	1	1	2					2	2	
	V	2			1						1	1	1
GWWF1	**S**		3					1					
	U	4	4		1	1		2					2
	V	4			1								2
KKS	**S**	1	2	1							1		
	U					1		2		1	3	2	1
	V	2						1		1	1		
RARG	**S**	1	2	1							1		
	U	1	4	2			2			1			
	V	2						1	1	1	1	1	
GWWF2, IBBS-1	**S**		3					1					
	U	4	4		1	1		2					2
	V	4			1								2
GWWF2, IBBS-2	**S**		3	1									
	U	4	5	4	1			1					2
	V	4		1	1								2

TABLE III. BANDWIDTH COST OF SCHEMES: GIVEN BANDWIDTH NEEDED FOR EACH PHASE, AND TOTAL BANDWIDTH FOR THE SIGNING PROTOCOL. THE NUMBERS OF RESULTS ARE GIVEN IN BYTES.

scheme	direction	cost:
ZK1	**C: S to U**	**32**
	B: U to S	**32**
	S: S to U	**32**
	P: U to BB	**64**
	total:	**160**
ZK2	**C: S to U**	**32**
	B: U to S	**32**
	S: S to U	**32**
	P: U to BB	**64**
	total:	**160**
HCW	**C: S to U**	**384**
	B: U to S	**32**
	S: S to U	**32**
	P: U to BB	**416**
	total:	**864**
GWWF1	**B: U to S**	**32**
	S: S to U	**96**
	P: U to BB	**96**
	total:	**224**
KKS	**C: S to U**	**384**
	B: U to S	**32**
	S: S to U	**32**
	P: U to BB	**64**
	total:	**512**
RARG	**C: S to U**	**384**
	B: U to S	**32**
	S: S to U	**32**
	P: U to BB	**64**
	total:	**512**
GWWF1,IBBS-1	**B: U to S**	**32**
	S: S to U	**96**
	P: U to BB	**96**
	total:	**224**
GWWF1,IBBS-2	**B: U to S**	**32**
	S: S to U	**96**
	P: U to BB	**96**
	total:	**224**

table, the reader must have in mind the cost of each of these operations. This can be done by looking at the performance results of any implemented library, for example [23, Subsection 3.3].

The second comparative table II is representing bandwidth cost - it gives an overview of what is sent in the interactive protocol of signing, such as the number of group elements, pairing elements and scalars.

The third comparative table III is again about the bandwidth cost. It is depicting the spent bandwidth in each step in the interactive protocol, and the total spent bandwidth after signing algorithm. The results are calculated based on Barreto-Naehrig family of pairing friendly elliptic curves [30]. The packed size of element in G_1 is actually just the x value of the point on the curve. That size is 32 bytes. In order to retrieve the full point (or unpack) we can just solve the elliptic curve equation to get the y value. The scalar element is also 32 bytes, and the pairing element is 384 bytes. These numbers can vary in different libraries for arithmetic in elliptic curves. We refer the reader to the Panda project [23] which has implemented base field arithmetic, group field arithmetic and fast pairing implementation under Barreto-Naehrig curves.

V. CONCLUSION

This paper rounds up significant propositions for ID-based blind signatures. We can see the inner algorithms of the proposed schemes side-by-side and draw conclusions for our benefit. Further more, we have given comparative results for the cost of implementation and running of these schemes. The detailed comparison of computation cost gives number of arithmetic operations in each phase of the interactive protocol of the signing algorithm. To the best of our knowledge, this is first paper which has given detailed bandwidth comparison of the phases in the signing algorithm.

The results of this paper can be used when choosing ID-based blind signature scheme. Depending on the overall e-voting scheme, computational resources and possible bottlenecks, the reader can easily find out the right scheme. If the e-voting scenario is using powerful administrator nodes, blind signatures can be chosen s.t. the burden of the signature is with the signer. This comes very well in scenarios of computationally-weak users (internet browsers, mobile phones). In some scenarios the bandwidth may be bottleneck, for example SMS e-voting. Or maybe the number of rounds are critical because the establishment of communicational channels is hard (or expensive). All these prerequisites can be used to easily find the appropriate blind signature scheme based on our results.

ACKNOWLEDGEMENT

This work is partially supported by the Faculty of Computer Science and Engineering at "Ss. Cyril and Methodius" University in Skopje, Macedonia.

REFERENCES

[1] A. Shamir, "Identity-based cryptosystems and signature schemes," in *Advances in Cryptology*, ser. Lecture Notes in Computer Science, G. Blakley and D. Chaum, Eds. Springer Berlin Heidelberg, 1985, vol. 196, pp. 47–53.

[2] A. Menezes, S. Vanstone, and T. Okamoto, "Reducing elliptic curve logarithms to logarithms in a finite field," in *Proceedings of the Twenty-third Annual ACM Symposium on Theory of Computing*, ser. STOC '91. New York, NY, USA: ACM, 1991, pp. 80–89.

[3] G. Frey and H.-G. Rück, "A remark concerning m-divisibility and the discrete logarithm in the divisor class group of curves," *Math. Comput.*, vol. 62, no. 206, pp. 865–874, Apr. 1994.

[4] A. Joux, "A one round protocol for tripartite diffie-hellman," in *Proceedings of the 4th International Symposium on Algorithmic Number Theory*, ser. ANTS-IV. London, UK, UK: Springer-Verlag, 2000, pp. 385–394.

[5] D. Boneh and M. K. Franklin, "Identity-based encryption from the weil pairing," in *Proceedings of the 21st Annual International Cryptology Conference on Advances in Cryptology*, ser. CRYPTO '01. London, UK, UK: Springer-Verlag, 2001, pp. 213–229.

[6] D. Boneh, B. Lynn, and H. Shacham, "Short signatures from the weil pairing," in *Proceedings of the 7th International Conference on the Theory and Application of Cryptology and Information Security: Advances in Cryptology*, ser. ASIACRYPT '01. London, UK, UK: Springer-Verlag, 2001, pp. 514–532.

[7] D. Chaum, "Blind signatures for untraceable payments," in *Advances in Cryptology*, D. Chaum, R. Rivest, and A. Sherman, Eds. Springer US, 1983, pp. 199–203.

[8] P. Horster, H. Petersen, P. Horster, and H. Petersen, "Classification of blind signature schemes and examples of hidden and weak blind signatures," 1994.

[9] A. Fujioka, T. Okamoto, and K. Ohta, "A practical secret voting scheme for large scale elections," in *Advances in Cryptology AUSCRYPT '92*, ser. Lecture Notes in Computer Science, J. Seberry and Y. Zheng, Eds. Springer Berlin Heidelberg, 1993, vol. 718, pp. 244–251.

[10] Y. Mu and V. Varadharajan, "Anonymous secure e-voting over a network," in *Computer Security Applications Conference, 1998. Proceedings. 14th Annual*, Dec 1998, pp. 293–299.

[11] B. Kharchineh and M. Ettelaee, "A new electronic voting protocol using a new blind signature scheme," in *Future Networks, 2010. ICFN '10. Second International Conference on*, Jan 2010, pp. 190–194.

[12] C.-T. Li, M.-S. Hwang, and Y.-C. Lai, "A verifiable electronic voting scheme over the internet," in *Information Technology: New Generations, 2009. ITNG '09. Sixth International Conference on*, April 2009, pp. 449–454.

[13] G. Gallegos-García, R. Gómez-Cárdenas, and G. I. Duchén-Sánchez, "Identity based threshold cryptography and blind signatures for electronic voting," *W. Trans. on Comp.*, vol. 9, no. 1, pp. 62–71, Jan. 2010. [Online]. Available: http://dl.acm.org/citation.cfm?id=1852381.1852388

[14] L. Zhang, Y. Hu, X. Tian, and Y. Yang, "Novel identity-based blind signature for electronic voting system," in *Education Technology and Computer Science (ETCS), 2010 Second International Workshop on*, vol. 2, March 2010, pp. 122–125.

[15] L. Lopez-Garcia, L. J. D. Perez, and F. Rodriguez-Henriquez, "A pairing-based blind signature e-voting scheme," *The Computer Journal*, 2013.

[16] F. Zhang and K. Kim, "Id-based blind signature and ring signature from pairings," in *Proceedings of the 8th International Conference on the Theory and Application of Cryptology and Information Security: Advances in Cryptology*, ser. ASIACRYPT '02. London, UK, UK: Springer-Verlag, 2002, pp. 533–547.

[17] Z. Fangguo and K. Kwangjo, "Efficient id-based blind signature and proxy signature," in *In Proceedings of ACISP 2003, LNCS 2727*. Springer-Verlag, 2003, pp. 312–323.

[18] Z. Huang, K. Chen, and Y. Wang, "Efficient identity-based signatures and blind signatures," in *Proceedings of the 4th International Conference on Cryptology and Network Security*, ser. CANS'05. Berlin, Heidelberg: Springer-Verlag, 2005, pp. 120–133.

[19] W. Gao, G. Wang, X. Wang, and F. Li, "One-round id-based blind signature scheme without ros assumption," in *Proceedings of the 2Nd International Conference on Pairing-Based Cryptography*, ser. Pairing '08. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 316–331.

[20] S. Kalkan, K. Kaya, and A. Selcuk, "Generalized id-based blind signatures from bilinear pairings," in *Computer and Information Sciences, 2008. ISCIS '08. 23rd International Symposium on*, Oct 2008, pp. 1–6.

[21] U. Rao B. and A. K. A., "An id-based blind signature scheme from bilinear pairings," in *International Journal of Computer Science and Security Volume (4) Issue (1)*, march 2010, pp. 98–106.

[22] W. Gao, G. Wang, X. Wang, and F. Li, "Round-optimal id-based blind signature schemes without ros assumption," *Journal of Communications*, vol. 7, no. 12, 2012.

[23] C. Chuengsatiansup, M. Naehrig, P. Ribarski, and P. Schwabe, "Panda: Pairings and arithmetic," in *Pairing-Based Cryptography – Pairing 2013*, ser. Lecture Notes in Computer Science, Z. Cao and F. Zhang, Eds. Springer-Verlag Berlin Heidelberg, 2013.

[24] M. Bellare and P. Rogaway, "Random oracles are practical: A paradigm for designing efficient protocols," in *Proceedings of the 1st ACM Conference on Computer and Communications Security*, ser. CCS '93. New York, NY, USA: ACM, 1993, pp. 62–73.

[25] D. Wagner, "A generalized birthday problem," in *Proceedings of the 22Nd Annual International Cryptology Conference on Advances in Cryptology*, ser. CRYPTO '02. London, UK, UK: Springer-Verlag, 2002, pp. 288–303.

[26] M. Bellare, C. Namprempre, D. Pointcheval, and M. Semanko, "The power of rsa inversion oracles and the security of chaums rsa-based blind signature scheme," in *In Financial Cryptography*, 2001, pp. 319–338.

[27] J. Camenisch, J.-M. Piveteau, and M. Stadler, "Blind signatures based on the discrete logarithm problem," in *Advances in Cryptology EUROCRYPT'94*, ser. Lecture Notes in Computer Science, A. Santis, Ed. Springer Berlin Heidelberg, 1995, vol. 950, pp. 428–432.

[28] P. Horster, M. Michels, and H. Petersen, "Meta-message recovery and meta-blind signature schemes based on the discrete logarithm problem and their applications," in *Advances in Cryptology ASIACRYPT'94*, ser. Lecture Notes in Computer Science, J. Pieprzyk and R. Safavi-Naini, Eds. Springer Berlin Heidelberg, 1995, vol. 917, pp. 224–237.

[29] F. Hess, "Efficient identity based signature schemes based on pairings," in *Selected Areas in Cryptography*, ser. Lecture Notes in Computer Science, K. Nyberg and H. Heys, Eds. Springer Berlin Heidelberg, 2003, vol. 2595, pp. 310–324.

[30] P. S. Barreto and M. Naehrig, "Pairing-friendly elliptic curves of prime order," in *Selected Areas in Cryptography*, ser. Lecture Notes in Computer Science, B. Preneel and S. Tavares, Eds. Springer Berlin Heidelberg, 2006, vol. 3897, pp. 319–331.

Homomorphic Encryption in the Cloud

Darko Hrestak and Stjepan Picek
Faculty of Electrical Engineering and Computing, University of Zagreb
Unska 3, Zagreb, Croatia
Email: darko.hrestak@fer.hr, stjepan@computer.org

***Abstract*—Since the first notions of fully homomorphic encryption more than 30 years ago, there has been numerous attempts to develop such a system. Finally, in 2009 Craig Gentry succeeded. Homomorphic encryption brings great advantages but it seems that, at least for now, it also brings many practical difficulties. Furthermore, in the last couple of years, several other fully homomorphic systems arose where each has its one advantages and drawbacks. However, with the developments in cloud computing, we need it more than ever to become practical for real-world usages. In this paper we are discussing the strengths and weaknesses of homomorphic encryption and we give a brief description of several promising fully homomorphic encryption systems. Next, we give a special attention to the homomorphic encryption systems for cloud computing. Finally, we discuss some recent developments by IBM and their open-source library for homomorphic encryption.**

I. Introduction

Today, cryptography plays important role in our lives, although many people are not fully aware of that fact. Whenever we buy something over the Internet or withdraw cash from ATM machine, cryptography protects our sensitive information.

One of the basic questions in cryptography concerns how it is possible to safely exchange information. If two parties, commonly called Alice and Bob, want to securely communicate, then they need a shared secret. This notion represents the basis of private key cryptography. In this setting Alice and Bob both have the same secret (called secret key) and they encrypt and decrypt data with it. All secure communication is limited to those that have that secret key. Of course, here we are simplifying the situation and assume that key and methods of encryption are appropriately chosen. If some third party, Eve, intercepts encrypted message it should be practically infeasible for her to decrypt it. Now, even from this abstract description, it is possible to observe one obvious drawback of private key cryptography. In order to be able to encrypt and afterwards decrypt the data, Alice and Bob must use the same key. If we add an additional condition that the key needs to be changed from time to time, question becomes, how to securely communicate (in order to exchange the key) before they can securely communicate to exchange the actual data. Almost 40 year ago, public key cryptography arose to solve that problem. In public key cryptography security relies on hard mathematical problems and each party needs a public key for encryption and a private key for decryption [1].

Additionally, Alice and Bob also wants for instance to be sure that the other party is he who says it is, and that the messages can not be changed while transmitted. For each of those constraints, appropriate solution is devised and implemented in practice. However, only a couple of months after the publication of the RSA algorithm paper [2], Rivest et al. asked the question whether it is possible to work with encrypted data, without the need for decrypting it first [3]. That question started the search for homomorphic encryption systems. It turned out that this question would pose unsolvable problem in following decades. Before going into more details about homomorphic encryption, we give a simple example why we would even need an option like that. If we have some information and we send it to the server for some additional processing, we may not want that the server can read our information. In the modern society, the need for privacy is present as never before and homomorphic encryption can help us in preserving it.

Homomorphic encryption allows us to perform operations on encrypted data without knowing the private key and without decrypting that data. When we decrypt the result of any operation, it is the same as if we carried out the calculation on the plain data. In the process of building a system with such capabilities, a plethora of systems were developed that offer partial functionalities. Those types of homomorphic encryption systems are known as partially homomorphic encryption systems. In 2009, Craig Gentry in his PhD thesis described the first fully homomorphic encryption system (FHE) [4]. From then, a number of systems were (and still are) developed that offer similar or better characteristics.

In Section 2 we give a brief introduction to the basic notions about homomorphic encryption. Several partially homomorphic encryption systems are described in Section 3. In Section 4 we describe fully homomorphic encryption as presented by Gentry and several other fully homomorphic encryption systems. Possible usages of fully homomorphic encryption with particular accent on cloud computing are discussed in Section 5. Finally, in Section 6 we give a conclusion.

II. Preliminaries

Here we give basic notions about homomorphic encryption and schemes presented. For a detailed introduction to general cryptographic notions we refer readers to [1].

A Homomorphic encryption is additive if:

$$Enc(x + y) = Enc(x) \cdot Enc(y). \quad (1)$$

A homomorphic encryption is multiplicative if:

$$Enc(x \cdot y) = Enc(x) \cdot Enc(y). \quad (2)$$

Next, we present building blocks of the homomorphic encryption system.

- Key generation is the algorithm that outputs public key P_k and secret key S_k.
- Encrypt function is the algorithm that takes public key P_k and a message m and outputs ciphertext c.
- Decrypt function is the algorithm that takes ciphertext c and a secret key S_k and outputs a message m.
- Evaluate function is an algorithm that takes evaluation key, circuit, a set of ciphertexts $c_1, ..., c_n$ and outputs a ciphertext c_e. Circuit represents a certain function realized with logical gates.

III. Partially homomorphic encryption systems

There are two different types of homomorphic encryption systems, fully homomorphic encryption systems and partially homomorphic encryption systems. Fully homomorphic encryption systems are defined over a ring and therefore support both operations, addition and multiplication. Partially homomorphic encryption systems are defined over a group and therefore support just a single operation on the encrypted data.

In last thirty or so years there have been several encryption systems that belong to the partially homomorphic encryption systems class. Those systems we can divide on a basis of the operation they support on additive homomorphic encryption systems and multiplicative homomorphic encryption systems. There are also systems that allow arbitrary many operations of one type and only a limited number (usually just one) of operations of other type.

A. Additive Homomorphic Encryption

Some well known examples of additive homomorphic encryption systems include Goldwasser-Micali system [5], Paillier system [7] and Damgard-Jurink's generalization of Paillier system [6].

The Paillier cryptosystem, invented by Pascal Paillier in 1999, is a probabilistic asymmetric algorithm for public key cryptography [7]. This cryptosystem is based on the decisional composite residuosity assumption (DCRA). DCRA assumption can be stated as the following: given a composite n ($n = p \times q$ for primes p and q) and an integer z, it is hard to decide whether z is a $n-$ residue modulo n^2 or not (i.e. it is hard to find out whether there exists y such that $z = y^n mod n^2$). Paillier cryptosystem computes encryption of $m_1 + m_2$ with public key and encryptions of m_1 and m_2. Short description of Paillier scheme is given in Algorithm 1. Implementation of that cryptosystem can be found online under the name "The Homomorphic Encryption Project" (thep) and it's main application is electronic voting where each vote is encrypted but only the sum is decrypted [8].

Suppose that we have two ciphers C_1 and C_2 such that following equations hold:

$$\begin{aligned} C_1 &= g^{m_1} \cdot r^n \ mod\ n^2, \\ C_2 &= g^{m_2} \cdot r^n mod\ n^2, \\ C_1 \cdot C_2 &= g^{m_1+m_2} \cdot (r_1 \cdot r_2)^n mod\ n^2. \end{aligned} \tag{3}$$

Algorithm 1 Paillier scheme

Key Generation KeyGen (p, q)
$Input : p, q \in P$
$Compute : n = p \cdot q, \lambda = lcm(p-1, q-1)$
$Choose : g \in Z^*_{n^2} \mid gcd(L(g^\lambda \ mod \ n^2), n) = 1 \ where \ L(u) = (u-1)/n$
$Compute : \mu = L(g^\lambda \ mod \ n^2)^{-1}$
$Output : (P_k, S_k)$
$Public \ key : P_k = (n, q)$
$Secret \ key : S_k = (\lambda, \mu)$
Encryption Enc (m, P_k)
$Input : m \in Z_n$
$Choose : r \in Z^*_n$
$Compute : c = g^m \cdot r^n \ mod \ n^2$
$Output : c \in Z_{n^2}$
Decryption Dec (c, S_k)
$Input : c \in Z_{n^2}$
$Compute : m = L(c^\lambda \ mod \ n^2) \cdot \mu \ mod \ n$
$Output : m \in Z_n$

It can be observed that Paillier cryptosystem realizes the property of additive homomorphic encryption since the product of two ciphertexts will decrypt to the sum of their corresponding plaintexts.

B. Multiplicative Homomorphic Encryption

When looking at the multiplicative homomorphic cryptosystem, good examples are previously mentioned RSA algorithm [3] and ElGamal encryption system [9]. Next we give a short description of the RSA scheme [3] as shown in Algorithm 2.

Algorithm 2 RSA scheme

Key Generation KeyGen (p, q)
$Input : p, q \in P$
$Compute : n = p \cdot q, \phi(n) = (p-1) \cdot (q-1)$
$Choose : e \mid gcd(e, \phi(n)) = 1$
$Compute : d \mid (e \cdot d) \equiv 1 \ mod \ \phi(n)$
$Output : (P_k, S_k)$
$Public \ key : P_k = (e, n)$
$Secret \ key : S_k = d$
Encryption Enc (m, P_k)
$Input : m \in Z_n$
$Compute : c = m^e \ mod \ n$
$Output : c \in Z_n$
Decryption Dec (c, S_k)
$Input : c \in Z_{n^2}$
$Compute : m = c^d \ mod \ n$
$Output : m \in Z_n$

Suppose that we have two ciphers C_1 and C_2 such that following equations hold:

$$\begin{aligned} C_1 &= m_1^e \ mod\ n, \\ C_2 &= m_2^e \ mod\ n, \\ C_1 \cdot C_2 &= (m_1 \cdot m_2)^e mod\ n. \end{aligned} \tag{4}$$

It can be observed that RSA as a homomorphic encryption system has some security weaknesses. For example, if we assume that two messages m_1 and m_2 correspond to ciphers C_1 and C_2, and if the client sends that pair (C_1, C_2) to the server, the server will perform multiplication of C_1 and C_2 and send it encrypted to the client. If the attacker intercepts ciphers C_1 and C_2 which are encrypted with the same private key, he will be able to decrypt all messages exchanged between the server and the client. This is due the homomorphic encryption multiplicativity where product of the ciphers equals cipher of the product.

C. Additive and multiplicative homomorphic cryptosystem

Each of the examples listed above allows homomorphic computation of only one operation (either addition or multiplication) on plaintexts. A step closer to the full homomorphic encryption systems represent the systems that allow arbitrary many operations of one type and limited number of operations of other type. Probably the best example of a cryptosystem which allows homomorphic computation of addition and multiplication but is not fully homomorphic systems is Boneh-Goh-Nissim cryptosystem [10] as shown in Algorithm 3. Boneh-Goh-Nissim cryptosystem supports evaluation of an unlimited number of additions but at most one multiplication. This encryption scheme is based on bilinear pairings on elliptic curves.

Algorithm 3 Boneh-Goh-Nissim scheme

Key Generation KeyGen (r)
$Input : r \in Z^+$
$Compute : \gamma(r)\ to\ obtain\ tuple\ (q_1, q_2, G, G_1, e)$
$Compute : n = p \cdot q$
$Choose : g, u \in Random$
$Compute : h = u^{q_2}$
$Output : (P_k, S_k)$
$Public\ key : P_k = (n, G, G_1, e, g, h)$
$Secret\ key : S_k = q_1$
Encryption Enc (m, P_k)
$Input : m$
$Choose : r \in Random$
$Compute : c = g^m \cdot h^r \in G$
$Output : c$
Decryption Dec (c, S_k)
$Input : c$
$Compute : m = dilog(c^{q_1})\ mod\ g^{q_1}$
$Output : m$

Here, $dilog$ represents discrete logarithm operation.

IV. Fully Homomorphic Encryption Systems

In this section we present several fully homomorphic encryption systems, starting with Gentry's systems and then covering several other constructions.

A. Gentry's FHE

The first construction of fully homomorphic scheme (based on ideal lattices which are ideals in algebraic number field) is described by Craig Gentry in 2009 [4]. This work is important not only because it represents the first fully homomorphic encryption system, but also since it served as a paradigm for later fully homomorphic encryption schemes as the one we describe in section below.

Gentry's scheme consists of several steps. First, one constructs a somewhat homomorphic encryption scheme (SWHE), which only supports a limited number of operations. In SWHE system there are usual Encrypt and Decrypt functions, but also an Evaluate function. Evaluate function conducts an operation on ciphertext where it is realized as a combinatorial circuit. However, the problem with homomorphic encryption is that ciphertext contains random noise where addition roughly doubles the noise and multiplication squares the noise. This noise becomes larger with successive homomorphic operations and only ciphertexts whose noise size remains below a certain threshold can be decrypted correctly. The origin of noise is in the probabilistic encryption process where we add some noise during the encryption, and remove that noise during the decryption. Because of this limitation, homomorphic encryption can not have arbitrary number of operations and is therefore called somewhat homomorphic encryption. Somewhat homomorphic encryption scheme is shown in Algorithm 4. To overcome this problem, Gentry devised *bootstrapping* method. In bootstrapping, Evaluate function is used to run Decrypt function. Since Evaluate works with encrypted data, the result is not the plaintext but a new encryption with reduced noise. This refreshed ciphertext can then be used in subsequent homomorphic operations. By repeated refreshing of the ciphertexts, the number of homomorphic operations becomes unlimited, and that results in a fully homomorphic encryption scheme. In order for system to work as described Decrypt function must not exceed the noise threshold. If this would not be the case, we would not be able to apply bootstrapping to the SWHE scheme. Indeed, in Gentry's original FHE scheme this condition was not satisfied. Therefore, he added a *squashing* procedure to the Decrypt function. Here, squashing means transforming SWHE scheme into one with same homomorphic capacity but with simpler Decrypt function [11]. The cost of the squashing method is that the key grows larger and more complicated.

B. FHE

After 2009 and Gentry's breakthrough result, many improvements have been made, introducing new variants, improving efficiency, and providing new features. Recently, Brakerski, Gentry and Vaikutanathan described a different framework where the ciphertext noise grows only linearly with the multiplicative level instead of exponentially, so that bootstrapping technique is no longer necessary to obtain a scheme supporting the homomorphic evaluation of any given polynomial size circuit. Currently, there are 3 main families of fully homomorphic encryption schemes:

Algorithm 4 Somewhat homomorphic encryption scheme

Key Generation KeyGen (R, B_i)
$Input : R, basis\ B_i$
$Compute : (B_{P_k}, B_{S_k}) = IdealGen\ (R, B_i)$
$Output : (P_k, S_k)$
$Public\ key : P_k = (R, B_i, B_{P_k}, S_{amp}$
$Secret\ key : S_k = B_{S_k}$
Encryption Enc (m, P_k)
$Input : m, P_k$
$Compute : e = S_{amp}(m, B_b i)$
$Compute : c = e\ mod B_{P_k}$
$Output : c \in \mathbb{R}\ mod\ B_{P_k}$
Decryption Dec (c, S_k)
$Input : c \in \mathbb{R}\ mod\ B_{P_k}$
$Compute : m = (c\ mod\ B_{S_k})\ mod\ B_i$
$Output : m$

- Gentry's original scheme based on ideal lattices [4].
- Van Dijk, Gentry, Halevi and Vaikuntanathan's (DGHV) scheme over the integers [12].
- Brakerski and Vaikuntanathan's scheme based on the Learning with Errors (LWE) and Ring Learning with Errors (RLWE) problems [13].

Since we already described Gentry's original scheme, now we give short descriptions of the other two schemes.

In DGHV scheme [12] a ciphertext has the form of

$$C = q \cdot p + 2 \cdot r + m, \tag{5}$$

where p is the secret-key, m is the bit plaintext (m is 0 or 1), q is a large random number and r is a small random number.

To decrypt, one calculates

$$m = (C\ mod\ p)\ mod\ 2. \tag{6}$$

It is easy to see that decryption works as long as the noise $2 * r$ is smaller than p.

Suppose that we have two ciphers C_1 and C_2 such that following equations hold:

$$\begin{aligned} C_1 &= q_1 \cdot p + 2 \cdot r_1 + m_1, \\ C_2 &= q_2 \cdot p + 2 \cdot r_2 + m_2, \\ C_1 + C_2 &= (q_1 + q_2) \cdot p + 2 \cdot (r_1 + r_2) + m_1 + m_2, \\ C_1 \cdot C_2 &= q_{12} \cdot p + 2 \cdot (2 \cdot r_1 \cdot r_2 + r_1 \cdot m_2 + r_2 \cdot m_1) \\ &\quad + m_1 + m_2. \end{aligned} \tag{7}$$

From these equations, we see that we can obtain the encryption of $m_1 + m_2$ $(mod\ = m_1 \oplus m_2)$ by computing $C_1 + C_2$. Similarly, encryption of $m_1 \cdot m_2$ can be obtained by computing $C_1 \cdot C_2$. However one gets a new ciphertext with noise roughly twice larger than in the original ciphertexts C_1 and C_2. Since the noise must remain below p, the number of permitted multiplications on ciphertexts is therefore limited. This is an example of somewhat homomorphic encryption scheme.

Scheme from Brakerski and Vaikuntanathan is known as "Fully Homomorphic Encryption without Bootstrapping" where their approach is based on Learning with Errors (LWE) and Ring Learning with Errors (RLWE) problems [13].

In RLWE problem for security parameter λ, let $f(x) = x^d + 1$ where $d = d(\lambda)$ is a power of 2. Furthermore, let $q = q(\lambda) \geq 2$ be an integer, $R = Z[x]/f(x)$ and $R_q = R/qR$. Let $\xi = \xi(\lambda)$ be a distribution over R. The RLWE problem is to distinguish two distributions, in the first distribution, one samples (a_i, b_i) uniformly from R_{q^2}. In the second distribution, one first selects $s < -R_q$ uniformly and then samples (a_i, b_i) $2R_q^2$ by sampling $a_i < -R_q q$ uniformly, $e_i < -\xi$ and setting $b_i = a_i s + e_i$. The RLWE assumption is that the RLWE problem is infeasible.

In this scheme the authors are able to obtain a direct construction of a bootstrappable encryption scheme without having to "squash the decryption circuit", and hence there is no need to apply a bootstrapping technique to achieve fully homomorphic encryption. As per Gentry's scheme, Brakerski and Vaikuntanathan started from somewhat homomorphic encryption which was constructed to work over a ring R, where R can be described as $R = Z$ for integers, or $R = Z[x]/(x^d+1)$ for polynomial ring with d being a power of 2. For integer q, R_q is used to indicate R/qR. The error distribution parameter ξ is introduced to work over the ring R, where it is set to be as small as possible. Brakerski and Vaikuntanathan somewhat homomorphic algorithms is presented in Algorithm 5.

Algorithm 5 Brakerski-Vaikuntanathan

Key Generation KeyGen (l^λ)
$Sample : S_k = s \Leftarrow R_q \mid polynomial,\ small\ coefficients \in error\ distribution\ \xi$
Encryption Enc $(S_k, \mu \in R_2)$
$Sample : a \Leftarrow R_q\ randomly, polynomiale \mid small\ coefficients \in errordistribution\ \xi$
$Compute : c = (a, a_s = 2 \cdot e + \mu)$
Decryption Dec $(c = (a, b), S_k)$
$Sample : b \Leftarrow R_q\ randomly$
$Compute : \mu' = b \cdot a_s \in R_q$
$Output : \mu = \mu' mod 2$

The key contribution of this scheme is the technique of managing noise so that it increases linearly with the multiplicative level instead of exponentially which was the case in previous FHE schemes.

V. Going to the Cloud

As Internet evolved, computing concepts changed from parallel to distributed and grid computing and recently to cloud computing. If we look for the definition of cloud computing we can find that "Cloud computing is a model for enabling convenient, on-demand network access to a shared pool od configurable computing resources that can be rapidly provisioned and released with minimal management effort or service provider interaction" [14]. Cloud computing is the most popular notion in IT today, where even an academic

report from Berkeley says “Cloud computing is likely to have the same impact on software that foundries have had on the hardware industry" [15]. They go on to recommend that “developers would be wise to design their next generation of systems to be deployed into cloud computing" [15]. Cloud computing brings many business advantages such as accessing and using configurable computing resources through sharing and virtualization techniques, and all that at a lower cost. Main disadvantages are security and privacy issues that come along with resource sharing, virtualization and possible lack of trust between business and cloud providers. Many privacy issues could be solved with using homomorphic encryption, for instance cloud user can encrypt data and store it in the cloud, where both the user and cloud provider are unable to carry on any computations on data before decrypting it. It becomes more usual even for individuals, let alone for the companies to use the computational power of cloud computing, but often they would like their data to remain private. It is not so unrealistic to envision a system that collects health information from users in real time and on a basis of that information give alerts to the users. However, in a setting like that, user privacy is important and therefore the data needs to be encrypted and the operations in the cloud must run on encrypted data. It is worth mentioning that there are many possible applications that require only a limited number of operations and therefore somewhat homomorphic encryption systems would suffice.

A. HElib

Craig Gentry was the creator of first fully homomorphic encryption scheme, which was inefficient for implementation, but in the last 4 years efficiency has rapidly improved. In 2013 IBM presented open source software library for homomorphic encryption Helib [16]. HElib is a software library that has implementation of Brakerski-Gentry-Vaikuntanathan scheme, along with many optimizations to make homomorphic evaluation faster. To be exact, in 2010 all available FHE schemes were so inefficient that they would simply not run on any hardware, for instance, at the 2012 Crypto conference Gentry, Halevi and Smart showed that it “only" takes 36 hours to evaluate AES in the encrypted domain [17]. Only one year later this was possible to do under 3 hours. HElib is written in C++ and uses NTL mathematical library, however, it does not have implementation of bootstrapping. We can say that HElib is a young library in a rather young field of cryptography.

VI. Conclusions

Homomorphic encryption systems evolved significantly in the last couple of years since the design of first fully homomorphic encryption system. However, all those systems are still to impractical for real-world applications. Nevertheless, we already know numerous practical scenarios where we could use them. Because of that, it makes it interesting problem not only from academic perspective, but also from the industrial one. This will lead to the design of more efficient homomorphic encryption systems in the following years. Naturally, since there are applications where somewhat homomorphic encryption systems would be powerful enough, it seems we are closer to that goal than it may look like.

References

[1] J. Katz and Y. Lindell, *Introduction to Modern Cryptography (Chapman & Hall/Crc Cryptography and Network Security Series)*. Chapman & Hall/CRC, 2007.

[2] R. Rivest, A. Shamir, and L. Adleman, “A method for obtaining digital signatures and public-key cryptosystems,” *Communications of the ACM*, vol. 21, pp. 120–126, 1978.

[3] R. L. Rivest, L. Adleman, and M. L. Dertouzos, “On data banks and privacy homomorphisms,” *Foundations of Secure Computation, Academia Press*, pp. 169–179, 1978.

[4] C. Gentry, “A fully homomorphic encryption scheme,” Ph.D. dissertation, Stanford, CA, USA, 2009.

[5] S. Goldwasser and S. Micali, “Probabilistic encryption,” *J. Comput. Syst. Sci.*, vol. 28, no. 2, pp. 270–299, 1984.

[6] I. Damgard and M. Jurik, “A generalisation, a simplification and some applications of paillierś probabilistic public-key system,” in *In proceedings of PKC ’01, LNCS series*. Springer-Verlag, 2001, pp. 119–136.

[7] P. Paillier, “Public-key cryptosystems based on composite degree residuosity classes,” in *Proceedings of the 17th International Conference on Theory and Application of Cryptographic Techniques*, ser. EUROCRYPT’99. Berlin, Heidelberg: Springer-Verlag, 1999, pp. 223–238.

[8] “thep - The Homomorphic Encryption Project - Computation on Encrypted Data for the Masses,” 2014. [Online]. Available: https://code.google.com/p/thep/

[9] T. El Gamal, “A public key cryptosystem and a signature scheme based on discrete logarithms,” in *Proceedings of CRYPTO 84 on Advances in Cryptology*. New York, NY, USA: Springer-Verlag New York, Inc., 1985, pp. 10–18.

[10] D. Boneh, E.-J. Goh, and K. Nissim, “Evaluating 2-dnf formulas on ciphertexts,” in *Proceedings of the Second International Conference on Theory of Cryptography*, ser. TCC’05. Berlin, Heidelberg: Springer-Verlag, 2005, pp. 325–341.

[11] V. Vaikuntanathan, “Computing blindfolded: New developments in fully homomorphic encryption,” in *FOCS*, 2011, pp. 5–16.

[12] M. van Dijk, C. Gentry, S. Halevi, and V. Vaikuntanathan, “Fully homomorphic encryption over the integers,” in *EUROCRYPT*, 2010, pp. 24–43.

[13] Z. Brakerski and V. Vaikuntanathan, “Efficient fully homomorphic encryption from (standard) lwe,” in *FOCS*, 2011, pp. 97–106.

[14] P. Mell and T. Gance, “The NIST Definition of Cloud Computing (Final) (NIST Special Publication 800-145) ,” 2011.

[15] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. H. Katz, A. Konwinski, G. Lee, D. A. Patterson, A. Rabkin, and M. Zaharia, “Above the clouds: A berkeley view of cloud computing,” Tech. Rep., 2009.

[16] “HElib An Implementation of homomorphic encryption,” 2014. [Online]. Available: https://github.com/shaih/HElib

[17] C. Gentry, S. Halevi, and N. P. Smart, “Homomorphic evaluation of the aes circuit,” in *CRYPTO*, 2012, pp. 850–867.

Secure and Privacy-enhanced E-Mail System based on the Concept of Proxies

Ioannis Kounelis[§], SeadMuftic[§], Jan Löschner[*]
[§]Royal Institute of Technology (KTH), Stockholm, Sweden
[*]Institute for the Protection and Security of the Citizen
Joint Research Centre (JRC), European Commission, Ispra (VA), Italy
{kounelis, sead}@kth.se, jan.loeschner@jrc.ec.europa.eu

Abstract - Security and privacy on the Internet and especially the e–mail, is becoming more and more important and crucial for the user. The requirements for the protection of e-mail include issues like tracking and privacy intrusions by hackers and commercial advertisers, intrusions by casual observers, and even spying by government agencies. In an expanding e-mail use in the digital world, Internet and mobile, the quantity and sensitivity of personal information has also tremendously expanded. Therefore, protection of data and transactions and privacy of user information is key and of interest for many users. Based on such motives, in this paper we present the design and current implementation of our secure and privacy-enhanced e-mail system. The system provides protection of e–mails, privacy of locations from which the e–mail system is accessed, and authentication of legitimate users. Differently from existing standard approaches, which are based on adding security extensions to e-mail clients, our system is based on the concept of proxy servers that provide security and privacy of users and their e-mails. It uses all required standards: S/MIME for formatting of secure letters, strong cryptographic algorithms, PKI protocols and certificates. We already have the first implementation and an instance of the system is very easy to install and to use.

Keywords—E-mail; Proxy Server; S/MIME; X.509 certificates; PKI

I. INTRODUCTION

During the last two decades, e–mail has evolved from a simple replacement of paper letters to a necessary and unavoidable means of Internet communication. It is more than common for the majority of the people to possess at least one e-mail account from completely different providers.

E-mails are not only used for communicating between users but as well between users and automated services. E–mail addresses often represent digital identities, which allow the owner to authenticate to online services, such as social networks. Moreover, accessing one's e-mail account does not occur from only one particular device, but on the contrary, often from many different ones, both in terms of capabilities and also in terms of privacy risks for the user. For example, it is very significant for someone's privacy to assure a confidential use of e-mail while moving in a public environment or with public transport, as already shown by our work [1].

The e-mail exchange became of interest for intelligent services to collect material about the user contents and activities. Such collection takes place on the server side and is not user specific [11].

In this paper we describe the design and current implementation of our secure e-mail system based on the use of a secure proxy. We explain our motivation behind this idea, describe our design, and provide details of our implementation.

In section 2 we review the relevant work in this area. In section 3 the design principles of the secure web-based e-mail system are described, while in section 4 the motivation of using a proxy server as a solution is explained. In section 5 we describe the current implementation of our secure and privacy–enhanced e–mail system, as well as the current drawbacks and potential solutions to them. Finally, in section 6 we outline conclusions and suggestions for future research and development.

II. RELEVANT WORK

The concept of using a proxy server in order to enhance security for an application or in particular for e–mail is not new. Already in 1999, Brown et. al [2] have implemented a proxy server called Enigma to add security to mail clients of that time. Their solution was using only the Post Office Protocol (POP), but more importantly, the proxy they developed was not located remotely, but was an extra component on the user's workstation. As a result, their proxy was a local application acting as an extension of an e–mail client, communicating internally with the e–mail client and over the network with the corresponding e–mail server.

Abbasi et al [4] proposed a secure e-mail system with a proxy server, which is logically located at the domain of the mail server and acts as a proxy by introducing S/MIME and other security mechanisms, such as spam filtering and address book protection. Their design is different from our implementation as it consists of two separate components, the secure e-mail server and the secure e-mail client, which both reside at the corresponding mail server. These components are responsible for performing security functions and act as a proxy on behalf of a user. Moreover, the two elements are actively interfering with the functionality of native e–mail servers in order to apply some of the extra security functionalities, to apply spam policies for example. Our implementation, on the other hand, is more flexible, i.e. it does not interfere at all with functions and operations of the native mail server, as we consider that the native mechanisms are more than capable of handling these kinds of threats. Moreover, the

mechanisms of the mail servers are updated regularly and as a result the adoption to each update would have been an important burden.

Another secure e-mail solution currently deployed and in testing phase, is Scramble [5]. Although providing end-to-end encryption to e-mails, it actually represents a new independent e-mail server and not an add-on solution. As a result, users cannot use it with their existing accounts, but would rather have to create a new, dedicated secure e-mail account. We consider that as a major obstacle for broader adoption and one of the distinguished features of our system is that users can keep their current e-mail addresses.

Jammalamadaka et al. [3] have implemented a proxy based architecture in order to enable secure browsing when using a public computer. Delegate, as their system is called, prohibits some known attacks, like hijacking, and prohibits the user of entering his/her credentials on an untrusted computer. The drawback of this implementation is the necessity of extra hardware, a mobile phone in this case. The user needs to use his/her mobile phone loaded with a specific mobile application in order to authenticate to the proxy in order to then initiate the secure communication with the untrusted computer.

Besides developing systems and designing architectures there have been also some more simple solutions, mostly focusing on a single e-mail server. For example the Gmail S/MIME [6] is a plugin for Firefox that adds S/MIME when accessing the Gmail server from the browser. The main drawback of this and other similar solutions is that the plugin is depending on the interface of the e-mail provider. As a result, when the GUI changes and functionalities are modified, the plugin is not functioning anymore.

Another example for Windows based systems is OpaqueMail [12]. This solution provides a local proxy that interacts with the mail client and provides S/MIME. The main differences with our solutions is that this solution requires local installation and thus, besides the OS requirements, needs to be installed on all users' devices.

III. SECURE E-MAIL DESIGN

When designing our secure e-mail system, we considered several important principles in terms of user–friendliness, usability, scaling, privacy of users, and security of e-mails.

The most important requirement is that all security and privacy extensions should be completely transparent to the end user. This principle is very important when designing and implementing new features, as it is common practice for users not to adopt new technologies that significantly change the way they interact with a standard service or a product. This is actually aligned with Saltzer and Schroeder's Psychological acceptability principle [7]. As they stated back in the 1970s "security mechanisms should not make the resource more difficult to access than if the security mechanisms were not present". The result of not applying this principle is that users will simply not use the extra security features.

As a result of this important aspect, in our design and implementation we have applied the above principle in two ways. First, we deploy all security extensions completely transparently to the user. Encryption and signing of e-mails, management of certificates, secure communication between the users and our proxy server, and secure communication between the proxy server and the native e-mail servers are not visible to the user at all. They are all performed automatically by our secure e-mail proxy and only appropriate indications are shown to the users as a means of informing and assuring them about the correctness and verification of the process. This approach is different from concepts like PGP where the user needs to establish the secure communication in a trusted way with each of his contacts by himself.

Second, our system is very simple to use. Our aim was not to differentiate the logic of our secure e-mail system from the current popular web-based implementations of e-mail systems. Even the login function to our system is integrated with the user credentials of the native e-mail system. So, the user needs to provide only his/her e-mail address and password of the native e-mail account and all necessary security operations are performed transparently in the background, automatically by our system.

Another important design principle that we followed was to implement the solution in such a way that no extra hardware or software is needed on the user side in order to access and use the system. As a result, the user is able to access the secure e-mail system with a computer, a mobile device or in general any device that can support a standard browser.

IV. PROXY AS A SECURITY SOLUTION

In our system, security functions are shifted from the user to the proxy server. As users are the most critical component of the entire system when it comes to security, our approach was to separate security functionalities from users and delegate them to the dedicated proxy server. Nonetheless, the user is the one that is still responsible for security, i.e. the proxy is not in the position to read any e-mails or security credentials of the user, without the input of the user's credentials. As explained in the next section, unique user's e-mail credentials represent the identifiers to generate, protect, store and use cryptographic keys, which are used for secure processing of e-mails.

By using the concept of a web proxy extra security functionalities are provided to the user. Our design and proposal can be used for other research areas, such as Internet of Things or secure m-commerce systems. Although these two areas are conceptually different, we use the same concept for both.

In our design the proxy server is responsible for establishing a secure link with the user and delegating as many interactions as possible from the user's workstation or mobile device. Our secure e-mail proxy implementation represents our first–level solution towards the use of this approach and technology and will also be used later as a paradigm for more concrete and complicated system designs.

PKCS #7 and S/MIME packaging are general mechanisms to secure e-mail systems. They can be used generally for all e-mail communications and are recommended to guarantee confidentiality and authenticity

in sensitive use cases such as the secure exchange of m-commerce objects, as described in [8].

When using a mobile device, the use of the security proxies becomes even more important. Mobile devices have limited capabilities and so a user can delegate to the trusted party the secure operations he/she would like to execute and be reassured that the contents he/she receives meet the desired security and data protection requirements. Moreover, since the proxy server functions act as a delegate for the more "heavy" computations in terms of processor and network speed, potential bottleneck on the mobile side will be avoided, especially when being in lack of network coverage (limited Internet speed) or low battery (avoid using heavy processor-related tasks).

V. IMPLEMENTATION

Our secure e-mail proxy is a Java web application using Java Server Pages (JSP) based on the PrimeFaces framework and Java Servlets. The security mechanisms are implemented with Java, using external cryptographic and e-mail security libraries. In particular, we used the Java Mail library to handle e-mail operations and OneSEC™ Java Security Platform for security functionalities [9]. Certificates are issued by the OnePKI™ Public Key Infrastructure system [10], but in principle any other PKI could also be used. The JSP pages and Java classes (the main elements of the application), along with other components of the project, are deployed in an Apache Tomcat server.

The architecture comprises two servers: the proxy server, which handles all the communication with the appropriate parties (users, CAs, e-mail servers) and the associated CA server. Users login to the proxy server using their e-mail address and password of the native e-mail account. The proxy server identifies the native e-mail server from the e-mail address and performs the authentication of the user with the native e-mail server. The protocols used are different, depending on the native e-mail server. The list of protocols and mail servers currently supported are given in Table 1.

Once the authentication is successful, the proxy fetches the list of e-mails from the native e-mail server and displays them to the user, exactly the same way as they would be displayed when using a normal web–based client. The user from this menu can have access to all normal mail functionalities, exactly as if he would use a desktop mail client. A screenshot of the secure email's inbox can be seen in Fig. 2.

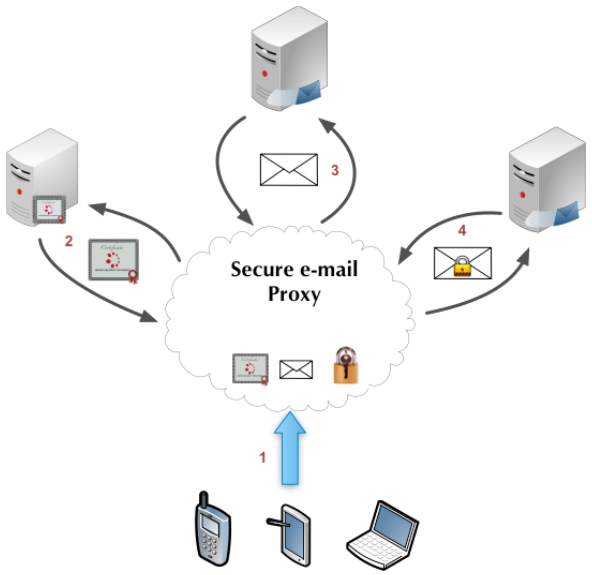


Figure 1. Secure e-mail proxy interactions

Step 1: The user logs into the secure e-mail proxy using the e-mail address and password of his/her native e-mail system; **Step 2:** For new users, the proxy generates a key pair, sends public key to the CA server, receives and stores the user's certificate; **Step 3:** The proxy fetches via IMAP the e-mails from the corresponding e-mail server. Upon double–click, each letter is cryptographically processed and displayed; **Step 4:** The user can now send signed and encrypted e-mails.

Finally, the connection between the proxy and the user is always secure. For this reason, the use of a standard SSL connection between the two parties is implemented. The architecture and the interactions between the parties in a normal workflow are shown in Fig. 1.

A. *Security Features*

When the user logs in for the first time, a RSA key-pair is automatically generated and a certificate is requested from the CA for his/her account. As the process is automated, the Common Name of the Distinguished Name (DN) is set to the user's e-mail address. They are both stored on the proxy and they are encrypted with the user's login credentials. As the standard secure e-mail approach

	Protocol	Incoming Server	Incoming Port	Outgoing Server	Outgoing Port
Gmail	imap	imap.gmail.com	993	smtp.gmail.com	465
Hotmail	pop3	pop3.live.com	995	smtp.live.com	587
Yahoo Mail	pop3	plus.pop.mail.yahoo.com	995	plus.smtp.mail.yahoo.com	465
KTH Mail	imap	webmail.kth.se	993	smtp.kth.se	465
JRC scientific Mail	imap	email.jrc.it	993	email.jrc.it	465

Table 1. Mail servers available with our secure e-mail system

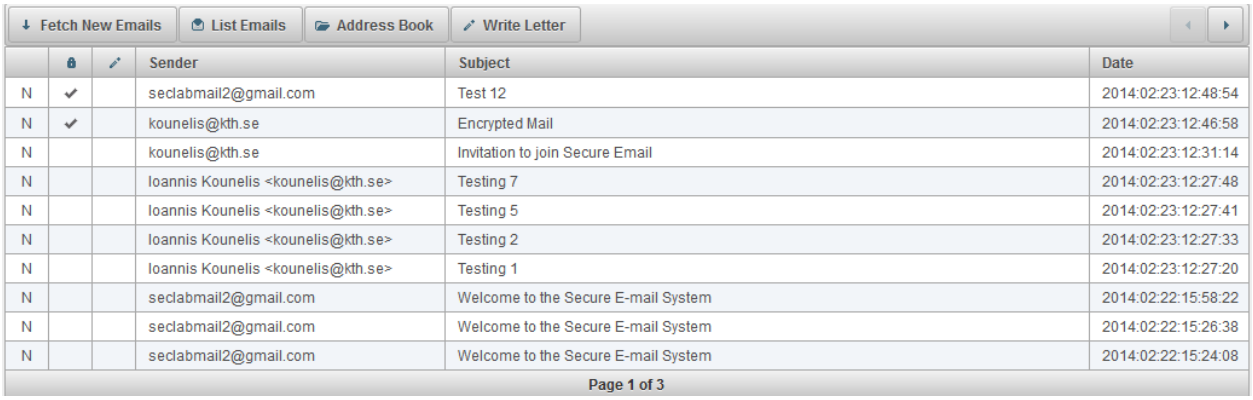


Figure 2. A screenshot of the secure email's inbox

requires, the private key is used to sign outgoing e-mails and certificates are used by recipients to verify the sender's signature.

When the user wants to send an encrypted or signed (or both) e-mail, the private key is decrypted on the fly (and kept in memory) and S/MIME formatting is per-formed with the original text of the e-mail. The user is not aware of any of these cryptographic processing details, as everything is performed transparently and by the proxy server. The user just selects, prior to sending an e-mail, the option to sign or/and encrypt the letter.

The certificate of the recipient is fetched automatically by the proxy from the local database of cryptographic keys. At that moment, the recipient(s) of the encrypted e-mail(s) must be registered in the same instance of our secure e-mail proxy, so that their certificate is available to the proxy. The e-mail address of the recipient is used as the unique identifier. As a result the recipient's certificate is found in the local database and the S/MIME packaging is performed according to the corresponding Internet standards.

When the recipient receives an encrypted e-mail, he/she will not be able to read it when fetching it and opening it directly from the native e-mail server. Instead when accessing the native e–mail server directly, he/she will just see an S/MIME attachment as a file with the standard .p7m extension and an empty body of the e-mail. However, when he/she uses the secure e-mail proxy after his/her login, the proxy will automatically understand that the e-mail is encrypted and will automatically perform decryption when the user clicks on the selected e-mail in order to read it. At the same time, he/she will see a notification that the letter is encrypted. The exact same procedure is performed with signed e-mails with the only difference being the attachment extension, which in this case is .p7s. A comparison of how an encrypted e-mail is viewed when using our secure email system and how it is viewed when accessing it directly from the native e-mail server (Gmail in this example) can be seen in Fig 3 and Fig. 4.

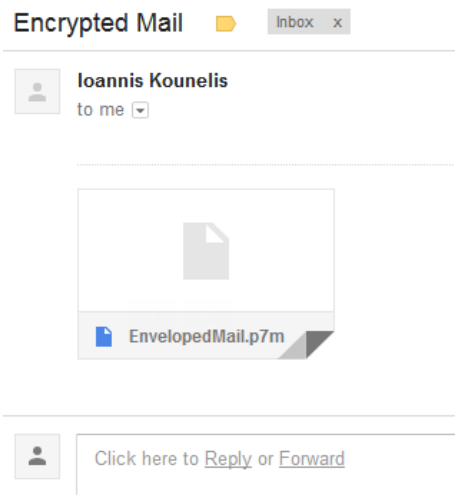


Figure 3. Encrypted e-mail when opened from the native mail server (Gmail)

Decryption is performed by recovering the encryption key packaged in the S/MIME format, using the private key of the recipient. The RSA private key of the user is already stored at the proxy encrypted with his/her password. When the user has already logged in, fetching and decryption of the private key is performed in order to be used for decrypting encryption keys contained in the incoming e-mails. A similar process is used to verify the signature of an e-mail, using the public key of the sender extracted from his/her certificate. The user should be made fully aware of the fact that the e-mail password is used as well for the encryption of the keys and as such he should be encouraged to apply a strong policy on his/her password management.

B. Drawbacks of Current Implementation and their Solutions

Besides security and privacy enhancements, which we introduced with our design and implementation, there are inevitably some minor drawbacks as well.

First, the proxy server represents a single point of failure. If it is attacked and not available to users, they will

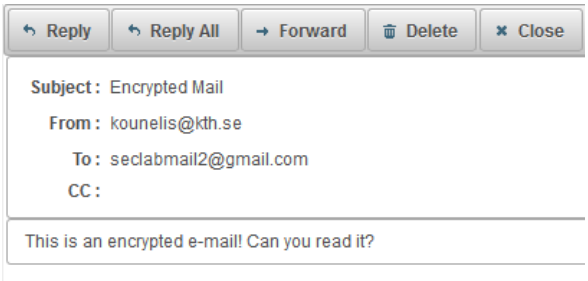


Figure 4. Encrypted e-mail as seen by using our SecureEmail

not be able to read their encrypted e-mails and verify signatures, as all the keys are stored there. As a result, even if they would be able to access their e-mails from the standard e-mail server, the non-availability of keys will not be possible to overcome. Nonetheless, a simple solution to this issue would be to have replica proxy servers that would serve as backups. All the encrypted key pairs and certificates would be saved there as well and a potential attack on the main proxy will not be able to compromise the users' keys. This solution is easily implementable in a scalable federated cloud environment.

Another implementation issue that is complicated to overcome is the native security policies of the different e-mail servers. Many of the most well-knows e-mail systems apply IP filtering specifically for each user and if they notice strange new login location such as the proxy location, they automatically block it. For example, Gmail by Google automatically blocks a user's login attempt over IMAP if it is occurring from a country that has never been previously used by the user. In this case the user will have to manually login to his/her native e-mail account and acknowledge that he/she allows the access from this third country. This security feature creates an extra burden for our implementation. As the proxy physically resides in a specific country (Sweden in our case), even if the user attempts to login from his/her country of residence to our proxy, the native e-mail server will detect the IP of the proxy and not of the user. As a result the above problem will occur.

The only way to deal with this issue is to have the user manually accept the remote access from the country where the proxy is located. An alternative would be to have proxies installed in every country, but even in this case the problem would remain when the user was travelling or when such security features would in the future not only be restricted to IPs, but also to the different devices that ask for remote e-mail access.

Moreover, a processing overhead, in terms of displaying e-mails and network traffic, is inevitable due to additional functions performed by the proxy. As all user traffic and data is deleted after the user logs off, so are all the e-mails that have been fetched from the server. As a result, each time the user logs into the secure e-mail proxy all his/her e-mails will need to be downloaded again. This issue is also present with native online mail servers and we do not consider it as an important overhead. Even more, as our intention in this phase was not to focus so much on the performance side, but rather on the security and privacy aspect of the secure e-mail proxy, we consider this aspect as an issue for future work in case our solution will be used on the market and not solely for research purposes.

Finally, it should be emphasized that encrypting an e-mail does not allow a third party to see the content of it. But, on the other hand, it also prohibits various advertisers, filters and spam tests to access the content of e-mails. As a result, malware and spam e-mail can also be encrypted and in that way bypass filtering performed by mechanisms applied on native mail servers. But this issue can be overcome by establishing security policies in which users may specify other users and mail servers from which they are willing to access e-mails.

VI. Conclusion & Future Research

In this paper we have described a secure and privacy–enhanced way to handle e-mails based on the concept of a secure proxy server. Our solution does not introduce any overhead to the user neither does it require any extra software or hardware. More importantly, all the operations are transparent to the end users, so they do not need to be concerned with any configuration or understanding of the underlying technology.

Future research will focus on improving even more user experience when using our secure e-mail proxy. We aim to create a new login layer to our platform that will allow the user with one account to handle all his/her e-mail accounts. The user will then be able to use the same certificates with all e-mail accounts that he/she chooses, instead of as today having a different certificate for each e-mail account.

Moreover, we plan to extend the system so that it provides federation of multiple proxies by distributing certificates between two or more instances of the Secure e-mail System. This means that for the next version we will introduce scalable federated architecture and utilization of the full PKI hierarchy. For example, if one proxy is deployed in Italy and the other in Sweden, a user that has already created a certificate for his/her account at the Italian proxy should not need to re-generate one if he/she uses the Swedish proxy. The two proxies should communicate and manage secure transfer of certificates in such a case.

The current system, being based on unreliable user e-mails, does not provide any assurance. One of the future extensions is to introduce versions with higher level of assurance in user identities and credentials. This will be achieved by linking the system to trusted and reliable identity and security providers and by using higher assurance authentication protocols, based on strong authentication, use of certificates, two–factor authentication and PIV smart cards.

Finally, future research will also focus on the use of secure e-mails and secure e-mail proxy in the area of m-commerce, as already described in the previous sections and on the hardening of the proxy itself, in terms of security. We aim to use the proxy technology and other security features for our future research and demo activities, by enhancing its functionality and capabilities.

References

[1] I. Kounelis, J. Loschner, D. Shaw, and S. Scheer, "Security of service requests for cloud based m-commerce," in 2012 Proceedings of the 35th International Convention MIPRO, 2012, pp. 1479 –1483.

[2] I. Brown and C. R. Snow, "A proxy approach to e-mail security," Software: Practice and Experience, vol. 29, no. 12, pp. 1049–1060, Oct. 1999.

[3] R. Jammalamadaka, T. Van Der Horst, S. Mehrotra, K. Seamons, and N. Venka-subramanian, "Delegate: A Proxy Based Architecture for Secure Website Access from an Untrusted Machine," 2006, pp. 57–66.

[4] A. Ghafoor, S. Muftic, and G. Schmölzer, "CryptoNET: Design and implementa-tion of the Secure Email System," in 2009 Proceedings of the 1st International Workshop on Security and Communication Networks (IWSCN), 2009, pp. 1–6.

[5] "Scramble" [Online]. Available: https://scramble.io/. [Accessed: 01-Oct-2013].

[6] “Gmail S/MIME.” [Online]. Available: https://addons.mozilla.org/en-US/firefox/addon/gmail-smime/. [Accessed: 01-Oct-2013].

[7] M. Bishop, Computer security : art and science. Boston: Addison-Wesley, 2003.

[8] I. Kounelis, G. Baldini, S. Muftic, and J. Loschner, “An Architecture for Secure m-Commerce Applications,” in 2013 19th International Conference on Control Systems and Computer Science (CSCS), 2013, pp. 519–525.

[9] SETECS, Inc., “OneSDK™ – Java Security Platform”

[10] SETECS, Inc., “OnePKI™ – Public Key Infrastructure”

[11] The Guardian, “The NSA Files”, [Online]. Available: http://www.theguardian.com/world/the-nsa-files [Accessed 30-Jan-2013]

[12] “Opaquemail” [Online]. Available: http://opaquemail.org/. [Accessed: 01-Apr-2014]

Requirements of practice in relation to the existing information technology and security management competencies

K. Arbanas* and D. Alagić**
* Paying Agency for Agriculture, Fisheries and Rural Development, Zagreb, Croatia
** NTH Media LTD, Varaždin, Croatia
krunoslav.arbanas@gmail.com, dalagic@foi.hr

Abstract - Until few years ago professional certifications were something that is good, but not necessary to have during employment process. That could be seen from the published job advertisements for variety of IT and information security professionals. In such job ads, there was first required specific experience in the subject area where no confirmation of that experience in form of professional certification was needed. Later, in addition of required experience in subject area, the possession of certain certificates was referred as advantage in job ads. Nowadays, it almost becomes the rule that the requirement for employment, in addition to having specific experience in the subject area, is also possession of a professional certificate which confirms stated experience. Professional certifications primarily demonstrate experience and knowledge in a particular area but also, due to the necessity of re-certification every few years, show commitment to professional development and constant training. The guiding principle of this study was to determine how many such certified professionals there is in Croatia and what kind of skills and competencies employers indirectly seek by requiring possession of certain certificates in job ads.

I. INTRODUCTION

With the rapid growth of information technology on a global scale, the necessary knowledge and skills required for IT professionals in different domains is constantly changing and expanding.

Today, there are significant changes in the responsibilities of IT and security professionals as expertise in one area is no longer guarantee of success in the IT profession. Moreover, employees are evaluated by their ability to perform multiple roles and it is increasingly trend that traditional business functions such as compliance, risk management and privacy compliance are assigned to such experts which means that they must master not only the technological requirements but also the business needs [1].

Assessment of the level of individuals' skills by the organizations that employ professionals in the field of information technology and security generally is based on their academic degree or professional certificate. The most important difference between academic education and professional certification is its purpose and the fact that professional certifications are focused on specific marketable skills and academic education on long-term employability and learning general theoretical model of problem solving.

II. REQUIREMENTS OF PRACTICE IN RELATION TO THE EXISTING COMPETENCIES

A. Literature review

Despite numerous frameworks and curriculum models for IT and security professionals competencies development, universities often do not provide their students with skills demanded by employers, which leads to the discrepancy between the level of skills expected by employers and the level of skills that students will have upon graduation, so many authors [2],[3],[1],[4] dealt with the issues of existing curricula and their appropriateness.

The main issue in the curriculum development is the fact that the curriculum cannot cover all potentially relevant material. Therefore, during its development one must consciously choose what will be included in the curriculum and how it will emphasize different sections, which should reflect the current needs of practice.

Curriculum models [1], [4] developed by ISACA international organizations adjust the academic community offering to the needs of professional practice and provide a framework for academic institutions and professional associations in the development of new and redesigning existing courses as they are based on the needs and expectations of the IT profession and rely on previous research of academics, practitioners, audit organizations and professional associations.

Thanks to the rapid changes and development of information systems and technology that drives them, as well as the necessary skills required, the curricula of information systems must be frequently updated to reflect the requirements of practice. For this reason, representatives of higher education should follow trends and act promptly to enhance curricula based on the industry needs, labour market statistics and defined necessary skills based on scientific research [2].

Research in which a comparison of business skills of academic and professional literature as well as those listed in the job ads was conducted [5], showed that the academic literature is lagging behind somewhat in terms of descriptions of specific technical skills, but provides the

richest understanding of the humanistic and business skills. Professional literature tends to provide views on the future of business skills, and contains the richest set of technical skills while the online job listings are fairly balanced in terms of technical, humanistic and business skills occurrences.

Many authors [5], [6], [7], [8], [9] point out that jobs within the information technology domain are becoming more strategic and business-focused and successful IT professionals today should not be satisfied with possessing just a technical or "*hard skills*" such as knowledge of hardware, software or network management, but also must have the "*soft skills*" i.e. personal, communication, organizational and interpersonal skills [7] which include good communication, leadership, interpersonal skills, motivation, negotiation skills, project and teams management, and sense for business and strategy [8]. In addition to technical and soft skills, some authors [6] points out that today's IT jobs also require "*critical skills*" that include cultural competencies and creativity.

In today's environment of constant and rapid change, a mix of different skills is essential for every IT professional and soft skills are becoming increasingly dominant. Having so far well served skills and competencies is necessary but not sufficient for successful progress [7] because all lead to a multidisciplinary approach where IT professionals are expected to specialize in two or more interconnected domains [6].

IT Managers should always have a high level of technology understanding and strengthening of core technical knowledge which remains a critical requirement for this position [10], and technical skills that are considered to be the base of the IT profession, some authors [11] split into three groups - foundational, operational and essential skills. The foundational skills group which represents the basis of this area consists of basic skills of programming, testing, support, databases, operating systems and telecommunication.

The operational skills group includes skills in operations, hosting, business continuity / disaster recovery and the mainframe / legacy area. Those skills, although specialized, commonly are not firm-specific. The third group of technical skills - essential skills group consists of skills necessary for development of IT solutions aligned with organizational goals. That includes systems analysis, systems design and IT architecture / standards [11].

On the other hand, non-technical skills can be seen as features that give professionals ability of understanding, developing and delivering effective solutions of organizational problems. Also divided into three groups those skills consist of project management skills (project leadership, project planning, project integration / program management, project risk management), problem skills (company-specific knowledge, functional area, process knowledge, industry knowledge, business process design, change management) and relationship skills (user relationship management, communications, negotiations, managing stakeholder expectations) [11].

The demand for workers with specialized skills puts considerable pressure on traditional educational strategies to ensure a skilled workforce. In response to the constant changes and advancements in the IT industry and slow response of traditional education, IT industry has established professional certifications as a preferred way of assessing skills and knowledge [12]. Thus, in the IT domain certifications become one of the criteria in hiring for many IT job roles and represent an indicator for Human Resource Managers that the condition of specific knowledge or competence has been met [13], [14]. Some authors [15] consider including IT certifications in the IT curricula as necessary while others [11] state that more important is to include basic knowledge of a defined group of technical and non-technical skills in curricula.

The reasons for IT professionals decide to gain the certification can be either self-development or career development (or both), whether it is about advancement opportunities in existing companies or to demonstrate their qualifications in terms of skills and knowledge [16] to potential new employers or clients.

As information security has become increasingly important for organizations, number of IT security certification has grown rapidly, making it difficult for stakeholders to decide which certificates to pursue or acquire [11]. There are identified four key stakeholders who have a special interest in security certificates and choices of them [14]. First of all, there are IT security professionals who are interested in getting the "best" security certificate for the purpose of employment, development expertise and income increase. Then, higher education institutions whose goal is to develop IT security education that prepares students for IT security professionals careers to whom can help a systematic approach to the assessment of IT security certificates. Third, there are employers who are looking for IT security professionals with the most relevant security certificates for their organization and finally, governments who can use a systematic approach to the evaluation of security certificates to determine the relevance of individual qualifications for accelerated immigration or regulation of the IT security profession [14].

Professional certifications, either vendor-neutral or vendor-specific, can confirm competencies and skills, but they cannot replace the experience or academic education. Education provides a broad general knowledge and skills by teaching students theoretical concepts, but there is a lack of skills for solving problems in information technology and security domain that can be reduced by acquiring professional certification from a specific domain [17]. Based on that, we can say that professional certifications, education and experience complement each other, not replace [15].

While vendor-specific certifications cover a highly specialized area, vendor-neutral certifications cover wider knowledge area, but at the same time focus on some job-specific components that are not there in formal education.

Research [18] results which included data from 299 IT professionals looking for professional certificates from various IT areas showed that the majority of professionals who are now working in IT is not certified in their field. Although there is considerable variation between different work areas of IT (from less than 14% of certified software

developers to nearly 70% of certified experts for network and security), a total of slightly less than 37% of IT professionals has certification.

On the other hand, the results of research [16] which included 240 participants (HR or IT manager) from the area of the South-eastern USA showed that hiring personnel perceive IT certification positively. Also, on question what certificate would they like their new IT employee to have over 85% of respondents chose at least one certificate from the list provided, while 14.6% put "None" and 52.1% of them reported that they would like job candidate has two or more certificates.

Professional certificates play an important role in today's IT industry mainly due to the fact they represent a standardized assessment method of knowledge and skills in a specialized field, they are globally recognized, and the certification schemes are directly mapped to defined set of required skills. Also, the benefits of the organizations that employ certified professionals are reflected in the increasing competitive advantages [12].

B. Professional certificates

There are several international organizations that provide certification of IT professionals, however authors' focus will be on two global organizations representing industry leaders in professional independent (vendor-neutral) certifications from a wide area of information technology - ISACA and $(ISC)^2$ [19], [20].

As an independent, non-profit, global organization, **ISACA** engages in the development, adoption and use of globally accepted, leading knowledge and practices for information systems. Formerly known as *Information Systems Audit and Control Association*, ISACA now uses only its acronym to emphasize a wide range of IT professionals it serves through the possibility of certification through one of the four certification programs [19]:

- *Certified Information Systems Auditor - CISA* (more than 103.000 certified professionals),
- *Certified Information Security Manager – CISM* (more than 22.500 certified professionals),
- *Certified in the Governance of Enterprise IT - CGEIT* (more than 5.600 certified professionals),
- *Certified in Risk and Information Systems Control - CRISC* (more than 16.800 certified professionals).

Obtaining the certification demonstrates that professional holds the key professional knowledge and experience in his/her profession, and each of ISACA's globally recognized certification refers to a specific area of expertise as is evident from Table I.

In order to earn ISACA's certification one must successfully pass the CISA/CISM/CGEIT/CRISC exam which consists of 200 (for CGEIT 120) multiple-choice questions in duration of 4 hours and meet the necessary requirements for specific work experience (Table I.).

ISC^2 - *International Information Systems Security Certification Consortium* with headquarters in the United States and offices in London, Hong Kong and Tokyo, is a global non-profit leader in educating and certifying professionals in information security during their careers, that currently offers the possibility of several certification programs [20] which of the most notably is the *Certified Information Systems Security Professional (CISSP)* certification. CISSP is designed for professionals in the field of information security management systems which currently holds more than 85.500 certified professionals.

In order to earn CISSP certification one must successfully pass the CISSP exam consisted of 250 multiple-choice questions in duration of 6 hours and meet the necessary requirements for work experience (Table I).

Comparative overview of observed certificates along with the minimum conditions necessary for a successful certification and profiles of professionals for which is particular certificate designed is shown in Table I.

In order to have CISA/CISM/CGEIT/CRISC/CISSP certificate in good standing it is necessary to have at least 20 hours of *Continual Professional Education – CPE* per year and 120 hours during 3-years period.

The objectives of continual education are [19], [20]:

- Maintain the competence of the individual by requesting the upgrade of existing knowledge and skills in the subject area,
- Provide the ability to distinguish between qualified certified professionals and those who did not meet the re-certification requirements,
- Provide a mechanism for monitoring the maintenance of professional competence in the subject area,
- Assist organizations in the development of valid functions of audit, control, risk management and information systems security by providing criteria for selecting quality employees.

There are numerous ways to acquire the required CPE credits from which some of them are: attending seminars as well as professional conferences, attending professional meetings and training courses either as a teacher or as a participant, publishing research papers, writing exam questions for ISACA/$(ISC)^2$ exams, reviewing books from the information technology domain, volunteering for ISACA/$(ISC)^2$ organization, etc.

III. RESEARCH

A. Methodology

After finding and analysis of relevant literature as well as previous studies on developed models of competencies management, knowledge and skills, types and characteristics of professional certificates, its comparison with formal education and other related topics in this area by using the analysis, synthesis, comparison and compilation methods and gaining some conclusions found in the literature, the authors began to collect the necessary data in order to answer the following research questions:

- Q1: How many registered certified professionals for information security (CISM, CISSP), information systems audit (CISA), IT risk (CRISC) and the management of enterprise IT (CGEIT) there is in Croatia?
- Q2: What is the certification trend in period of 10 years in Croatia in terms of certified professionals who possess the observed certificates?
- Q3: (a) To what extent organizations in Croatia are asking new employees for possession of the observed certifications when posting job ads, and (b) whether the possession of the certificate is mandatory or an advantage?
- Q4: To what extent employers in Croatia are asking for: (a) formal education, (b) technical and/or soft skills and competencies of new employees when posting job ads?

In order to find answers to research questions Q1 and Q2, professional organizations ISACA and (ISC)^2 were contacted and a request was sent by e-mail for the submission of quantitative data on the number of certified professionals over the period of 10 years in Croatia. For ISACA it was information about CISA, CISM, CGEIT and CRISC certified professionals and for ISC^2 information about CISSP certified professionals.

Simultaneously with waiting for a response from ISACA and ISC^2 to above inquiry, the authors began to collect data in order to be able to answer to research questions Q3 and Q4. This part of the research was conducted by identifying and analyzing all active job ads published on three main sites for online job searching in Croatia (job portal MojPosao - www.mojposao.net and job portal Posao.hr - www.posao.hr that also publishes jobs from the Croatian Labor Exchange - www.hzz.hr) during the period 01.07.2012 to 01.07.2013. This period was selected on purpose because that represents the last year before Croatian accession to the European Union. The first step was to identify all the jobs in IT domain advertised in reporting period, after which only those ads were filtered in which employers are seeking for security, computer, audit, information technology and other related experts in the domain of information technology, and which are related to the observed certifications. After acquiring data needed, descriptive statistics was used and results were presented in respective tables.

B. Results Analysis and Discussion

Based on the received data on the number of certified professionals in Croatia by observed certifications over the last 10 years tabular presentation was made as shown in Table II.

As seen in Table II., the answers to research questions Q1 and Q2 were partially obtained. That means that, regarding Q1, in 2012 Croatia had 90 certified professionals who possess CISA certification, 56 CISM certified professionals, 18 CGEIT certified professionals and same amount (18) of CRISC certified professionals.

TABLE I. CHARACTERISTICS AND REQUIREMENTS FOR ACQUIRING THE OBSERVED INTERNATIONAL CERTIFICATION

Certifi-cate	***Targeted audience***	***Waivers and substitutions***	***Minimum certificate requirements***
CISA[a]	IS/IT: -Auditors -Consultants -Audit Managers Security Professionals Non-IT Auditors	YES - up to 2 years based on formal education and/or work experience (working as IS or non-IT auditor or university professor)	Passed CISA exam + 3 yrs work experience in the subject area
CISM[a]	Information Security Managers Aspiring Information Security Managers IS/IT Consultants	YES - up to 2 years: • 2 years - valid CISA/ CISSP OR postgraduate degree in InfoSec or related field • 1 year – exp. in IS/ InfoSec management OR security certificates based on skills OR completed InfoSec program complied with ISACA model curricula	Passed CISM exam + 3 yrs work exp. in InfoSec mgmt in 3 or more practical business domain
CGEIT[a]	IS/IT: -Directors -Managers -Consultants -Executives IT Governance Professionals	NO - must have 5 or more years of experience managing, serving in an advisory or oversight role, and/or otherwise supporting the governance of the IT-related contribution to enterprise	Passed CGEIT exam + 5 yrs work exp. (with at least 1 yr related to frame-work for IT mgmt (CGEIT domain 1).
CRISC[b]	Professionals - in IT - in Risk - in Control - in Compliance Business analysts Project managers	NO - must have 3 or more years of cumulative work exp. performing the tasks of CRISC professional across at least 3 CRISC domains	Passed CRISC exam + 3 yrs work exp. in at least 3 CRISC domains
CISSP[a]	Security - Consultants - Managers - Auditors - Architects - Analysts - Engineers IT Managers CISOs Network Architects	YES - up to 1 year based on completing a four-year study or additional certificates (certificates of other organizations) from a list approved by the (ISC)²	Passed CISSP exam + 4 yrs direct experience working full-time in a minimum of two CISSP domains

a. 5 years of required work experience

b. 3 years of required work experience

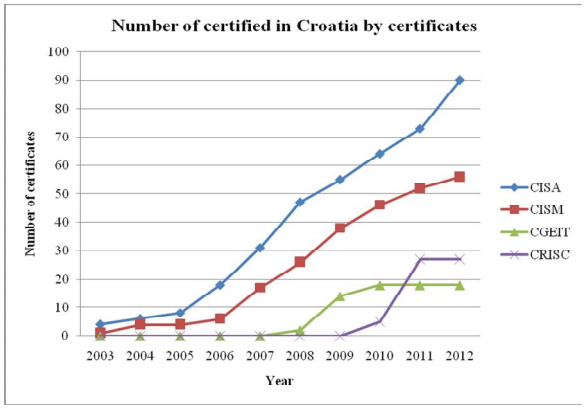


Figure 1. Number of certified professionals in Croatia by certificates

Official data for the CISSP certification was not available because ISC[2] did not provide authors with requested data with an explanation that these data are for internal use only.

Also, our data indicate a positive trend over the period of last 10 years (Q2), which one can best see from the graphical representation shown in Figure 1. of the upward trend in the number of observed certifications in Croatia in the last 10 years based on data in Table II. provided by ISACA organization.

CISA, as being the first and oldest ISACA's certificate which celebrated its 35th anniversary, is logically most represented observed certification in Croatia with the largest number of certified professionals. The next oldest certification is CISM followed by two "youngest" certifications CGEIT and CRISC. It is interesting to mention that, as one can see in Table II. and Figure 1., there has been a sudden increase of CGEIT certified professionals from 2008 to 2009 and from 2010 to 2011 for CRISC certified professionals after which the number of certified professionals becomes constant. The reason is that in the initial period for these certifications, CGEIT and CRISC certification could be obtained by the "grandfathering programs" and professionals with enough years of IT and business experience were able to apply for obtaining these certificates without taking an exam.

Results of the second part of this research related to the research questions Q3 and Q4 are as follows. In the observed period of one year on job portals MojPosao.net and Posao.hr authors found total of 30 active job ads published from which one ad was published on both portals. Of the remaining 29 ads, four ads were repeated once more in the reporting period, which brings us to a total of 25 different ads. Those ads can be grouped into seven groups as shown in Table III. The group "Other" consists of one company from catering industry and one German company for which no name nor industry or business activity was stated in published ad. Only 3 ads were published for work outside the Croatia.

If we look at the results from Table III., we can see that the observed certifications were stated in 36% of ads from which none of them as mandatory requirement, but as an advantage (Q3). The mentioned certifications were CISA, CISM, CISSP and in 1 ad CRISC. CGEIT was not mentioned in any ad probably because it is relatively new certification and it is not custom for Heads of IT in

TABLE II. NUMBER OF OBSERVED INTERNATIONAL CERTIFICATES IN CROATIA IN THE LAST 10 YEARS

Year	Number of certified in Croatia by certificates				
	CISA	*CISM*	*CGEIT*[a]	*CRISC*[b]	*CISSP*[c]
2003	4	1	-	-	N/A
2004	6	4	-	-	N/A
2005	8	4	-	-	N/A
2006	18	6	-	-	N/A
2007	31	17	-	-	N/A
2008	47	26	2	-	N/A
2009	55	38	14	-	N/A
2010	64	46	18	5	N/A
2011	73	52	18	27	N/A
2012	90	56	18	27	N/A

a. CGEIT certification program began in 2008

b. CRISC certification program began in 2010

c. (ISC)[2] organization did not provide data

Croatia. There were no bigger surprises in the results, except for the fact that the group "Audit and Advisory Services" did not ask for CISA (or any other) certification in its advertisement. However, it was ad for apprentices.

Of the total number of ads, in 20 ads (80%) in formal terms a university degree was required, in 3 ads (12%) bachelor's degree and in 2 ads (8%) there was not required any formal qualifications (Q4a). Required technical skills were listed in all 25 ads (100%), while the various soft skills were mentioned in 18 ads (72%). Top 5 requested soft skills were communication skills, responsibility, motivation, flexibility and organizational skills (Q4b).

IV. CONCLUSION

IT has become the driving force of business, and is an area in need of continuous education and training. The knowledge and skills that IT professionals gain today in a

TABLE III. PUBLISHED ADVERTISEMENTS GROUPED BY BUSINESS ACTIVITIES AND REQUIREMENTS OF CERTIFICATE POSSESSION[A]

Business activity	Number of published ads	Certification possession	
		Advantage	*Mandatory*
Production and sales	5	0	0
Consulting, development and implementation of IT solutions	12	6	0
Audit and Advisory Services	2	0	0
Banking Institutions	2	1	0
Insurance Companies	1	1	0
Government Agencies	1	1	0
Other	2	0	0
TOTAL	25	9	0

few years becomes obsolete and so professional certifications are becoming more important.

Possession of an internationally recognized industry-neutral certification certainly enhances professional recognition and credibility of IT professionals and often is accompanied with better salary than professional colleagues who do not have proper certification.

This research resulted with number of certified professionals in information security, IT audit, IT risk management and governance of enterprise IT in Croatian organizations showing the positive trend over the last 10 years. Also, there were identified and analyzed active job advertisements published during the period of the last year before the Croatian accession to the European Union. The results showed that the Croatian employers partially follow the world's trend of requiring professional certification in order to use certification to differentiate between job candidates with similar levels of education and experience. However, no organization stated in published advertisements possession of a professional certification as a mandatory condition of employment in such way like as for academic education.

The recognized limitation of this study is primarily the inability to obtain necessary data from the $(ISC)^2$ organization on the number of CISSP certified professionals in Croatia for the selected period which prevented comparison of CISSP with the other four observed certifications. Another limitation is the relatively short period of only one year to examine trends in required certifications in published job ads, and finally the lack of a sufficiently large sample of published job ads for a variety of reasons, of which the dominant are recession in Croatia and the lack of new jobs.

The possible new research could take into account the greater time span of published advertisements in order to be able to draw some useful conclusions on trends before and after the Croatian accession to the European Union. Of course, more interesting and more detailed study would be to compare current trends of professional certificates recognition with other EU countries.

REFERENCES

[1] ISACA, "ISACA Model Curriculum for Information Security Management, 2nd Edition." Rolling Meadows, Illinois, pp. 1–33, 2012.

[2] B. Khoo and P. Harris, "Enhancing The Graduate Information Systems Curriculum: A Career Skills Oriented Approach," *Am. J. Bus. Educ.*, vol. 2, no. 2, pp. 65–80, 2009.

[3] F. Girard, B. Meunier, D. Hua, and E. Dubois, "A Rigorous Approach to the Definition of an International Vocational Master's Degree in Information Security Management," *Int. J. Secur. Softw. Eng.*, vol. 1, no. 4, pp. 1–17, 2010.

[4] ISACA, "ISACA Model Curriculum for IS Audit and Control, 3rd Edition." Rolling Meadows, Illinois, pp. 1–30, 2012.

[5] H. Huang, L. Kvasny, K. D. Joshi, E. M. Trauth, and J. Mahar, "Synthesizing IT job skills identified in academic studies, practitioner publications and job ads," in *Proceedings of the special interest group on management information system's 47th annual conference on Computer personnel research - SIGMIS-CPR '09*, 2009, pp. 121–128.

[6] S. Ang, D. Joseph, and S. A. Slaughter, "IT Professionals and the IT Profession," *Encycl. Manag. Inf. Syst.*, pp. 1–13, 2013.

[7] D. Allison, "The future CIO: Critical skills and competencies," *ECAR Research Bulletin 15*, Boulder, Colorado, pp. 1–14, 2010.

[8] M. E. Swinarski, D. H. Parente, and K. Noce, "Are all IT professionals created equally?," *J. Behav. Stud. Educ.*, vol. 2, pp. 1–11, 2010.

[9] D. Joseph, S. Ang, R. H. L. Chiang, and S. A. Slaughter, "Practical intelligence in IT: assessing soft skills of IT professionals," *Commun. ACM*, vol. 53, no. 2, pp. 149–154, 2010.

[10] M. T. Dlamini, J. H. P. Eloff, and M. M. Eloff, "Information security: The moving target," *Comput. Secur.*, vol. 28, no. 3–4, pp. 189–198, May 2009.

[11] K. Gallagher and T. Goles, "A Typology of Requisite Skills for Information Technology Professionals," in *Proceedings of the 44th Hawaii International Conference on System Sciences*, 2011, pp. 1–10.

[12] V. Fedák, P. Chlebana, I. Sivý, F. Jakab, J. Varnham, and P. Belko, "IT Industrial Certifications in Practice," in *Emerging eLearning Technologies and Applications (ICETA), 2011 9th International Conference*, 2011, pp. 51–56.

[13] J. Coelho, "Competence certification as a driver for professional development: A IT-related case-study," in *Education Engineering (EDUCON), 2010 IEEE*, 2010, pp. 81–88.

[14] N. J. A. Tate, S. Lichtenstein, and M. J. Warren, "IT security certifications: Stakeholder evaluation and selection," in *19th Australasian Conference on Information Systems*, 2008, pp. 991–1001.

[15] M. A. Rob, "The Value of IT Certification: Perspectives from students and IT," *Issues Inf. Syst.*, vol. 14, no. 1, pp. 153–161, 2013.

[16] D. Hunsinger and M. Smith, "IT certification used by hiring personnel," *J. Comput. Inf. Syst.*, vol. 50, no. 2, pp. 71–82, 2009.

[17] T. Lu, L. Chang, and Z. Lee, "Integrating Security Certification with IT Education," in *Proceedings of 2011 International Conference on System Science and Engineering*, 2011, no. June, pp. 582–587.

[18] B. A. Reinicke and T. Janicki, "Who Needs Certifications? A Survey of Certifications in the IT Industry," in *Proceedings of the Information Systems Educators Conference*, 2013, pp. 1–8.

[19] "ISACA." [Online]. Available: www.isaca.org.

[20] "ISC2." [Online]. Available: www.isc2.org.

Development of Users' Information Security Awareness Questionnaire (UISAQ) - Ongoing Work

T. Velki*, K. Solic** and H. Ocevcic***
* J.J. Strossmayer University/ Faculty of Teacher Education, Osijek, Croatia
** J.J. Strossmayer University/ Faculty of Medicine, Osijek, Croatia
*** J.J. Strossmayer University/ Faculty of Electrical Engineering“, Osijek, Croatia
tena.velki@gmail.com, kresimir@mefos.hr, ocevcic@gmail.com

Abstract - The user is still weakest link regarding information security matters, but studies on this subject are rare. The aim of this work is to develop general Users' Information Security Awareness Questionnaire (UISAQ). Development consists of selecting suitable items for which is assumed that measure the level of security awareness and testing impact of each item in measurement. Questionnaire consisted of 4 parts with total of 37 items. Results showed that first part of questionnaire, that examine the common user's risk behavior, should consist of 17 items (3 items had low factor loadings) separate in 3 subscales. Second part of questionnaire, which consisted of 6 items that measured the level of user's information security, had high internal consistency (k=6, α=0.89) and a satisfactory factor loadings. Third part of questionnaire, which consisted of 5 items that measured the level of user's beliefs about information security, should consist of 3 items (2 items significantly disrupted internal consistency) with high factor loadings and good internal consistency (α=0.76). Descriptive statistics showed that all the questions (n=6) in the fourth part of the questionnaire, which had examined the password quality and security, had a full range of answers and that normal distribution wasn't significantly violated. Although developed questionnaire requires more work and validation, first results showed that UISAQ has potential to become a good and reliable measure of users' security awareness in the future.

I. INTRODUCTION

Through the years, user of information system is still its weakest link regarding information security matters [1, 2]. Information system's user can, with his potentially risky behavior, significantly influence on overall system's security [3 - 5] and all hacker attacks usually combine social engineering with technical hacking skills [6]. However scientific studies on this subject are rare and there is need for universal measurement instruments [7, 8]. These instruments should enable measurement of user's influence on overall systems' security for state analysis and future studies.

Some previous solutions are proposed, but are partial and not universal enough [9, 10]. Actually most of the previous research on user's behavior is focused only on examining password usage and password quality and strength [11 -14].

The aim of this work was to develop reliable universal instrument which will measure level of information system's users' awareness on security matters, as general as possible, the Users' Information Security Awareness Questionnaire (UISAQ).

Development of this kind of questionnaire comprises selection of suitable items and testing impact of each item. Impact of each item which is assumed that measures the level of security awareness among users, is measured by using descriptive statistics, factor analysis and reliability analysis. Results of those analyses will exclude items with low impact and point out items with higher impact that present well defined questions [15].

With internationally validated questionnaire it should be possible to gain general conclusions about user's security awareness and potentially risky behavior. Results of those kinds of studies will enable concrete improvements of existing [16 - 18] and development of new information security solutions focused on user's education.

II. METHOD

A. Participants

Participants in this study were students (N=135) on second year of undergraduate study, from three different faculties of J.J. Strossmayer University of Osijek: Faculty of Teacher Education (N=41), Faculty of Medicine (N=51) and Faculty of Electrical Engineering (N=43). Proportion of mail students was 47.6% while proportion of female students was 52.4%. The average age of participants was 19.85 +/- 0.58 (arithmetic mean +/- SD).

B. Procedure

During regular classes students were asked to voluntarily give some general information about self (age and gender) and to fill out the UISAQ. Filling out the questionnaire lasted for approximately 30 minutes. Survey

was done on all three groups of participants during one week period.

C. Instruments

For the purpose of this research authors created UISAQ consisting of four parts with total of 37 items collected from different security guidelines and results of previous studies [19 - 24]. Each item is a question in the UISAQ presenting variable for the statistical analysis.

The four parts of UISAQ are as follows:

- First part of UISAQ consisted of 20 items measuring computer users' potentially risky behavior.
- Second part of questionnaire consisted of 6 items that measured the level of user's information security awareness.
- Third part of questionnaire consisted of 5 items which measured the level of user's beliefs about information security.
- The last part of UISAQ consisted of 6 questions that examined the quality and security of passwords.

TABLE I. STRUCTURE MATRIX FOR THE FIRST PART OF UISAQ EXTRACTION (METHOD: PRINCIPAL COMPONENT ANALYSIS; ROTATION METHOD: OBLIMIN WITH KAISER NORMALIZATION)

Items	Factors		
	1	*2*	*3*
sc1	.404		.521
sc2	.518	-.337	.519
sc3			.687
sc4			.799
sc5			.653
sc6		.666	
sc7		.741	
sc8		.810	
sc9	.598		
sc10			
sc11	.427		
sc12	.677		
sc13	.624		
sc14	.410		
sc15		.343	
sc16	.646		
sc17		.405	
sc18		.561	
sc19			
sc20			

TABLE II. RELIABILITY ANALYSIS: ITEM - TOTAL STATISTICS FOR THE FIRST PART OF UISAQ EXTRACTION

Items	Analysis results			
	Scale Mean if Item Deleted	*Scale Variance if Item Deleted*	*Corrected Item - Total Correlation*	*Cronbach's Alpha if Item Deleted*
1. factor (subscale)				
sc9	8.1756	6.084	.405	.539
sc11	8.7643	7.044	.274	.592
sc12	9.0153	6.569	.444	.532
sc13	8.8397	6.659	.423	.540
sc14	9.1832	7.997	.265	.602
sc16	7.6260	5.051	.365	.586
2. factor (subscale)				
sc6	15.4809	21.944	.471	.599
sc7	15.7176	22.866	.473	.600
sc8	15.5725	21.585	.594	.556
sc15	15.5038	25.929	.223	.688
sc17	16.6718	24.930	.293	.664
sc18	14.6031	25.703	.373	.637
3. factor (subscale)				
sc1	5.3806	3.410	.432	.609
sc2	5.2239	3.002	.515	.565
sc3	5.7015	4.136	.502	.580
sc4	5.8507	4.248	.417	.611
sc5	5.9925	5.060	.321	.658

The participants had to evaluate to which extend each statement refers to him/her on a scale from 1 to 5. This implies that each variable, for each examinee, can have only one value in range from 1 to 5.

Unanswered questions were not included in the statistical analysis.

III. RESULTS

Development of questionnaire consisted of selecting suitable items for which there was assumption that measure the level of potentially risky behavior of computer users, level of information security awareness, users' believes about safety or password quality issues. Using descriptive statistics, factor analysis and reliability analysis we tested if selected items were good measure of hypothesized construct.

For the first part of UISAQ Exploratory factor analysis (method principal components, oblimin rotation) was used. Analyses have shown extraction of 8 factors (using the Guttman-Kaiser criterion) with eigen values larger than 1 and explanation of 66.11 % of overall variance. Given factor structure shown only 3 dominant factors (each explained more than 10% of variance and 37.42% of overall variance) and other factors had smaller eigen

TABLE III. MEASURES OF SENSITIVITY FOR THE FIRST PART OF UISAQ EXTRACTION

Items	Analysis results					
	Min	*Max*	*Range*	*Mean*	*Std. Deviation*	*Test of normality*
sc1	1.00	5.00	4.00	1.674	.937	.330[a]
sc2	1.00	5.00	4.00	1.822	.984	.264[a]
sc3	1.00	4.00	3.00	1.348	.615	.434[a]
sc4	1.00	5.00	4.00	1.187	.627	.497[a]
sc5	1.00	5.00	4.00	1.045	.365	.526[a]
sc6	1.00	5.00	4.00	3.201	1.685	.227[a]
sc7	1.00	5.00	4.00	3.007	1.549	.225[a]
sc8	1.00	5.00	4.00	3.142	1.513	.199[a]
sc9	1.00	5.00	4.00	2.134	.916	.284[a]
sc11	1.00	5.00	4.00	1.552	.771	.332[a]
sc12	1.00	5.00	4.00	1.303	.730	.454[a]
sc13	1.00	4.00	3.00	1.478	.723	.373[a]
sc14	1.00	3.00	2.00	1.149	.434	.517[a]
sc15	1.00	5.00	4.00	3.195	1.612	.218[a]
sc16	1.00	5.00	4.00	2.699	1.273	.228[a]
sc17	1.00	5.00	4.00	2.038	1.600	.422[a]
sc18	1.00	5.00	4.00	4.120	1.273	.303[a]

a. $p < 0.01$

values and very small proportion (less than 7%) of explained variance. First factor explained 14.30 % of overall variance, second factor explained 12.92% of overall variance and third factor explained 10.19 % of overall variance. Than, Confirmatory factor analysis was used in order to test 3 hypnotized factors which were extracted from previous Exploratory factor analysis. The saturation (factor loadings) was defined as larger than 0.3 which interpreted the three mentioned factors.

The factor structure of the first part of UISAQ is shown in Table 1. As shown in table, 3 items had factor loading lower than 0.3 on all of three factors and they were suppressed and thereby not shown in table (items sc10, sc19 and sc20) so they were excluded from further

TABLE IV. COMPONENT MATRIX FOR THE SECOND PART OF UISAQ EXTRACTION (METHOD: PRINCIPAL COMPONENT ANALYSIS)

Items	Factor
sc1	.694
sc2	.731
sc3	.853
sc4	.880
sc5	.868
sc6	.768

TABLE V. RELIABILITY ANALYSIS: ITEM - TOTAL STATISTICS FOR THE SECOND PART OF UISAQ EXTRACTION

Items	Analysis results			
	Scale Mean if Item Deleted	*Scale Variance if Item Deleted*	*Corrected Item - Total Correlation*	*Cronbach's Alpha if Item Deleted*
sc1	13.7615	20.540	.582	.888
sc2	14.6385	20.000	.620	.882
sc3	13.7308	17.795	.771	.858
sc4	13.9538	17.192	.806	.852
sc5	14.1692	18.064	.793	.855
sc6	13.9769	19.030	.662	.876

analysis. Final version of first part of UISAQ should consist of 17 items separate in 3 subscales: first subscale measures risky behavior of computer users (k=6), second subscale measures maintenance of computer systems (k=6), and third ones measures using other users' data (k=5). Than reliability analysis was done for three new scales (Table 2). First subscale had little bit lower internal consistency (k=6; Cronbach α=0.61), but all items contributed significantly to good internal consistency which implies that this form of subscale should be kept as finale one. Second subscale had a satisfactory internal consistency (k=6; Cronbach α=0.67), as well as third one (k=5; Cronbach α=0.66) which implies that both of these two forms of subscales should be kept as finale ones.

Results of sensitivity test of new formed questionnaire are shown in Table 3. Only 3 items did not have full range (sc3, sc13, sc14) which implies about good sensitivity of new formed scales. Distribution of results was not normal (Kolmogorov-Smirnov Statistic was significant for all items), which was expected. For first and third subscales means were at lower part of subscale (positive asymmetry) meaning less risky behavior of computer users and for second subscale means were at higher part of subscale (negative asymmetry) meaning more risky behavior of computer e.g. low level of users' maintenance of personal computer systems.

The Exploratory factor analysis (method principal

TABLE VI. MEASURES OF SENSITIVITY FOR THE SECOND PART OF UISAQ EXTRACTION

Items	Analysis results					
	Min	*Max*	*Range*	*Mean*	*Std. Deviation*	*Test of normality*
sc1	1.00	5.00	4.00	3.083	.946	.233[a]
sc2	1.00	5.00	4.00	2.203	.975	.283[a]
sc3	1.00	5.00	4.00	3.121	1.126	.252[a]
sc4	1.00	5.00	4.00	2.893	1.172	.207[a]
sc5	1.00	5.00	4.00	2.667	1.068	.213[a]
sc6	1.00	5.00	4.00	2.872	1.076	.189[a]

a. $p < 0.01$

TABLE VII. COMPONENT MATRIX FOR THE THIRD PART OF UISAQ EXTRACTION (METHOD: PRINCIPAL COMPONENT ANALYSIS)

Items	Factor
u7	.391
u8	.796
u9	.856
u10	.758
u11	.428

components) was also used for the second part of UISAQ, which have shown extraction of 1 factors and explanation of 64.37 % of overall variance. In table 4 is shown factor structure of the second part of UISAQ which consisted of 6 items that measured the level of user's information security awareness, and which had a satisfactory factor loadings for all items on one factor.

Reliability analysis (Table 5) had shown high internal consistency (k=6, Cronbach α=0.89) and a satisfactory factor loadings which implies that it should be kept in their original form. All items had full range of response which implies good sensitivity (Table 6) of a new formed scale. Distribution of results was not normal (Kolmogorov-Smirnov Statistic was significant for all items), which was expected. Means were at higher part of subscales (negative asymmetry) meaning low level of user's information security awareness.

For the third part of UISAQ, Exploratory factor analysis (method principal components) was also appropriate. Analysis had shown extraction of 1 factors and explanation of 45.55 % of overall variance. In table 7 is shown factor structure of the second part of UISAQ which consisted of 5 items that measured the level of user's beliefs about information security and which had a satisfactory factor loadings for all items on one factor.

Reliability analysis (Table 8) had shown lower internal consistency (k=5; Cronbach α=0.60) with two items significantly violating internal consistency (items u7 and u11) so those items were excluded from further analysis and form with 3 items which had high factor loadings and good internal consistency (k=3; Cronbach α=0.76) was kept. All items had full range of response (Table 9) which was a measure of good sensitivity of a new formed scale.

TABLE VIII. RELIABILITY ANALYSIS: ITEM - TOTAL STATISTICS FOR THE THIRD PART OF UISAQ EXTRACTION

Items	Analysis results			
	Scale Mean if Item Deleted	*Scale Variance if Item Deleted*	*Corrected Item- Total Correlation*	*Cronbach's Alpha if Item Deleted*
u7	8.5564	10.158	.205	.698
u8	9.0977	11.422	.510	.496
u9	9.1880	10.654	.611	.445
u10	8.9699	10.772	.484	.492
u11	9.4511	12.825	.200	.626

TABLE IX. MEASURES OF SENSITIVITY FOR THE THIRD PART OF UISAQ EXTRACTION

Items	Analysis results					
	Min	*Max*	*Range*	*Mean*	*Std. Deviation*	*Test of normality*
u8	1.00	5.00	4.00	2.218	1.003	.188[a]
u9	1.00	5.00	4.00	2.128	1.040	.233[a]
u10	1.00	5.00	4.00	2.346	1.175	.197[a]

a. $p < 0.01$

Distribution of results was not normal (Kolmogorov-Smirnov Statistic was significant for all items), which was expected. Means were at lower part of subscale (positive asymmetry) meaning high level of user's beliefs about information insecurity.

On the last, fourth part of UISAQ was applied Descriptive statistics as this part consisted of different types of questions with different possible answers. Results (Table 10) showed that all the questions (k=6) in this part of the questionnaire, which had examined issues regarding quality and safety of passwords, had a full range of answers for all items which implies good sensitivity. Although, test of normality of distribution (Kolmogorov-Smirnov Statistic) was significant, normal distribution wasn't significantly violated (asymmetry coefficients Skewness and Kurtosis were not greater than - / + 2).

IV. CONCLUSION

Although this questionnaire requires more work, first results look promising. Results show that UISAQ has potential to become a good and reliable instrument for measurement of users' information security awareness. After validation it may become first international measurement tool of its kind as basis for ongoing professional and scientific research.

With the UISAQ IT professionals will be able to analyze information systems' users in order to identify issues with low security level, while scientists will be able to generally categorize information systems' users regarding level of their information security awareness. By analyzing enough samples of all kinds of information system's users it should be possible to gain some general

TABLE X. DESCRIPTIVE STATISTICS FOR THE FOURTH PART OF UISAQ EXTRACTION

Items	Analysis results					
	Min	*Max*	*Range*	*Mean*	*Std. Deviation*	*Test of normality*
p1	1.00	5.00	4.00	3.531	1.576	.286[a]
p2	1.00	5.00	4.00	3.712	.993	.250[a]
p3	1.00	5.00	4.00	2.909	1.723	.274[a]
p4	1.00	5.00	4.00	2.015	1.680	.452[a]
p5	1.00	5.00	4.00	3.692	1.201	.187[a]
p6	1.00	5.00	4.00	2.977	1.479	.216[a]

a. $p < 0.01$

conclusions about user's potentially risky behavior, correlation with level of security awareness and identification of most insecure kinds of users.

As future work, authors will repeat collecting data analyzing them and that way improving UISAQ as many times as needed in order to develop as-good-as-possible questionnaire. The end of development process should be international validation of this questionnaire.

REFERENCES

[1] M. A. Sasse, S. Brostoffand, and D. Weirich, "Transforming the 'weakest link' - a human/ computer interaction approach to usable and effective security", BT Technology Journal, vol. 19, pp. 122-131, July 2001.

[2] S.J. Lukasik, "Protecting Users of the Cyber Commons", Communications of the ACM, vol. 54, pp. 54-61, September 2011.

[3] H. Thompson, "The Human Element of Information Security", IEEE Security&Privacy, vol. 11, pp. 32-35, January-February 2013.

[4] K. Solic and V. Ilakovac, "Security perception of a portable PC user (The difference between medical doctors and engineers): a pilot study", Medicinski glasnik Dobojsko-Tuzlanskog kantona, vol. 6, pp. 261-264, August 2009.

[5] M.E. Johnson and S.L. Pfleger, "The Human Side of Risk Management", IEEE Security&Privacy, vol. 9, pp. 51, January-February 2011.

[6] K. D. Mitnick, The Art of Deception - Controlling the Human Element of Security, John Wilwy & Sons, 2002.

[7] R. E. Crossler, A. C. Johnston, P. B. Lowry, Qing Hu, M. Warkentin and R. Baskerville, "Future directions for behavioral information security research" Computers&Security, vol. 32, pp. 90-101, June 2013.

[8] Kim-Kwang and R. Choo, "The cyberthreat landscape: Challenges and future research directions", Compures&Security, vol. 30, pp. 719-731, June 2011.

[9] K. Solic, B. Tovjanin and V. Ilakovac, "Assessment Methodology for the Categorization of ICT System Users Security Awareness", Proc. IEEE 35th MIPRO, pp. 1560-1564, May 2012.

[10] K. Solic, F. Jovic and D. Blazevic, "An Approach To The Assessment Of Potentially Risky Behavior Of ICT System's Users", Technical Gazette, vol. 20, pp. 335-342, February 2013.

[11] M. Dell'Amico, P. Michiardi and Y. Roudier, "Password Strength: An Empirical Analysis", Proceedings IEEE INFOCOM, (San Diego, CA) pp. 1-9, March 2010.

[12] A.G. Voyiatzis, C.A. Fidas, D.N. Serpanos and N.M. Avouris, "An Empirical Study on the Web Password Strength in Greece", 15th Panhellenic Conference on Informatics, (Kastonija Greece), pp. 212-216, September-October 2011.

[13] Ma Wanli, J. Campbell, D. Tran and D. Kleeman, "Password Entropy and Password Quality", 4th International Conference on Network and System Security, (Melbourne, VIC), pp. 583-587, 1-3, September 2010.

[14] P.G. Kelley, S. Komanduri, M.L. Mazurek, R. Shay, T. Vidas, L. Bauer, N. Christin, L.F. Cranor and J. Lopez, "Guess Again (and Again and Again): Measuring Password Strength by Simulating Password-Cracking Algorithms", IEEE Symposium on Security and Privacy, (San Francisco, CA), pp. 523 - 537, May 2012.

[15] C. Jackson, Psihologijsko testiranje. Jastrebarsko: Naklada Slap, 2003.

[16] C. Paulsen, E. McDuffie, W. Newhouse and P. Toth, "NICE: Creating a Cybersecurity Workforce and Aware Public", IEEE Security&Privacy, vol. 10, pp. 76-79, March-April 2012.

[17] I. Kirlappos and M.A. Sasse, "Security Education against Phishing: A Modest Proposal for a Major Rethink", IEEE Security&Privacy, vol. 10, pp. 24-32, March-April 2012.

[18] S. Furman, M.F. Theofanos, Yee-Yin Choong and B. Stanton, "Basing Cybersecurity Training on User Perceptions", IEEE Security&Privacy, vol. 10, pp. 40-50, March-April 2012.

[19] K. Solic and I. Horvat, "Weaknesses of global free-of-charge email services – analysis and recommendations", MEDIX, vol. 17, pp. 212-214, December 2011, URL: http://www.kardio.hr/images/stories/medix/m97/212-214.pdf

[20] International ISO/IEC standard 27001, Second edition 2013.10.01, URL: http://www.27000.org/iso-27001.htm

[21] System and Information Integrity family publication, NIST, URL: http://www.nist.gov/publication-portal.cfm

[22] IT Security Guidelines. Federal Office for Information Security, Bonn, Germany, URL: http://www.bsi.bund.de/DE/Themen/ITGrundschutz/ITGrundschutzInternational/intl.html

[23] ENISA publications, URL: http://www.enisa.europa.eu/publications

[24] CNiL Security of Personal Data. France, URL: http://www.cnil.fr/fileadmin/documents/en/Guide_Security_of_Personal_Data-2010.pdf

Question based user authentication in commercial environments

K. Skračić, P. Pale, B. Jeren
Faculty of Electrical Engineering and Computing, University of Zagreb
Unska 3, 10000 Zagreb, Hrvatska
E-mail: kristian.skracic@fer.hr

Abstract - The aim of this paper is to analyze question based user authentication methods with regard to their potential use in commercial environments and the limitations they impose. We provide an overview of current work related to question based authentication and describe the difference between static and dynamic questions in user authentication. Additionally, we give an introduction into some common practices used in user authentication in commercial environments. Based on existing research we suggest that only dynamic questions should be used in commercial environments and propose a set of requirements a data source used for creating authentication challenges should meet. Also, we categorize the types of data that can be leveraged for authenticating users.

I. INTRODUCTION

Textual passwords present the dominant knowledge based authentication method for authenticating users. It relies on a shared secret which is known only to the user and the authenticator. Also, the method relies on users being able to recall their password during authentication. We propose that an authentication method should inherently be able to remind the user what the answer is, while at the same time it should not leek any information to third parties. Alternative authentication methods have been researched quite extensively in the past years in the hopes of replacing textual passwords [1].

Question based authentication presents a way of authenticating users by posing questions to which only a legitimate user would know the answer. However, authenticating users by allowing them to answer questions has some practical limitations. The user would have to provide answers that are private or confidential in nature and others, close to the user, may know the correct answer.

Also, the strength of the authentication process relies on the information used to construct authentication questions which could have implication on the security and integrity of the authentication process.

In this paper we give an overview of some of the existing question based authentication methods and categorize the existing types of questions available for authentication. Also, we conduct a feasibility study of their use in commercial environments based on related work and their practical limitations and weaknesses.

Question based authentication is a form of authentication that relies on a user answering a specific question which only he knows the answer to. The main premise of question based authentication is that a legitimate user will easily give the right answer to a question, while an imposter would not. However, this is rarely the case since most authentication questions are based on information that is publicly available or well known and easily guessable. In this paper we analyze the various aspects of question based authentication and propose a set of requirements that could help such authentication methods to become viable solution for commercial systems. We choose to focus our study on commercial systems since they have high security requirements. By commercial systems we mean any information system that either handles sensitive or private information or financial transactions.

II. RELATED WORK

A. Question based authentication

Question based authentication represent a way of authenticating users by posing questions to which only a legitimate user would know the answer to. It can be divided in two main groups, static and dynamic questions [2]. The main difference with static and dynamic questions is that with dynamic questions, the user is not aware of what the question will be until the login attempt is made. With static questions, the question and answer are usually chosen during registration and do not change over time unless the user wishes to change them. As described in [3], there are three types of static challenge questions: fixed, open and controlled questions. Examples of static questions are a user's mother's maiden name or the name of the street they grew up on. In contrast, a dynamic questions could be the number of calls the user made today.

In the following sections we give an overview of related work that describes the use of static and dynamic questions for user authentication.

1) Static Questions

The research in [5] proposes an additional authentication mechanism to the classical username and password. It proposes the use of a profile based authentication framework (PBAF), which is a multi-modal authentication approach, for securing online examination.

During the learning process, users are given various profile questions that are used to extend and refine the user's individual profile. In essence, the user is asked to

answer several questions such as a friends surname or a favorite academic subject. Before being allowed to attempt an online examination, the student has to provide the same answers that were given during the first questioning.

In essence, the research is based on multiple static questions. We argue that the proposed solution has the same drawbacks as solutions that only use static questions. As illustrated in [6], static questions are not immune against attacks that include automated data-mining of public records. Also, we argue that online examination has different security requirements with regards to authentication. The user (student) is more likely to want to cheat by allowing another user to take the exam in his place. This differs from e-banking systems where a user does not want another user accessing their account.

The research in [7] analyses human relationships as a basis for user authentication. More precisely, social authentication is based on the premise that a user would be able to recognize a friend or acquaintance, while a stranger could not. The research is based on an authentication method proposed by Facebook, which relies on a user recognizing his friends from various images in his album and identifying them. The research proposes that social authentication is not secure against close enemies. The paper also argues that face recognition algorithms are becoming more and more sophisticated and could be used to automate guessing attacks. Also, the data on social networks is exposed to a wide array of possible attacks and could be used to subvert such an authentication method. The paper also suggests that an attack may be derived by leveraging name statistics in challenges where a user has to name the person on a picture. Bonneau et al [8] showed that many personal knowledge questions related to names are vulnerable to trawling attacks.

The paper proposes a community-based challenge selection, in which a user is shown various pictures that come from different communities. For example, a community could be the users family, work colleagues, childhood friends etc. The result of the research paper suggests that each user has 3 to 4 communities. This does not present a viable solution for systems with medium or high security requirements. Surveys [9] and [10] show that 80% of users do not change their privacy settings from the default ones, allowing their data to be seen by various individuals. Also, a critical observation is that many attackers are usually in the user's friend circle and thus have an intimate understanding of the user.

As stated by [3], question based authentication challenges are often based on information that is already known to the user, while passwords are constructed using various rules and are often hard to memorize. Thus, question based challenges are often used for credential recovery.

The research under [11] describes a similar authentication method based on personal information available from social networks such as Facebook. The conclusion of the research is that attackers can use publicly available information to subvert some question based authentication challenges. For example, the users favorite pets name or their mother's maiden name are examples of weak authentication challenges that can be answered by using publicly available resources like search engines and social networks.

The research in [15] proposes a new authentication method called query-directed passwords (QDP). The research makes a distinction between push and pull passwords. Push passwords are standard password authentication systems where a user has to memorize a new piece of information. Pull passwords are question based authentication challenges as described in [3]. The pull type is further categorized in the following sections:

- Selectable question and selectable answer – a question based system where the system provides both the question and the answers. The user chooses both and needs to memorize them
- Selectable question and open answer – a system where the user has to select a predefined question and responds with his own answer. This present the most common type of static question authentication. This is referred to as fixed questions in [3].
- Open questions and open answer – a question based system where the user has to supply both the question and the answer. This is referred to as open questions in [3].

QDP uses a combination of selectable questions and selectable answers with selectable questions and open answers. However, the answers are restricted to numbers which is called the QDP PIN (personal identification number). When the user tries to authenticate, the PIN derived from the current answers Is compared to the stored PIN.

As stated in [15], QDP achieves optimal security results when used infrequently as there is a finite number of PIN available.

2) Dynamic questions

The research under [4] proposes an authentication method based on leveraging personal history data. The security of the proposed solution is based on the volume of information known in detail only to the authentic user and is too large for an imposter to learn, rather than relying on a shared secret.

The research proposes that the majority requirement for a trusted-third-party (TTP) authentication is the information that the user knows as it is portable and is available at a lower cost. Thus, the method relies on information stored in the user's calendar. The questions are constructed based on various appointments that are in the user's calendar. Examples of such questions are:

- On 27th of June 2003 Friday, was "DMC" scheduled at 14:00-15:00? -Where the answer is either yes or no;
- On the 21st of May 2003 Wednesday, when was it scheduled to meet "Steven". -Where

the answer is a time interval (e.g. 9:00-10:00, 12:00-13:00, 15:00-16:00).

The research results show that such a method is not suitable for real-world authentication, as stated by the authors. The reason for this is that the chosen data source (a users calendar) provides guessable data. Being a feasibility study, it does prove that such an authentication method is possible. Also, the research is focused on questions relating to events. We argue that these are not the only types of questions that can be used for authentication.

Similarly to [7], the research in [12] describes a form of question based authentication aimed at group authentication. The research is based on the work of Basu and Muylle [13], which outlined that there are systems where the affiliations with a group is more important than an individual's identity. The method outlined in this research assumes that an outsider will cause almost negligible damage in case granted access. Thus, it is not a viable solution for commercial systems. However, the work outlines a new form of data that is available and could be leveraged for authentication. The method relies on some form of initial knowledge that is formed within the group and is shared between the group members. The authentication pilot experiment assumes that the members of the group should be able to identify their group knowledge whereas an outsider would find it hard to identify group knowledge. Also, it assumes that the group knowledge accurately reflects the group characteristics.

The results of the experiment showed an almost perfect authentication score in which insiders could recognize group knowledge but outsiders could not. The experiment was also repeated for virtual groups (e.g. chat rooms) where the results were similar to the first experiment. In order to automate the question generation, the Social Network Analysis (SNA) technique is used which was developed in [14]. Using structural and relational properties provided by the SNA technique, the system can generate various types of questions asking about group norms, objectives, decision making, activities and language. For example, it could create a question based on who made a particular decision within the group and who the detractors were. In the pilot experiment, the following three questions were used to authenticate group users:

- Who is the most active member
- Which message topic has recently been sent
- Which is the latest group picture

We note that the answers to these questions change with a frequency of one week to a few months.

The research in [16] proposes the use of user's medical history as a biometric trait. The medical history data is constructed in hierarchical format and then used to construct challenges in form of questions a user must answer in order to authenticate. It relies on the premise that only the authentic user has the secret knowledge about their past medical conditions. The research is focused on solving two problems. The first is related to the speed of the authentication process. The research proposes that a quick method for extracting relevant data from medical history is needed. Secondly, the research addresses the aspect of privacy when leveraging medical data for authentication. The research suggests a solution to both problems by arranging the medical history data in a hierarchical format using h-Data and attribute value taxonomies (AVT). By abstracting the data, the challenges can guard a user's privacy, and at the same time allow for a quick traversal through the data.

The research acknowledges the possibility of using medical data in both direct and implicit questioning. An example of direct questioning would be to simply ask the user if he has a certain condition or has had a specific condition in the past. Also, an example of direct questioning would be to visually inspect the user and see if some medical condition is present. This may be less effective when a human validator is not present. An example of implicit questioning would be to ask a patient suffering from type II diabetes about his daily diet or if he experiences hyperinsulinemia.

Although direct questioning allows for a quick physical inspection of possible answers, it is argued to be inappropriate because users may feel reluctant to answer truthfully in the presence of a human validator. Instead, general questions such as lifestyle and dietary habits could be asked for authentication. The research lacks in defining how an answer, or group of answers, are to be measured for correctness. Also, one if left with the impression that all questions are fundamentally equal with regard to their guess ability and security.

Based on this research, we could conclude that data relating to medical history is related to the user in a very specific way. More precisely, the challenges created by leveraging medical history data could be different than challenges created by user's financial data. We could argue that other user specific data are equally independent and need to be viewed separately. In other words, the way a challenge is created with medical data, may differ from the way data is used with financial data.

B. Authentication in commercial environments

The paper under [17] classifies Internet banking authentication methods based on their resistance to offline and online credential-stealing attacks. Offline attacks take place on an insufficiently protected computer where an attacker may install various malware to steel sensitive and confidential information from the user. Also, attackers can poison the users local DNS cache and redirect traffic to a malicious server he controls in order to launch additional attacks, like Phishing.

Online attacks are defined as attacks that take place during the user's interaction with the target information system. An attacker tries to insert himself in the communication stream and use or steal relevant information to gain more access to the target system. The paper proposes that one-time passwords may be a viable solution for Phishing and similar attacks on user's credentials. Also, the paper proposes that all transactions should be signed by the user's private key. However, such an option relies on a PKI infrastructure and on the notion

that a user would keep their private key safe. Other research has proven that users tend to dislike such methods of authentication [18]. The research results in [18] show that users tend to prefer authentication methods that are based on a single factor. Such preferences show that users are more focused on usability than on security. This could be interpreted as a need for creating a new authentication method that will be both usable and secure.

The research under [19] proposes a new authentication method that removes the need for passwords altogether. The motivation lies in the idea that most phishing attacks could be reduced if users do not need to enter passwords during authentication. By using one-time passwords, phishing becomes much more complex since a stolen password can only be used once. Also, with time sensitive one-time passwords stolen passwords can only be used in a limited time period.

The paper under [20] describes various methods of authentication that are commonly used to authenticate users in e-banking environments. The evaluation shows that passwords are the most insecure method of authentication, while tokens that generate one-time passwords are the safest. The analysis does not include question based authentication.

Since users tend to dislike tokens [18], some approaches to one-time passwords have been focused on delivering them via SMS messages on a user's mobile phone [21]. This form of authentication greatly depends on the security of the underlying network and may include additional overhead costs. For example, in some countries it is possible to launch a SIM swap attack which could be used to attack SMS based user authentication methods.

III. USE OF QUESTION BASED AUTHENTICATION IN COMMERCIAL ENVIRONMENTS

The research in [3] provides criteria for building and evaluating question based challenges. The rules are based on three criteria: privacy, security and usability. Privacy is an important criterion as it does not limit the challenges, but the authenticator. The research proposes that in order to preserve user privacy, the authentication system should store a limited amount of information locally. The security criterion requires that the challenge is not easily guessable or observable from different sources. Usability is a well documented requirement in user authentication [18]. Thus, the usability criterion is concerned with providing a user-friendly experience with regard to applicability, memorability and repeatability. The work in [22] expands the criteria to include 25 properties that are beneficial for an authentication system. The criteria are generic and not related only to question based challenges.

From the above analysis we conclude that commercial environments tend to gravitate towards one-time passwords and two factor authentication [23]. We argue that the reason for this is the short lifespan of a credential used for authentication. Also, this presents one of the requirements outlined in [24]. We argue that in order for a question based authentication method to be a viable option in commercial environments, these requirements need to be met.

As described in [4], a crucial part of the research relevant to question based authentication is to understand how do humans remember and what are the parameters that can affect the human memory. Also, it is crucial to define the types of questions that are difficult for others to answer. In order for a question based authentication method to reach the required level of transiency, it has to use a data source with sufficient information from which the questions can be created.

The limitations of question based authentication methods for use in commercial systems are listed below.

- The questions could be answered by third parties
- Legitimate users could not provide correct answer
- Challenges and answers are known to be static in nature. They do not change over time which makes them susceptible to stealing

Based on the above analysis, we conclude that questions need to be dynamic in order to present third parties from reusing a recorded credential. Furthermore, we propose that a user must know the answer to the dynamic question with high probability. At the same time, the intruder should have a low probability of guessing the answer.

We propose that further research in question based authentication should first define a set of requirements for data sources used to construct questions. Secondly, that there is a need for categorizing information and its properties as well as define the semantic value and user relation to information. The expected outcome of such research should be a set of structures that can be used to construct questions. In the following sections we make a first step towards these goals.

A. Users relation to information

We propose that in order to leverage various types of user information, it is necessary to establish how a user is tied to the information he creates. Further research should involve analyzing how humans remember various types of information. For example, how humans remember facts, relations and values.

We suggest a categorization of user information that can be leveraged for authentication.

- Personal information - this category includes financial and medical data related to the user, as well as other personal information,
- Events - data related to various events,
- Social data - this category includes data that a user generates as part of a group of people or through social interaction,
- Habits - personal routines,
- Preferences & opinions - includes the way a user feels or thinks about a certain topic,

- User generated data - includes various sources where a user generates information that are not necessarily personal. For example, information a user has posted on a forum or group chat, or sent via email.

B. Dynamic questions

We propose that question based authentication can reach the level of security dictated by current standards, such as one-time password tokens, by leveraging large data sources and by creating dynamic questions.

IV. DATA SOURCE REQUIREMENTS

In this section we propose a set of requirements that should be met for a question based authentication method to be viable for use in commercial environments.

1) Dynamic questions

The questions should be created and evaluated on-the-fly, during a user authentication request by the system in charge of the data source. We propose that the authentication system should use its local data source and construct challenges based on its own data source. The questions should be created dynamically and should not be static in nature. This way, the challenges come as close to one-time passwords as possible.

2) Data source availability

The data source should not be publicly available in any way or form. As demonstrated by [11], public data sources are not a viable solution for creating challenges. Since attackers have the same access rights to public data sources as both the user and the authenticator, they can leverage this information for giving correct answers. We propose that a data source should be private and that no other entity should have access to it.

Furthermore, we propose that a question based on any form of user data should have this characteristic: the sum of information is large and well-known only to the individual and also that the volume is too large for imposters to learn. Then it should be possible to collate and query personal information automatically for the purposes of user authentication.

3) Data duplication

No other data source should have the information from which the question was created. Or, at least, if the secondary data source exists, it should have restricted access. If the data source is not public, there is still a possibility of an internal attacker who has access to the data source to use the available information to pose as another user. In order to minimize this threat, we propose that the data in the data source should not exist in other data sources which are publicly available.

V. CONCLUSION

This paper proposes that only dynamic questions can be used to authenticate users in commercial environments. We propose that challenges in question based authentication methods can be improved by using a data source that meets the requirements outlined in the previous section. Additionally, we suggest that legitimate users could be more likely to remember the correct answer if they are asked in a specific way. Although this paper does not provide insight into how this could be achieved, we propose that the key lies in the relationship between the user and the data used for constructing challenges. Thus, we propose a categorization of the data that can be used for authentication.

Further research should focus on dynamic questions and how they can be used in authentication systems. Also, we propose that representative data sources should be studied with the aim of defining methods for authentication question generation. With the goal of hopefully yielding configurable general purpose tools or algorithms for question generation from a variety of appropriate data sources.

REFERENCES

[1] S. M. Furnell, I. Papadopoulos, and P. Dowland, "A long-term trial of alternative user authentication technologies," *Inf. Manag. Comput. Secur.*, vol. 12, no. 2, pp. 178–190, Apr. 2004.

[2] S. Pannifer, "Alternative Authentication – What does it really Provide?," in *ISSE 2013 Securing Electronic Business Processes*, H. Reimer, N. Pohlmann, and W. Schneider, Eds. Springer Fachmedien Wiesbaden, 2013, pp. 121–130.

[3] M. Just, "Designing Authentication Systems with Challenge Questions," in *Security and Usability: Designing Secure Systems That People Can Use*, 1st ed., O'Reilly Media, 2005, pp. 143–160.

[4] A. Nosseir, R. Connor, and M. D. Dunlop, "Internet authentication based on personal history - a feasibility test," in *Proceedings of Customer Focused Mobile Services Workshop at WWW2005*, J. Hjelm, A. Hayrynen, N. Wei, and R. Jana, Eds. ACM Press, 2005.

[5] A. Ullah, Hannan Xiao, and M. Lilley, "Profile based student authentication in online examination," *Inf. Soc. -Soc. 2012 Int. Conf.*, pp. 109–113, Jun. 2012.

[6] V. Griffith and M. Jakobsson, "Messin' with texas deriving mother's maiden names using public records," in *Proceedings of the Third international conference on Applied Cryptography and Network Security*, Berlin, Heidelberg, 2005, pp. 91–103.

[7] H. Kim, J. Tang, and R. Anderson, "Social Authentication: Harder Than It Looks," in *Financial Cryptography and Data Security*, vol. 7397, A. Keromytis, Ed. Springer Berlin Heidelberg, 2012, pp. 1–15.

[8] J. Bonneau, J. Anderson, R. Anderson, and F. Stajano, "Eight friends are enough: social graph approximation via public listings," in *Proceedings of the Second ACM EuroSys Workshop on Social Network Systems*, Nuremberg, Germany, 2009, pp. 13–18.

[9] A. Acquisti and R. Gross, "Imagined communities: awareness, information sharing, and privacy on the facebook," in *Proceedings of the 6th international conference on Privacy Enhancing Technologies*, Cambridge, UK, 2006, pp. 36–58.

[10] B. Krishnamurthy and C. E. Wills, "Characterizing privacy in online social networks," in *Proceedings of the first workshop on Online social networks*, Seattle, WA, USA, 2008, pp. 37–42.

[11] A. Rabkin, "Personal knowledge questions for fallback authentication: security questions in the era of Facebook," in *Proceedings of the 4th symposium on Usable privacy and security*, Pittsburgh, Pennsylvania, 2008, pp. 13–23.

[12] A. Nosseir, R. Connor, and K. Renaud, "Question-based group authentication," in *Proceedings of the 18th Australia conference on Computer-Human Interaction: Design: Activities, Artefacts and Environments*, Sydney, Australia, 2006, pp. 277–283.

[13] A. Basu and S. Muylle, "Authentication in e-commerce," *Commun ACM*, vol. 46, no. 12, pp. 159–166, 2003.

[14] D. F. G. Calvin L. Streeter, "Social Network Analysis," *J. Soc. Serv. Res.*, vol. 16, pp. 201–222, 1992.

[15] L. O'Gorman, A. Bagga, and J. Bentley, "Query-directed passwords," *Comput. Secur.*, vol. 24, no. 7, pp. 546–560, 2005.
[16] S. Fong and Y. Zhuang, "Using Medical History Embedded in Biometrics Medical Card for User Identity Authentication: Data Representation by AVT Hierarchical Data Tree," *J. Biomed. Biotechnol.*, vol. 2012, pp. 1–11, 2012.
[17] A. Hiltgen, T. Kramp, and T. Weigold, "Secure Internet banking authentication," *Secur. Priv. IEEE*, vol. 4, no. 2, pp. 21–29, Apr. 2006.
[18] C. S. Weir, G. Douglas, T. Richardson, and M. Jack, "Usable security: User preferences for authentication methods in eBanking and the effects of experience," *Interact. Comput.*, vol. 22, no. 3, pp. 153–164, May 2010.
[19] C.-Y. Huang, S.-P. Ma, and K.-T. Chen, "Using one-time passwords to prevent password phishing attacks," *Adv. Top. Cloud Comput.*, vol. 34, no. 4, pp. 1292–1301, Jul. 2011.
[20] P. Hanacek, K. Malinka, and J. Schafer, "e-banking security - A comparative study," *Aerosp. Electron. Syst. Mag. IEEE*, vol. 25, no. 1, pp. 29–34, Jan. 2010.
[21] S. Indu, T. N. Sathya, and V. Saravana Kumar, "A stand-alone and SMS-based approach for authentication using mobile phone," *Inf. Commun. Embed. Syst. ICICES 2013 Int. Conf.*, pp. 140–145, Feb. 2013.
[22] J. Bonneau, C. Herley, P. C. van Oorschot, and F. Stajano, "The Quest to Replace Passwords: A Framework for Comparative Evaluation of Web Authentication Schemes," *Secur. Priv. SP 2012 IEEE Symp.*, pp. 553–567, 20.
[23] "Banks move towards two-factor authentication," *Comput. Fraud Secur.*, vol. 2005, no. 7, p. 20, Jul. 2005.
[24] K. Skracic, P. Pale, and B. Jeren, "Knowledge based authentication requirements," in *2013 36th International Convention on Information Communication Technology Electronics Microelectronics (MIPRO)*, 2013, pp. 1116–1120.

Technique for Detection of Cooperative Black Hole Attack using True-link in Mobile Ad-hoc Networks

Gayatri Wahane[1], Ashok M. Kanthe[2], Dina Simunic[3]
[1, 2] Sinhgad Institute of Technology, Lonavala, Pune, India
[2, 3] Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia
gayatri.wahane@gmail.com[1], ashokkanthe@gmail.com[2], dina.simunic@fer.hr[3]

Abstract - Mobile Ad-hoc Network (MANET) is a collection of communication devices or nodes that wish to communicate without any fixed infrastructure and pre-determined organization of available links. Security is a major challenge for these networks owing to their features of open medium, dynamically changing topologies. A black hole is a malicious node that falsely replies for any route requests without having active route to specified destination and drops all the receiving packets. Sometimes the black hole nodes cooperate with each other with the aim of dropping packets. These are known as cooperative black hole attack. This proposed work suggests the modification of Ad-hoc on Demand Distance Vector (AODV) routing protocol. We used a technique for detecting as well as defending against a cooperative black hole attack using True-link concept. True-link is a timing based countermeasure to the cooperative black hole attack. This paper shows the performance of MANET decreases for end-to-end delay, normalized routing overhead and increases throughput and packet delivery ratio.

I. INTRODUCTION

A Mobile Ad hoc network (MANET) is a self - configuring network [1, 2] that does not require any fixed infrastructure, which minimizes their cost as well as deployment time. As each node in this network is free to move that makes the network to change its topology continuously. These infrastructure-less mobile nodes in ad-hoc networks dynamically create routes among themselves and form own wireless network on the fly. Because of the dynamic nature, these networks are more vulnerable to attacks so security is an important as well as serious issue in MANET. One of the most critical problems in MANETs is the security vulnerabilities of the routing protocols. A set of nodes in a MANET may be compromised in such a way that it may not be possible to detect their malicious behavior easily. Such nodes can generate new routing messages to advertise non-existent links, provide incorrect link state information, and flood other nodes with routing traffic.

One of the most widely used routing protocol in MANETs is the Ad hoc on-demand distance vector (AODV) routing protocol. AODV is vulnerable [2] to the well known black hole attack. Most authors have proposed a solution to identify single black hole attack. However in their proposed solution many of them found multiple black hole nodes. Some author have suggested solution for detecting cooperative black hole attack but due to multipath routing it require more end to end delay and more routing overhead. The proposed technique works with modified AODV protocol and routing information table for searching trustful node with the help of tru-link.

This paper is organized as follows. In Section II related work for detecting cooperative black hole attack has been discussed. Section III provides AODV and its work in which discuss overview of AODV protocol with the description of black hole attack characteristics. Section IV describes the proposed solution for detecting cooperative Black hole attacks in mobile ad hoc networks and shows the working of the algorithm with the help of an example. Section V shows the simulation results of the proposed work. We conclude plan for future work in section VI.

II. RELATED WORK

In this section we will discuss research work has been done by various authors.

J. Sen, S. Koilakonda and A. Ukil [3] proposed a mechanism for defending against a cooperative black hole attack. The mechanism modifies the AODV protocol by introducing two concepts, (I) data routing information (DRI) table and (II) cross checking. In the DRI scheme, two bits of additional information are sent by the nodes which respond to the RREQ message of a source node during route discovery process. Each node maintains an additional data routing information (DRI) table. In the DRI table, the bit 1 stands for "true" and the bit 0 stands for "false". The first bit "From" stands for the information on routing data packet from the node (in the Node filed), while the second bit "Through" stands for information on routing data packet through the node. The process of cross checking the intermediate nodes is a one-time procedure which should be affordable for security. The cost of crosschecking the nodes can be minimized by allowing the nodes to share the DRI table of their trusted nodes with each other.

J. Eriksson, S. V. Krishnamurthy, M. Faloutsos [4] proposed an algorithm to detect a wormhole attack by

using True-link concept. True-link is a timing based countermeasure to the wormhole attack. Using True-link, a node can verify the existence of a direct link to an apparent neighbor. Verification of a link operates in two phases. In the rendezvous phase, the nodes exchange nonce. This is done with tight timing constraints, within which it is impossible for attackers to forward the exchange between distant nodes. In the authentication phase, i and j transmit a signed message, mutually authenticating themselves as the originator of their respective nonce. True-link is meant to be used together with a secure routing protocol. True-link is virtually independent of the routing protocol used. This work shows that True-link provides competent protection against potentially devastating wormhole attacks.

S. Banerjee [5] proposed a mechanism capable of detecting and removing the malicious nodes launching these two types of attacks. Their approach consists of an algorithm which works as follows. Instead of sending the total data traffic at a time we divide the total traffic into some small sized blocks. So that malicious nodes can be detected and removed in between the transmission of two such blocks by ensuring an end-to-end checking. Source node sends a prelude message to the destination node before start of the sending any block to alert it about the incoming data block. Flow of the traffic is monitored by the neighbors of the each node in the route.

L. Tamilselvan and V. Sankar Narayana [6] proposed a solution with the enhancement of the AODV protocol which avoids multiple black holes in the group. A technique is used to identify multiple black holes that are cooperating with each other and discover the safe route. It was assumed in the solution that nodes are already authenticated and therefore can participate in the communication. It uses fidelity table where every node that is participating is given a fidelity level that will provide to that node. Any node having 0 value is considered as malicious node and is eliminated.

H. Weerasinghe [7] proposed the solution which discovers the secure route between source and destination by identifying and isolating cooperative black hole nodes. This solution adds on some changes in the solution proposed by the Ramaswamy to improve the accuracy. This algorithm uses a methodology to identify multiple black hole nodes working collaboratively as a group to initiate cooperative black hole attacks. This protocol is slightly modified version of AODV protocol by introducing Data Routing Information (DRI) table and cross checking using Further Request (FREQ) and Further Reply (RREP).

C. Wu Yu, Tung-Kung, W. ReiHeng, Cheng and S. Chao Chang [8] proposed a distributed and cooperative procedure to detect black hole node. In this each node detect local anomalies. It collects information to construct an estimation table which is maintained by each node containing information regarding nodes within power range. This scheme is initiated by the initial detection node which first broadcast and then it notifies all one-hop neighbors of the possible suspicious node. They cooperatively decide that the node is suspicious node.

III. AODV AND ITS SECURITY PROBLEMS

A. AODV and Black Hole Attack

1. Overview of Ad hoc On Demand Distance Vector (AODV)

AODV is a reactive [9,10] routing protocol that does not require maintenance of routes to destination nodes. As its name indicates AODV is an on-demand routing protocol that discovers a route only when there is demand from mobile node.

2. Cooperative Black Hole Attacks

A black hole attack is a kind of Denial of service (DoS) attack in mobile ad hoc networks. In this attack [11], a malicious node sends a fake RREP packet to the source node that has initiated a route discovery, in order to show itself as a destination node or an intermediate node to the actual destination node. In such a case the source node would send all of its data packets to the malicious node. The malicious node then absorbs all the packets and drops them fully or sometimes partially. As a result source and destination node will not be able to communicate with each other as shown in figure 1.

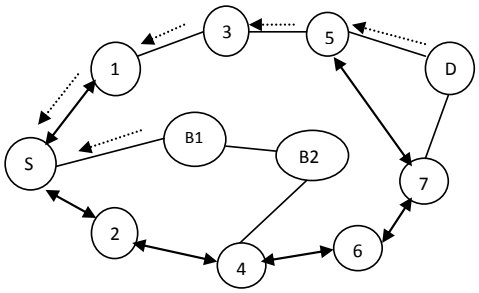


Figure 1. Detection of cooperative black hole attack

Consider the case where S is the source node, D is the destination node and B1 is the malicious node, here node S starts with the route discovery process then the node B1 advertises itself as having a valid shortest route to the destination, here the route is false with the purpose of intercepting packets. Moreover a malicious node does not need to check its routing table when sending a false message; its response is more likely to reach the source node first. This makes the source node think that the route discovery process is complete, ignore all other reply messages and begin to send data packets. As a result, all the packets through the malicious node are simply absorbed discarded and then lost. The malicious node could be said to form a black hole in the network. Sometimes these malicious nodes cooperate with each other with the same aim of dropping packets, these are known as cooperative black hole nodes and the attack is known as cooperative black hole attack.

IV. THE PROPOSED WORK

We propose a methodology for identifying multiple black hole nodes cooperating as a group with slightly

modified AODV protocol by introducing Data Routing Information (DRI) Table and Cross Checking with True-link concept shows in the system architecture of figure 2.

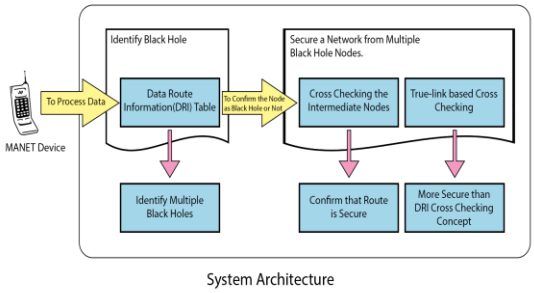


Figure 2. System Architecture

A. Data Routing Information Table

Figure 1 shows the cooperative black hole attack. Figure 1 illustrates this topology where black hole node B1 and B2 are malicious and cooperate with each other. They are attracting intermediate nodes (IN) for shortest path towards a destination and drop the packets.

The solution to identify multiple black hole nodes acting in cooperation involves two bits of additional information from the nodes responding to the RREQ of source node S. The first bit "From" stands for Information on routing data packet from the node (in the Node field) while the second bit "Through" stands for information on routing data packet through the node (in the Node field). Each node maintains an additional Data Routing Information (DRI) table. In the DRI table I, 1 stands for 'true' and 0 for 'false'.

TABLE I ADDITIONAL TABLE OF DATA ROUTED FROM, AND ROUTED TO NODES MAINTAINED BY NODE 4.

Node	**Data Routing Information**	
	From	***Through***
5	1	0
6	1	1
B2	0	0
2	1	1

Consider DRI table I which is maintained for node 4 in which the entry 1 0 for node 5 shows that node 4 has routed data packets from 5, but has not routed any data packets through 5 (before node 5 moved away from 4). The entry 1 1 for node 6 shows that, node 4 has successfully routed data packets from and through node 6. The entry 0 0 for node B2 shows that, node 4 has NOT routed any data packets from or through B2. As an example, let's consider the network in figure 3.

When node B1 responds to source node S with RREP message, it provides its next hop node B2 and DRI for the next hop (i.e. if B1 has routed data packets through B2). Here the black hole node lies about using the path by replying with the DRI value equal to 0 1. Upon receiving RREP message from B1, the source node S will check its own DRI table to see whether B1 is a reliable node. Since S has never sent any data through B1 before, B1 is not a reliable node to S. Therefore, *S* sends further request to *B2* via alternative path *S*-5-6-*B2* and asks *B2* about three things: if B2 has routed any data from B1, who is B2's next hop, and if B2 has routed data packets through B2's next hop. Since B2 is collaborating with B1, it replies positively to all the three requests and gives node 6 (randomly) as its next hop. When the source node contacts node 6 via alternative path S-5-6-D to cross check the claims of node B2, node 6 responds negatively. Since node 6 has neither a route to node B2 nor has received data packets from node B2, the DRI value corresponding to B2 is equal to 0 0 as shown in figure 4. Based on this information, node S can infer that B2 is a black hole node. If node B1 was supposed to have routed data packets through node B2, it should have validated the node before sending it. Now, since node B2 is invalidated through node 6, node B1 must cooperate with node B2. Hence both nodes B1 and B2 are marked as black hole nodes and this information is propagated through the network leading to their listing as black holes, and revocation of their certificates. Further, S discards any further responses from B1 or B2 and looks for a valid alternative route to D.

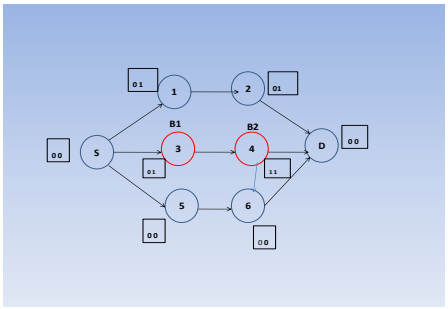


Figure 3. Detection of cooperative black hole attack by using DRI table

B. Cross Checking with True-link

Crosschecking-True-link is a timing based mechanism (time-based approach) against cooperative black hole attacks. Black hole enters mostly through network layer. So, we include Mac layer with the network layer integration (i.e cross layer) attacker will be confused that which node having the data and direction of forwarding the data. This assumption will helpful for detecting multiple black holes. By using this True-link we are integrating DRI table information with True-link concept. First we collect the black hole information using DRI table after that make list of malicious nodes and checking malicious nodes using true-link concept. True-link consists of rendezvous phase. The rendezvous phase is similar to RTS-CTS-DATA-ACK exchange, here nodes a and b will exchange two randomly generated numbers, called values. The assumption is that the IEEE 802.11 standard makes it extremely difficult for an attacker to successfully relay the exchanged frames. In the

authentication phase, i and j each transmit a signed message (α_i, β_j) mutually authenticating themselves as the originator of their respective values.

To verify the existence of a link, i initiate a link verification exchange. The exchange can be done using rendezvous phase. The rendezvous phase is managed by strict timing constraints, making it extremely difficult for an attacker to fake a link. The rendezvous phase is implemented as a single RTS-CTS-DATA-ACK exchange. We describe the operation of the protocol for each of these frames. The rendezvous operation includes following steps:

Request to Send (RTS) : Node i called initiator, calls init ( ). Node i sends an RTS to node j.

Clear to Send (CTS (α_j)): After receiving the RTS, node j calls HandleRTS ( ). A locally generated value α_j is included in the CTS, which is sent after a delay of one Short Interframe Space (*SIFS*), is the small time interval between the data frame and its acknowledgment.

DATA (βi): Having received the CTS, i generate value β_i. After a SIFS delay, the nonce is sent as the packet payload, together with a header, identifying the packet as a rendezvous packet.

ACK: When node j receives the payload packet containing values β_i, the HandleRTS ( ) function sends the ACK frame, after a SIFS delay. The received packet and the locally generated values α_j are handed to the upper layer for processing. When node i receives the ACK, the init () function at i returns the pair (α_j , β_i) to its caller.

RREQ: Route request packet
RREP: Route reply packet
SN:Source node
DN: Destination node
IN: Intermediate node
α,β: nonce key
DRI: Data Routing Information
Step 1: SN broadcasts the RREQ packet
Step 2: SN receives RREP
Step 3: if RREP is from any IN then
Step 4: SN verify the RREP of the node which sends RREP
Step 5: SN verify updates entry of DRI value of the intermediate node that initialized RREP
Step 6.then sendRTS();
Step 7.# recvCTS() times out after $t_{timeout}$ = SIFS
Step 8.α=recvCTS();
Step 9.if α != nil then
Step 10.β->nonce();
Step 11: # success = true if ACK received
Do reactive Link verification and routing method
Step 12 success=sendPacket([VerifyLink,β]);
Step 13.if success == true then
Step 14: Consider the route to be safe and start routing the data packets
Step 15: else
Step 16: Mark the node as malicious node and build a blackhole list
Step 17: end if

Figure 4. Proposed algorithm with True-link concept

The authentication phase does not have any strong timing constraints. Nodes send the authentication messages as soon as they have finished computing the cryptographic signature. Node i sends the message (α_j , β_i)i (signed by i) to j, thereby proving that it was the one sending the nonce β_i and that it received the correct α_j. This proves the existence of the link i to j to node j. Similarly, node j sends the message (α_j , β_i)j to i. This proves the existence of the link to node i.

C. Link Verification Maintenance and Routing

Verifying the existence of a link is important in itself, but in a dynamic network, this knowledge needs to be maintained and properly updated as the topology changes. We present two methods of link verification maintenance with True-link.

1. Proactive Link Verification

With this method, a node periodically transmits beacon messages to find its' current apparent neighbor set. It maintains knowledge about the actual neighbor set by periodically probing each neighbor with a link verification exchange. The proactive method is well suited to proactive routing protocols, or reactive protocols with local hello-messages. With these protocols, periodic beaconing maintains a consistent record of the apparent neighbors of a node, and it requires moderate extra effort to verify the links to each of these neighbors. When a new neighbor is found, a link verification exchange is initiated before the routing layer is notified of the neighbor discovery. Moreover, after an expiry period expiry interval has passed, each link is re-verified, to ensure that it still exists. The value of expiry interval is flexible, but should not be smaller than the beaconing interval.

2. Reactive link verification

Nodes using this method perform no link verification until the use of a link is needed. The reactive method is best suited to reactive routing protocols. As in the proactive case, a cache of recently verified neighbors is maintained, with an expiry interval (texp). In the reactive link-verification process for a broadcast route request a node waits until it receives a broadcast packet across a previously unverified link. At this time, it initiates a link verification exchange with the sender of the packet. To avoid a potential "implosion" effect, where many receivers try to initiate an exchange at the same time, the receiver waits for a small randomly selected time interval before initiating the link verification exchange.

D. Overhead vs. Security Tradeoff:

During the time interval expiry interval, a node may believe that it has a link with a neighbor, when in fact the neighbor has moved away. This provides a window of opportunity for a black hole attack faking the previously existing link. Thus, the value of expiry interval provides a tradeoff between high overhead due to frequent verifications and attack vulnerability, due to potentially obsolete information. In our simulations, we use expiry interval values between 10 and 120 seconds, with 30 seconds being the standard choice.

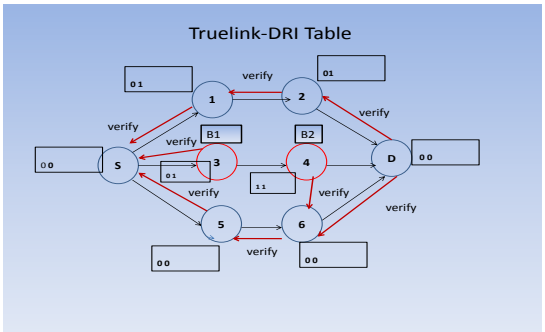


Figure 5. True-link with DRI table

V. SIMULATION RESULTS

The proposed algorithm has been carried out using the network simulator *ns-2* [12]. The 802.11 MAC layer implemented in ns-2 is used for simulation. We have used improved version of random waypoint model. Performances of the three protocols are evaluated: (I) standard AODV protocol, (II) AODV with two malicious nodes cooperating in a black hole attack, and (III) AODV with the proposed algorithm. The scenarios developed to carry out the tests using two parameters: (I) the mobility of the nodes and (II) the number of node in the network. The chosen parameters for simulation are shown in Table II. Following metrics are chosen to evaluate the impact of the black hole attack on the simulated network: (I) Throughput (II) Packet delivery ratio (III) Normalized Routing overhead and (IV) Average end-to-end delay. By using these metrics we are going to check the performance of the proposed scheme as compared to previous scheme.

TABLE II SIMULATION PARAMETERS

Sr. No.	Parameter	Value
1	Simulator	NS 2.35
2	DoS Attack	Black hole
3.	Channel Type	Wireless channel
4	Antenna Type	Omni directional
5	Protocol used	AODV
6	Underlying MAC Protocol	IEEE 802.11
7	Propagation Model	Two-ray ground
8	Mobility Model	Random way point
9	Queue	Droptail/Priqueue
10	Area(meter square)	1000*1000
11	Simulation time	200 sec
12	Pause time	1.0 sec
13	Speed(mps)	1,5,10,15,20,25,30
14	Traffic Type	CBR
15	MAC type	802.11
16	Link layer type	LL
17	CBR rate	1pkt/sec
18	Packet Size	512 bytes/packet
19	Number of sources/Destination connections	3
20	Number of Nodes	50
21	Number of malicious node detected	Two or more nodes which cooperate each other for dropping packet

Effects of node mobility vs performance metrics

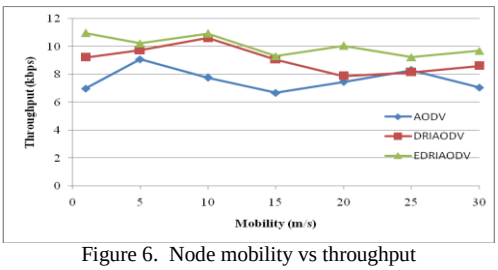


Figure 6. Node mobility vs throughput

Figure 6 to figure 9 shows the effects of mobility vs performance metrics like throughput, packet delivery ratio, normalized routing overhead and average end-to-end delay.

Figure 6 shows simulation result of node mobility vs throughput. Green color line shows proposed work in which we have achieved high throughput when node mobility is increasing as compared to AODV and DRIAODV means it improves the performance of proposed algorithm.

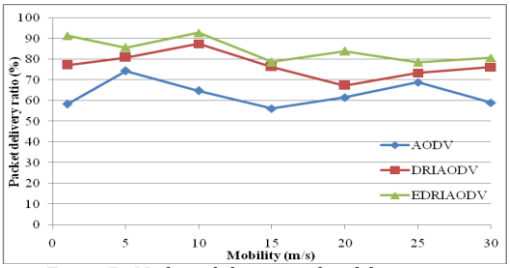


Figure 7. Node mobility vs packet delivery ratio

Figure 7, the metric packet delivery ratio is plotted against the node mobility. As the total number of mobile nodes in the network is 50, the node mobility is 1 to 30 mps. It achieves maximum delivery ratio. This is improvement over previous mechanism. It is observed that AODV performs better for lower node mobility rates. The delivery rate starts dropping with increasing mobility of the nodes. The proposed algorithm increases the delivery ratio to 72%.

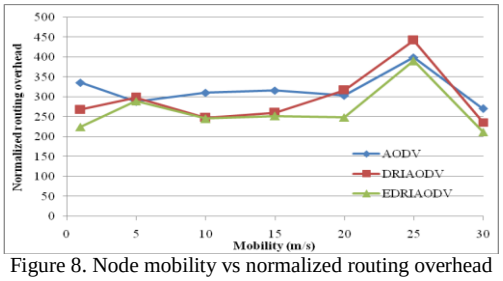


Figure 8. Node mobility vs normalized routing overhead

Figure 8 shows the simulation result with respect to Node mobility vs Normalized routing overhead. It achieves less normalized routing overhead. When mobility is high at that time routing overhead is also high but if we compare with normal AODV protocol it is 50% less. This proposed work achieves higher performance as compared to existing work.

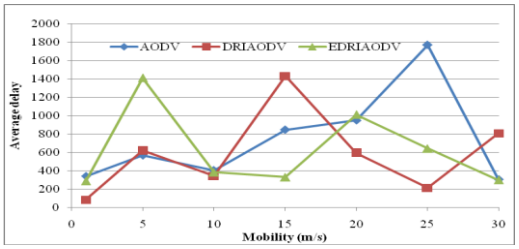


Figure 9. Node mobility vs average end-to-end delay

Above figure shows the node mobility vs average end-to-end Delay. Average delay should be less while achieving performance of AODV. In this, delay is 25% less as compared to existing system because of link verification is performed while sending packet on the network.

Effects of number of nodes vs performance metrics

Figure 10 to figure 13 shows throughput, packet delivery ratio, normalized routing overhead and average end-to-end delay.

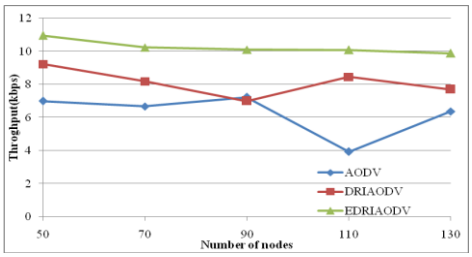


Figure 10. Number of nodes vs throughput

Figure 10 shows that when the number of nodes increases, the throughput increases for EDRIAODV protocols since source nodes can find more and more routes to destinations which can avoid black hole nodes (some normal nodes can reply before the black hole nodes).

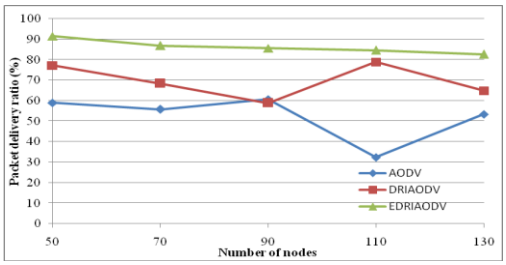


Figure 11. Number of nodes vs packet delivery ratio

Figure 11 shows that Packet delivery ratio is high as compared to other protocol. Because true-link performs link verification on selected black hole node means it will take less time as compared to other protocol when number of nodes are more. The delivery rate starts dropping with decreasing number of nodes.

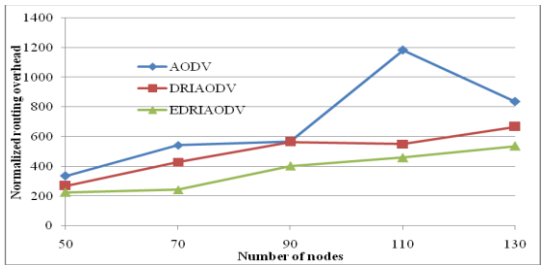


Figure 12. Number of nodes vs normalized routing overhead

Figure 12 shows that normalized routing overhead is less as compared to other protocol. Because AODV makes the lowest overhead when black holes are present since AODV uses less route request when black hole reply quickly for route request.

Figure 13 shows the impact of the number of nodes on the end-to-end delay. First, the end-to-end delay does not depict much difference when the number of nodes increases because the source and destination pairs of all CBR applications are the same for each scenario regardless of the number of nodes. But for less number of nodes, all protocols take more time since alternative routes are limited. Second, our protocol has little bit of more end-to-end delay than others since it takes more time to find out a secure route.

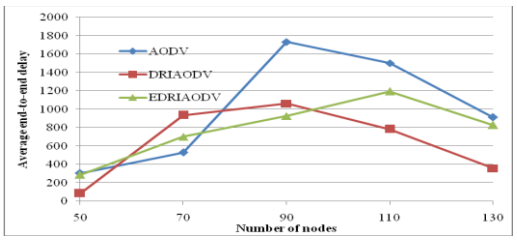


Figure 13. Number of nodes vs average end-to-end delay

VI. CONCLUSION

In this proposed work, the issue of cooperative black hole attack and its affect on the AODV-based routing protocol has been discussed. The route discovery process in the AODV is susceptible to Cooperative black hole attack and therefore, it is vital to have an efficient security method built into the AODV protocol in order to mitigate the effect of such attacks. True-link-crosschecking method is designed to isolate and mitigate the effect of black hole attacks in MANET. True-link-

crosschecking enhances AODV protocol to improve the network performance by improving routing update condition. The enhancement only involves a minimum modification in DRI based cross checking with True-link rendezvous phase by changing the existing AODV protocol scheme. This solution reduces routing overhead and delay. It achieves maximum throughput when number of nodes and pause time more. In proposed work we have reduced end to end delay as well as routing overhead. In future work, we are planning to reduce routing overhead by making nonce more secure and timestamp in link verification.

REFERENCES

[1] R. Prasad, S. Dixit, R. Van Nee, "Globalization of Mobile and Wireless Communication", March 2011, Springer, p. 335.

[2] L. Gavrilovska, R. Prasad, "Ad Hoc Networking Towards Seamless Communications", Springer 2006, p.284.

[3] J. Sen, S. Koilakonda and A. Ukil, "A Mechanism for Detection of Cooperative Black Hole Attack in Mobile Ad Hoc Networks". Second international conference on intelligent system, modeling and simulation, Kolkata, 2011, 25-27 .

[4] J. Eriksson, S. V. Krishnamurthy, M. Faloutsos, "TrueLink: A Practical Countermeasure to the Wormhole Attack in Wireless Networks", 2011. The Communications and Networks Consortium sponsored by the U. S. Army Research Laboratory under the Collaborative Technology Alliance Program.

[5] S. Banerjee, "Detection/Removal of Cooperative Black and Gray Hole Attack in Mobile Ad-Hoc Networks". The World Congress on Engineering and Computer Science 2008.

[6] L. Tamilselvan and V Sankar Narayana, "Prevention of Black Hole Attack in MANET". Journal of Networks, 2008, Volume 3, Number 5, pp 13-20.

[7] H. Weerasinghe, "Preventing Cooperative Black Hole Attacks in Mobile Ad Hoc Networks: Simulation Implementation and Evaluation", Proceedings of the Future Generation Communication and Networking, 2007,Volume 2, pp 362-367.

[8] C. Wu Yu, Tung-Kuang, W. ReiHeng, Cheng and S. Chao Chang, "A Distributed and Cooperative Black Hole Node Detection and Elimination Mechanism for Ad Hoc Networks". PAKDD International Workshop, Nanjing, China,2007, pp 538-549

[9] C. Perkins, E. B. Royer and S. Das, "Ad hoc On Demand Distance Vector (AODV) Routing, Internet Draft", RFC 3561, IETF Network Working Group, July 2003.

[10] C. Perkins, E. B. Royer and S. Das, "Ad hoc On-Demand Distance Vector Routing", Proceeding of the 2nd IEEE Workshops on Mobile Computing System and Applications (WMCSA), pp. 90-100, 25-26 Feb.,1999.

[11] A. M. Kanthe, D. Simunic, M. Djurek, "Denial of Service (DoS) Attacks in Green Mobile Ad-hoc Networks", MIPRO 2012, IEEE Conference, Proceedings of the 35th International Convention, ISBN:978-1-4673-2511-6, May21-25, 2012, Opatija,Croatia

[12] The network simulator-ns 2.35. http://www.isi.edu/nsnam/ns/1997

The Effects of Combined Application of SOM, ANFIS and Subtractive Clustering in Detecting Intrusions in Computer Networks

Dr.Sc Zikrija Avdagić
Computer Science Department
Faculty of Electrical Engineering
Zmaja od Bosne bb., Kampus
71 000 Sarajevo, Bosnia and Herzegovina
Phone: + 387 33 250 737
Fax: + 387 33 250 725
E-mail: zikrija.avdagic@etf.unsa.ba

M.Sc Admir Midžić
IT Services Monitoring and Security Department
Join Stock Company BH Telecom Sarajevo
Obala Kulina banSa 8, 71 000 Sarajevo, Bosnia and Herzegovina
Phone: + 387 37 229 929
Fax: + 387 33 776 403
E-mail: admir.midzic@bhtelecom.ba

Abstract - Building a system for the detection and prevention of intrusions into computer networks is a major challenge. Huge amounts of network traffic that process these systems are characterized by diversity and the data are described by a number of attributes. In addition, input data are often changing in a relatively short period of time, creating a completely new traffic patterns. This significantly complicates the identification of potentially unwanted network traffic. The aim of this paper is to present and analyze the effects of combined application of Self Organizing Map (SOM), Adaptive Neuro Fuzzy Inference System (ANFIS), Subtractive Clustering (SC) and Voting Mechanism (VM) in building systems for intrusion detection in computer networks in order to maintain an acceptable level of efficiency of data processing and increased system adaptivity.

Keywords: Self Organizing Map, Adaptive Neuro Fuzzy Inference System, Subtractive Clustering, Voting Mechanism, Intrusion Detection and Prevention.

I. INTRODUCTION

The development of information technology in the field of processing, collection and distribution of data is further accelerated by the needs of modern business. Modern business is increasingly based on various forms of electronic commerce, which is further stimulated by the expansion of the Internet. This way of doing business brings new risks to the security of information systems. New threats appear daily by individuals and organizations who attack and abuse information systems. Accordingly, for the achievement of business goals of the company it is essential to ensure the availability and proper functioning of computer networks and to ensure that information systems are provided with information when they are needed. This creates the conditions for proper decision-making at all levels of management and decision-making system. One of the main requirements for the designers of computer networks in the past was to enable their fast and efficient work. One of the main objectives today is to prevent intrusions into computer networks through preventive action, or detect them when they happened. There is a trend in the world to theoretically and practically explore ways of detecting and preventing intrusions in computer networks [3]. The research in this paper is based on the use of SOM (neural network based on competitive learning), ANFIS, SC and VM in process of intrusion detection in computer networks. This paper is organized as follows. Section II presents related work in past several years. Section III presents data set that was used for system development, training and testing. Section IV present architecture of system describing its components using block diagrams and mathematical expressions. Section V presents proposed algorithm and reasons for combining SOM, ANFIS and SC. Section VI presents and discuss experimental result of algorithm execution. Section VII concludes the paper.

II. RELATED WORK

The researchers have been widely used SOM for data clustering [4]. However, SOM have some drawbacks. For instance, the network architecture of SOM has to be established in advance and it requires knowledge about the problem domain [5]. That kind of knowledge is difficult to achieve, especially with data of high dimensionality as it is case with intrusion detection and prevention data set [3]. Moreover, the hierarchical relations among input data are difficult to represent.

Several dynamic algorithms were proposed to overcome those drawbacks. The growing hierarchical SOM (GHSOM) is one algorithm that tries to face these problems. It is an neural network which consists of independent growing SOM arranged in layers [7]. The next is Enhanced Dynamic SOM (EDSOM). EDSOM is based on GHSOM. It is a dynamic node growth structure, but instead of using rectangular grid, a closed circle structure is applied for outpout node structure [6]. A Large Scale Memory Storage and Retrieval (LAMSTAR) network is another approach that is used for intrusion detection by combining SOM modules and statistical decision tools. It was specifically developed for application problems involving very large memory that relates to many different attributes [5]. The logistic regression model is used for data analysis concerned with describing the relationship between a response variable and one or more predictor variables. A logistic regression model with a binary response variable is a particular case of the general multinomial logistic regression model. Its dependent variable can have more than two choices that are coded categorically and in that case, one of the categories is taken as the reference category. A study on KDD CUP 99 that was published [3] identified predictors that are statistically significantly associated with the response variable.

III. KDD CUP 99 DATASET

Comparing various Intrusion Detection Systems (IDS) is extremely difficult because it is hard to find publicly avaliable data set. Fear of violating privacy is one of the main reasons why most companies are unwilling to provide their data to other institutions. Even if some of those data are available, it is very difficult to denote data as normal or as those signifying attack, since this would require the participation of a large number of experts in this area, and also a significant amount of time. Constant changes of network traffic include also new types of cyber attacks. It also brings a new definition of "normal" behavior. Because of that, it is no surprise that only a small number of shared data sets are available. IDS researchers have mostly used the KDD CUP 99 dataset in their work, thus the same data set was used in this research. It is composed of basic attributes, such as connection duration, protocol type, service type, status indicator, total bytes sent, total bytes received, whether the source and target address are same or not, number of wrong fragments, and number of urgent packages. In addition to these attributes, each connection is described by further thirty-two attributes that may be classified into the following three categories: contents attributes, time-based traffic attributes and host traffic based attributes. Each entry is composed of forty-one attributes plus target value. The target value represents the type of attack. The number of records was 4898430. 972780 records represented normal traffic, and the remaining records were attacks, with the following distribution: DoS (Denial of Service) 3883370, testing 41102, compromising attacks 1178 with 52 in relation to U2R (User to Root), and 1126 R2L (Root to Local). Duplicate records were removed from the dataset. After deleting duplicate records, the traffic was distributed as it is shown in Table 1.

TABLE 1. AN OVERVIEW OF TRAFFIC DISTRIBUTION

TRAFFIC	records	percentage	records	percentage
normal	812813	75,611	60593	19,481
probe	13860	1,289	4166	1,339
DoS	247267	23,002	229853	73,901
U2R	52	0,005	228	0,0733
R2l	999	0,093	16189	5,205
Total:	1074991	100	311029	100

Most of classification methods are not able to process the data in such a format. It is therefore necessary to perform pre-processing the data into a format acceptable by the classification algorithms. Examples of records from the KDD CUP 99 data set before pre-processing of data are as follows:

a) 0 tcp other S0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 130 13 1.00 1.00 0.00 0.00 0.10 0.05 0.00 255 13 0.05 0.05 0.00 0.00 1.00 1.00 0.00 0.00 neptune.

b) 0 udp private SF 105 147 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0.00 0.00 0.00 0.00 1.00 0.00 0.00 255 253 0.99 0.01 0.00 0.00 0.00 0.00 0.00 0.00 normal.

For the purposes of this research attributes that describe each record used in the training and test data are labeled with the following symbols: $A_1, A_2, A_3, A_4, A_5, A_6, A_7, A_8, A_9, A_{10}, A_{11}, A_{12}, A_{13}, A_{14}, A_{15}, A_{16}, A_{17}, A_{18}, A_{19}, A_{20}, A_{21}, A_{22}, A_{23}, A_{24}, A_{25}, A_{26}, A_{27}, A_{28}, A_{29}, A_{30}, A_{31}, A_{32}, A_{33}, A_{34}, A_{35}, A_{36}, A_{37}, A_{38}, A_{39}, A_{40}, A_{41}, A_{42}$ a refer to the value of: DURATION, PROTOCOL TYPE, SERVICE, FLAG, SRC BYTES, DST BYTES, LAND, WRONG FRAGMENT, URGENT, HOT, NUM FAILED LOGINS, LOGGED IN, NUM COMPROMISED, ROOT SHELL, SU ATTEMPTED, NUM ROOT, NUM FILE CREATIONS, NUM SHELLS, NUM ACCESS FILES, NUM OUTBOUND CMDS, IS HOST LOGIN, IS GUEST LOGIN, COUNT, SRV COUNT, SERROR RATE, SRV SERROR RATE, RERROR RATE, SRV RERROR RATE, SAME SRV RATE, DIFF SRV RATE, SRV DIFF HOST RATE, DST HOST COUNT, DST HOST SRV COUNT, DST HOST SAME SRV RATE, DST HOST DIFF SRV RATE, DST HOST SAME SRC PORT RATE, DST HOST SRV DIFF HOST RATE, DST HOST SERROR RATE, DST HOST SRV SERROR RATE, DST HOST RERROR RATE, DST HOST SRV RERROR RATE and ATTACK TYPE respectively.

IV. SYSTEM ARCHITECTURE

Building a system that is used in this work was being done in phases using multiple components (Fig. 1). Each component has its own role and responsibilities in process of intrusion detection in computer networks. Components are as follows:

1) Component for data pre-processing
2) Component for forming a training and test data sets

3) Compoment for data attribute selection (SC Voting Mechanism)
4) ANFIS component
5) SOM component

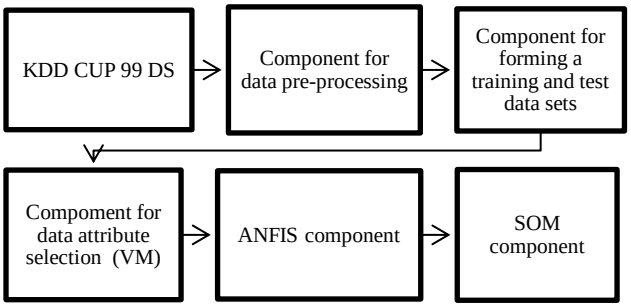


Figure 1. System architecture

Purpose of the component for data pre-processing is to prepare input data (training and test data set) for next component. This activity is performed in a manner to carry out the conversion of the symbolic value into numerical data and it is necessery to perform normalization on it (Fig. 2). This means that it is necessary to perform data pre-processing to obtain a format acceptable to classification algorithms [2]. Symbol attributes, such as the attack type (twenty-three symbols), protocol type (three symbols), service (seventy symbols), and flag (eleven symbols) are expressed as numeric values (whole numbers) with the range of (0-(N-1)), where N is the number of symbols. Each of the symbols thus expressed were scaled linearly into the following range: [0.0,1.0] using following equation:

$$B = A / \max(A) \quad (1)$$

where A is original symbol value, and B is the result of scaling. Then, all attributes with whole number fields, with a narrow value range were scaled linearly into [0,1] interval. It was not necessary to scale the remaining attributes, that are Boolean, or within the [0,1] range. Attributes with the wide vale range (0-1.3 billion) were scaled logarithmically:

$$B = 9.14 \times \log(A) / \log(\max(A)) \quad (2)$$

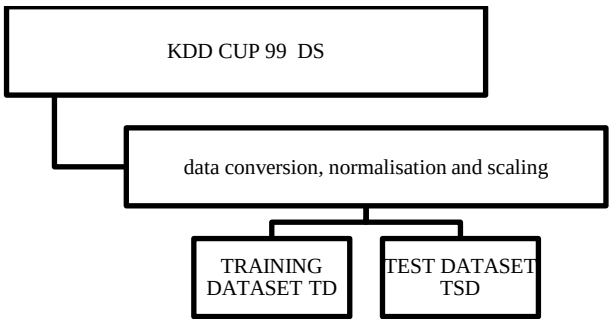


Figure 2. Training and testing data set conversion, normalisation and scalling

The fact that processing of data (the training and test data sets) is time and performance consuming process [4], it was challenge to made component of the system for forming a training and test data sets [5]. The objective of this component is to form smaller groups of training and test data sets (Fig. 3, and 4.), which are used in the construction of the system and later testing. The number and size of these data sets is determined based on the following steps:

- number of rows = size of training and test matrix
- random chosen rows= permutation (number of rows) (3)
- training and test matrix = random chosen rows (1:1000);



Figure 3. Building training data set group

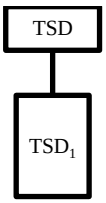


Figure 4. Building test data set group

Component for data attribute selection for each attribute of the input training data sets using the SC determines whether the attribute should participate in building the ANFIS module. The criterion according to which selection of attribute is done is the number of clusters. If the number of clusters for the attributes in the input training data is greater than one, then that attribute will participate in the construction of the module (Fig. 5). Otherwise, the attribute is not taken into consideration.

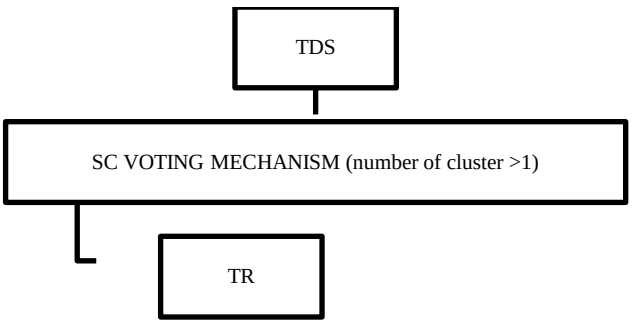


Figure 5. SC voting mechanism

Thus, for the selected training data sets attributes verifiyng is done using SC Voting Mechanism, according to previously described condition.

The result of these activities is the creation of a training matrix TR_1, TR_2, .., TR_n with reduced number of attributes. In other words the dimensionality of the input data is reduced (Fig. 6).

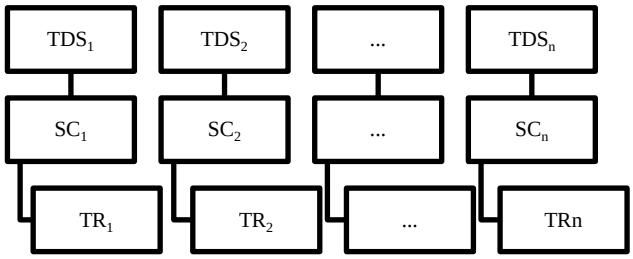


Figure 6. Training matrix dimensionality reduction

ANFIS modules are then trained with training groups of data, so that AN_1 modul is trained with TDS_1 training data group which is processed with SC_1 (TR_1), AN_1 modul with training data group TDS_2 which is processed with SC_2 (TR_n), and all remaining ANFIS modules including AN_n (TRN) (Fig. 7).

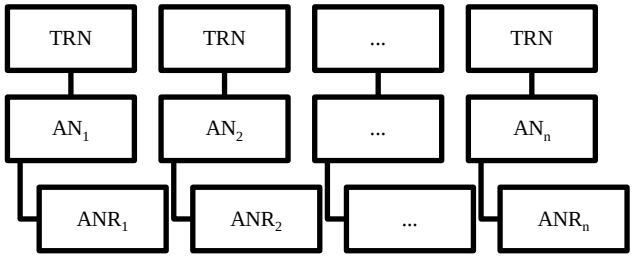


Figure 7. ANFIS module evaluation with new training data set

As input for prepared ANFIS modules (AN_1, AN_2,..., AN_n) we introduce second training data group (matrix TRN). Attributes that are marked in previous step are selected for every module from this new training group (i.e. those attributes that were involved in ANFIS module building The results obtained in this phase ANR_1, ANR_2, ..., AN_n are the input SOM component of system (Fig. 8).

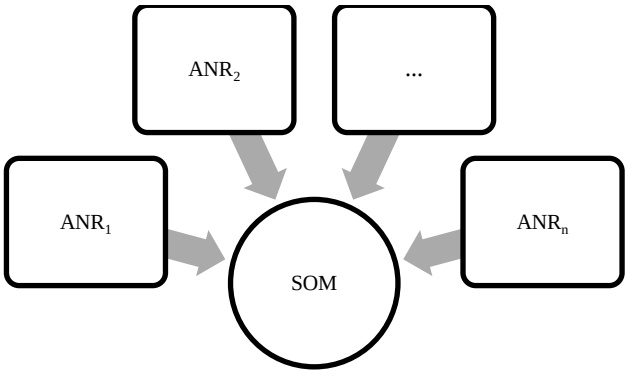


Figure 8. SOM training

SOM component of the system is then trained with these data by building a network with five neurons.

V. ALGORITHM PROPOSAL

Each of the methods and techniques that were used in the construction of the system has certain advantages that are used in the proposed algorithm [1]. Thus, the SC is used in the process of voting to reduce the dimensionality of the training and test data. ANFIS is used to build the module because the expected output for the training data is already known. And because of the fact that the random sample is taken from a small percentage of the training data of SOM is used for clustering the results obtained simultaneously by ANFIS modules and final classification of test data set.

The proposed algorithm is shown below:

1) Loading and processing KDD CUP 99 dataset training and test data in the training *TRD* matrix and test matrix *TSD* respectively

2) The formation of input training matrix TDS_k ($k=1,..,3$) with dimension [m,n] that contains data generated from training data where $m=1000$ and $n=42$. Matrix are formed as a random selection of rows from the training matrix *TRD*.

3) Applying SC on all attributes ($A_1, A_2, ..., A_{n-1}$) for each of training matrix t_k and determination of number of clusters for all matrix attributes. If the number of clusters is equal to one, this attribute is not taken into consideration when building the ANFIS module AN_k. That is how new training matrix TR_k are formed .

4) ANFIS modul AN_k training using new training matrix TR_k .

5) The formation of new *TRN* training matrix with dimension [a,b] that contains data generated form training data set where $a=1000$ and $b=41$.

6) ANFIS modul evaluation AN_k taking into consideration those matrix attributes positions TRN_k that where participated in the making of these modules respectively

7) The formation TSD_1 test matrix with dimension [c,d] that contains randomly generated data from test matrix *TSD* where $c=1000$ and $r=41$.

8) TSD_1 evaluation using the previosly described model. In other words, using ANFIS modul (for attributes defined in step 3.) and using results from ANR_k as input for SOM modul.

VI. EXPERIMENT RESULTS AND DISCUSSION

For the purpose of performing the experiments Matlab software was used. All simulations were performed on HP Compaq 6710b with Intel Core2 Duo T7100 1,8 GHz and 2048 MB RAM. During training and testing main parameters values were: for SC a range of influence of 0,5 is specified for all data dimensions, for ANFIS module SC is used for generating FIS and for SOM module five neurons for network size were used.

The following training matrices are generated: TDS_1, TDS_2, TDS_3 i TRN. All these matrix are generated from the training data set. Distribution of data in a matrices is as shown in Table 2.:

TABLE 2. AN OVERVIEW OF TRAINING DATA DISTRIBUTION

-	normal	Probe	DoS	U2R	R2L	Total
TDS_1	750	10	239	0	1	1000
TDS_2	771	15	211	0	3	1000
TDS_3	793	14	192	0	1	1000
TRN	765	6	228	0	1	1000

Applying SC Voting Mechanism on TDS_1, TDS_2, TDS_3 matrix attributes that satisfy the criterion (that is the number of clusters is larger than one) are separeted. Using these attributes new training matrix TR_1, TR_2, TR_3 are generated.

These matrices had significantly reduced dimensionality, so that the new training matrix, TR_1, TR_2 were formed with thirteen attributes (instead of the previous forty-one), while the matrix TR_3 is formed with twelve attributes.

The following table (Table 3.) is a representation of the attributes used in the new training matrix.

TABLE 3. ATTRIBUTE LIST FOR TRAINING DATA AFTER SC VOTING

Training matrix	Featured attributes of the matrix training group - TDS_k
TR_1	$A_3, A_4, A_5, A_6, A_{12}, A_{25}, A_{26}, A_{29}, A_{32}, A_{33}, A_{34} A_{38}, A_{39}$
TR_2	$A_3, A_4, A_5, A_6, A_{12}, A_{25}, A_{26}, A_{29}, A_{32}, A_{33}, A_{34} A_{38}, A_{39}$
TR_3	$A_3, A_4, A_5, A_{12}, A_{25}, A_{26}, A_{29}, A_{32}, A_{33}, A_{34} A_{38}, A_{39}$

Matrices TR_1, TR_2, TR_3 are used for training of ANFIS modules i.e. building AN_1, AN_2, AN_3. Next figure (Fig. 9) represent surface for AN_3.

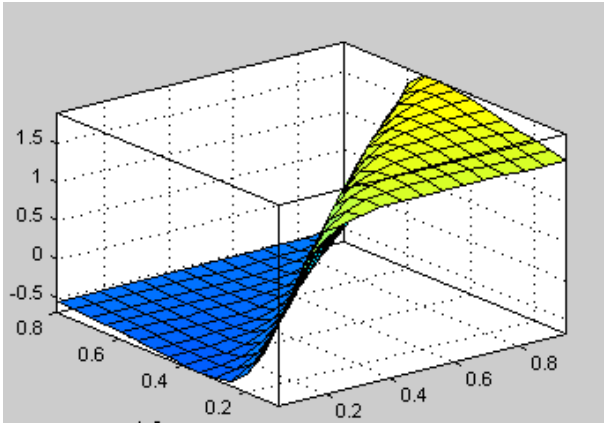


Figure 9. AN_3 surface

After completed training of ANFIS modules as their inputs TRN attributes were brought - new training matrix that is generated from the training data set. Output from this modules were used as inputs for training SOM network with five neurons.

This trained network is then used to classify the data from the test data group. Test data group TSD_1 s then brought as input data for the system, in order to carry out the evaluation. Data classification result for TSD_1 that are output from the last component of system (SOM) is graphically represented on next figure (Fig. 10).



Figure 10. SOM data classification result

The following table (Table 4.) shows the results that are obtained together with the expected results of the classification in order to perform a comparison of the results.

TABLE 4. EXPERIMENT RESULTS

	normal	Probe	DoS	U2R	R2L	1000
SC, ANFIS, SOM	859	48	91	0	2	1000
TSD_1	924	18	56	0	2	1000

The results shown that the portion of traffic that is actually classified as a normal is classified as Probe and DoS. This generates false alarms. One reason for this is the fact that volume of sample data used for training and testing is not so big and traffic is not distributed equally. On the other side, experiment shown the possibilities of the proposed algorithm on a smaller sample of data.

The comparation to other research (Table 5.) shown that proposed algorithm can give acceptable level of the intrusion detection accuracy even using this small data sample.

TABLE 5. EXPERIMENT RESULTS COMPARATION TO OTHER RESEARCH

	normal	Probe	DoS	U2R	R2L
GAUSSIAN MIXTURE [4]	98,97	93,03	88,24	22,80	9,60
RBF IDS [4]	99,07	91,31	75,10	7,01	5,6
SOM [4]	93,98	64,30	96,10	21,49	11,70
BINARY TREE [4]	96,43	77,94	96,45	13,59	0,44
ART [4]	97,19	98,48	97,09	17,98	11,29
LAMSTAR IDS [4]	99,69	98,48	99,21	28,94	41,20
MULTINOMIAL LOGISTIC MODEL [3]	98,30	85,60	97,20	25,90	11,20
EDSOM [6]	95,20	92,30	96,20	38,40	43,60
SC, ANFIS, SOM	92,66	100,00	100,00	-	100

VII. CONCLUSION

This paper demonstrates the effects of combined application SOM, ANFIS, SC and Voting Mechanism in detecting intrusions in computer networks. This was done by comparing the results of execution of the ANFIS, SC and SOM implemented in the same software package Matlab on KDD CUP 99 data that were processed using various methods. Pre-processing of KDD CUP 99 dataset included removing duplicated data from training and testing data. This paper proposes algorithm which combines the previously listed methods and techniques, in order to take full advantage of them. Thus, the proposed algorithm has effect of reducing the complexity of the input data by reducing the number of attributes involved in the construction of ANFIS models. with no significant lost on its efficiency. For the selected amount of data, ANFIS module with parallel execution contributes in increasing the speed of execution of this algorithm. The presented algorithm will be modified in the further research by adding one more parameter. It will be the distance between clusters (in the case when the number of clusters is larger than one). In addition, the algorithm will be further tested with higher number of training samples taking the advantage of parallel execution. The proposed algorithm result in this paper shows that combining different methods (SOM, ANFIS, SC and Voting Mechanism) can give acceptable level of the intrusion detection accuracy.

REFERENCES

[1] Z. Avdagic, Vještačka inteligencija & fuzzy-neuro-genetika, Sarajevo, Grafoart 2003

[2] A. Midzic, *The Significance of Data Pre-processing in Desting and Developing Intrusion Detection Systems, 32nd International Convention on Information and Communication Technology, electronics and microelectronics, Vol 5, Mipro, 2009*

[3] Y. Wang, *Statististical Techniques for Network Security: Modern Statistically-Based Intusion Detection and Protection*, Information Science Reference - IGI Global, 2009

[4] V. Venkatachalam S.Selvan, *Intrusion Detection using an Improved Competitive Learning Lamstar Neural Network*, International Journal of Computer Science and Network Security, Vol.7, No.2, 2007

[5] V. Venkatachalam S.Selvan, *An Approach for Reducing the Computational Complexity of LAMSTAR Intrusion Detection Using Pricincipal Component Analysis*, International Journal of Computer Science and Network Security, Vol.2, No.1, 2006

[6] Li Feng, Li-Quan Sun, *Embeded Dynamic Self-organizing Maps For Data Cluster*, International Technology Journal 12(2) Asian Network for Scientifiic Information, 2013

[7] Esteban J. Palomo, Enrique Domínguez, Rafael M. Luque and Jose Munoz, *An Intrusion Detection System based on Hierarchical Self-Organization*, Journal of Information Assurance and Security, 2009

Customer and User education as a tool to increase security level

Ivan Sedinić, Zrinka Lovrić, Tamara Perušić
Croatian Telecom, Cyber & Data Security Work Unit
ivan.sedinic@t.ht.hr, zrinka.lovric@t.ht.hr, tamara.perusic@t.ht.hr

Abstract - In activities to assure security of IT/NT systems and Data, where there is increasing number of attacks and where are more and more sophisticated cyber criminals, with all modern technical protection solutions in place, people are still the weakest link in the chain. And unless policy, education and training, awareness, and technology are properly employed to prevent people from accidentally or intentionally damaging or losing information, they will remain the weakest link. In case of telecom operators this are not only company employees but also employees of vendors and partners, as well as customers. To improve such situation, education is a cornerstone of activities. In this paper will be discussed different ways how to educate each of mentioned groups, what prerequisites for such education are and how technology could help in educating people. Additionally, based on examples from telecom operator, results of such approach will be shown.

Introduction

"Security policy is one thing but reality is another." These words are used very often by employees in organizations worldwide to describe everyday situation. Such statement indicates that in the certain organization a culture of security is underdeveloped. Security culture cannot be created or installed as software, but intentionally shaped and directed through mindset of people who are part of an organization and those who cooperate with organization: customers, vendors, partners, other stakeholders and the society at large. That kind of culture must be "implanted", nurtured and accepted gradually with changes in mindset, the outlook and attitudes of the people. Changes could be led by company management but cannot be imposed on people. Anyhow, senior management is the key component and the vital force for a successful implementation of an information security program [5]. Ways to achieve such changes are education, training, awareness campaigns and communication about security in general. With those activities different type of security cultures, in sense how strict it will be, could be build. There are some organizations as national intelligence agencies or banks which require strict security culture and other where alignment with the business as a whole will require less strict security culture [7]. According to ISO 27001:2013 the following control objectives are crucial [6]:

A5: Information security policies

A6: Organization of information security

A7: Human resource security

A8: Asset management

A9: Access control

A10: Cryptography

A11: Physical and environmental security

A12: Operations security

A13: Communications security

A14: System acquisition, development and maintenance

A15: Supplier relationships

A16: Information security incident management

A17: Information security aspects of business continuity management

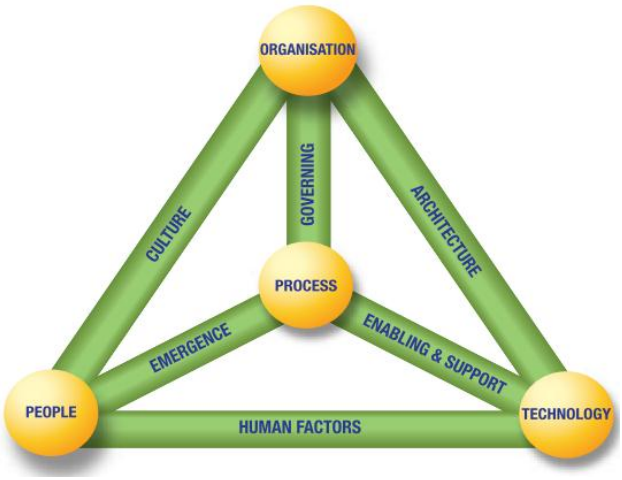
A18: Compliance

In this article focus will be on the control "A.7.2.2. Information security awareness, education and training" described with control "All employees of organization, and where relevant, contractors shall receive appropriate awareness education and training and regular updates in organizational policies and procedures, as relevant for their job function" [7]. This paper is based on the experience in Telco industry and customers are also included since they are integral part of information security.

Position of education in security

When we are talking about security in this article we have in mind certain business model for information security as shown in figure 1 [1].

Figure 1. Business Model for Information Security



Source: ISACA, The Business Model for Information Security, USA 2010

In a model presented above, it is visible that people are one of model cornerstones. When we have in mind the nature of human being, it is also clear that this cornerstone needs additional set of tools to be able to perform its designed role. Those tools could be obtained through education, training, awareness and motivation that are regular part of Information Security Process Model as it is shown in Figure 2. In this model, mentioned activities are included in group of "Run Processes", subgroup: "Interaction Processes", activities: "Communications and Relationship Management" [2].

Figure 2. Gartner Information Security Process Portfolio Model



Source: Gartner, Creating a Security Process Catalog, USA 2013

Communications and relationship management — The activities that ensure everyone knows what is allowed and what is forbidden under current security policies, and why. It includes reporting on current state, performance and proposed changes regarding security (organization, policy, processes, technology, and so on) to all relevant parties. The potential target audience includes executives, business unit management, IT staff, end users, contractors, consultants, service providers, vendors and customers. Specific subprocesses include [2]:

- Security awareness marketing, culture building and training
- Security metrics and measurement
- Security service-level measurement and reporting
- Relationship management with external service providers.

Security education is everything we do to enable people to successfully carry out their roles in security processes and it happens all the time, in all sorts of settings and ways. We also need to have in mind that education and awareness is not the same thing. People could be aware of something but may not be enough educated to take proper action. Education is this foundation on which people can plan and execute most proper action in certain circumstances.

Definition of education groups

Technical controls are very important to the information security program, and are necessary to address the aspects such as antivirus, encryption, firewalls, security mechanisms, physical security, authentication, and monitoring, but given that technical controls cannot fully address the end-user behaviors, security awareness training must be in place to reduce the risk [4].

Without organization's investment in personnel and personal involvement and commitment by people, an organization's information security management capability will be very limited in its ability to provide proper support to business goals.

Information security in general is not only the responsibility of security practitioners but of all employees. Security education is also responsibility of each individual involved in process which needs to be secured. To ensure a proper influence of security education, it needs to be tailored to specific need of certain people group and overall security program. Additionally, education needs to be continuous. Education frequency needs to be developed based on organization's areas of concern, having in mind potential threats and vulnerabilities, local, regional and worldwide news with impact on organization security, and employee's personal duties and responsibilities with impact on security. In general, security education could be grouped in [3]:

- Initial: minimal education for everyone included
- Periodical/Refresher: renewal of initial education with new content and additional topics
- Specific: extended education on specific topics needed just for certain group of people.

Methods to deliver security education could vary from standard to innovative: formal trainings, briefings, conferences, videos, e-learning, bulletins, desktop instructional materials, posters, flyers, etc. Relationship between quality and cost for different education methods is shown in figure 3, based on experience in Telco organization. Quality of one particular method is defined in comparison with other methods and possible outcomes are: High, Medium and Low. Cost parameter was defined in the same way.

Figure 3. Different education methods

Method	Quality	Cost
Training	High	High
Briefing	Medium	Medium
Conference	Medium	High
Video	Medium	Low
E-learning	High	Low
Web Portal	Medium	Low
E-mail	Low	Low
Bulletin	Low	Medium
Instructional materials	Low	Medium
Posters	Low	Medium
Flyers	Low	Medium

Each security education is supposed to do the following [3]:

- Advise: advise a personnel of effect that can result from unauthorized data disclosure
- Indoctrinate: indoctrinate personnel in the ways for protection of information
- Familiarize: familiarize personnel with procedure to ensure that security implications are considered before decision is taken
- Inform: inform personnel about techniques to exploit security weakness, obligation to report such attempts and consequences of violation security regulation
- Instruct: instruct personnel about obligations of data owners and data users.

During education we need to be careful to avoid trap of using fear, uncertainty and doubt as methodology to reach our goals. While this could give us short-term gains, the long-term outcome typically results in security personnel being avoided and mistrusted. Instead, other techniques should be used. One of them is presenting threat and vulnerability analysis. This allows the intended audience to have a better understanding of why a concern exists and possibility to form their own opinion on problem.

Based on all mentioned before, there are several groups of people for which education tailored according to their specific needs should be provided:

- Managers
- Security practitioners
- IT/NT experts
- Partners
- Regular employees
- Contracted workers
- Customers

For each of these groups messages in education need to be prepared in different format as shown in figure 4.

Figure 4. Education format for different groups

Group	Format
Managers	Brief, to-the-point, strategic, graphic
Security practitioners	Detailed, strategic and tactical, benefits
IT/NT experts	Thorough, tactical, showing benefits
Partners	Thorough, showing benefits
Regular employees	Graphic, explanatory, showing benefits
Contracted workers	Graphic, explanatory, obligations
Customers	Graphic, personal, little explanation

EXECUTION OF EDUCATION

First point when starting execution of education is to define areas which will be covered. These are:

- Physical and Personal Security
- Technology Security, Data Protection, Classified Information
- Fraud, Incident, Crisis Management, Compliance
- Business Continuity, Risk and Insurance Management

The next step is to recognize appropriate approach in which education can be delivered:

- Web
- E-mail
- Paper
- Human interaction.

The most important thing is to define the content for planned education.

In the next part few examples how education is established and which main topics should be covered in an organization will be shown.

A. Initial education:

Participants: all new employees

- Topics: Security organization and responsibilities, compliance, anticorruption, role of security in processes and services, security policies and requirements, practical advices and basic security rules, social engineering, security incidents, "security by design" process, contact information
- Method: Briefing

Participants: Partners

- Topics: Security policies and requirements, security incidents, contact information

- Method: Instructional materials

Participants: Customers

- Topics: Basic security rules and obligations
- Method: Instructional materials , Flyers

B. Periodical education:

Participants: Managers, Security practitioners, IT/NT experts, Regular employees, Contracted workers

- Topics: Security portal with most important information on security issues (strategy, security organization, policies, requirements, procedures, guidelines, other documentation, key projects, trainings, e-learnings), Security Department's current activities, Basic security principles, Physical security, IT/NT security, Security in contact centers, Social Media security, "Card of the Month" campaign - one security topic is presented every month
- Method: Web Portal, E-learning

Participants: Partners

- Topics: Changes in Security policies and Security Requirements
- Method: E-mail

Participants: Customers

- Topics: Short, easy to understand practical guide for using services in secure way: PC Safety and Basic Protection, Choosing a Safe Password, Online Shopping, Using Smartphones Safely, Using Apps Safely, Leaving No Traces on the Net, Phishing, Social Engineering, The Safety Barometer, Smart Use of Social Networks, Cloud Computing,
- Method: Web Portal

C. Specific education:

Participants: Security practitioners, IT/NT experts,

- Topics: Security incident management and CERT, Privacy and Security by Design
- Method: E-learning, Training

Participants: Managers

- Topics: Security trends, activities and plans
- Method: Briefing

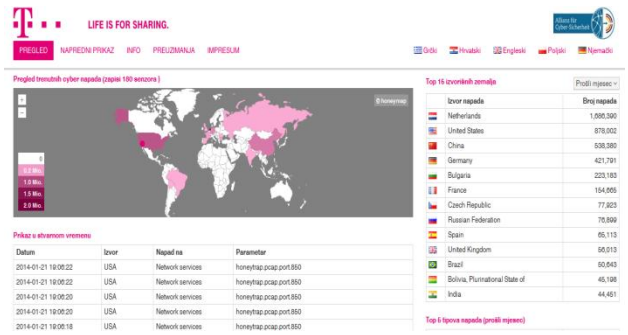
It was recognized that most efficient and cost effective ways for security educations are e-learning and Web Portal, so they are most used for this purpose. Except for education, Security Portal is also used for awareness campaign and publishing of different security information such as shown in following figures:

Figure 5. DT Strategic Threat Radar



Source: DT AG, Data Privacy and Data Security Report 2013, Germany 2014 [8]

Figure 6. DT honeypot map – cyber-attacks on HT



Source: http://www.sicherheitstacho.eu/

Conclusion

Security is all around. It accompanies people from the very first minute every day at work. It is not only about procedure but also about way people act at work. Procedures, application and systems cannot work without people. Every single person is responsible for security.

Everyone is security.

Every person needs to understand and to apply security policies, requirements, procedures and instructions. Everyone should understand role of security in his/hers company. At the end everyone is responsible for risks and security in his area of work. And to be able to take this responsibility, education of each person is crucial.

From table in figure 3 it is clear that most cost efficient education method is e-learning with following advantages:

- Easy access for everyone

- Easy way to verify security knowledge among employees
- Interactive mode of education
- Student could choose time of education according to personal preferences
- Education could be done in few different time slots
- Education could be done from own office or even from home
- No travel and similar expenses.

To make e-learning even more efficient it is recommended to apply the following rules:

- Don't use long topics, divide one big in few smaller
- Keep descriptions short
- Describe it in interesting way using real life examples
- Be graphical
- Make knowledge verification at the end and provide explanation for answers
- Provide links to more detailed explanations for those interested in

With proper education of all employees in organization it is possible to reach status that security is integral part of product development process and is built in "by design" in products and services. In case which is subject of this paper, security organization was involved in early development phase of 100% of security relevant projects and formal confirmation of security requirements fulfillment was done for 99% of projects. Also, by educating customers, it could be achieved that most of them have perception that security is important or very important. Market research "Security and Data Protection" organized by HT in December 2013 shows that 85% of customers consider data privacy and data security very important or important. And in telecommunication industry with such customers' approach it is easier to improve security of overall telecommunication network and services.

References

[1] ISACA: The Business Model for Information Security, USA 2010

[2] Tom Scholtz: Creating a Security Process Catalog, Gartner, USA 2013

[3] Carl Roper, Joseph Grau and Lynn Fischer: Security Education, Awareness and Training: From Theory to Practice, Butterworth-Heinemann, USA 2006

[4] Fitzgerald, Todd: Information Security Governance Simplified: From the Boardroom to the Keyboard, CRC Press, USA, 2011

[5] Whitman, M.E., Mattord, H.J.: Principles of Information Security, Cengage Learning, USA 2011

[6] International Organization for standardization ISO: ISO/IEC 27001:2013, Switzerland 2013

[7] ISACA: Creating a Culture of Security, USA 2011

[8] DT AG: Data Privacy and Data Security Report 2013, Germany 2014

Izradba zaštitnog profila pomoću ISO 27k kontrola

Z. Kummer*, V. Lebinac**
* Financijska agencija, Zagreb, Hrvatska
** Veleučilište Velika Gorica, Hrvatska

Sažetak - Svaki proizvod prije korištenja treba biti provjeren sa sigurnosnog stajališta i izrađen njegov zaštitni profil. Zaštitni profil se koristi kao dio procesa certiciranja proizvoda. Utvrđivanje sigurnosti proizvoda uključuje provjeru rada samog proizvoda i njegov utjecaj na sustav u kojem radi. Standardno se zaštitni profil izrađuje pomoću Općih kriterija (*Common Criteria*). U radu se za izradbu zaštitnog profila sustava za centralizirano upravljanje korisnicima koristi standard ISO 27k. Sustav za centralizirano upravljanje korisnicima izrađen je s *openssh*, *openssl* paketima uz dodatni softver. ISO 27k standard je izabran radi njegove implementacije u cijelom sustavu.

I. UVOD

Početak rada na računalu započinje predstavljanjem identiteta registriranog korisnika u sustavu. Provjera identiteta vrši se u slijedećem koraku kada korisnik s tajnom (lozinkom) potvrđuje taj identitet. Nakon pozitivne provjere identiteta, korisnik je potvrđen, dodjeljuju mu se ovlasti, aktivira se sjednica i korisnik može započeti rad na sustavu. Dodjela ovlasti vrši se u skladu s implementiranom politikom.

Standardni protokol za rad koji koriste prvenstveno administratori, za sigurnu uspostavu konekcije i sigurni rad na udaljenim sustavima je ssh protokol. Programska podrška koju koristi ssh protokol (korisnički i poslužiteljski dio) postoji praktički na svim platformama i dostupna je kao komercijalna i kao slobodna i jednostavna je za uspostavu i korištenje.

Registracija korisnika na računalu sastoji se od određivanja UID-a (*User Identication Number*) i GID-a (*Group Identification Number*) korisnika. Vrijednosti UID-a i GID-a zapisuju se u sistemske tablice s čime je korisnik poznat sustavu. Na različitim sustavima isti korisnik ne mora imati isti UID i GID. To omogućuje da se na aplikativnoj i globalnoj razini barata imenom korisnika, a na sustavnoj, odnosno lokalnoj razini s UID i GID aktiviranog korisnika.

Korisnički profil uz UID i GID sadrži još minimalno lozinku i ovlasti. Korisnik je uspostavljen na sustavu kada se on može predstaviti tom sustavu, provjeriti se njegovo postojanje, mogu mu se dodijeliti ovlasti, može raditi i može se pratiti njegov rad na tom istom sustavu.

Postojanje više računala u domeni i veći broj korisnika koji mogu raditi na njima, dovodi do problema održavanja korisničkih profila. Korisnik može raditi samo na dijelu računala iz domene i na svakom računalu imati drugačije ovlasti. Radi toga korisnik mora biti prepoznat na svakom računalu pojedinačno i na svakom od njih se po želji može prijaviti na rad i aktivirati. Kao problem ističe se: kako održavati korisničke profile.

Zahtjevi održavanja korisnika trebaju ispuniti da se korisnik samo jednom registrira, a sva ostala manipulacija s korisnikom obavlja se na jednom mjestu i postaje poznata i aktivna na svim sustavima na kojima korisnik želi raditi.

Svi navedeni postupci su kritični u funkcioniranju sustava, jer ako imaju slabosti mogu dovesti do neovlaštenog rada na sustavu. Pri izradbi rješenja, provjera rješenja vrši se usklađivanjem s standardom. Za to se uobičajeno koriste opći kriteriji (Common criteria) [1]. U radu će biti prikazano kako se za usklađivanje može koristiti ISO 27k standard [2].

Pri projektiranju sustava kreće se od mehanizama koji implementiraju kontrole koje zahtjeva politika (od dolje prema gore). Kontrola može zahtijevati neki mehanizam, ali ga ne definira dovoljno određeno što znači da svaki mehanizam treba biti usklađen s kontrolom ili kontrolama i imovinom na koju se odnosi. Tek tada je sustav usklađen s standardom čije smo kontrole odbrali. Nakon toga možemo odrediti (napisati) politiku [3].

Zaštita je proces koji treba osigurati sigurnost, odnosno stanje u kojem je sustav siguran [4].

Rad je strukturiran kako slijedi. U poglavlju II opisan je motiv za izradbu programskog rješenja za centralizirano upravljanje korisnicima koji koriste za rad ssh protokol. Poglavlje III opisuje poteškoće koje mogu nastati u radu distribuiranog sustava radi lokalne uspostave korisnika na svakom od sustava gdje korisnik želi raditi. Poglavlje IV opisuje korisnički račun kao središnji objekt potreban za I&A i uspostavu korisnika na računalu. Poglavlje V prikazuje konceptualni model rješenja. Poglavlje VI prikazuje pretpostavljene slabosti sustava s posljedicama. Poglavlje VII pokazuje korelaciju kontrola ISO27K standarda i potrebnih i ostvarenih mehanizama koji podupiru navedena kontrole.

II. MOTIV I DISTRIBUIRANI SUSTAVI

Najvažnije stanje (postupka) u procesu rada korisnika na sustavu je autentikacija, odnosno provjera identiteta. Za autentikaciju se može koristiti lozinka koja je lokalno definirana na svakom sustavu pojedinačno. Motiv za stvaranje sustava za centralizirano upravljanje korisnicima je mogućnost jednostavnog i sigurnog održavanja korisnika. Praktički se postupak svodi na održavanje identiteta i autentifikatora korisnika. Dok je identitet lokalno definiran, autentifikator je definiran globalno za cijelu domenu.

Korisnici koriste za rad na računalu ssh protokol koji omogućuje da korisnik radi u tekstualnom i grafičkom načinu rada.

Centralno upravljanje korisnicima (CUM - *Centalized User Management*) mora biti usklađeno sa sustavom za upravljanje informacijskom sigurnošću (ISMS - *Information Security Management System*). ISMS možemo definirati kao[5]:

- Skup politika i procedura za sustavno upravljanje osjetljivim podacima.
- Skup politika, procesa i sustava koje organizacija oblikuje, izrađuje, implementira i održava.

Distribuirani sustavi sastoje se od kooperativnih računala na kojima su registrirani korisnici. Korisnici se moraju prijaviti za rad na svakom od računala gdje žele obaviti neki posao i koristiti računalnu opremu. U standardnoj distribuciji poslužitelja, korisnik se uvijek registrira lokalno i po prijavi je poznat samo na tom računalu. Korisnik se po želji može prijaviti na određeni broj računala.

Poteškoće nastupaju kada imamo više računala jer održavanje korisnika i sinkronizacija stanja u kojim se oni nalaze na svakom od računala postaje složena. Poteškoće i zahtjevi u radu takovog sustava su slijedeće:

- Postoji veliki broja nepovezanih poslužitelja koje koristi isto tako veliki broj korisnika koji ne moraju biti svi registrirani na svakom računalu.
- Računala održava relativno mali broj administratora.
- Administratori su privilegirani korisnici čija prijava se mora nadzirati, pratiti i analizirati.
- Administrator mora biti uspostavljen na svakom računalu pojedinačno radi dodjele ovlasti korisnika i održavanja poslužitelja.
- Administratori koriste sekundarne komunikacijske kanale, odnosno komunikacijske kanale koje ne koriste regularni ili produkcijski korisnici.

- Uspostava administratora na velikom broju računala je nespretna, a u slučaju distribuiranih sustava jako otežana.
- Postupak autentikacije mora biti sto je moguće jednostavniji i standardiziran.
- Postupak autentikacije mora biti pod kontrolom vlasnika sustava. Razina kontrole podrazumijeva da vlasnik ima potpuni uvid u izvorni kod rješenja iz kojeg se može stvoriti samostalno produkcijsko rješenje.
- Vlasnik sustava mora po potrebi moći provjeriti rješenje.
- Nepovjerenje u zatvorena rješenje radi sumnje u postojanje „stražnjih vrata“, prikrivenih kanala, mogućnosti komunikacije s nepoznatim odredištima i postojanja „*snake oil*“ kripto zaštitnih algoritama.
- Potreba za robusnim postupcima provjere identiteta koji koriste isto tako jake zaštitne mehanizme.
- Potreba usklađenosti rješenja s standardima za sigurnost.
- Korištenje i održavanje rješenja mora biti sto je moguće jednostavnije.
- Rješenje se može oblikovati, izraditi i uspostaviti s pristupačnim alatima, po mogućnosti „otvorenim“.
- Rješenje ne treba zahtijevati posebno specijalističko obrazovanje, dovoljna su standardna znanja.
- Rješenje se mora moći jednostavno i efikasno koristiti i održavati.

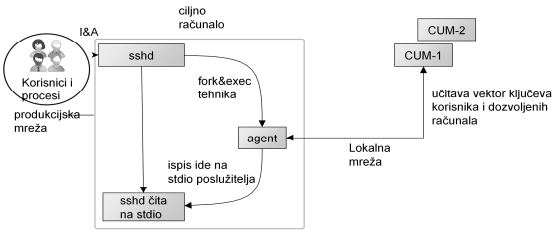
III. KORISNIČKI RAČUN

Osnovni objekt koji se koristi u fazi identifikacije i autentikacije je korisnički račun čiji sastavni dio je korisnički profil. Kreiranje računa je upis korisnika u sustavne tablice. Pri tome se korisniku daje obično sustavno određen UID i GID. Nakon toga, korisnik je poznat sustavu. Račun može biti u slijedećim stanjima:

- Registracija korisnika na računalu kreira račun kojem su pridruženi sustavni resursi. Registracija je zapis korisničkog imena koje se koristi u fazi identifikacije. Račun je još onemogućen i ne može se koristiti.
- Uspostava lozinke omogućuje autentikaciju sustavu poznatog identiteta korisnika i aktiviranje računa. Aktiviran račun dodjeljuje ovlasti i uspostavlja sjednicu za taj račun i u korist korisniku koji ga koristi.
- Račun može biti suspendiran. Korisnik je poznat sustavu. Korisnik ne zna lozinku i ne može provesti autentikaciju.
- Račun može biti opozvan. Sustav ne pozna korisnika i ne može se provesti identifikacija.
- Korisnik može sam mijenjati svoju lozinku kada to zaželi.
- Ako se koristi drugi mehanizam i podaci za autentikaciju, promjenu podataka za autentikaciju korisnika, može vršiti samo ovlaštena osoba.

Centralizirano upravljanje modificira gore navedeni postupak na slijedeći način:

- Registracija korisnika na računalu kreira račun i dodjeljuje sustavne resurse. Registracija korisnika je zapis korisničkog imena koje se koristi u fazi identifikacije. Račun je još onemogućen i ne može se koristiti.
- Registracija korisnika u CUM-u. Pri tome se upisuje javni ključ, korisnički identitet i lista računala na koje se korisnik može prijaviti. Korisnik se može prijaviti za rad samo na računala koja su navedena u listi odobrenih računala.
- Račun može biti suspendiran. Korisnik postoji na lokalnom računalu i CUM-u. Javni ključ je modificiran tako da se ne može obaviti postupak uparivanja s privatnim ključem.
- Račun je opozvan. Korisnik postoji na lokalnom računalu, a na CUM-u je obrisan. Javni ključ ne postoji i ne može se obaviti postupak uparivanja s privatnim ključem. Račun se mora ponovno kreirati na CUM-u da bi se mogao koristiti
- Korisnik ne može sam mijenjati zapise u CUM-u, nego to radi ovlaštena osoba



Slika 1. Konceptualni model

Preduvjet: korisnički račun nema lozinku. Korisnik ne može mijenjati lozinku i ne može se uspostaviti novi mehanizam i materijal za autentikaciju osim ovlaštenog.

IV. MODEL I POTEŠKOĆE ZAŠTITE SUSTAVA

Poslovni zahtjevi na sustav mogu se odrediti slijedećim pravilima:

- Korisnik za prijavu za rad na računalu mora obaviti I&A (I&A - identifikacija i autentikacija).
- Korisnik se mora moći prijaviti na svako računalo na kojem je registriran. Za autentikaciju korisnik može koristi lozinku ili neki drugi materijal koji ili zna ili ima. Veću razinu zaštite (povjerljivosti) imamo kada koristimo token na kojem je materijal za autentikaciju. Uobičajeno se koristi token (USB token koji koristi čip kao pametna kartica). Token dozvoljava korištenje i brani uvid u povjerljivi materijal.
- Prijava za rad mora biti kontrolirana i moguća samo registriranim korisnicima.
- Prijava mora biti preventivna kontrola koja brani neovlašten pristup vrijednostima korisnika. Uobičajeno je u ISMS-u koristiti pojam imovina. Ona opisuje vrijednost i njenog vlasnika pa ćemo u značenju imovine koristiti samo njenu vrijednost.
- Ovlasti po prijavi korisnik dobiva u skladu s politikom.

Sustav se sastoji od ssh poslužitelja na ciljnom računalu i korisnika koji se mogu prijaviti na njega radi uspostave sjednice. Svi korisnici trebaju biti prethodno registrirani kako bi se mogla provesti identifikacija koju provodi ssh poslužitelj. Za provedbu autentikacije, ssh poslužitelj koristi lokalnog agenta kojeg on pokreće u izvođenje s standardnom sustavnom naredbom. Argument u naredbi je ime korisnika ili ime korisnika u čiju korist se pokreće neki od pomoćnih programa (scp). Agent je standardni program čiji se integritet provjerava prije pokretanja u izvođenje. Tako je zajamčena izvornost programa i izbjegnuto podmetanje zlćudnog koda. Ovlast za pokretanje u izvođenje agenta ima samo server i onemogućeno je neovlašteno pokretanje agenta od strane bilo kojeg korisnika ili drugog procesa.

Za dohvat javnih ključeva može se koristiti bilo koji protokol za dohvat tereta. U primjeru se javni ključevi dohvaćaju pomoću http protokola. Zapis s javnim ključevima ima dodan naziv korisnika kojem pripada javni ključ i vektor hostova na koje se korisnik prijaviti. Na svakom od tih hostova korisnik može prethodno biti registriran. Moguća je anonimizacija da ime korisnika ne bude ime realnog korisnika nego neko od generičkih imena (admin0, admin1, ...) koja se dodjeljuju realnom korisniku i tako simuliraju uloge. Registracija inicijalnih korisnika može se obaviti u fazi instalacije sustava. Svaki korisnik ne smije imati lozinku i ne smije koristiti program za upravljanje lozinkama i tako neovlašteno uspostaviti lozinku. Za prijavu korisnika ne koristi se lozinka nego digitalni certfikat koji je spremljen na tokenu. Pristup tokenu i aktiviranje/korištenje certifikata je lokalno zaštićeno, odnosno omogućeno s PIN-om (PIN - *Personal Identification Number*).

U slučaju korištenja pomoćnih programa sustav javno/privatnih ključeva je također razdvojen tako da korisnik ima samo jedan od ključeva dok je drugi zaštićen i nedostupan. To je osigurano tako da ssh server dozvoljava pristup korisniku i provodi I&A sam. Materijal za I&A server dobiva od korisnika i agenta, odnosno iz dva razdvojena

izvora s čime je osigurana separacija. Poslužitelj (ssh demon) na obradu dobiva samo javni ključ korisnika koji treba obaviti I&A, dok su mu javni ključevi ostalih korisnika nedostupni.

Svaki IT sustav treba biti dostupan kada treba, pouzdan i siguran (vrijedan povjerenja)[6]. To se može osigurati preko tri osnovna načela koja sustav treba ispuniti kao siguran sustav. On mora biti povjerljiv, cjelovit i dostupan. Cjelovitost se mora proširiti uz teret i na način manipulacije. Tada je definicija cjelovitosti uz ovlaštena promjenu vrijednosti proširena i na ovlašteni način.

Slabosti su sastavni dio IT sustava. Njih napadač koristi kao ranjivosti za vrijeme napada i tako nanosi štetu vrijednostima sustava[7]. Narušavanje povjerljivosti, cjelovitosti i dostupnosti odnosi se i na prijetnje koje se žele otkloniti. Standardne prijetnje koje narušavaju osnovna načela sigurnosti su:

- Neovlašteni uvid što narušava povjerljivost.
- Prekidanje odvijanja, što narušava dostupnost i integritet. Ako se prekine neka transakcije može doći do nesigurnog i neovlaštenog odvijanja transakcije. Isto se događa ako se obriše neki objekt sustava s čime se narušava dostupnost i integritet sustava.
- Promjena postojećeg ili podmetanje novog sadržaja što narušava najprije integritet i nakon toga povjerljivost, jer imamo podmetnutu neistinu kojoj vjerujemo sa svim posljedicama toga. Posebno je opasno ako je cilj napada dnevnik sustava.

Da bi se otklonile slabosti a s time i ranjivosti sustava mora se zadovoljiti slijedeće pravilo:

„Sustav je zaštićen i siguran samo ako su inicijalna stanje sigurna i prijelaz iz jednog stanja u drugo stanje siguran, s time da je stanje u kojem se nalazi sustav sigurno i stanju u koje prelazi sustav isti tako sigurno. U početnom stanju sustav može biti samo jednom."

Da bi osigurali sustav trebamo definirati zaštitnu politiku i mehanizme koji politiku podupiru. Politika se može opisati kao niz zahtjeva, a mehanizmi kao ostvaritelji i osiguravatelji zahtjeva. Zaštitni profil pokazuje da li sustav omogućuje zaštitnu politiku s podupirućim mehanizmima.

Bitni zaštitni mehanizmi za CUM su:

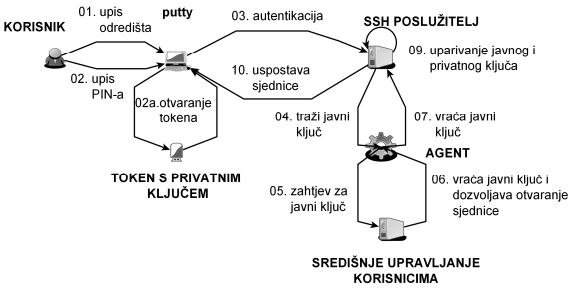
- Enkripcija što osigurava povjerljivosti i pomaže integritet sustava.
- Identifikacija i autentikacija što osigurava inicijalno sigurno stanje.
- Nadzor i praćenje što osigurava brzu detekcija eventualnog nesigurnog stanja i minimiziranje štete.

Prijetnje sustavu za središnje upravljanje korisnicima dane su u slijedećoj tablici 1.

TABLICA 1. VRIJEDNOSTI SUSTAVA I POSLJEDICE

Vrijednosti sustava	Posljedice narušavanja zaštite
Kompromitiranje vrijednosti	Neovlašteno narušavanje povjerljivosti i integriteta vrijednosti. Posljedica napada ili nestabilnosti sustava. Ništa nije detektirano, šteta je učinjena.
Zaobilaženje sigurnosne funkcije	Mogu se prevariti čvorovi (računala) i ostvariti I&A, prevariti komunikacijski protokoli i podmetnuti lažni teret (informacijski podaci).
Dohvat sistemskih tablica	Može se dohvatiti podatke koje koristi CUM i utjecati na njihov integritet, povjerljivost i zloupotrijebiti ih.
Loše priređen token	Priprema tokena može biti loša, tako da se može zloupotrijebiti KZ funkcija tokena i par ključeva koji su isto tako loši.
Fizički prodor	Neovlašteni korisnik može fizički doći do poslužitelja, CUM-a i promijeniti pravila za I&A.
Lažno predstavljanje	Neovlašteni korisnik može se lažno predstaviti kao regularni korisnik.
Upis u sistemske tablice	Može se zloupotrijebiti komunikacijski kanal za održavanje sustava ili produkciju, upisati neovlaštene informacije koje dozvoljavaju narušavanje sheme I&A, registrirati korisnika na poslužitelju i CUM-u, importirati ključ korisnika na CUM, importirati podatke koji narušavaju funkciju I&A i aktiviraju nestabilno/neodređeno) stanje i dozvoljavaju neovlašteni I&A.
Nesigurno stanje sustava	Poslužitelj, agent i CUM mogu doći u nesigurno stanje i producirati krivi rezultat ili otkazati rad.
Presretanje podataka	Neovlašteni korisnik može presresti povjerljivu komunikaciju između poslužitelja i CUM-a.
Loše (zloćudno) funkcioniranje	Agent i poslužitelj mogu biti modificirani i pretvoreni u zloćudni kod.
Promjena sistemskih podataka	Može se modificirati javni ključ i prouzročiti postavljanje poslužitelja u neodređeno stanje koje može prouzročiti štetu.
Neovlašteno kreiranje korisnika	Može se kreirati neovlašteni korisnik s nepoznatim ovlastima što neminovno omogućuje napad/štetu.
Otkrivanje informacija potrebnih za I&A	Neovlašteni korisnik može dostupiti do poslužitelja tako da ga napadne i dobije podatke potrebne za I&A.
Opovrgavanje	Korisnik može opovrgnuti I&A i maskirati praćenje, nadzor i pregled.
Rezidualni sistemski podaci	Sve informacije koje se koriste za I&A su u stanju potrebe zaštićene od neovlaštenog dostupa i sačuvan je njihov integritet i povjerljivost. Odmah po korištenju su sve informacije obrisane i s time nedostupne.
Neovlašteni dostup do agenta	Neovlašteni dostup do ssh-agenta i zapisa koje koristi CUM. CUM kao SPOF mora biti uvijek dostupan. (Dostupnost CUM-a osigurava se dovoljnim kapacitetom i prostornom raspodjelom - više CUM-ova koji se raspodijeljeni u prostoru).
Lažno predstavljanje	Može se poslati krivi (lažni) javni ključ koji se prihvaća kao pravovaljani. Zloćudni kôd može preko komunikacije neovlašteno doći do vrijednosti (objekata subjekata) i narušiti CIA cijelog sustav. Prijava neovlaštenog korisnika čiji identitet je lažan (sustav ga prepoznaje, ali identitet koji je prijavljen ne pripada realnom korisniku).

V. REALIZIRANI MODEL



Slika 2. Ostvareni model

Opća pravila koja se ne mijenjaju u vremenu i prostoru:

- Sustav radi kao transakcijski sustav.
- Svi podaci u prijenosu, obradi i pohrani su zaštićeni i osigurana je povjerljivost, cjelovitost i dostupnost.

- Krajnji korisnik može jedini za prijavu na ssh poslužitelj koristiti token na kojem je privatni ključ.
- ssh poslužitelj uparuje korisnikov javni i privatni ključ za provjeru izvornosti identiteta korisnika.
- Korisnik mora biti registriran na svakom poslužitelju na koji se želi prijaviti za rad.
- Korisnik mora biti registriran na CUM-u.
- Registraciju korisnika vrši ovlaštena osoba.
- UID i GID korisnika određuje sustav pri registraciji korisnika.
- Postupke instalacije i održavanja softvera (ssh poslužitelj, ssh-agent, CUM) vrši ovlaštena osoba.
- Instalacija, održavanje i korištenje softvera pripada ulogama koje mogu izvršiti samo djelomične postupke (onaj tko može registrirati i aktivirati korisnika ne smije biti korisnik koji koristi sustav radi separacije dužnosti).
- Osjetljivim materijalom (naziv korisnika, javni ključ) rukuju samo ovlaštene osobe.
- Ovlasti koje dobiva korisnik po prijavi određene su Politikom ovlasti koja se aktivira nakon obavljene I&A i nije predmet ovog rada.

VI. USKLAĐENOST S ISO 27K STANDARDOM

Usklađenost s ISO 27k treba osigurati da zaštita djeluje s mehanizmima na vrijednosti sustava koje se sastoje od objekata i subjekata. Objekti, subjekti i pripadajuće ISO 27k kontrole su navedeni su slijedećim tablicama:

TABLICA 2. OBJEKTI (VRIJEDNOSTI) SUSTAVA

Objekt sustava	Funkcije i namjena
PIN	Identifikator koji pozna samo korisnik. PIN je ekvivalent lozinke. Utječe na povjerljivost i dostupnost.
Identitet korisnika	Identitet korisnika zabilježen je u identifikatoru korisnika na svakom računalu gdje se korisnik želi prijaviti za rad. Utječe na integritet.
Privatni ključ	Privatni ključ uparuje se s javnim ključem. Privatni ključ nalazi se na tokenu i nikada nije izravno vidljiv. Utječe na integritet i dostupnost.
Javni ključ	Javni ključ služi za uparivanje s privatnim ključem. Javni ključ se nalazi na CUM-u. Služi za uspostavu sjednice i aktiviranje ovlasti korisnika. Utječe na integritet i dostupnost.
Lista računala gdje korisnik smije raditi	Lista računala povezana je s javnim ključem korisnika. U slučaju podudarnosti poslužitelj uspostavlja sjednicu s korisnikom. Utječe na dostupnost i integritet.

TABLICA 3. MEHANIZMI

Cjeline i kontrole	Funkcije i namjena
Korisnički SW	SW rješenje za prijavu korisnik na poslužitelja. Ima biblioteke koje koristi token.
Token A.11.3.1* A.15.1.2 A.11.5.2	Uređaj na kojem je generiran par javni/privatni ključ i spremljen x.509 certifikat. Služi u postupku identifikacija i autentikacije korisnika pri njegovoj prijavi za rad na računalu. Utječe na povjerljivost i dostupnost. Token je komercijalni produkt koji ima autorski zaštićene pripadajuće biblioteka (pkcs#11) i koje se ne smiju neovlašteno diseminirati i koristiti.
ssh poslužitelj	ssh poslužitelj. Provodi I&A korisnika i korištenjem agenta i vlastitog koda.
Agent A.10.1.3 A.12.2.1 A.12.2.2 A.12.2.3 A.12.2.4	Agent služi za komunikaciju između poslužitelja na koji se spajaju korisnici i CUM-a. Razdvajanje poslužitelja kojeg vidi korisnik i CUM-a kojeg ne vidi korisnik smanjuje moguću štetu na CUM-u Agent kontrolira dozvoljeni format i teret ulazne poruke.
CUM A.10.9.2	U ostvarenom rješenju koristi se http protokol s pripadajućim poslužiteljem.

* *Oznake A.xx.xx.xx odnose se na kontrole 27k standarda*

TABLICA 4. KORISNICI I POSLOVI

Korisnik i kontrole	Sto radi
Krajnji korisnik	Koristi token za I&A (prijavu) na poslužitelj.
Administrator ssh poslužitelja A.11.4.4	Administrira ssh poslužitelj. Registrira korisnika na poslužitelju. Onemogućuje njegov račun, tako da se ne može prijaviti ni na jedna drugi način osim s tokenom. Za administriranje poslužitelja koristi se poseban povjerljivi put s pripadajućim portom. Administratorski i produkcijski komunikacijski put su razdvojeni i ne mogu se koristi ni na koji drugi način osim dozvoljenog.
Administrator CUM-a A.11.4.4	Administrira CUM. Registrira korisnika na CUM-u. Koristi povjerljivi put.
Administrator tokena	Generira par javni/privatni ključ i PIN.

TABLICA 5. FIZIČKE PRETPOSTAVKE

Subjekti i kontrole	Uvjeti
CUM, ssh A.9.1.5 A.9.2.1 A.11.3.2	CUM i poslužitelj na koji se prijavljuje korisnik, su fizički odvojeni od korisnika i nalaze se u štićenom prostoru u koji mogu ući samo ovlaštene osobe. Svi poslovi na CUM-u moraju se raditi u štićenom prostoru.

TABLICA 6. PRETPOSTAVKE RADA SUSTAVA

Kontrole	Pretpostavke uvjeta i načina rada
A.10.9.2	Sustav radi kao transakcijski sustav (upućuje upit na koji očekuje odgovor). Izvorište mora sigurno znati da je upit postavljen željenom odredištu, došao do njega i da je pripadajuću odgovor isto tako došao od tog odredišta. Sav teret koji se prenosi od izvorišta do odredišta i obratno mora imati sačuvanu povjerljivost i integritet.
A.8.1.1	U sustavu moraju biti definirane uloge koje se dodjeljuju korisnicima ili procesima koji rade u korist korisnika.
A.10.1.3	Dužnosti/funkcije svake uloge moraju biti razdvojene i ne smiju se preklapati radi mogućnosti uspostave neovlaštene funkcije s čime bi se mogla narušiti povjerljivost, integritet i dostupnost sustava.
A.11.2.2	Svaka uloga mora imati pripadajuće ovlasti potrebne za obavljanje posla.
A.11.4.5	Poslovi za koje su korisnici ovlašteni, određeni su ulogom koju oni obnašaju u sustavu.
A.11.5.4	Sustavni programi koje može koristiti korisnik ovise o ovlasti koja je korisniku dodijeljena. Potrebno je obratiti pozornost na separaciju dužnosti radi onemogućavanja neovlaštenog korištenja sustavnih programa i narušavanja povjerljivosti, integriteta i dostupnosti.
A.12.5.4	Sustav ne smije dozvoliti curenje informacija, odnosno narušavanje povjerljivosti.

TABLICA 7. KRIPTO ZAŠTITA

Kontrole	Uređaj
A.12.3.1 A.12.3.2	Token koristi kriptografske funkcije kod uparivanja ključeva i korištenja privatnog ključa. Kripto zaštita mora biti određena politikom i ne može se proizvoljno koristiti. Manipulacija s ključevima (generiranje, promjena) vrsi ovlaštena osoba u strogo kontroliranim uvjetima i odvojeno od produkcijskog sustava.

TABLICA 8. KOMUNIKACIJA

Kontrole	Uvjeti i način rada
A.10.6.1 A.10.6.2	Identifikacija uređaja koji podupiru komunikaciju mora biti uspostavljena. Usmjeravanje mora biti vjerodostojno i ne smije doći do narušavanja ruta. Svi podaci u prijenosu moraju imati sačuvanu povjerljivosti i cjelovitost. Mrežni servisi moraju imati osiguranu dostupnost i kontrolu ovlaštenosti dostupa i korištenja.

TABLICA 9. PRAĆENJE I BILJEŽENJE

Kontrole	Postupci
A.10.10.6	Sve aktivnosti se bilježe u dnevnike sustava i moraju imati vjerodostojni vremenski atribut (točno vrijeme kada su nastale). Svaka zabilješka mora biti zabilježena u realnom vremenu.
A.15.3.1 A.15.3.2	Sustav sistemskih zabilješki treba biti osiguran od neovlaštenog dostupa, promjene zapisa i korištenja alata za pregled sistemskih dnevnika.

VII. ZAKLJUČAK

Projektiranje zaštite je postupak identičan postupku koji se provodi pri projektiranju sustava. Razlika u pristupu je jako velika. Projektiranje funkcije sustava uzima u obzir da se ništa loše neće dogoditi, dok projektiranje zaštite mora osigurati da se ništa loše ne smije dogoditi i ako se dogodi, da je funkcija sustava osigurana i nenarušena.

Zaštitu, radi njenog utjecaja na funkcioniranje sustava je najpovoljnije početi projektirati istovremeno s oblikovanjem funkcije sustava. Svaka druga faza u razvoju sustava samo otežava implementaciju zaštite. ISO 27k standard je postao praktički opće prihvaćen i usklađenost s njime podrazumijeva da je je sustav zaštićen i s time siguran.

Rad je pokušaj projektiranja zaštite istovremeno s oblikovanjem sustava i pokazao je svoju opravdanost. U radu su prikazani odgovori na prvo pitanje „što treba napraviti“. Detaljniji tehnički dio koji nije predmet rada dao bi odgovore na pitanje „kako treba napraviti“.

Rad je pokazao da je metoda od dna prema vrhu povoljnija od obratne [8][9]. Ako bi se oslonili pri odabiru kontrola samo na zakonske i ostale propise i najbolju praksu rezultat bi bio općenit. Drugi put od razvoja i izradbe rješenja koje mora biti vrlo konkretno, obuhvaća veliku sliku od potreba, mehanizama do usklađivanja. To znaci da je postupak od definiranje sigurnosne politike do tehničkih kontrola nepovoljniji od obrnutog puta. Nakon što je određeno „što treba napraviti“ s „kako treba napraviti“ i međusobno usklađeno moguće je puno jednostavnije napisati sigurnosnu politiku koja je vršni dokument zaštite.

LITERATURA

[1] http://www.commoncriteriaportal.org/files/ccfiles/ccpart1v2.3.pdf, ccpart2v.2.3, ccpart3v2.3.pdf (10.03.2014.)

[2] ISO/IEC 27002:2005 Code of practice for information security controls

[3] R. Focardi and R. Gorrieri (Eds.): FOSAD 2000, LNCS 2171, pp. 137– 196 , 2001. Springer-Verlag Berlin Heidelberg 2001

[4] http://www.ncix.gov/publications/policy/docs/CNSSI_4009.pdf (10.03.2014.)

[5] http://en.wikipedia.org/wiki/Information_security, (20.05.2014.)

[6] http://csrc.nist.gov/publications/nistpubs/800-32/sp800-32.pdf (10.03.2014.)

[7] https://cwe.mitre.org/about/faq.html#A.2, (2.04.2014)

[8] http://cert-mu.gov.mu/English/Documents/Guidelines/2011/ Guideline on Information Security Policy.pdf (10.02.2014.)

[9] Dejun Mu; Wei Hu; Baolei Mao; Bo Ma, A bottom-up approach to verifiable embedded system information flow security, Information Security, IET, issue 1, 2014

Djela protiv intelektualnog vlasništva, kod kojih se kao sredstvo izvršenja krivičnih djela javljaju računari, računarske mreže, kao i njihovi proizvodi

Feriz Gabela * i Adnan Gabela**
Dioničko društvo BH Telecom Sarajevo
Obala Kulina bana broj 8 Sarajevo
Bosna i Hercegovina
Telefoni: 387 – 33 586 508; 387 – 61 507 530 e-mail: feriz.gabela@bhtelecom.ba

Sažetak - Porast upotrebe informacionih tehnologija, odnosno globalne računarske mreže, stvorile su mogućnosti novih oblika krivičnih djela u vezi napada na intelektualno vlasništvo, što je zahtjevalo da zakonodavaci iz mnogih država i međunarodnih organizacija izmjene legislativu kako bi mogli pratiti ove promjene. Unifikacija i efikasna međunarodna saradnja, osnovne su pretpostavke za bolju koordinaciju nadnacionalnih napora za suzbijanje ove vrste krivičnih djela. U radu je dat uporedni pregled sistema zaštite softvera u Sjedinjenim Američkim Državama i državama Evropske unije. Izložen je i pregled propisa koji su na pravnoj snazi u Bosni i Hercegovini, a odnose se na suzbijanje krivičnih djela u vezi napada na intelektualno vlasništvo. Ono što je neophodno u suzbijanju kršenja autorskih i njima sličnih prava i industrijskog vlasništva jeste stabilan pravni sistem sa jakom zakonskom regulativom.

I. UVOD

Razvoj informacionih tehnologija i interneta prouzrokovao je povećanje različitih oblika ugrožavanja intelektualnog vlasništva, odnosno autorskog i srodnih prava i industrijskog vlasništva. Gotovo neograničena dostupnost književnih djela, naučnih radova, muzičkih radova, video radova i drugih autorskih djela u digitalnom obliku uzdrmala je iz temelja tradicionalni koncept autorskog i srodnih prava. On je morao da doživi ograničenja i promjene pod uticajem novih koncepata: copyleft, softvera, licenci otvorenog sadržaja, slobodne razmjene informacija, itd. Bosna i Hercegovina je dio međunarodnog sistema zaštite intelektualnog vlasništva, jer je najveći broj međunarodnih konvencija o pravu intelektualnog vlasništva ratifikovala, te su odredbe tih konvencija postale dio nacionalnog prava. Takve su Bernska konvencija za zaštitu književnih i umjetničkih djela, Konvencija o izdavanju evropskih patenata, Ugovor o trgovinskim aspektima prava intelektualnog vlasništva (TRIPS) i dr.

Računarske mreže su stvorile nove oblike kriminala, koji se vrše na računaru ili kod kojih je računar sredstvo izvršenja. U krivičnopravnoj literaturi za cyber kriminal upotrebljavaju se termini kao što su: zloupotreba računara, računarski kriminal i delikti uz pomoć računara. Tipovi cyber kriminala su brojni, a najčešći su:

- djela protiv intelektualnog vlasništva, imovine i pravnog prometa kod kojih se kao objekat ili sredstvo izvršenja krivičnih djela javljaju računari, računarske mreže, računarski podaci, kao i njihovi proizvodi;
- djela gdje se računari javljaju kao sredstvo izvršenja krivičnih djela, kod zloupotreba platnih kartica na Internetu, zloupotreba u elektronskoj trgovini i bankarstvu;
- neovlašten pristup računaru, računarskoj mreži i elektronskoj obradi podataka.

II. ZLOUPOTREBE INTELEKTUALNOG VLASNIŠTVA

Kršenje autorskih prava je kopiranje i mijenjanje računarskih programa, osim:

- kada to radi sam vlasnik programa, ili
- kada se posjeduje licenca (dozvola) vlasnika prava na program.

Vrste zloupotrebe su slijedeće:

- „*warez*'' kršenje autorskih prava: „*warez*“ je sajt na Internetu koji omogućava ljudima da preuzmu ilegalne kopije. Softver će obično imati svoja „oštećenja“, pa se ona nazivaju „*warez* kopijama“. Ljudi koji prave „*warez*“ kopije, ljudi koji vode „*warez*“ sajtove i ljudi koji preuzimaju te kopije i koriste ih su prekršioci autorskih prava;

- pravljenje nezakonite kopije softvera na CD-ovima, radi davanja nekom drugom, za komercijalnu svrhu i prodavanje istih pod izgovorom da su to zakonite kopije;
- kopiranje softvera od prijatelja ili preko interneta, gde licenca ne dozvoljava ovo;
- zloupotreba dijeljenja: gde je softver licenciran „za određene svrhe“, to je kršenje autorskih prava;
- kupovina softvera za sniženu cijenu pretvarajući se da radite u obrazovnoj ustanovi.

Internet piraterija podrazumjeva upotrebu Interneta za ilegalno kopiranje i distribuciju neautorizovanog softvera. Internet može da se koristi za reklamiranje, ponudu, nabavku i distribuciju piratizovanog softvera. Međunarodna organizacija vodećih proizvođača i prodavača softvera (BSA), procjenjuje da postoji milion Internet sajtova koji prodaju nelegalni softver.

III. MEĐUNARODNA ZAŠTITA INTELEKTUALNOG VLASNIŠTVA

Međunarodna zaštita autorskog prava regulirana je Bernskom konvencijom koja je donijeta 1886. godine, i više puta je revidirana (Pariz 1896., Berlin 1908., Bern 1914., Rim 1928., Brisel 1948., Štokholm 1967. i Pariz 1979.[1] Ovom konvencijom se štite književna i umjetnička djela, uz uslov da su originalna. A kao jedan od nedostataka ove konvencije bili su njeno nereguliranje Interneta i informaciono-komunikacijskih tehnologija, što je rezultiralo osnivanjem Svjetske organizacije za intelektualno vlasništvo (World Intellectual Property Organization, WIPO) 1996. godine. Konvencija o osnivanju Svjetske organizacije za zaštitu prava intelektualnog vlasništva na snazi je od 28.10.1979. Ujedinjene nacije su svrstale ovu organizaciju kao agenciju u sastavu UN. Osnovna aktivnost WIPO je razvoj međunarodnih normativa i standarda intelektualnog vlasništva.

Svjetska trgovinska organizacija kao jedan od uslova za članstvo zahtijeva od budućih članica da usvoje odredbe Bernske konvencije koje su sastavni dio Sporazuma o trgovinskim aspektima prava intelektualnog vlasništva (TRIPS), kojim je predviđeno da se programi računara štite kao književna djela prema Bernskoj konvenciji. Efektivan od 1. januara 1995. godine, TRIPS pokriva prava kopiranja i srodna prava kao što su izvođačka prava i prava muzičkih producenata i medijskih organizacija.

Prema Bernskoj konvenciji djela su zaštićena samim činom svog stvaranja, odnosno nije potrebno vršiti njihovu registraciju kod nekog nadležnog organa, što važi i za računarski program. Pored Bernske za zaštitu intelektualnog vlasništva značajne su Rimska Konvencija za zaštitu prava izvođača, proizvođača fonograma i organizacija za radiodifuziju iz 1961. godine i Ženevska Konvencija o zaštiti proizvođača fonograma od neovlaštenog umnožavanja iz 1971. godine.

A. Zaštita autorskog djela i patentna prava

Gotovo sva nacionalna zakonodavstva propisuju pravnu zaštitu slijedećih autorskih djela:

- računarskih programa,
- književnih djela,
- muzičkih djela,
- umjetničkih djela
- fotografskih djela,
- filmskih djela, i dr.

Oblici pravne zaštite računarskog softvera su:

- zaštita autorskim pravom,
- zaštita patentnim pravom i
- zaštita ugovorom.

Patentna prava se ne stiču nastankom pronalaska ili obavještavanjem javnosti o tome. Ova prava se stiču poslije sprovedenog postupka za priznanje patenta koji se pokreće podnošenjem prijave patenta. Prava važe od datuma podnošenja prijave patenta instituciji koja je nadležna za priznavanje patentnih prava. Patent se izdaje onda kada pronalazak zadovoljava uslove patentibilnosti. Da bi mogao biti zaštićen patentom pronalazak mora predstavljati tehničko rješenje određenog problema koje je novo, koje predstavlja rezultat stvaralačkog rada. Pronalazak mora biti tehničke prirode, odnosno djelo koje se štiti autorskim pravom. Što se tiče novosti, potrebno je da pronalazak ne bude identičan onim tehničkim rješenjima koja su bila dostupna javnosti u pismenom ili usmenom obliku, ili na neki drugi način. Pronalazak se smatra primjenljivim ako ga je moguće upotrijebiti ili proizvesti.

Patent garantuje nosiocu patenta slijedeća prava:

- ako je predmet patenta proizvod: zabranu trećim osobama da bez njegove dozvole izrađuju, upotrebljavaju i nude na prodaju zaštićeni proizvod,
- ako je predmet patenta postupak: da zabrani trećim osobama da bez njegove dozvole upotrebljavaju postupak ili nude na prodaju proizvod koji je dobijen direktno tim postupkom.

Nosioci patenta imaju pravo da putem ugovora ili nasljeđivanjem prenose prava i zaključuju ugovore o licenci. Međutim, postoje pronalasci koji se ne mogu štititi patentom. Jedno od takvih izuzeća su programi računara i primjenjuje se u svim državama svijeta, uključujući i BiH.

[1] Bernska konvencija ima 166 države potpisnice.

A. *Zaštita softvera u Sjedinjenim Američkim Državama*

Zaštita softvera u Sjedinjenim Američkim Državama regulirana je Dekretom o patentima iz 1952. godine. Zavod za patente i žigove SAD je 1996. godine donio Metodologiju ispitivanja pronalazaka vezanih za računare, gdje se navodi da upotrebljivost pronalaska mora biti u oblasti tehnologije, te da pronalazak vezan za računare jeste predmet pronalaska u smislu zakona, odnosno da predstavlja pronalazak koji ima praktičnu primjenu u oblasti tehnologije i kao takav zadovoljava navedeni uslov. Međutim, iako u pojedinim stručnim krugovima nije osporena patentibilnost softvera, ipak je postavljeno pitanje novosti i inventivnosti ovakvih pronalazaka.

B. *Zaštita računarskih programa u Evropskoj Uniji*

U okviru Svjetske organizacije za intelektualno vlasništvo prihvaćen je princip po kojem računarski programi treba da budu zaštićeni autorskim pravom, dok bi postupci korišteni u računarskim softverima trebalo da budu zaštićeni patentom. Autorsko pravo štiti originalnu implementaciju i način prikaza neke ideje, izvorni i izvršni kod i dokumentacija u digitalnom ili pisanom obliku. Patentno pravo za razliku od autorskog prava koje štiti prezentaciju ideje i oblik izražavanja štiti samu ideju i matematičke postupke programa.

Da bi harmonizovala propise svojih članica koji se odnose na zaštitu računarskih programa od neovlaštenog umnožavanja, Vijeće Evrope je donijelo direktivu 91/250/EEC od 14.05.1991. godine, o pravnoj zaštiti računarskih programa. Ova direktiva obavezuje države članice da štite računarske programe putem autorskog prava kao književna djela u smislu Bernske konvencije. Računarski program se štiti ako je originalan u tom smislu da je autorova intelektualna tvorevina. Pri tome se autorom smatra pravna ili fizička osoba ili grupa fizičkih osoba koja ga je stvorila.

Autor ima prava da daje ovlaštenje trećim osobama da:

- vrše umnožavanje računarskog programa;
- prevode ili mijenjaju računarski program i
- vrše iznajmljivanje i distribuciju računarskog programa ili njegovih kopija.

Prema ovoj direktivi smatraju se kažnjivim slijedeće radnje ako ih vrši osoba koja nema ovlaštenje autora:

- stavljanje u promet nelegalne kopije računarskog programa;
- posjedovanje nelegalne kopije računarskog programa iz komercijalnih razloga i
- stavljanje u promet ili posjedovanje sredstava koja su namijenjena za uklanjanje tehničkih sredstava za zaštitu računarskog programa.

Direktivom 93/98/EEC određeno je da zaštita računarskih programa autorskim pravom traje za života autora i još 70 godina posle njegove smrti, ili ako je više autora, posle smrti poslednjeg od njih. Da bi se razumjelo koje su mogućnosti za zaštitu računarskih programa kroz patentni sistem u državama EU treba imati u vidu odredbe Konvencije o evropskom patentu. Ovom Konvencijom propisano je izuzeće od zaštite patentom. Kao jedno od izuzeća u člana 52. navode se računarski programi. Ova odredba je zasnovana na uvjerenju da računarski programi nemaju tehničku prirodu, te da kao takvi nisu patentibilni. Iz ovoga je proizašlo shvatanje da patentibilnost softvera zavisi od toga da li pronalazak koji je u vezi sa softverom ima tehničku prirodu.

Ovo je precizirano u EPO Guidelines for Examination (Metodologija ispitivanja EPO) gde se navodi da se računarski program ne može zaštititi patentom. Ovdje se navodi da se smatra da pronalazak koji karakteriše program i koji predstavlja tehnički proces ima tehnički doprinos i da je zato patentibilan. U praksi EPO su priznavani patenti za računarske programe u formi postupaka za upravljanje procesom ili uređajem. EPO priprema reviziju Evropske konvencije o patentu kojom bi se ukinula odredba kojom se računarski programi izuzimaju iz zaštite patentom. Na osnovu toga se može zaključiti da je EPO praktično prihvatio praksu Zavoda za patente i žigove SAD. Međutim, ovo je samo djelimično tačno, jer još uvek postoje razlike između njih. Najveću od njih predstavlja tretman prijava patenata za softvere namijenjene za obavljanje poslova. Naime, metode za obavljanje poslova su u skoro svim državama svijeta isključene iz zaštite patentom. Stoga bi softveri za obavljanje poslova kao takvi bili isključeni iz zaštite patentom po dva osnova. Međutim, nakon naprijed navedenih izmjena svoje regulative SAD su počele da izdaju patente i za ovakve pronalaske. Sa druge strane, stav EU se nije promijenio i ovakve prijave patenta podnijete Evropskom zavodu za patente se odbijaju. U ovom trenutku još nije jasno koji koncept će preovladati.

C. *Konvencija o cyber kriminalu*

Evropska konvencija o cyber kriminalu donesena 23.11.2001. godine u Budimpešti, pod okriljem Vijeća Evrope (u daljem tekstu Konvencija) je prvi cjelovit formalni potpisani i široko prihvaćeni multilaterarni sporazum usmjeren na probleme cyber kriminala. Konvencija o cyber kriminalu do sada je potpisalo 46 država, a ratifikovalo 30. Konvencija je prvi međunarodni ugovor kojim se sankcionišu krivična djela i druge vrste kriminala izvršene preko Interneta i drugih računarskih mreža i veoma je značajna za zaštitu intelektualnog vlasništva. Sadržajem odredaba člana 10. Konvencije svaka država potpisnica je obavezna da usvoji takve legislativne mjere, da bi

se u njenom nacionalnom pravu, kao krivično djelo, okvalifikovalo kršenje autorskih prava definisanih u zakonima te članice koji se odnose na obaveze koje je ona preuzela po Praškom aktu od 24.06.1971. godine, Konvenciji iz Berna o zaštiti književnih i umjetničkih djela, Međunarodnoj konvenciji iz Rima o zaštiti izvođača, proizvođača fonograma i emisionih organizacija, Sporazumu o aspektima prava na intelektualno vlasništvo i WIPO ugovoru o autorskim pravima.

IV. PRAVNI OKVIR KRIVIČNIH DJELA PROTIV INTELEKTUALNOG VLASNIŠTVA U BOSNI I HERCEGOVINI KOD KOJIH SE KAO SREDSTVO IZVRŠENJA JAVLJAJU RAČUNARI

A. Zakon o autorskom i srodnim pravima

Cilj Zakona o autorskom i srodnim pravima BiH je postizanje potpune usklađenosti tog propisa sa pravnim poretkom EU i važećim međunarodnim konvencijama i ugovorima koje je ratificirala BiH [Bernska konvencija o zaštiti književnih i umjetničkih djela, Međunarodna konvencija za zaštitu umjetnika izvođača, proizvođača fonograma i organizacija za radiodifuziju (Rimska konvencija), Konvencija o zaštiti proizvođača fonograma od neovlaštenog umnožavanja njihovih fonograma (Fonogramska konvencija), Konvencija o distribuciji signala satelita (Satelitska konvencija), Ugovor WIPO-a o autorskom pravu, Ugovor WIPO-a o izvođenjima i fonogramima, te sa Sporazumom STO - TRIPS-a]. Odredbama Zakona o autorskom i srodnim pravima, računarski programi su određeni kao vrsta pisanih autorskih djela, bez obzira na njihov pojavni oblik – u tom smislu je kažnjivo neovlašteno raspolagati na bilo koji način skriptom računarskog programa, kao i mehanički kopirati CD/DVD na kojem se taj program nalazi u elektronskom obliku.[2] Zakon o autorskom i srodnim pravima BiH je usklađen sa sljedećim direktivama EU:

- Direktiva o pravnoj zaštiti računarskih programa, Direktiva Vijeća 91/250/EEC od 14.05.1991.
- Direktiva o pravu davanja u zakup i na poslugu i o određenim pravima umnožavanja u području intelektualnog vlasništva, 92/100/EEC od 19.11.1992, i kodificirana verzija od 12.12.2006.
- Direktiva o pravilima autorskog i srodnih prava koja se primjenjuju na satelitsko emitiranje i kablovsko reemitiranje, 93/83/EEC od 27.09.1993.
- Direktiva o harmonizaciji rokova trajanja autorskog i srodnih prava, 93/98/EEC od 29. oktobra 1992, i kodificirana verzija od 12.12.2006.
- Direktiva o pravnoj zaštiti baza podataka, Direktiva 96/9/EC od 11.03.1996.
- Direktiva o harmonizaciji određenih aspekata autorskog i srodnih prava u informatičkom društvu, 2001/29/EC od 22.05.2001.
- Direktiva o pravu preprodaje u korist autora originalnog djela 2001/84/EC od 27.09.2001.
- Direktiva o provođenju prava intelektualnog vlasništva, 2004/48/EC od 29.04.2004.
- Direktiva o roku trajanja autorskog i određenih srodnih prava, 2006/116/EC od 12.12.2006.

A. Krivična djela protiv intelektualnog vlasništva u Krivičnim zakonima u Bosni i Hercegovini

Bosna i Hercegovina je ratificirala Konvenciju o cyber kriminalu i Dodatni protokol uz Konvenciju 2006. godine. Nakon stupanja na snagu Konvencije u odnosu na BiH slijedila je implementacija odredbi krivičnog materijalnog i procesnog prava Konvencije u državni i nedržavne krivične i krivično procesne zakone. Sadržajem odredbi Krivičnog zakona BiH za razliku od ostala tri nedržavna Krivična zakona (Federacije BiH, R Srpske i Brčko Distrikta BiH) nisu implementirane preporučene odredbe krivičnog prava već je to „dogovorno prepušteno“ u nadležnost nedržavnim krivičnim zakonima F BiH, RS i Brčko Distrikta.

Krivična djela protiv intelektualnog vlasništva, kod kojih se kao sredstvo izvršenja krivičnih djela javljaju računari i računarske mreže, regulirana su:

- Krivičnim zakonom Federacije BiH i
- Krivičnim zakonom Republike Srpske.

Krivična cyber djela uvedena su Krivičnim zakonom Federacije BiH, u Glavi 32, gdje su predviđena „krivična djela protiv sistema elektronske obrade podataka" (čl. 393-397). Krivičnim zakonom Federacije BiH u Glavi 32, propisane su sankcije za krivična djela kod oštećenja računarskih podataka i programa. U članu 393. propisuje se krivično djelo Oštećenje izmjena ili uništenje tuđih računarskih podataka ili računarskih programa, u članu 394. propisano je krivično djelo računarskog krivotvorenja, u članu 395. propisano je kazneno djelo računarske prevare, u članu 396. propisuje krivično djelo Ometanje rada sistema i mreže elektronske obrade podataka, u članu 397. propisuje krivično djelo Neovlašteno uključenje u sistem i mrežu elektronske obrade podataka. Krivično djelo iz člana 6. Konvencije odnosno člana 7. Prijedloga Direktive propisano je u odredbama člana 393.

Krivični zakon RS (2003. g.) predviđa krivično djelo upad u računarski sistem u Glavi 16. A kao „krivična djela protiv sigurnosti računarskih podataka" (čl. 186a-čl.186e).

Krivična djela koja se odnose na povredu autorskog i srodnih prava iz člana 10. Konvencije, su odredbe krivičnih djela iz člana 242. (Zloupotreba autorskih prava), člana 243

[2] Svrstavanje računarskih programa u pisana autorska djela nije originalnost bosanskohercegovačkog Zakona o autorskom i srodnim pravima – slično rješenje primijenjeno je u zakonskim rješenjima u Japanu, Nemačkoj i Francuskoj, dok je Evropska unija 1991. izdala direktivu kojom je računarske programe izjednačila sa književnim djelima u smislu Bernske konvencije.

(Nedozvoljeno korištenje autorskih prava), člana 244. (Nedozvoljeno korištenje prava proizvođaća zvučne snimke), člana 245. (Nedozvoljeno korištenje prava radiodifuzije) i člana 246. (Nedozvoljena distribucija satelitskog signala) Krivičnog zakona BiH, koje obuhvataju i krivična djela počinjena pomoću računarskog sistema.

Analizom sadržaja odredbi državnog i nedržavnih zakona se može zaključiti da se obaveza preuzeta ratifikacijom Konvencije u pogledu krivično matrijalnog prava ispoštovala i da su krivična djela predviđena Konvencijom ugrađena u odredbe nedržavnih krivičnih zakona u BiH.

V. ZAKLJUČAK

Razvoj informacionih tehnologija i njihova masovna primjena u poslovne i privatne svrhe, uslovili su napade na autorska djela i njima slična prava. S obzirom da su načini zloupotrebe informacionih tehnologija svakim danom sve složeniji i savršeniji za otkrivanje, i da je teško ići u korak sa tim kriminalnim aktivnostima, potrebno je ulagati napore u to da javnost bude svjesna sa kakvim se fenomenom današnje društvo suočava, te je potrebno konstantno raditi na što adekvatnijem odgovoru na različita kriminalna djelovanja u ovoj oblasti. Mnoge države su postale svjesne ove pojave i u svojim krivičnim zakonima su predvidjele kršenje autorskog i njima sličnih prava kao posebna krivična djela.

U svijetu su prisutna dva sistema zaštite softvera. Prvi od njih tradicionalno predstavlja autorsko pravo i prisutan je još od pojave prvih računarskih programa. Drugi od njih, koji se naknadno pojavio predstavlja zaštitu softvera patentima. U ovom radu je analizirana zaštita softvera u SAD i državama EU, te su ukazani potencijalni smijerovi razvoja u budućnosti.

Na osnovu stanja zakonodavstva u BiH, a imajući u vidu činjenicu da je napad na intelektualno vlasništvo nov oblik kriminala, možemo da konstatujemo da su država i entiteti usvajanjem odgovarajućih pravnih propisa omogučili krivični progon počinioca krivičnih djela u vezi sa napadom na autorsko i njima slična prava.

LITERATURA

[1] Krivični zakon BiH (“Sl. glasnik BiH”, broj: 3/03, 32/03, 37/03, 54/04, 61/04, 30/05, 53/06, 55/06, 32/07, 8/10).

[2] Krivični zakon Brčko Distrikta BiH (“Sl. glasnik Bd BiH”, broj: 10/03, 6/05, 47/11, 52/11).

[3] Krivični zakon Federacije BiH (“Sl. novine Federacije BiH”, broj: 36/03, 37/03, 21/04, 69/04, 18/05, 42/10, 42/11).

[4] Krivični zakon RS (“Sl. glasnik RS”, broj: 49/03, 108/04, 37/06, 70/06, 73/10, 1/12).

[5] S. Šarboh, Zaštita softvera patentima – regulativa i iskustva SAD i država EU, Zavod za intelektualnu svojinu, Beograd 2011.

[6] Zakon o autorskom i srodnim pravima, („Službeni glasnik BiH”, broj 63/10).

[7] Zakon o krivičnom postupku BiH (“Sl. glasnik BiH”, broj: 3/03, 32/03, 36/03, 26/04, 63/04, 13/05, 48/05, 46/06, 76/06, 29/07, 32/07, 53/07, 76/07, 15/08, 58/08, 16/09,93/09).

[8] Zakon o krivičnom postupku Brčko Distrikta BiH (“Sl. glasnik Bd BiH”, broj: 10/03, 48/04, 06/05, 06/05 (2), 12/07,14/07, 21/07).

[9] Zakon o krivičnom postupku Federacije BiH (“Sl. novine FBiH”, broj: 35/03, 37/03, 56/03, 78/04, 38/05, 55/06, 27/07, 53,07, 09/09, 12/10).

[10] Zakon o potvrđivanju Konvencije o cyber kriminalu Vijeća Evrope i Dodatni protokol uz Konvenciju o kibernetičkom kriminalu („Sl. glasnik BiH - Međunarodni ugovori“ broj 06/06).

[11]http://www.microsoft.com/scg/pres/24.01.2014.

[12]http://w.w.w.wipo.int/treaties/en/ip/berne/index 20.02.2014.

[13]http://w.w.w.wipo.int/treaties/en/convention/trt docs 22.02.2014.

[14]http://conventions.coe.int/Treaty/en/Treaties/Html/185.htm 22.02.2014.

miproBIS

International Conference on

BUSINESS INTELLIGENCE SYSTEMS

Steering Committee

Chairs: Boris Vrdoljak, University of Zagreb, Croatia
Mirta Baranović, University of Zagreb, Croatia
Matteo Golfarelli, University of Bologna, Cesena, Italy
Roberto Sandri, Invenio Consulting Ltd., Zagreb, Croatia

Members: Alberto Abello Gamazo, Universitat Politècnica di Catalunya, Barcelona, Spain
Marko Banek, University of Zagreb, Croatia
Mirta Baranović, University of Zagreb, Croatia
Ladjel Bellatreche, Poitiers University, Futuro Scope Chasseneuil, France
Ljiljana Brkić, University of Zagreb, Croatia
Alfredo Cuzzocrea, University of Calabria, Rende, Italy
Jamel Feki, Multimedia, Information, Systems and Advanced Computing Laboratory, Sfax, Tunisia
Dragan Gamberger, Ruđer Bošković Institute, Zagreb, Croatia
Matteo Golfarelli, University of Bologna, Italy
Igor Mekterović, University of Zagreb, Croatia
Damir Pintar, University of Zagreb, Croatia
Zoran Skočir, University of Zagreb, Croatia
A Min Tjoa, Vienna University of Technology, Austria
Mladen Varga, Faculty of Economics and Business, Zagreb, Croatia
Boris Vrdoljak, University of Zagreb, Croatia
Robert Wrembel, Poznan University of Technology, Poland

INVITED PAPER

Log-driven user-centric OLAP

Patrick Marcel Université François Rableais Tours
Laboratoire d'Informatique
Blois, France
Email: patrick.marcel@unv-tours.fr

***Abstract*—While OLAP (On-Line Analytical Processing) is now a mature, efficient technology, very little attention has been paid to the effectiveness of the analysis and the user-friendliness of a technology often considered tedious of use. This work summarizes various contributions to developing user-centric OLAP, focusing on the use of former queries to enhance subsequent analyses.**

We first introduce a logical model to describe and manipulate queries and logs, including a language parametrized by binary relations over analytical sessions. In particular, these relations can be specialization relations or based on similarity measures tailored for OLAP queries and sessions.

We then describe how logs can be mined for knowledge representing the user behavior. This knowledge includes simple preferences, navigational habits and discoveries made during former explorations. We show how it can be used in various query personalization or query recommendation approaches, that vary in terms of formulation effort, proactiveness, prescriptiveness and expressive power.

Finally, we remark that the models and approaches introduced can be seen as a starting point for assessing the effectiveness of analytical sessions, and bridging the gap between OLAP and exploratory search.

I. Introduction

Data warehousing and OLAP are now mature technologies, having attracted a lot of attention from the academic and industrial community, and benefiting from the maturity of relational databases. Noticeably, the major part of this attention has been devoted to the efficient implementation and querying of the warehouse, and very little attention has been paid to the quality of the analysis and the ease of use of such technologies. Recent studies showed that BI tools, although powerful, are often underused by decision makers, who frequently rely on static dashboards [1], [2] instead of using the full potential of the tools.

To answer these limitations, it is now commonly agreed that BI should benefit from a combination with Web 2.0 approaches (a focus on user empowerment, social networks, and community collaboration), a trend often referred to as BI 2.0 [3]. Achieving user empowerment without asking for too much formulation effort, while remaining proactive, can be done using approaches based on the history, i.e., the traces left by the users during former activities. Actually, it has already been pointed out the necessity to come up with flexible, powerful means for analyzing the issued queries, and decompose, store and handle them in a dedicated subsystem in order to better support any decisional task with the knowledge captured in the analytical queries [4]. Evidences of this need are given in [5], where the authors carried out tests demonstrating that browsing through past SQL query sessions helped the users by speeding up query composition. It turns out that user-centric approaches in databases, like query personalization and query recommendation, are indeed getting more and more attention. Using preferences to personalize queries has been investigated for the last ten years [6], and recommending queries [7], [8] is emerging as a promising way of supporting the user browsing large databases.

Strangely enough, although highly relevant to BI in general and OLAP in particular, user-centric approaches, especially the ones based on leveraging past user queries, have attracted relatively little attention from the BI/data warehouse community. This paper summarizes research recently led at the University of Tours to develop user-centric OLAP, focusing on the use of former queries to enhance subsequent analyses. Section II briefly surveys personalization and recommendation approaches in databases. Section III introduce the query models we use to reason about queries as well as a technique to obtain OLAP logs. Section IV illustrates how such logs can be manipulated using a simple language inspired by the relational algebra. Section V describes the user-centric approaches. Section VI concludes and draw perspectives.

II. User-centric approaches in databases

This section briefly recalls two popular user-centric approaches adopted by the database community, namely personalization and recommendation. A more detailed presentation can be found in [9].

Query personalization in databases include two types of approaches:

- The use of explicit preference operators in queries to specify profile elements. The most representative works in this category include Winnow [10], Preference SQL [11] and Skyline [12]. This type of approaches requires high formulation effort, is not prescriptive, not proactive, but is highly expressive.
- The rewriting (expansion) of regular database queries based on a profile. The most representative work is initiated in [13]. This approach requires a low formulation effort, is prescriptive, not proactive and has low expressiveness.

Note that in these two categories, the profile is only given in terms of the constants, attributes and tuples of the database. Thus, using the criteria introduced above, all the approaches are current-state approaches.

The basic definition of the operator computing dominating tuples is the following. Given a relation r with schema $sch(r)$

and a preference C over $sch(r)$ defining a preference relation $>_C$, the Winnow operator [10], denoted w_C, is defined by: $w_C(r) = \{t \in r | (\nexists t' \in r)(t' >_C t)\}$. This operator can be used to order the query answer. Indeed, the answer to a query q can be partitioned according to C, i.e., $q = w_C(q) \cup w_C(q - w_C(q)) \cup \ldots$ meaning that the answer can be presented by displaying $w_C(q)$ first, then $w_C(q - w_C(q))$, etc. This operator has been used with various syntaxes for expressing models of preferences. The work of Kiessling [11] enrich the SQL syntax with a PREFERRING clause that enables the user to specify in each query the model of preference, through the use of of specific preference constructors. Each constructor can be used to express a specific atomic preference. Preferences can be composed with Pareto or prioritization. A restricted form of Kiessling's SQL extension is the addition of the Skyline operator to SQL [12]. This is a restriction in the sense that, originally, Skyline queries feature only numerical attributes and Pareto composition.

In the absence of a dedicated preference operator, a regular user query can be processed and transformed with preferences, resulting in another regular query that is typically a subquery of the initial one. We use the work of [13] to illustrate this approach. In this work, preferences come from a user profile which is modeled as a graph of atomic quantitative preferences of the form *(selection condition, score)*, where *selection condition* is a regular selection predicate (that may be used to join two tables) and *score* is a real between 0 and 1 indicating the intensity of the preference. Atomic preferences are composed using a very simple principle: composition of preference $(s1, v1)$ with preference $(s2, v2)$ results in preference $(s1 \wedge s2, v1 \times v2)$. Query expansion is performed as follows. First, given a query, the k most relevant preferences are selected from the profile. Then the selected preferences are added as hard constraints to the query, and the query is finally executed.

Recommender systems [14] are now well established as an information filtering technology, used in a wide range of domains, from E-commerce to web search to social networks, to suggest items to users. Recently, to our knowledge, there has been only two attempts to formalize database query recommendations for exploration purpose [15], [7]. All the approaches presented below require a low formulation effort, are proactive (a query is suggested) and prescriptive (no soft constraint are added to the suggested query) and offer a low expressiveness. In [7], the authors propose two techniques to recommend queries based on the database instance and/or the current query answer. The first technique, called local analysis, analyzes the answer to the user query to discover patterns and uses these patterns to recommend. The second technique, called global analysis, extends this principle to the entire database instance. The instance would have to be analyzed off-line, for example to discover correlations among attribute values.

For [15], the problem of query recommendation is viewed as a sessions $\times$ tuples matrix. With this approach, a query is represented as a vector whose arity is the number of tuples of the database instance. The basic approach considers that each component of such a query is either a 1, if the query used the tuple, or a 0 otherwise. A session is also represented by a binary vector which is the logical OR of the vectors of the queries of the session. Consider a particular session S_c called the current session. Recommendations for S_c are computed as follows. First, sessions similar to S_c are found, using some vector similarity measures like e.g., cosine. For those sessions closest to S_c, the queries they contain are also compared to S_c using the same similarity measure. Finally, the queries of those sessions that are the most similar to S_c are recommended. In subsequent works [8], the authors focus on fragments (attributes, tables, joins and predicates) of queries and consider thus a sessions $\times$ query fragments matrix. In this work, the matrix is used column-wise in the sense that recommendation computation relies on fragment similarity instead of session similarity.

To conclude this section, we introduce the requirements for user-centric approaches in the context of OLAP analysis of data warehouses. As evidenced by e.g., [16], [17], [18] basic peculiarities of typical data warehouses can be summarized by:

1) A data warehouse is a read-mostly database and its instance has an inflationist evolution (data are added, never or very seldom deleted). It is for instance likely that a user issues periodically *similar sequences of queries* more than once, in the sense that *queries may not be fully identical*.
2) A data warehouse is a database shared by *multiple users, mostly executives, whose interests are diverse and may vary over time*. It is argued in [19], [18], [20], [21] that *user preferences are of particular importance in data warehouse exploration*. It would for instance be important to *issue recommendations computed from other users' habits* (e.g., in a collaborative filtering fashion) and at the same time *respecting the user interests*.
3) A data warehouse has a particular schema that reflects a known topology, often called the lattice of cuboids, which is systematically used for navigation [22]. *Roll-up and drill-down operations that allow to see facts at various levels of detail are very popular in this context*.
4) A typical analysis session over a data warehouse is a sequence of queries having an analytical goal, each one written based on the past results of the session. They may be expressed in a *dedicated query language (like e.g., MDX*[1]*)*, may produce *large results that are usually visualized as cross-tabs*. Moreover, *the session has a sense w.r.t. some expectations*. For instance, the user may assume a uniform distribution of the data [23], [24] or that two populations follow the same distribution [25]. Sessions (as sequences of queries) are of particular importance in this context since *with this sequence of queries, the user navigates to discover valuable insights w.r.t. her expectations or assumptions*.

III. Modeling and constructing logs

This section describes the models for queries, sessions and logs that were used in our works [19], [26], [27], [28], [29],

[1] http://msdn.microsoft.com/en-us/library/ms145506.aspx

[30], to reason about queries. In these works, three main ways are used for modeling queries. A multidimensional query can be modeled as a set of fragments (group by sets, sets of selection predicates, sets of measures), extracted from the query expression. A query can be modeled as a set of tuples, obtained with a partial evaluation of the query expression. For instance, a multidimensional query over a star schema can be evaluated only over the dimension tables, resulting in a set of references. Note that obtaining this set of references from a query expression can be computed efficiently when dimension tables fit in main memory. In that case, a query can be expressed intentionally as a tuple of sets, one set of members in each dimension. The cross-product of these sets is a set of references, which forms the query model. Finally, a query can be modeled as the set of its answer when evaluated over a database instance. It can be defined as follows: Given a cube instance and a query expression, the query models is the set of facts extracted from the data cube by the query expression.

Intuitively, these three models can be seen as three different focuses on how the user interacts with the cube, where the importance is either the way the query is written (unevaluated queries), or the part of the cube queried (partially evaluated queries), or the values retrieved (fully evaluated queries). It is also worth noting that both the effectiveness and efficiency of the user-centric approach depends on the query model used. Indeed, a trade-off exists, that can be expressed as follows. On the one hand, modeling queries as their results allows to leverage the knowledge extracted with the query, but requires important resources to store or recompute the query answer, especially in a data warehousing context. On the other hand, using only information derived from the query expression may give rise to efficient approaches, but not knowing the data extracted with the query may prevent valuable deductions, like two different query expressions lead to identical results on a given instance.

Regarding logs and sessions, OLAP logs are modeled as sets of sessions, sessions being modeled as sequences of OLAP queries. Obviously, not all query logs produced by an OLAP system, especially those based on the ROLAP technology, naturally respect this model. OLAP queries can be expressed under the form of SQL (or relational) expressions. The log may simply be a long sequence of such expressions that may not be organized as a set of sessions. And even though the queries are organized as sessions, they may not be logically connected toward the exploration of an analytical goal. Consequently, OLAP semantics must be extracted from each expression, under the form of OLAP operations, and sessions must be identified. [29] introduces a technique to obtain an OLAP log from an ordered set of SQL queries. The techniques consists of normalizing multidimensional expressions, and checking for all pairs of successive such expressions, if a small number of OLAP operations can be used to transform the output of the former into the output of the latter. If the queries are not expressed in the multidimensional algebra, they are translated into it.

IV. Manipulating and comparing logs

This section describes the manipulation language introduced in [31], that is dedicated to query logs, and stemming from the relational algebra. While the (extended) relational algebra enables the manipulation of relations, i.e., sets of tuples, that have a schema, the language presented here manipulates logs, i.e., sets of sessions, that does not have a schema. The schema can therefore not be used in the expressions formed with this language. Instead, the operations of the language are parametrized by a binary relation over sessions θ, that may for instance be used to compare sessions. In particular, θ can be a specialization or a similarity relations over sessions. A study of similarity measures for OLAP sessions can be found in [32].

For the sake of simplicity, the names and symbols of the operations are the same as the ones used in the relational algebra. The language features 5 operations. Two of them are unary operations: selection σ and group by and aggregation π. Three of them are binary: join $\bowtie$, union $\cup$ and difference $\setminus$. In what follows, let L and L' be two logs, f be a binary functions outputting a session, agg be a function aggregating a set of sessions into a session, s be a session and θ be a binary relation over sessions.

The selection operation enables to select from a log those sessions satisfying a given condition. This condition is given under the form of a relation to another session. Formally, $\sigma_{\theta,s}(L) = \{s' \in L | \theta(s, s')\}$

As logs are defined as sets of sessions, the set union and set difference operations can be used to manipulate logs. Formally, $L \cup L' = \{s \in L \text{ or } s \in L'\}$ and $L \setminus L' = \{s \in L | s \notin L'\}$.

The join operation enables to combine the sessions in two logs, that are related through a relation θ. Formally, $L \bowtie_{\theta,f} L' = \{f(s, s') | s \in L, s' \in L', \theta(s, s')\}$. Note that intersection can be simulated with the join operation.

Finally, the grouping and aggregate operator allows to group sessions that are related to one another through relation θ, and to aggregate them using an aggregation function agg. Formally, $\pi_{\theta,agg}(L) = \{agg(\{s' \in L | \theta(s, s')\}) | s \in L\}$.

This language allows to manipulate logs in a flexible way, and can be used to search, filter, compare, combine, modify logs, and express advanced manipulations like summarization or query recommendation [31]. The language is closed under composition, and is complete. The language is not minimal since σ can be simulated with $\bowtie$. It can easily be seen that removing σ makes the language minimal.

V. Log-driven user-centric analysis

This section describes informally the approaches we proposed for user-centric OLAP using query logs.

A. Extracting profile information from the log

This section describes how relevant knowledge can be extracted from an OLAP query log to subsequently support OLAP analyses. Extracted knowledge include (i) simple preferences over dimensions, members and measures [19], (ii) correlations between query fragments [30], and (iii) investigations representing what has been seen during past sessions

for a pair of cells where measures differ significantly [27]. Unsurprisingly, it can be seen that, the more precise the query model used (with respect to the database instance), the more complex the information extracted is.

We defined simple preferences over multidimensional data to be: a partial order over the dimensions, dimension-wise partial orders over members, and an order over measures. The order over dimensions is inferred from the former queries' projections, over levels or measures, and the orders over members are inferred from the former queries' selections. Note that an order over levels in a dimension can be easily computed using this principle.

A navigational habit is defined by an association rule of the form $X \rightarrow Y$, with the intuitive meaning that whenever fragment X is used in a query expression, then fragment Y tends to be present as well. Our algorithm for extracting navigational habits uses any classical association rule extractor, and adjusts automatically support and confidence thresholds [30].

If the query expressions in the log are grouped into sessions and are modeled as unevaluated queries (i.e., represented by their results), the log can be processed to discover what was the goal of the user during the analysis session. The key idea of processing the log to discover what was the goal of the user, is to detect the difference between measure values that queries investigated. More precisely, the log is examined to discover the pairs of cells whose measures differ significantly, to retain the most general ones (the most general difference pairs, $mgdp$) as well as the queries that contains them (the most general difference queries, $mgdq$). For such pairs, a structure called investigation is created that records the set of $mgdq$ and, at a lower level of detail, the queries that confirm the difference (their drill-down differences), and the queries that contradicts the difference (their exceptions).

B. Personalizing queries with a single user log

In this section, we present two approaches for personalizing multidimensional queries, one based on the use of dedicated operators and one based on query expansion. Importantly, these approaches are consistent with the requirements stated in Section II. In particular, (i) they rely on user navigational habits extracted from past analytical sessions, (ii) the profile elements used respect the multidimensional model, and (iii) they can take advantage of the user's past discoveries in the data.

The first approach [30] deduces from the log the preference constructs to be used with a user query, assuming the use of the MYMDX preference language [33]. The profile consists of navigational habits R extracted from the log. The approach consists of three steps:

1) *Rule selection*. When a user formulates an MDX query q, a subset $R_q \subseteq R$ of rules is selected. Each rule in R_q is *pertinent* w.r.t. q, meaning that its antecedent matches with q, and *effective*, meaning that the preference it would be translated into can actually induce an ordering on the facts returned by q. Then, let a positive integer *personalization degree* α be chosen by the user to express the desired preference complexity. A query F_α is generated from R_q in such a way that the final preference expression of the MYMDX statement has α base constructors, where α is user-defined.
2) *Fragment translation*. Each fragment in F_α is translated into a base constructor; the resulting base constructors are then coalesced and composed using the Pareto operator into a preference expression p.
3) *Querying*. Query q is annotated with p, translated into MYMDX, and executed. As shown in [33], the user can effectively explore query results by visually interacting with a graph-like structure that emphasizes the better-than relationships induced by p between different sets of facts. Preferred facts are then displayed in a multi-dimensional table.

The second approach uses query expansion to enrich the user query with preferences, without relying on dedicated operators. The principle is to reduce the answer to a user query by computing the preferred subquery in the classical sense of query inclusion. The principle introduced in [13] can be directly applied: the simple preferences extracted over members can be used to generate a user profile consisting of weights on selection conditions over members. This profile could be exploited using the algorithm described in [13]. Alternatively, if the profile consists in the navigational habits, the principle used in the first approach can be used. Instead of deriving preference constructors from the relevant habits, these habits are used to enforce hard constraints, or to discard group by sets.

C. Collaborative recommendations with a multi-user log

Finally, this section describes various query recommendation techniques using information extracted from an OLAP query log. They are aligned with the requirements stated in Section II, being tailored for the multidimensional model and the ways users analyze a cube. The approaches either rely on the log and on the database instance, and in that case the recommended queries indicate potential discoveries made by other users analyzing a related set of data, or they exploit past unevaluated queries only, by looking for similar user intensions or navigational habits.

The first approach [27] uses both the information extracted from the log and the current state of the database, more precisely, the content of the fully evaluated query. In this approach, the profile is a set of investigations extracted from a multi-user log. Recommendations are computed on the basis of the differences discovered in the log. The key idea is to detect the difference that the current query is investigating and to recommend the queries in the log that investigated the same difference. An investigation records the set of $mgdq$ (the most general difference queries, i.e., the queries containing a difference pair at the coarsest granularity level in the log) and, at a lower level of detail, the queries that confirm the difference (the $mgdq$'s drill-down differences), and the queries that contradicts the difference (their exceptions). Recommendations are computed online each time a current query is added to the current session by the current user. The current query is

analyzed to detect to which investigations it corresponds (this query may be itself a $mgdq$, a drill-down difference, or an exception of what is detected in the log). Then a navigation plan (a set of queries arranged in a graph) is proposed for the current user to see drill-down differences or exceptions to the $mgdq$, by using the queries of the investigations.

The next query recommendation approaches use only the history of the database, that is the unevaluated queries of the log. The second recommendation approach relies on a general framework for computing history-based recommendations [34], where a query is recommended by using both the server log, i.e., the set of sessions already posed to navigate the cube, and the sequence of queries of the current session. Note that this framework can be used whatever the query model (unevaluated, partially evaluated, fully evaluated), by adapting the similarity used for queries (and thus sessions). The framework consists in three steps: (i) Matching, i.e., searching the log for sessions related to the current session, (ii) Predicting, i.e., generating candidate recommendations by predicting what could be the next query, and (iii) Ranking, i.e., ordering the candidate recommendations to present the queries to the user in interest order. This framework is flexible in the sense that it does not lay down a specific method for matching, predicting or ranking. Instead of that, each of these steps is parametrized with functions and relations. By changing these parameters, the way recommendations are computed changes. Instantiations of this framework are given in [26].

Finally navigational habits extracted from a multi-user query log can be exploited to derive recommendations, using an approach similar to that of [30]. In this case, the set of rules is searched for pertinent rules, i.e., whose antecedent matches the user query. The main difference is that the rules need not be effective, since the goal is not to reduce the answer set. Then pertinent rules' consequents are used to derive the recommendation.

VI. Conclusion

This paper summarizes various contributions to developing user-centric OLAP, focusing on the use of former queries logged by an OLAP server to enhance subsequent analyses. Logs are modeled as sets of sessions, sessions being modeled as sequences of OLAP queries. Three main approaches are presented for modeling queries: as unevaluated collections of fragments (e.g., group by sets, sets of selection predicates, sets of measures), as sets of references obtained by partially evaluating the query over dimensions, or as query answers. Such logs can be constructed even from sets of SQL query expressions, by translating these expressions into a multidimensional algebra, and bridging the translations to detect analytical sessions. Logs can be searched, filtered, compared, combined, modified and summarized with a language inspired by the relational algebra and parametrized by binary relations over sessions. In particular, these relations can be specialization relations or based on similarity measures over queries and sessions. Logs can be mined for various hidden knowledge, that, depending on the query model used, accurately represents the user behavior extracted. This knowledge can be used for query personalization, i.e., coping with a current query too few or too many results, or query recommendation, i.e., suggesting queries to pursue an analytical session.

To assess the effectiveness of the contributions summarized here, subjective test involving users and user traces will have to be run. To conduct and simulate such tests, a perspective is the development of a platform for assessing the quality of an analytical session over a cube. Indeed, although data quality is a well studied area of data management, the quality of the querying process has not yet been investigated. Data quality is modeled according to quality dimensions (like e.g., accuracy.), each of them grouping quality factors (like e.g., semantic correction, or syntactic correction), each of them measured by various metrics [35]. A similar approach could be used to model the quality of analytical queries and the quality of analytical sessions. In addition, such a model would allow the development of a platform for assessing and validating user-centric approaches in the context of OLAP. Notably, in domains like Information Retrieval, benchmarks are used [36] to validate approaches aiming at supporting browsing and exploratory search results. In the domain of data warehousing, the existing benchmarks are exclusively focused on performance (see for instance [37]), and effectiveness of the querying process in terms of how successful the analytical session is, is simply not addressed. So far, validating personalization or recommendation approaches is based on synthetic query logs [38] or the processing of existing logs with no guarantee of being realistic [15]. A platform for assessing user-centric approaches should allow to measure to which extent approaches are effective in that answers found are relevant, the effort spent to conduct the analysis is reduced, etc. Such a platform could benefit from previous effort to develop a benchmark to validate personalization approaches in databases [39].

Finally, an even longer term perspective is extending OLAP to exploratory search. Exploratory search can be defined by an elaborated activity that goes beyond lookup and involves learning and investigating [40]. Modeling the overall decision making process under the form of exploratory search will allow to take advantage of how the user understanding of data evolves along her/his interaction with the data warehouse.

References

[1] T. B. Pedersen, "How Is BI Used in Industry?: Report from a Knowledge Exchange Network," in *Data Warehousing and Knowledge Discovery, 6th International Conference, DaWaK 2004, Zaragoza, Spain, September 1-3, 2004, Proceedings*, 2004, pp. 179–188.

[2] H. Berthold, P. Rösch, S. Zöller, F. Wortmann, A. Carenini, S. Campbell, P. Bisson, and F. Strohmaier, "An architecture for ad-hoc and collaborative business intelligence," in *Proceedings of the 2010 EDBT/ICDT Workshops, Lausanne, Switzerland, March 22-26, 2010*, 2010.

[3] J. Trujillo and A. Maté, "Business intelligence 2.0: A general overview," in *Business Intelligence*, ser. Lecture Notes in Business Information Processing, M.-A. Aufaure, E. Zimányi, W. Aalst, J. Mylopoulos, M. Rosemann, M. J. Shaw, and C. Szyperski, Eds. Springer Berlin Heidelberg, 2012, vol. 96, pp. 98–116.

[4] N. Khoussainova, M. Balazinska, W. Gatterbauer, Y. Kwon, and D. Suciu, "A Case for A Collaborative Query Management System," in *CIDR 2009, Fourth Biennial Conference on Innovative Data Systems Research, Asilomar, CA, USA, January 4-7, 2009, Online Proceedings*, 2009.

[5] N. Khoussainova, Y. Kwon, W.-T. Liao, M. Balazinska, W. Gatterbauer, and D. Suciu, "Session-Based Browsing for More Effective Query Reuse," in *Scientific and Statistical Database Management - 23rd International Conference, SSDBM 2011, Portland, OR, USA, July 20-22, 2011. Proceedings*, 2011, pp. 583–585.

[6] K. Stefanidis, G. Koutrika, and E. Pitoura, "A survey on representation, composition and application of preferences in database systems," *ACM Trans. Database Syst.*, vol. 36, no. 3, p. 19, 2011.

[7] K. Stefanidis, M. Drosou, and E. Pitoura, ""You May Also Like" results in relational databases," in *Proceedings International Workshop on Personalized Access, Profile Management and Context Awareness: Databases*, Lyon, France, 2009.

[8] G. Chatzopoulou, M. Eirinaki, S. Koshy, S. Mittal, N. Polyzotis, and J. S. V. Varman, "The QueRIE system for Personalized Query Recommendations," *IEEE Data Eng. Bull.*, vol. 34, no. 2, pp. 55–60, 2011.

[9] P. Marcel, "Olap query personalisation and recommendation: An introduction," in *Business Intelligence*, ser. Lecture Notes in Business Information Processing, M.-A. Aufaure, E. Zimányi, W. Aalst, J. Mylopoulos, M. Rosemann, M. J. Shaw, and C. Szyperski, Eds. Springer Berlin Heidelberg, 2012, vol. 96, pp. 63–83.

[10] J. Chomicki, "Preference formulas in relational queries," *ACM Trans. Database Syst.*, vol. 28, no. 4, pp. 427–466, 2003.

[11] W. Kießling, "Foundations of Preferences in Database Systems," in *VLDB 2002, Proceedings of 28th International Conference on Very Large Data Bases, August 20-23, 2002, Hong Kong, China*, 2002, pp. 311–322.

[12] S. Börzsönyi, D. Kossmann, and K. Stocker, "The Skyline Operator," in *Proceedings of the 17th International Conference on Data Engineering, April 2-6, 2001, Heidelberg, Germany*, 2001, pp. 421–430.

[13] G. Koutrika and Y. E. Ioannidis, "Personalization of Queries in Database Systems," in *Proceedings of the 20th International Conference on Data Engineering, ICDE 2004, 30 March - 2 April 2004, Boston, MA, USA*, 2004, pp. 597–608.

[14] G. Adomavicius and A. Tuzhilin, "Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 6, pp. 734–749, 2005.

[15] G. Chatzopoulou, M. Eirinaki, and N. Polyzotis, "Query Recommendations for Interactive Database Exploration," in *Scientific and Statistical Database Management, 21st International Conference, SSDBM 2009, New Orleans, LA, USA, June 2-4, 2009, Proceedings*, 2009, pp. 3–18.

[16] S. Chaudhuri and U. Dayal, "An Overview of Data Warehousing and OLAP Technology," *SIGMOD Record*, vol. 26, no. 1, pp. 65–74, 1997.

[17] S. Sarawagi, R. Agrawal, and N. Megiddo, "Discovery-Driven Exploration of OLAP Data Cubes," in *Advances in Database Technology - EDBT'98, 6th International Conference on Extending Database Technology, Valencia, Spain, March 23-27, 1998, Proceedings*, 1998, pp. 168–182.

[18] S. Rizzi, "OLAP preferences: a research agenda," in *DOLAP 2007, ACM 10th International Workshop on Data Warehousing and OLAP, Lisbon, Portugal, November 9, 2007, Proceedings*, 2007, pp. 99–100.

[19] L. Bellatreche, A. Giacometti, P. Marcel, H. Mouloudi, and D. Laurent, "A personalization framework for OLAP queries," in *DOLAP 2005, ACM 8th International Workshop on Data Warehousing and OLAP, Bremen, Germany, November 4-5, 2005, Proceedings*, 2005, pp. 9–18.

[20] M. Golfarelli and S. Rizzi, "Expressing OLAP Preferences," in *Scientific and Statistical Database Management, 21st International Conference, SSDBM 2009, New Orleans, LA, USA, June 2-4, 2009, Proceedings*, 2009, pp. 83–91.

[21] S. Rizzi, "New Frontiers in Business Intelligence: Distribution and Personalization," in *Advances in Databases and Information Systems - 14th East European Conference, ADBIS 2010, Novi Sad, Serbia, September 20-24, 2010. Proceedings*, 2010, pp. 23–30.

[22] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2000.

[23] S. Sarawagi, "Explaining Differences in Multidimensional Aggregates," in *VLDB'99, Proceedings of 25th International Conference on Very Large Data Bases, September 7-10, 1999, Edinburgh, Scotland, UK*, 1999, pp. 42–53.

[24] ——, "User-Adaptive Exploration of Multidimensional Data," in *VLDB 2000, Proceedings of 26th International Conference on Very Large Data Bases, September 10-14, 2000, Cairo, Egypt*, 2000, pp. 307–316.

[25] C. Ordonez and Z. Chen, "Evaluating Statistical Tests on OLAP Cubes to Compare Degree of Disease," *IEEE Transactions on Information Technology in Biomedicine*, vol. 13, no. 5, pp. 756–765, 2009.

[26] A. Giacometti, P. Marcel, and E. Negre, "Recommending Multidimensional Queries," in *Data Warehousing and Knowledge Discovery, 11th International Conference, DaWaK 2009, Linz, Austria, August 31 - September 2, 2009, Proceedings*, 2009, pp. 453–466.

[27] A. Giacometti, P. Marcel, E. Negre, and A. Soulet, "Query Recommendations for OLAP Discovery-Driven Analysis," *IJDWM*, vol. 7, no. 2, pp. 1–25, 2011.

[28] J. Aligon, P. Marcel, and E. Negre, "Résumés et interrogations de logs de requête OLAP," in *Extraction et gestion des connaissances (EGC'2011), Actes, 25 au 29 janvier 2011, Brest, France*, 2011, pp. 239–250.

[29] O. Romero, P. Marcel, A. Abelló, V. Peralta, and L. Bellatreche, "Describing Analytical Sessions Using a Multidimensional Algebra," in *Data Warehousing and Knowledge Discovery - 13th International Conference, DaWaK 2011, Toulouse, France, August 29-September 2,2011. Proceedings*, 2011, pp. 224–239.

[30] J. Aligon, M. Golfarelli, P. Marcel, S. Rizzi, and E. Turricchia, "Mining Preferences from OLAP Query Logs for Proactive Personalization," in *Advances in Databases and Information Systems - 15th International Conference, ADBIS 2011, Vienna, Austria, September 20-23, 2011. Proceedings*, 2011, pp. 84–97.

[31] J. Aligon, H. Li, P. Marcel, and A. Soulet, "Towards a logical framework for OLAP query log manipulation," in *PersDB 2012, 6th International Workshop on Personalized Access, Profile Management, and Context Awareness in Databases (invited paper)*, 2012.

[32] J. Aligon, M. Golfarelli, P. Marcel, S. Rizzi, and E. Turricchia, "Similarity measures for OLAP sessions," *Submitted*, 2012.

[33] P. Biondi, M. Golfarelli, and S. Rizzi, "Preference-based datacube analysis with MYOLAP," in *Proceedings of the 27th International Conference on Data Engineering, ICDE 2011, April 11-16, 2011, Hannover, Germany*, 2011, pp. 1328–1331.

[34] A. Giacometti, P. Marcel, and E. Negre, "A framework for recommending OLAP queries," in *DOLAP 2008, ACM 11th International Workshop on Data Warehousing and OLAP, Napa Valley, California, USA, October 30, 2008, Proceedings*, 2008, pp. 73–80.

[35] L. Berti-Equille, I. Comyn-Wattiau, M. Cosquer, Z. Kedad, S. Nugier, V. Peralta, S. S.-S. Cherfi, and V. Thion-Goasdoué, "Assessment and analysis of information quality: a multidimensional model and case studies," *IJIQ*, vol. 2, no. 4, pp. 300–323, 2011.

[36] N. I. for Standards and T. (NIST), "Text retrieval conference (trec) home page," http://trec.nist.gov/, June 2012.

[37] T. T. P. P. C. (TPC), "Tpc benchmark ds (tpc-ds): The new decision support benchmark standard," http://www.tpc.org/tpcds/, April 2012.

[38] C. Mishra and N. Koudas, "Interactive query refinement," in *EDBT 2009, 12th International Conference on Extending Database Technology, Saint Petersburg, Russia, March 24-26, 2009, Proceedings*, 2009, pp. 862–873.

[39] V. Peralta, D. Kostadinov, and M. Bouzeghoub, "Apmd-workbench: A benchmark for query personalization," in *Workshop on Contextual Information Access, Seeking and Retrieval Evaluation (CIRSE)*, 2009.

[40] G. Marchionini, "Exploratory search: from finding to understanding," *Commun. ACM*, vol. 49, no. 4, pp. 41–46, 2006.

PAPERS

Combining domain and business ontologies in a modular construction method: EDF study case

Kevin Royer*, Ladjel Bellatreche† and Stéphane Jean†
*EDF R&D, Moret sur Loing, France
†LIAS/ISAE-ENSMA, Poitiers University, Futuroscope, France
kevin.royer@edf.fr, {ladjel.bellatreche,stephane.jean}@ensma.fr

***Abstract*—Ontologies are seen by companies such as EDF[1] as an answer to the data heterogeneity problem. However the design and management of consensual and complete ontologies in new and evolving domains remain important challenges that must be tackled for the widespread adoption of this notion. In this paper we present the approach we have developed to overcome these issues in the context of a data warehouse project with EDF concerning the electric mobility development. We first propose an incremental and modular design of ontologies. Then we suggest to formalize not only the domain knowledge as an ontology but also the business processes followed to exploit this knowledge. As a result, this method allows engineers to use the ontology as an interface for manipulating heterogeneous data as well as for reusing existing business processes. Finally we report the results and benefits of our approach for the project conducted with EDF.**

Keywords*—*Domain Ontology, Business Ontology, Study Case

I. Introduction

Initially defined by Gruber as an explicit specification of a conceptualization [1], the notion of ontology has been widely studied by the research community. As a result, standard ontology languages such as RDFS [2], OWL [3] or PLIB [4] have been defined. These ontology languages are equipped with tools to edit ontologies (e.g., Protege) [5] and to manage large amount of ontology instances (e.g, Oracle Semantic Technology [6] or OntoDB [7]). Thanks to these developments, ontologies are currently spreading into the industrial world. They are considered as practical solutions to the problem of data heterogeneity which is an important challenge for companies as most of them evolve in distributed settings.

If companies such as EDF are convinced of the interest of ontologies, their usage in an industrial context raises several challenges. One of them is the design of ontologies in new and evolving domains. This is particularly the case of the electric mobility domain in which EDF is an important actor. Indeed, during the past decade there has been a huge development of the electric vehicle (EV). This evolution is closely studied by EDF, as it is an electricity producer and supplier. The electric mobility domain concerns everything linked to EV, such as infrastructures, events, equipment, partners, services, etc.

EDF is involved in multiple EV experiments, with multiples partners across Europe, which consist in using and testing various equipment. We can cite the Kleber experiment in Strasbourg with Toyota, the SAVE experiment with Renault near Paris, or the ongoing trans-boundary CROME experiment between Germany and France. The aim of EDF is to design a data warehouse to gather the data of those experiments. As the data heterogeneity is important in those experiments, it suggests to use ontologies. However, as this domain is new and evolving, reaching a consensus among the different partners is difficult. To solve this problem we propose in this paper an incremental and modular ontology design methodology. Moreover, as the same data analyses are frequently performed on the data warehouse, we propose to formalize them in a *business process ontology* to reuse this knowledge. Then, the business process ontology is linked to the domain ontology and queries are written on the business process ontology to analyze the data. The interest of our approach is shown by describing the results of our project with EDF and the benefits they have obtained from this experiment.

This paper is organized as follows. In Section 2 we present the basic notions needed to understand our proposition. We also discuss the related work and introduce a motivating example. Section 3 presents our proposed design methodology for the incremental construction of domain ontologies. Section 4 describes our approach to define the business process ontologies linked to the domain ontologies. Then section 5 details the result of our project with EDF. Finally section 6 concludes this paper and introduces future work.

II. Preliminaries

A. Ontology Notion and OntoDB/OntoQL

As an extension of the definition of Gruber, we define an ontology as a *formal and consensual dictionary of categories and properties of entities of a domain and the relationships that hold among them* [8]. This definition highlights three main characteristics of an ontology:

- *Formal:* an ontology is based on a formal theory.
- *Consensual:* an ontology is shared by a community.
- *Capability to be referenced:* each concept of an ontology has a universal identifier.

Two main types of ontology are distinguished. While *linguistic ontologies* describe a domain as the set of terms that are used in this domain and the relationships between these terms (e.g., synonymy), *conceptual ontologies* are composed of classes and properties organized in a hierarchy [4]. In our

[1]EDF: Électricité De France, the French major electricity supplier.

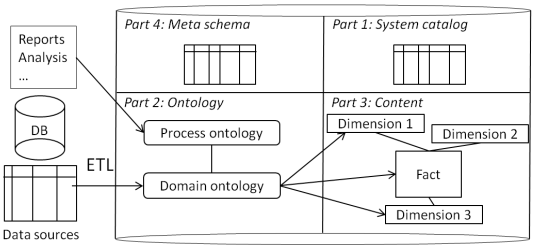


Fig. 1. Our proposition implemented on the OntoDB architecture

work we focus on conceptual ontologies that can be defined with most ontology languages such as RDFS, OWL or PLIB.

Specialized databases have been designed to store ontologies and their instances. We use OntoDB in our work [7]. Its architecture, depicted in Figure 1, is composed of four parts. Part 1 is the *system catalog*, part 2 (*ontology*) and 3 (*content*) store respectively the ontologies and their instances and part 4 (*metaschema*) records the ontology model into a reflexive meta model (description of concepts and properties, available data types, supported languages, etc.). OntoDB is equipped with the OntoQL language to manage these parts. It keeps a syntax close to SQL.

B. Related Work

In our project with EDF we had to design an ontology for the electric mobility domain. Two recent surveys have been made on ontologies construction methodologies [9], [10]. We have identified the following methods to be the closest to our needs.

The first method is the older one: a direct construction of a global ontology [11], [8]. These *global methods* require to describe the whole domain at once. As our domain is rapidly evolving, this kind of method is not adapted to our needs. Even if we proceed by an iterative process, this global method would require time to reach a consensus on new concepts and to revise the older ones.

The second closest method consists in using a common vocabulary allowing each contributors to build a local ontology on the shared vocabulary [10]. This approach permits to quickly build and work with an ontology. However its main drawback is that the shared vocabulary does not ensure that the local ontologies build on it will be compatible. As a result, some work are required to ensure the compatibility of local ontologies. Without it, the heterogeneity issue will remain.

As we can see, the proposed ontology design methodologies are not adapted to recent and evolving domains. Moreover, there are numerous work on domain ontologies but few on business process ontologies. As a consequence, we propose a method to facilitate the creation of an ontology and to make it less time consuming. Then we suggest to complete this method by building a process ontology connected to the domain ontology. With this connection the two ontologies allow engineers to capitalize knowledge efficiently in a data warehouse. Figure 1 illustrates this proposition, in part 3 the instances are stored according to the star schema.

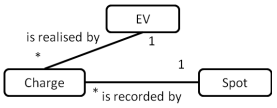


Fig. 2. Domain ontology sample

C. Motivating Example

To demonstrate the industrial interest of our method we propose a concrete case study. During an experiment involving the deployment of vehicles, EDF wants to store data generated by vehicles and charging spots. If we want to study a driver behavior, we need to record every charge done by his vehicle on any charging spot. A *charge* consists in charging the *EV*'s battery which is recorded by the charging *spot*. This description leads to the ontology depicted in Figure 2. For readability, the properties of the classes are not shown. The *charge* class is described by several dates (e.g., *connection date* or *starting and ending charge dates*), energy consumption data (e.g., *number of phases* or *charging speed*) and other information (e.g. if it is *booked*). The *EV* is mingled with its battery so we use only one class to represent it. This concept is described with several properties such as the battery's capacities or the types of charging modes accepted (fast or slow). The *spot*'s properties concern location, plugs available, etc.

The described ontology gives an extract and a high level view of the electric mobility domain. However, to define a complete ontology on this domain we need to describe precisely the different types of vehicles, charges and spots. The main difficulty is that these domains are quickly evolving and numerous vehicles, charges and spots can be added, changed and removed. Thus we have designed a particular ontology design methodology to support these changes.

III. Proposed Ontology Design Methodology

Instead of waiting for a general consensus on a domain, our methodology proposes to build a first draft of an ontology and to design the associated tools so that changes in the ontology can easily be taken into account. Moreover, as this domain is composed of different fields of expertise, we propose to decompose the ontology in several parts called *bricks* that can be switched if necessary.

A. Modularity: The Notion of Brick

A brick is a set of classes and properties that describes a field of expertise in a given domain. Each brick can be designed without reaching a consensus. Thus, several bricks can be defined on the same field of expertise by different experts. As a consequence, we view a brick as a local ontology in an integration system. Once the different bricks on the domain have been defined, they are combined to define the global domain ontology. The consensus that must be reached consists in deciding which bricks must be combined to obtain the global ontology. Then each brick can be extended or switched with another brick designed by another expert as long as it keeps the relationships with the other bricks.

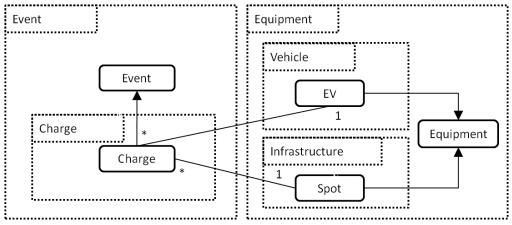


Fig. 3. Defining bricks

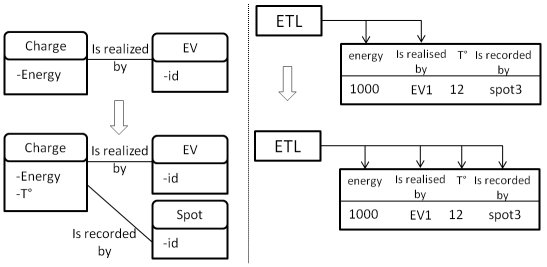


Fig. 4. ETL flexibility at work

As our objective is to be able to easily switch between different bricks on the same field of expertise, we have designed two types of bricks:

- *high level bricks* represent abstract concepts on the domain that are considered to be stable, they may contain other bricks. The link between the different fields of expertise should be done at this level if possible.
- *low level bricks* contain the most specific concepts of the ontology. The main classes of a low level brick must inherit from the classes of higher level bricks.

Example. Figure 3 presents the decomposition of our previous ontology example (see Figure 2) in several bricks. This ontology is composed of two main high level bricks: *Event* and *Equipment*. The *Event* brick defines a generic class *Event* and contains a low level brick *Charge*. The *Charge* class inherits from the class *Event*. In the same way, the *Equipment* brick defines the generic class *Equipment* superclass of *EV* and *Spot*.

In this example, the *Charge* brick is composed of a unique class. But if an equipment partner is not satisfied by our brick, he can add another concept in this brick such as *electrical properties* (e.g., *charging current* or *energy voltage*) because he may have a more precise definition of a charge. To use the rest of the ontology with his brick the only operation he has to do is to switch the actual brick with the one of his own and ensure that the connections with the other bricks are correct.

B. Brick Organization

If we consider a whole project that uses the proposed design methodology, the result is a collection of bricks. This collection should be organized regarding the domains concerned by the ontology from an engineering point of view. By combining the brick-design and the business knowledge of the domain each stakeholder of the ontology design process can propose bricks from their field of expertise. Each stakeholder can develop the bricks he needs sharing the construction only with a stakeholder that uses the same bricks. The flexibility of our design methodology relies on the fact that any brick can be either switched with an other brick or specialized if needed as long as relationships between bricks are maintained. Next section describes how we support this flexibility.

C. Supporting the Modularity

1) Ontology Consistency: Having a modular design allows experts to use a more accurate local ontology on their domains of expertise. To ensure the consistency of the ontology after modifying some bricks we propose the following mechanism. Each brick contains one main concept which defines the links with other bricks. *Vertical* links between a main brick's concept and the bricks within are ensured by using the inheritance mechanism. *Horizontal* links between distinct bricks are defined by the brick's main concept which acts like a gateway.

2) Flexible ETL Process: In order to provide a generic ETL we have designed it so that changes in the ontology are dynamically exploited: *for each mapped ontological class load the available properties* and *If a property is a relation then request the primary key*. With this process, the ontology can be changed without modifying the ETL.

To formalize the ETL process we use the algebra proposed in [12] and more precisely, the following operators.

- *Retrieve*: $Source \rightarrow RawData$. It retrieves raw data from the sources.
- *Convert & Join*: $RawData * Ontology \rightarrow Instances$. The raw data (e.g., CSV or XML files) are converted into class instances by matching the raw data properties with the ontological class properties.
- *Filter*: $Instances * Ontology \rightarrow Instances$. The instances obtained by the previous operator are filtered using the knowledge described in the business process ontology (described in section IV).
- *Store*: $Instances * Ontology \rightarrow TargetedSGBD$. This operator generates the OntoQL statements required to insert the selected instances.

As our ETL process takes as input the domain ontology, we have enriched this algebra with the operator *Retrieve meta data* to retrieve the ontology from the OntoDB platform and the source's scheme. The ETL process starts by retrieving the source's data model before retrieving the data. Then the ETL reads the ontology and stores it in its cache. Thus, the ETL process is not defined on a hard-coded ontology.

Example. Figure 4 illustrates the ETL capacities. In the first case the ETL has to read the *energy* value. *Energy* could have been a synonymous known by the ontology such as *chargedEnergy*. For the reference to an *EV*, the ETL will request the *EV*'s table for the corresponding primary key. In the second case, if we need to perform an infrastructure analysis, we add in the ontology the temperature (T) property and a reference to the spot where the charge happens. As the ETL reads the modified ontology, it can automatically retrieve the corresponding data.

IV. Business Process Ontology

A. The Need of a Business Process Ontology

In the project with EDF, the goal is not only to collect data for maintenance and supervision purposes but also to get a feedback on EV usage such as the location and period of peak demands or the reliability of the supply network. This feedback is given by engineers who perform data analyses.

By studying the current practice, we have observed that the same analyses are performed on different dataset as EDF is involved in several experiments. Moreover, as the electric mobility domain is emerging, engineers are specifying new indicators to understand and describe EV usage. Thus the knowledge used by engineers for crafting an analysis is built on personal experience and insights. As this knowledge is not recorded, engineers do not share their experience on the indicators that should be used to answer a specific business question. Yet, engineers may need to reuse indicators developed by others experts on the same sort of dataset. Moreover, there is a risk to lose this knowledge if an engineer leaves its expert role. As a solution to these problems we propose to create a business process ontology to automatize business analysis and to capitalize and share engineers personal knowledge.

B. Business Process Ontology Design

The business analyzes conducted by engineers on the electric mobility domain are rapidly expanding and evolving. As a consequence, we have adopted the same design methodology to build the business process ontology as the one used for the domain ontology. Thus the business process ontology is composed of a set of bricks. A high level brick *Knowledge* includes all the other bricks. This brick is decomposed in several sub-bricks that characterize the different forms of analyses: sorting treatments, indicators, pattern recognition, statistical reports, etc. Low level bricks describe specific analyzes tasks by defining their computation, business concepts and entities concerned by it, etc.

Example. One of the main issue of the electric mobility concept is that it contributes to the energy peak demand around 7 pm when everybody is getting home. The EV usage is a new source of energy consumption that must be added to the peak demand. Therefore, daily energy consumption is a key indicator that should be studied. To compute this indicator we had to ensure data quality, which led us to define sorting knowledge. Figure 5 illustrates this part of the knowledge brick.

The knowledge ontology is essentially used for description and sharing, the indicators are computed without the ontology with views, or batch, and loaded with the ETL. We have represented *sorting*: the physical one describes, for example, the electrical values of a charge (positive energy, mean power value, etc.) and the quality one, the relations of the charges with other concepts; the daily load curve is our main indicator built on a selection of charges (depending on the season, drivers, infrastructures, etc.). Its computation is simple, for each charge we determine the mean power: P_m used between the charge start date t_S and the charge end date t_E. Then we split a day into ten minutes intervals, and we increment the intervals between t_S and t_E by P_m.

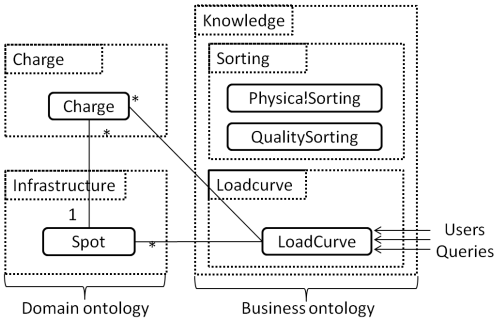


Fig. 5. Knowledge bricks sample

C. Link between Domain and Business Process Ontologies

The connection between the domain and business process ontologies is rather intuitive. The business ontology describes what is needed to generate an analysis including the data. The data needed is described by the domain ontology. So we can link the knowledge concept to the domain concepts it requires. Now the two ontologies are connected as shown in Figure 5.

In the EDF project, the domain and business process ontologies are used as follows. If an engineer identifies a useful indicator, (s)he describes it in the business process ontology and links it to the domain ontology concepts. Then anyone working on these concepts will see an annotation referring to the knowledge brick. This process permits a better knowledge sharing and it pushes engineers to formalize their knowledge. It also encourages engineers to be active contributors to the ontology, the brick design allowing them to easily integrate new indicators and/or concepts.

V. The EDF Project Experiences and Impact

A. Data Warehouse Implementation

As said previously, we have chosen the OntoDB platform for implementing our data warehouse. We have made this choice for two main reasons (1) ontologies can be stored in OntoDB and (2) the meta-schema part of OntoDB can be used to manage ontologies defined with different ontology languages. We have used OntoDB as depicted in Figure 1. The domain and business process ontologies are stored in the ontology part while the data are stored in the content part. A star schema is used in the content part where the fact tables store all the events (e.g., *EV charges* or *energy consumption*, etc.) and the dimensions are picked in the high level bricks of the ontologies (e.g, *Vehicle*, *Spot* or *Location*).

The OntoDB platform uses PostgreSQL as the underlying database management system. Thus we have been able to use SQL views to compute simple indicators defined in the business process ontology. For complex indicators, like sorting, we have implemented them with batch programs. New analysis to perform will first appear as concepts in the business process ontology. The implementation will differ depending on the

analysis: adding a dimension or a fact table with the current, or new, dimensions for a new analysis, expanding a dimension or a fact table with new properties.

B. Complete Use Case

1) Goal: EDF is an electricity supplier and charging an EV is a new electricity usage so, naturally, EDF is studying the different types of user behaviors to define new adapted contracts. The goal of this analysis is (1) to establish mean user profiles and (2) to identify the type of users that contributes the most to the consumption peak.

2) Data sources: the analysis core concerns the charges. They are extracted from monthly reports (CSV files) provided by our partners responsible for providing equipment and gathering data. These files provide several attributes for a charge such as the charge number, the charged energy, the charge starting and ending date, the vehicle id, etc.

3) Ontologies: the high level bricks of the domain ontology contain several concepts corresponding to our data sources: the charged energy, the charge starting and ending date and the vehicle id (EV concept). A 1-1 mapping is established between these concepts and the corresponding attributes in the data source. This mapping can be implicit if the attribute and the property in the ontology have synonymous names.

4) The ETL process: it reads the ontology, searches for all concepts mapped to the data sources and loads the (possibly transformed) data in the tables corresponding to these concepts. Only the attributes mapped to the ontology are loaded in the data warehouse. But we can extend the ontology. For example, if we decide to extend the definition of a charge with the property *air temperature* to study weather impact on charging speed, the ETL will automatically retrieve the corresponding data and data from older sources will be updated. Once the sources describing the EV fleet are loaded, we dispose of the necessary row data for our study.

5) Business analyzes: the main indicator for this study is the normalized mean vehicle load curve (virtual typical day). This indicator is built for each vehicle on its charges (see Figure 6 for an example) and eventually on a selection of spots.

This indicator is described in the business process ontology so the OntoDB platform automatically creates a table corresponding to it. This description includes the name of batch programs that must be launched to fill the indicator table with the new data. At the end of this batch process we have access to the indicator with a simple request: *Select * From LoadCurve*. With these data, we can draw the mean daily usage depicted Figure 6 (X axis represents a day and Y axis the charging power). This figure also displays the national electricity consumption curve from RTE[2]. Both curves are normalized for a better shape comparison.

With this knowledge, we can finally reach our goal by launching machine learning algorithms such as K-means on the indicators to identify homogeneous groups of users. The result of this step is a set of groups of users and mean profiles. These profiles and the list of users attached to it constitutes a high level indicator which will be recorded in the business intelligence brick.

[2]RTE: Réseau de Transport de l'Électricité, the French electricity distribution network operator.

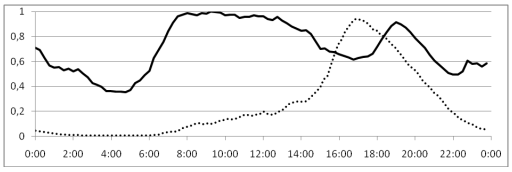


Fig. 6. Plain line: 02/14/2014 French electricity production; Dotted line: an homogeneous group of user daily load curve

An other interesting result is the identification of the groups of users that need a special focus regarding contract definition. For example in figure 6 the consumption peak occurs in a low production activity which is satisfying. But about one third of the peak surface does contribute, which can be problematic (cold weather, sudden massive energy demand...). Also, we observe that EV are fully charged around midnight. As this group is composed of companies, we can assume that EVs are not used between 12am and 6am. EDF might be able to propose a specific incentive contract to make this group use shifted charges.

6) Trans-disciplinary analyzes: EDF teams from other discipline have conducted interviews with users and fleet managers. In order to create some elements for contract creation, we cross our results with them. As we dispose of a precise list of users belonging to the 'problematic' groups, we were able to interpret the profiles we had. This interpretation gave EDF precious insights on which incentive they could propose to the users through contracts.

7) Knowledge integration: after the trans-disciplinary work we have enhanced our business intelligence bricks with more indicators to be more accurate. And we are able to generate them on every incoming data. This is possible thanks to our ontology design methodology.

C. Benefits of our Approach

1) Comparison with previous methods: before our project, EDF had several dedicated databases, one for each experiment. This situation was the result of the sources heterogeneity from one experiment to another. For example, smaller experiments used spreadsheets while bigger experiments used databases. As a result the sorting processes had to be defined and executed for each experiment. Our work helped to improve this situation. For older experiments we had to define mappings between the sources and the ontology but for the upcoming experiments, sources will comply to the ontology. Thus, our approach provides a simplified and unified data warehouse with a single flexible ETL process (see Figure 7).

The same problematic situation occurred at the business process level. The analyses conducted on data had to be developed for each experiment. This work was expensive and tiresome. Moreover, the knowledge generated during these experiments were only recorded as reports. As complete as these reports were, the re-utilization of this knowledge represented a lot of work. Using our approach, the business processes are recorded

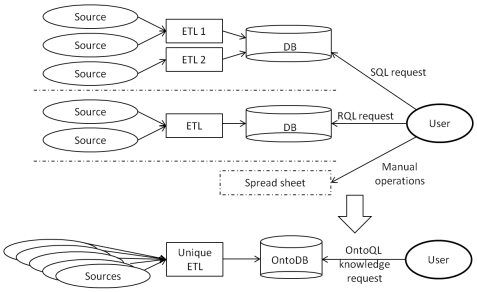


Fig. 7. Request simplification

in the data warehouse and the computation of an indicator is fully automated. Finally, the communication between research teams was not sufficient as it is not reasonable to gather all teams frequently. Therefore the cross analyzing results could not be as relevant as expected. In our approach, as every team describe their results according to the ontology, the communication between these teams is improved and it's easier to cross the results using OntoQL queries.

2) Industrial's benefits: our approach is an interesting solution for companies, like EDF, where engineers are not all familiar with database management systems. It allows several teams working on different aspects of a domain to share and capitalize knowledge and sources descriptions. In our project, several teams are working on the electric mobility domain: energy engineers, network specialists, sociologists, etc. and each of them are benefiting from the other. For example, crossing sociologists analyses with infrastructure reports showed interesting results. Some public charging spot weren't used, EV usage analysts thought that was the spots fault (plugs problems, unclear indications about the charge). In fact, people were not aware of the very existence of those charging spots.

3) Engineers' benefits: an engineer forced to re-write the same request for different logical models is a waste of time and money. Or when an engineer perform an analysis without knowing that an other already did. Moreover, the mastery of different languages and database management systems can be a burden for engineers. In our approach, the ontology is the only interface between engineers and the database. And only one requesting language is used (OntoQL). Its mastery is simplified because this language is close to SQL and because it uses the ontology terms. Thus, as the ontologies are made by engineers on their fields of expertise, they can manipulate a complex storage system with simple requests. At the current state of the project, engineers role only consists in maintaining the ontology and they can now focus on developing new knowledge and adding new data sources.

VI. Conclusion

The context of the work presented in this paper is a project with EDF aiming at designing a semantic data warehouse for the electric mobility domain. We have identified that existing ontology design methodologies are not adapted to this domain as it is rapidly evolving and expanding. As a consequence, we have designed a modular and incremental ontology design methodology. This methodology consists in decomposing the domain in several *bricks* representing specific fields of expertise. The design of a brick does not require a consensus with the other partners. The global ontology is then obtained by composing a set of bricks. The only consensus that must be reached consists in choosing the bricks, that are used knowing that the bricks can then be switched or extended if necessary. Observing that the business process analyzes are often the same, in addition to the domain ontology, our methodology suggests to design a *business process ontology*. This ontology records the different indicators used during a business analysis and their computation methods. As these indicators are also evolving, the business process ontology is defined, as the domain ontology, with a set of bricks. The link between this two ontologies consists in defining the required data for a given analysis.

The proposed methodology has been successfully applied to our project with EDF. The benefits identified during our project are: (1) to save a lot of engineer time which is a strong argument for a company (2) to ease the communication between trans-disciplinary teams by using the ontologies as an interface and (3) to be able to quickly extend or modified the ontology and (4) to simplify the exploitation of the data warehouse by using a single language that uses the terms of the ontologies. As future work, we plan to build new tools to support our design approach.

References

[1] Gruber and Thomas, "A translation approach to portable ontology specifications," 1993.

[2] D. Brickley and R. V. Guha, *RDF Vocabulary Description Language 1.0: RDF Schema*. W3C, 2004.

[3] S. Bechhofer, F. v. Harmelen, J. Hendler, I. Horrocks, D. L. McGuinness, P. F. Patel-Schneider, and L. A. Stein, *OWL Web Ontology Language Reference*. W3C, Feb. 2004.

[4] G. Pierra, "Context representation in domain ontologies and its use for semantic integration of data," *Journal Of Data Semantics (JODS)*, vol. X, pp. 34–43, 2007.

[5] "Protege." [Online]. Available: http://protege.stanford.edu

[6] Z. Wu, G. Eadon, S. Das, E. I. Chong, V. Kolovski, M. Annamalai, and J. Srinivasan, "Implementing an inference engine for RDFS/OWL constructs and user-defined rules in oracle." ICDE, 2008.

[7] H. Dehainsala, G. Pierra, and L. Bellatreche, "OntoDB: an ontology-based database for data intensive applications," in *DASFAA*, ser. Lecture Notes in Computer Science, vol. 4443. Springer, 2007, pp. 497–508.

[8] S. Jean, G. Pierra, and Y. A. Ameur, "Domain ontologies: A database-oriented analysis," in *WEBIST 2005 and WEBIST 2006. Revised Selected Papers*, 2007, pp. 238–254.

[9] M. Bergman, "A brief survey of ontology development methodologies," Aug. 2010. [Online]. Available: http://www.mkbergman.com/906/a-brief-survey-of-ontology-development-methodologies/

[10] R. Subhashin and J. Akilandeswari, "A survey on ontology construction methodologies," *International Journal of Enterprise Computing and Business System,(Online)*, vol. 1, no. 1, 2011.

[11] N. F. Noy and D. L. McGuinness, "Ontology development 101: A guide to creating your first ontology," Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880, 2001.

[12] N. Berkani, L. Bellatreche, and S. Khouri, "Towards a conceptualization of ETL and physical storage of semantic data warehouses as a service," in *Cluster Computing*, 2013.

Data Modeling Styles in Data Warehousing

V.Jovanovic[*], D. Subotic[**], S. Mrdalj[***]
[*]Georgia Southern University, Statesboro, USA,
[**]University of Rijeka/Department of informatics, Rijeka, Croatia
[***]Eastern Michigan University, Ypsilanti, USA
vladan@georgiasouthern.edu, dsubotic@inf.uniri.hr, stevan.mrdalj@emich.edu

Abstract - The paper presents a coordinated set of data modeling styles relevant for data warehouse design in the context of relational databases. The scope of presented models covers: a) entity relationship models of existing relational DB sources, b) logical data vault model for integrated enterprise data warehouses as a system of records, c) dimensional fact model for analysis leading to query prototyping and dimensional models, and d) dimensional model for data marts. There is no claim of sufficiency, uniqueness, and/or universality of the selected styles set except utility. The principal contributions of the paper are: definition of data modeling styles as distinct modeling mechanisms, and initial coordination of selected complementary styles.

I. INTRODUCTION

The scope of our interest regarding data modeling in data warehouse (DW) design encompasses following practices: modeling of existing database (DB) sources, modeling enterprise data warehouse (EDW) as system of records (back-end DW tier used for compliance and for feeding presentation side of DW i.e. front-tier data marts), dimensional modeling for data marts (DM), as well as dimensional fact modeling (DFM) in user requirements driven analysis of data relationships aimed at exploring querying requirements. The last practice fills an important and frequently missing link, valuable in preparing detail star shaped models for DM.

From a general viewpoint, any system that can be described as a set of observable entities, their relationships and observable properties of those entities and relationships, can be represented by a DB organized according to system's entity relationship (ER) data model.

What is a Data Modeling Style? A data modeling style can be defined by patterns and constraints that produce models of recognizable form [18]. A style is distinct from a modeling language (and/or its visual formalism). A number of languages (notations) for entity-relationships (ER) modeling is in widespread use. Representative sources on using ER modeling in DB design/integration context are: for IDEF1X [5], for UML ([9], [24]), and for some other ER variations ([3], [25], [2], [8], [26], [27] and [28]). We do not consider notational variations to be essential in defining a data modeling style, and clearly separate modeling styles from modeling techniques [15].

Integrating modeling styles means at the minimum: keeping complementary models directly representable in common format (we selected relational DB and accomplished it by restricting IDEF1X notation), and amenable to model transformations. The paper presents complementary styles whose integration is in progress. The Figure 1 represents precedence among selected styles. The styles are intended to model: existing DB sources, integrated EDW, dimensional hierarchies in detail, and Data Marts. The three models (ER for DB, DV for EDW, and dimensional model for DM) are intended for realization using relational DB engines. The aim of the fourth model, the DFM, is to improve design, verification, and exploitation processes.

A visual observation of four kinds of models, side by side with a relational model (relational DB engines provide typical implementation of DB, EDW, and DM) provided initial motivation for considering it not only as a set of complementary representative models but as a coordinated set of modeling styles forming a grounded basis for a novel methodology of comprehensive DW design. Along the same line of observations is a longer term goal, an expectation of elaboration of formal reasoning leading to generalization and automation of transformations between such models (using relational model as an underlining basis). We expect to achieve acceleration of data integration/mapping during the user requirements analysis activities by a comprehensive model-based design using four complementary styles.

Our aim, in this paper, is to present those four complementary styles by an example that would set a stage for a fuller integration. Organization of the paper follows selected styles. The section II introduces our baseline example and an ER style for representation of existing data sources in a relational DB. The section III illustrates DV style for EDW based on integrating existing sources, the section IV introduces DFM style for conceptualization of user requirements, and section V dimensional model for DMs. The Dimensional Fact Model and the Data Vault model are relatively less known and thus will require more justification and basic illustrations.

A business case we will use in this paper deals with a DW to be designed in support of rapid introduction of 3D printing in discrete manufacturing (with a high rate of change in assembly specifications). A general expectation is that by providing analytical capabilities of DMs and a compliant permanent system of records, a desired DW solution can help with early adoption of 3D printing. Investment in technology is time sensitive and to recoup costs support from a contemporary DW system is needed.

All data model examples are shown as diagrams in the case tool ERwin 9.5 using appropriate style-related variations based on the IDEF1X (see Figures 2 to 7). Note also that even simplified examples of DBs (Figure 2) are structurally rich configurations providing an opportunity to fairly illustrate selected modeling styles as a complementary set. The rest of the paper elaborates, in a separate section, each of the four styles, all followed by a conclusion in a form of summary of research directions.

II. RELATIONAL DB ORIENTED ER STYLE

The example in Figure 2 shows a logical ER model (i.e. no data types and sizes from a source DB are shown) to simplify the analysis and design work in the paper. The example uses two data sources, a Sales DB and a fairly simplified Manufacturing Design DB. The Sales DB is a classic sale of ITEMs to CUSTOMERs placing ORDERs for multiple items (one per ORDERLINE). Probably more interesting is a DB focused on a Bill of Material (BOM). The assembly specification BOM_PART, is a resolution of a Many to Many relationship between a PART used (in assembling) and a one (to be assembled) i.e. ASSDEMBLED_PART.

To reverse engineer DB schemas closeness of IDEF1X to the relational model (RM) is an important advantage, and is also helpful in explaining models to developers. Here is a simple demonstration, just by looking at Figure 2 box-frames representing Entities, one can directly observe the intended content of every relational table. Specifically, primary key (PK) area is visually separated, all foreign keys (FK) are noted as such and traceable via relationship lines to related entities (tables). Role names for primary keys (PK) serving as FK are used only when necessary (assembledItem, instead of partID, in ASSEMBLED_PART). The redundancy of showing foreign keys (FK) as attributes is an advantage during model reviews. Subtypes are either complete or incomplete specializations, in intentional sense, i.e. one instance of a PART can be PRINTABLE or not. An incomplete case is represented by ASSEMBLED-PART and MATERIAL implying the possibility of additional sibling subtypes, and a complete case by PRINTABLE. In IDEF1X one entity type (i.e. PART) can have several independent criteria for specialization ('type' and 'printable') and all subtypes within the same specialization criteria are mutually exclusive i.e. one instance of a PART can be either an ASSEMBLED_PART or a MATERIAL, or neither. In ER modeling, as a rule, we eliminated all M:M relationships and as a result all ER IDEF1X models are directly representable in the RM (DDL prototyping using case tools is automatic). Our models use intersection entities (instead of M:M relationships) between a pair of entities, with two 1:M relationships depicted using solid lines representing identification dependencies (BOM-PART). As a pragmatic rule names of entities are shown in uppercase i.e. PART.

For each of the remaining styles a visual Template can be defined (describing relevant IDEF1X restrictions and convenient visualization conventions). In this paper we will only give some textual specification as examples models will have all visual elements displayed.

III. LOGICAL DATA VAULT MODELING STYLE

The ER style data modeling practice in a DB context (as well as dimensional modeling style in the context of DM design) are well established, but a few arguments can help make a case for the DV style now in a DW context.

First, it clearly makes sense to separate concerns and design data loading oriented EDW as a system of records for compliance (preservation of history and provenance i.e. to enable tracking the lineage of data back to the source) as logically and physically separate from the user query heavy, analysis and presentation oriented DMs.

Second, a DV approach [21],[22],[16], [10] and [23] was developed as an approach to a sustainable EDW design, and represent a style of modeling for DW that can be characterized as dependencies minimization aiming at flexibility and performance. The main advantages of DV over traditional 3NF EDW designs, in a style advocated by ([12], and [11]), are:

a) Inserts, deletes, or updates of rows are implemented only as additions (nothing ever get lost/overwritten),

b) Structural changes of and in data sources results in model expansion, principally by new links and without structural reconstruction of existing DW elements (a holy grail of architectural stability), and

c) Enable rapid parallel data loads [22].

The DV modeling style is recognizable by its major concepts i.e. Hubs, Links and Satellite entities. In terms of entity relationship models a Hub entities holds identities of lasting entities (typically business keys and their surrogates), a Link entities are representations of foreign key references (to two or more business components i.e. Hubs). A Satellite entity holds history of attributes for a related Hub or a Link. In Figure 3 we illustrate use of a DV style template, in IDEF1X, with coloring schema of: blue background for Hubs (H_CUSTOMER, H_ORDER, H_ITEM), red for Links (SAME_AS, L_BOM, L_CO, L_O-OL_I) and yellow for Satellites (S_CUST_PHONE, S_QTY, S_PRICE, S_NAME, S_BOM_EFFECTIVITY etc.). Note that all DV entities are implementable as tables. All entities are shown with standard DV metadata (a **LDT**-load date time stamp, and a **source**, a unique id for data source system- typically a DB). Note that grouping of attributes into different Satellites is driven by record data source and frequency of changes, for example S_R1 (name, creditLimit) and the S_2 are grouping attributes expected to change with different frequencies.

The source data ER models are to be lightly integrated i.e. source data mapped to a DV without a loss of information. The DV modeling starts with the identification of stable business keys representing permanent identities for major business entities, and treats

them as Hub candidates. During this modeling step a soft integration can be performed, for example PART and ITEM from the two source systems (from Figure 2) can be recognized as somehow related (if they shared the business key than they would be represented as one Hub under some generalized name). In our example where several ITEMs for sale are referring to the same product (PART) a (SAME_AS) linking needs to be explicitly established. One of the expectation from 3D printing is to offer more products while reducing the number of parts, so we decided to use PRODUCT name for current (and future) PARTs. Once Hubs are selected, we have to model Links, and in case of existing DB sources we use FK as clues. It is also very important to eliminate dependencies and deal appropriately with subtypes in DV modeling. Originally subtypes were justified in ER models when only a subset of an entity instances could possess certain attributes and/or share a relationship with other entities. Both reasons disappear in DV, restricted attributes from a subtype form Satellites (S2_ASSEMBLED, S_3D). Also, in Figure 2, (ASSEMBLED_PART) subtype associated with relationship (BOM_PART) became a role of a Hub (H_PRODUCT) in a Link (L_BOM). Such elimination of subtypes in logical and physical DV modeling results in simpler models. Any M:M relationship, which we already represented in ER style with an intersection entity is a DV Link (L_BOM). Note that a history entity (PRICE_HISTORY) is really a DV Satellite (S_PRICE) for its Hub (H_ITEM) of the base entity (ITEM). The logical DV modeling deals with composition (part of) dependencies as well, and while some authors ([22], [10], [23]) differ in their recommendations, the best option seems to be the general one of severing dependencies and treating all dependent entities as Hubs (with concatenated keys) while passing Hub key to a Link down the dependency line i.e. H_ORDER in Figure 3 where hubs H_ITEM and H_OL form a link L_OL_I_O. Note that logical DV modeling follows pragmatic and data source oriented recommendations that differ from those recommended for conceptual DV modeling [17] in situations with no existing data sources.

Using DV allows incremental integration of data sources into DW, and at the same time simplifies loading of DW from multiple sources. The primary author of architecture DW 2.0, W. Inmon, recognized DV as a DW 2.0 compliant data model. The DV approach aims at preserving history of data in a system of records manner thus leaving complex transformations for downstream processing- towards DMs. We omit discussion of additional DV design steps in data integration (related to application of common business rules) here. Authors who recommend DV approach ([21], [10], [23]) as a solution to long term flexibility and performance problems of first generation DWs also accept dimensional models for reporting oriented DMs. Note that using DFM as a stepping stone to requirements oriented DMs design makes the job of designing extract-transform-load (ETL) from DV to DMs substantially easier.

IV. DIMENSIONAL FACT MODELING STYLE

While industry leaders and authors ([19], [1]) advocate direct use of dimensional modeling for DM design, we decided to follow agile practice [6] and analyze requirements beforehand specifically using a Dimensional Fact Model (DFM) for visualization. The DFM is a conceptual model created specifically to support DM design [7] where one DFM fact schema leads to one DM star schema. Furthermore, a documented DFM provides visual representation of detailed dimensional structure helpful in verifying that standard reporting requirements are captured, is quite useful in specifying dimensions in SQL DDL, and is of lasting value in preparing complex, multidimensional SQL queries by programmers. Each DFM is showing model facts as a set of measures in the context of dimensional attributes, with explicit hierarchical relationships shown, to facilitate expression of any available granularity of potential queries. Facts model set of 'business' process events (ordering) or possibly process/state specification (assembly). Two DFM schemas are shown, in Figures 4 and 5 respectively for orders and assembly. DFMs in our example are derived each from one of the source systems from Figure 2. A requirement to track price changes in time (per ITEM) is set aside as it can be answered from ORDERED fact.

We are treating Figure 4 as a template for drawing DFM schemas in IDEF1X for convenience only i.e. using visualization as a conceptual model not to be realized directly. Detailed derivation processes can use operations such as pruning, grafting etc. [7] but those steps are not shown in the interest of brevity and to maintain the simplicity of our case.

DFM schema Template recommendation for IDEF1X representation utilizes simultaneously three levels of presentation: fully attributed Entity level (i.e. ORDERED), Entity key level (i.e. CUSTOMER), and Entity only (i.e. year). A FACT entity show in attribute view with entity name in capital letters, i.e. ORDERED. Model shows all measures as attributes in lowercase i.e. as 'quantity' and 'price' in ORDERED. Names for each dimension are represented as root entities with its grain as the key attribute, (CUSTOMER, DATE, and ITEM), lists the name of the most frequent value i.e. lowest level of granularity as a root key attribute; this means roots are shown at key level presentation in a DFM diagram. For each dimension (i.e. DATE, ITEM, CUSTOMER) represent every attribute as an entity but with names in lowercase (month, year, day, week, etc.). Whenever relationships between attributes represents a (hierarchical) identification dependency we use solid line pointing to lower level, as between state and city (1 state to M-many cities). Whenever attribute to attribute association cardinality is 1:M or 1:1 we represent such relationships using dotted line, and in cases of optional 1:1 or 1:M relationships we use dotted line with a diamond icon (holiday to date).

V. DIMENSIONAL MODELING STYLE

The Dimensional Model is practically universally used for DM design presentation [19]. Mapping of DFM into DM is almost trivial, simply collapse i.e. de-normalize all

hierarchies into corresponding dimensions, as illustrated by Figures 5 and 6. The real world complication is standardizing such dimensions into conforming dimensions used as master references for otherwise incremental development, and as such is outside the scope of this overview. The dimensional modeling style is distinguished by star schema designs where fact (set of measurement), for a lowest granularity of data needed to be tracked, is surrounded with a rich context of dimensions. In practice there are some additional complications (snow flaking of dimensions, bridge tables etc.) but because dimensional modeling is very well represented in reliable sources such as ([19], [1]) we will not elaborate.

The DV approach delays complex transformations downstream, speeding up extracting and loading of data from the sources. A soft integration based on lasting business keys (hubs) leaves to user's views to match partial requirements in DMs. Having a clearly justified DM designs (via DFM) enables incremental ETL design amenable to automation given the standard patterns of DV [22]. One point of the paper is bridging the gap between an integrated DV and individual DMs, and it is where the four styles approach converge, enabling evolving solutions for DW.

VI. CONCLUSION

In this paper we presented a set of complementary data warehouse design styles which may enable well integrated DW solutions for relational DB implementations. Furthermore, this is achieved by using a set of data modeling styles based on a common notation (IDEF1X) and in a single design tool. We are researching opportunities for further integration including elements of automation of model transformations: a) from ER to DV in a manner of extract-load and minimally transform (ELT)- with only simple and reversible transformations, b) from ER to DFM (supporting graph manipulation operations for creative i.e. essentially semi-automatable requirements analysis), c) from DFM to a dimensional model (a de-normalization by collapsing attribute hierarchies into respective dimensions), and d) from DV to DM in the classic ETL manner (including now irreversible transformations). It is worth noting that modeling redundant data aggregations i.e. materialized views (MVs) in DMs using DFM to analyze and recommend (visualize) starting set of views is easy [7], but was not treated here especially as DB engines collecting usage statistics are already recommending MV automatically. The physical design aspect as well as end-user presentation forms such as scorecards and dashboards are also outside the scope of this paper. Some of the current limitations of the approach are also research directions for future work. Specifically needed are investigations of compatible modeling for sources of semi-structured and unstructured data, formalization of models and model transformations for highly distributed environments, and development and empirical evaluation of pertinent learning aids (essential in order to support dissemination of practices proposed). Several students are already engaged inspired by work of ([13], [22], [14], [4], [20], [16] etc.).

REFERENCES

[1] C. Adamson, Star schema- the complete reference, McGraw Hill 2010,

[2] C.Batini, S. Cheri, S. Navathe, Conceptual database design, 1992,

[3] R. Barker, CASE*Method: tasks and deliverables, Addison Wesley, 1989,

[4] A. Berson, L. Dubov, Master Data Management and Data Governance, 2nd edition, McGraw Hill 2011,

[5] Bruce T. (1992) Designing Quality Databases with IDEF1X Information Models, Dorset House,

[6] L. Corr, J. Stagnitto, Agile data warehouse design- collaborative dimensional modeling from whiteboard to star schema, Decision Press 2011,

[7] M. Golfarelli, S. Rizzi, Data warehouse design, McGraw Hill 2009,

[8] D. Hay, Data model patterns- conventions of thought, Dorset House 1996,

[9] D. Hay, Enterprise model patterns, Technics Publications, 2011,

[10] H. Hultgren, Modeling the Agile Data Warehouse with Data Vault, New Hamilton, 2011

[11] C. Imhoff, N. Galemmo, J. Gregor, Mastetring data warehouse design- relational and multidimensional techniques, John Wiley 2003,

[12] W. Inmon, Building the data warehouse, 4th edition, John Wiley 2005,

[13] W. Inmon,, D. Strauss, G. Neushloss, DW 2.0- the architecture for the next generation of data warehousing, Morgan Kaufmann, 2008,

[14] B. Jiang, Constructing data warehouses with metadata-driven generic operations, DBJ Publishing, 2011,

[15] V.Jovanovic, A. Gardiner, L. Cupic, Taxonomy of data modeling techniques, Proceedings of the Southern Association for Information Systems Conference, Jacksonwile 2007,

[16] V. Jovanovic, I. Bojicic, Conceptual Data Vault Model, Proceedings of the Southern Association for Information Systems Conference, Atlanta, 2012,

[17] V.Jovanovic, S. Bojicis, C. Knowles, M. Pavlic, Persisitent staging area models for data warehouses, Issuses in Information Systems, V13, N1, 2012,

[18] V.Jovanovic, S. Benson, Aggregated data modeling style, Proceedings of the Southern Association for Information Systems Conference, Savannah, 2013,

[19] R. Kimball, M. Ross, The data warehouse toolkit, 3rd ed. John Wiley 2013,

[20] R. Lans, Data Virtualization for Business Intelligence Systems, MK 2013,

[21] D. Linstedt, K. Grazianno, H. Hultgreen, The business of data vault modeling, 2nd edition, 2009,

[22] D. Linstedt, Supercharge your data warehouse, Create Space Independent Publishing, 2011,

[23] D. Mangano, The Integrated Data Hub- Next Generation Data Warehouse, 2013,

[24] R. Muller, Database design for smarties, MK 1999,

[25] A. Scheer, Enterprise-Wide Data Modeling, Springer-Verlag 1989,

[26] B Schmitdt., Data modeling for information professionals, Prentice Hall, 1999,

[27] L. Silverston, The data model resource book- volume 2, a library of universal data models by industry type, revised edition. John Wiley, 2001,

[28] G. Simsion, G. Witt, Data modeling essentials, 3ed., MK, 2005.

Appendix

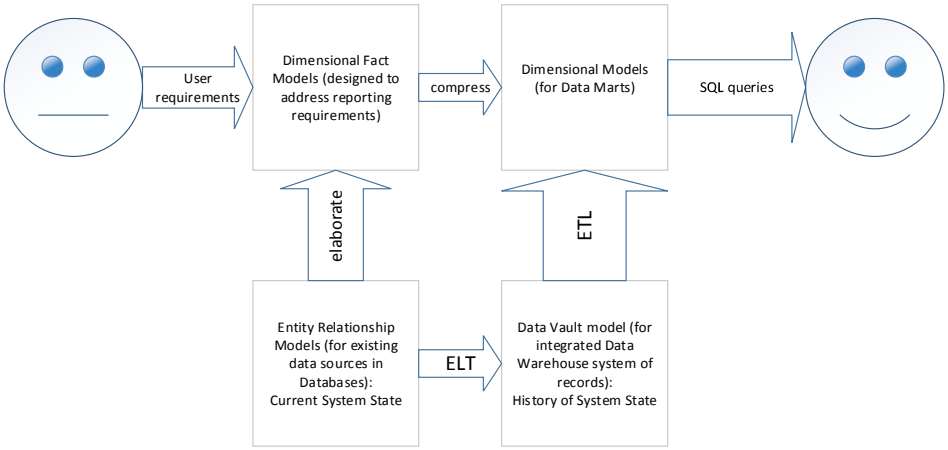


Figure 1. Data Modeling Styles in Data Warehouse design

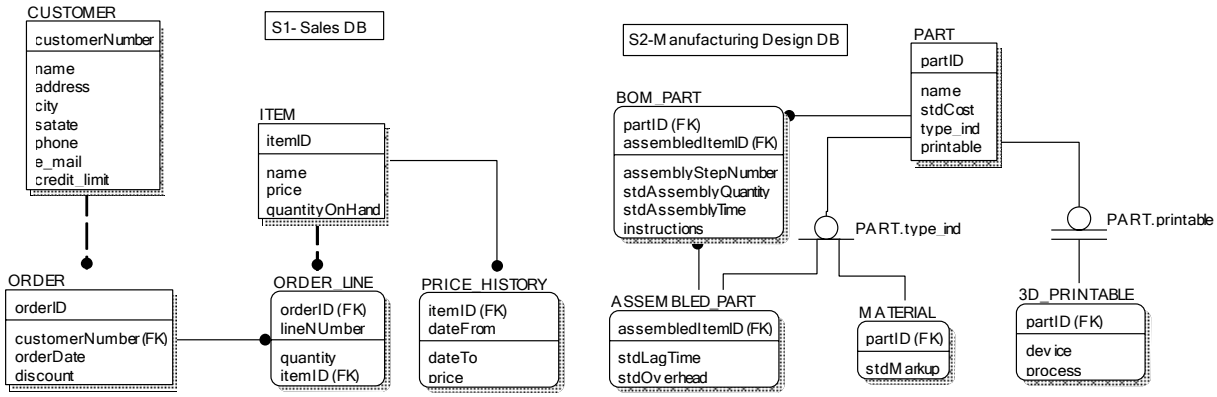


Figure 2. Data Models of source databases

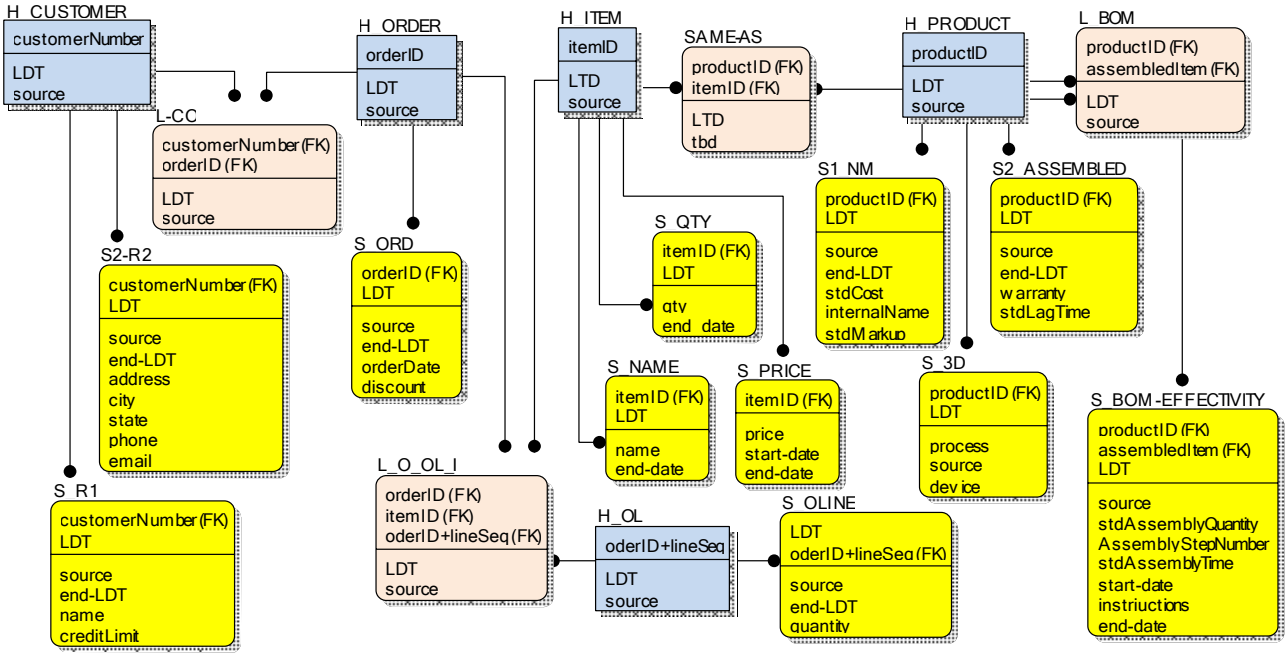


Figure 3. Data Vault model integrating data sources, with color codes H-blue, L-read, and S-Yellow

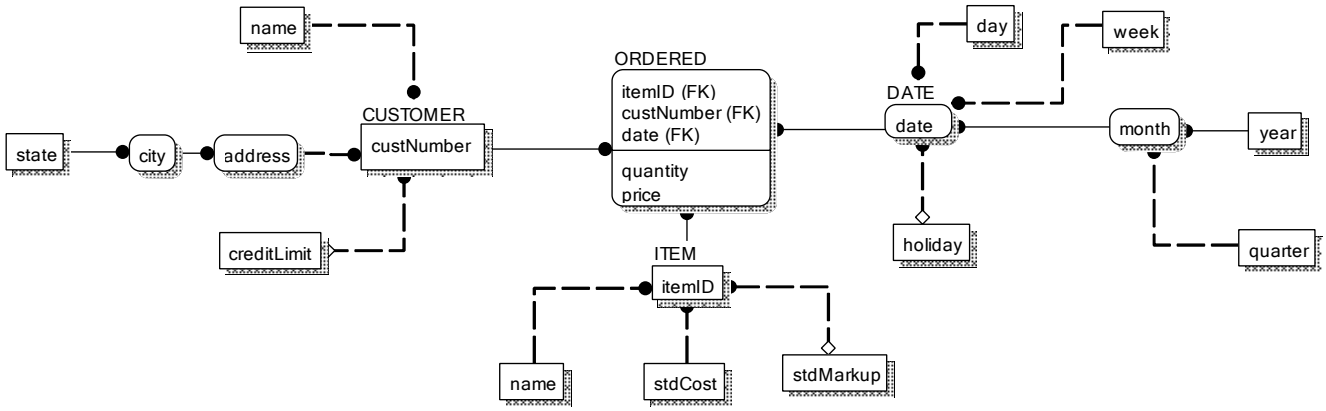


Figure 4. Dimensional Fact Model for ORDERED events

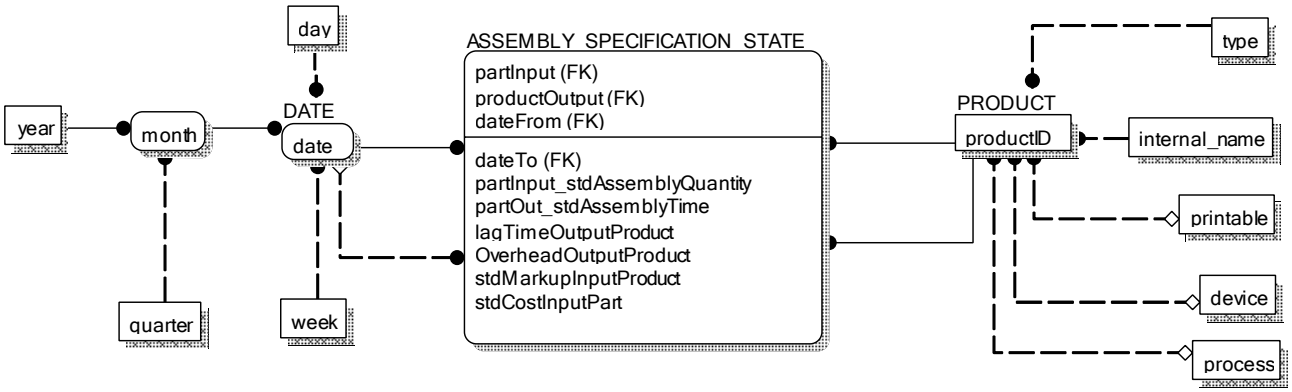


Figure 5. DFM starter model of the assembly process specification

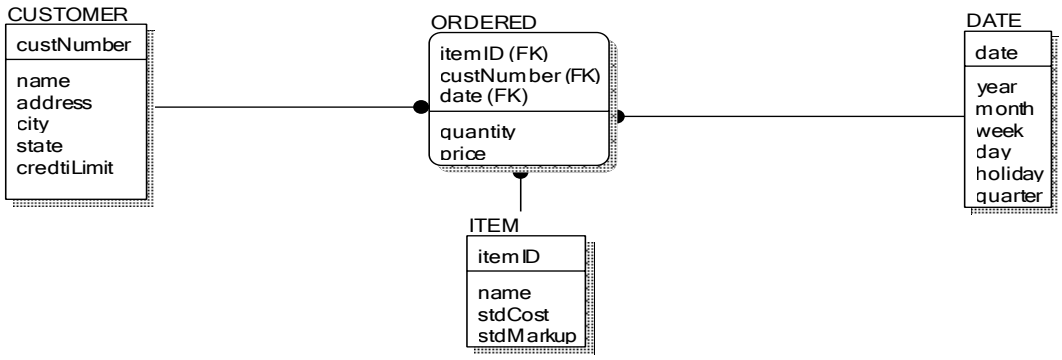


Figure 6. Dimensional Model for Orders Data Mart

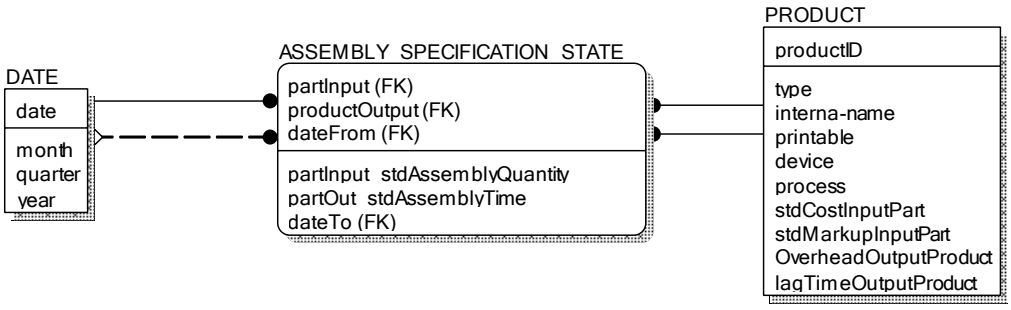


Figure 7. Dimensional Model for Assembly Specification Data Mart

Big Data and Sentiment Analysis using KNIME: Online Reviews vs. Social Media

Ana Mihanović, Hrvoje Gabelica, Živko Krstić
Poslovna inteligencija d.o.o., Zagreb, Croatia
{ana.mihanovic, hrvoje.gabelica, zivko.krstic}@inteligencija.com

Abstract - Text analytics and sentiment analysis can help an organization derive potentially valuable business insights from text-based content such as word documents, email and postings on social media streams like Facebook, Twitter and LinkedIn. The system described here analyses opinions about various gadgets collected from two different sources and in two different forms; online reviews and Twitter posts (tweets). Sentiment analysis can be applied to online reviews in easier and more detailed way than to the tweets. Namely, online reviews are written in clear and grammatically more accurate form, while in tweets, internet slang, sarcasm and allegory are often used. System described here explains methods of data collection, sentiment analysis process for online reviews and tweets using KNIME, gives an overview of differences and analysis possibilities in sentiment analysis for both data sources.

I. INTRODUCTION

Text mining or sentiment analysis [1] is analysis of data contained in a natural language text, which deals with the computation of opinion, sentiment and subjectivity in text. Sentiment analysis refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information from the text documents. Basic task of sentiment analysis is to determine the polarity of a given texts.

Tasks of dictionary making and sentiment analysis process are done by the means of KNIME [2], which is a user-friendly graphical workbench capable of entire analysis process. KNIME uses six different steps to process texts: reading and parsing documents, named entity recognition, filtering and manipulation, word counting and keyword extraction, transformation and visualization. Following workflows and tasks are developed and executed using KNIME:

- Retrieving data from database
- Dictionary development and implementation
- Review scoring

II. DATA COLLECTION

In gathering online reviews and tweets about gadgets, focus was set on few gadget manufacturers such as Apple, Samsung, Nokia, Nexus, LG and on the webpages that contain online reviews about gadgets. Total number of collected online reviews and tweets which were collected during few hours for the purposes of this paper was:

- 812176 tweets
- 419624 online reviews

The system presented here handles crawling, extracting gadget reviews and storing them for analysis. Collected unstructured text is prepared for text mining and sentiment analysis. System presented gives the analysis results for every single review or tweet.

Online reviews were crawled from webpages using Apache Nutch [3] crawler, highly extensible and scalable open source web crawler which traverses the web site starting from a given set of URLs and follows the links matching a given pattern to a certain depth. Tweets were collected with in-house developed Java package used for streaming posts from Twitter. Both online reviews and tweets were collected within few hours. Since tweets are much shorter than online reviews and collecting these posts takes less time than crawling online reviews, the number of tweets is much bigger than the number of crawled online reviews.

Crawled online reviews and tweets and stored into HBase tables on Apache Hadoop [4] server. Hadoop is an open source software project that enables the distributed processing of large data sets across clusters of commodity servers and it is popular for developing large-scale data-intensive applications. Hadoop is used as a storage environment, and HBase table, as a component of Hadoop, is used to store information about gadgets and gadgets reviews. HBase [5] is non-relational, distributed database written in Java which efficiently holds unstructured data in large tables and can be concurrently and randomly accessed. Hadoop and HBase were used because of their main functionalities, which are storage of large quantities of unstructured, textual data and their possibility of reading and writing data in the real time.

III. DATASET DESCRIPTION

Data collected into HBase tables slightly differs for online reviews and tweets.

Online reviews are described with following attributes and structured in the following way:

- Key - unique key indicator for each gadget review created before review was stored into database
- PID (Product group ID) – name of the gadget, such as Samsung Galaxy s4 or iPhone 5
- Review Date – date when online review was stored into HBase table
- Review text – text of the review about specific gadget
- Lang – language of the review text

Tweets are described with following attributes and structured in the following way:

- Key - unique key indicator for each gadget review created before tweet was stored into database
- userScreenName – Username of Twitter user
- CreationDate – date when tweet was stored into HBase table
- text – tweet text about specific gadget
- keyword – name of the gadget, such as Samsung Galaxy s4 or iPhone 5, analogous to PID field in online reviews table
- Lang – language of the tweet

Key and PID value (keyword vale for tweets) allowed aggregated analysis on the level of each review or tweet, or on the level of the specific gadget. Date values allowed aggregated analysis for the defined period of time.

Described data is loaded into KNIME with HBase Reader node and processed. In this phase, only online reviews and tweets in English language were collected. Language of the text is set to English and all texts that have different language values are filtered out, because English dictionary applied on reviews and posts written in other languages would not give results.

IV. DICTIONARY BUILDING

A. *Online reviews*

Online reviews are analyzed on the category level. Seven different categories are introduced for the gadget reviews:

- Accessibility/usage – ease of usage, availability of applications and software
- Value for money – is the price of the product reasonable or not, is the price justified by quality
- Content/composition – materials that product is made of, all hardware components, accessories (handset, bluetooth handset, keypad, etc.)
- Quality – quality of materials product is made of, quality of hardware components, manufacturer quality
- User experience – experience of the user when buying and using product (issues, would recommend, wouldn't recommend, would buy again, etc.)
- Look/appearance – physical appearance of the product (color, design, size, weight, etc.)
- Service/Support – seller's attitude and kindness, service center help support, warranty of the product

Service/support category is the category with the least number of the recognized phrases and for that reasons it is not included in chart visualizations. Redefiniton of this category is being considered.

Dictionary building for detailed sentiment analysis implies making an initial list of adjectives and nouns which are normally used when describing a specific product. Initial list is collected by analyzing existing opinions and reviews and manually extracting the words which appear the most in those reviews and opinions or those words that seem to be most important. Grade scope is defined from 0 to 5 (0 meaning extremely dissatisfied, and 5 meaning extremely satisfied). Null means no rating.

Since terms and phrases extracted from reviews need to have category and grade, nouns are holders of categories, and adjectives are holder of grades. That means that every noun on the list had to be categorized into one of seven categories, and every adjective on the list has to be graded with one grade. Once the initial list of nouns and adjectives has been collected, nouns and adjectives are used in regular expressions which bind them together and recognize complete phrases in reviews consisted of all combinations of adjectives and nouns that are in the list. Regular expressions are composed for the most used grammatical forms in English language. List of regular expressions used for dictionary development:

- Adjective + noun and negations („not“ + adjective + noun)

This form is used to detect most simple phrases of English language, such as „amazing battery“, „bad screen“ or „great apps“. For the negation detection, regular expressions are written in the form „not“ + adjective + noun.

- Noun + „to be“ + adjective and negations (noun + „not to be“ + adjective)

With these forms of regular expressions, more complicated phrases are detected, such as „battery doesn't last“, „screen is great“, or „camera is good“. To make phrases detection more accurate and successful, all forms of verb „to be“ must be included, including grammatical mistakes which are people likely to make.

- Noun + intensifier/downtoner + adjective and negations (noun + „not to be“ + intensifier/downtoner + adjective)

Intensifies or amplifiers are words in English that increase the effect of the verb and include such words as „completely“, „totally“, „undoubtedly“, „absolutely“, etc. Downtoners are words that decrease the effect of the verb

and include such words as „kind of“, „not so much“, „sort of“, and so on.

Another important feature is to correctly grade phrases that contain negations and/or modifiers because these type of words change total grade and meaning of a phrase. If phrase contains some of defined words, grade is calculated using different “if...else if” rules. These exceptions are handled as follows:

- If the phrase contains negation, the grade is reversed and replaced by its module (e.g. „camera is great“=4, „camera isn’t great“=1)
- If the phrase contains intensifier, the grade is increased by one (e.g. „good processor “=3, „really good processor“=4). If the phrase contains negations along with intensifier, the grade module is calculated (e.g. „not really good processor“ =1)
- If the phrase contains downtoner, the grade is decreased by one (e.g. „display is good“=3, „display is merely good“=2). Negations used along with downtoners are rare and therefore not analyzed

B. Twitter posts

Dictionary built for sentiment analysis of tweets contains of key words graded only as positive or negative. Scoring or sentiment analysis of tweets is done on the positive-negative level, because tweets don’t contain clear phrases such as those that can be found in online reviews. Therefore, tweets are impossible to categorize into numerous categories as online reviews, and they need to be analyzed on the word level, giving each word positive or negative polarity.

Dictionary built for sentiment analysis of social media posts consist of most used words in those posts. Because of the different structure when comparing online reviews and tweets, dictionary needs to include jargon words, internet slang, smiley icons, abbreviations, hashtags and similar. These words and symbols are of great importance, since tweets are usually not rich with useful phrases and terms. Example of such words included into dictionary are: #loveit, #horror, :-(, :-*, OMG, JK and similar.

For this task, publicly available MPQA subjectivity lexicon was used as a starting point for recognizing contextual polarity [6], which was expanded with Twitter specific words mentioned above. Existing dictionary containing of approximately 8000 words is expanded to fit the needs for gadget analysis in a way that initial portion of tweets are collected, which are separated into single words with Bag of Words processing. Unnecessary words such as symbols or web URLs are filtered out, and all useful, social media specific words are graded and added to the dictionary.

V. DATA SCORING

Sentiment analysis of online reviews and tweets differs greatly. Online reviews are easier to objectively analyze because of clearer written form and bigger amount of meaningful sentiments, but they have to be analyzed on a more detailed level than tweets.

Tweets often contain internet slang, sarcasm and allegory which are often used. It could be said that grading online reviews is easier, but the dictionary is more complicate to make, and that grading tweets is harder, but the dictionary is easier to make.

Online reviews are analyzed on the phrase and category level, giving phrase a grade for one of seven categories. Tweets are analyzed on the word level giving a positive or negative grade for each term. Text analytics of online reviews is accomplished simply with phrases counters and mean calculations, while analytics of tweets is frequency-driven.

Big quantity of reviews or posts are loaded into KNIME to be graded. Two separate wokflows are built, one for grading onine reviews based on a grade and category, and other one for positive-negative grading.

A. Online Reviews Grading

First input of sentiment analysis workflow developed in KNIME are online reviews read from the database, and second, parallel input is the dictionary made of phrases recognized with regular expressions. Phrases from the dictionary are recognized from the dictionary file and tagged in Review Texts using Dictionary Tagger node.

Online reviews are processed and graded with the term presence method [7], rather than term frequency method. Term presence method gives a binary value which simply indicates does the term or phrase occur in the text (value 1) or not (value 0). Every term or phrase has also a grade and category joined and binary values of terms are summed on the level of each review, giving term counters and grade sum for each review.

Sum of grades for every category is divided with counter number for every category, giving final grade for each online review and for each category. The results were written back to the Hadoop database (Fig. 1), and as a result of Hadoop aggregation, the average grade for every category on the level of single gadget is calculated.

ReviewText	ValueForMoney	Content_Quality	Accessibility_Usage	Look_Appearance
"4.5 stars out of 5 Niall I	5	4	4	4	5
"> In reply to jasz @ 20	4	4	5	5	3
"Great phone, with exce	4	3	4	5	5
"I bought this phone 5 d	4	4	3,5	3,5	3

Figure 1. Example results of analyzed online reviews

B. *Twitter Posts Grading*

Tweets are analyzed only on the positive-negative level. The result can be one of the three; positive, negative and neutral. Tweets for gadgets are not analyzed on the category level because these kind of text documents are often much shorter and not so descriptive as online reviews. Tweets are processed in similar way, but this approach uses more preprocessing steps and uses frequency-driven approach. In the preprocessing steps, various filters are applied on the text documents, such as stopwords filter, n-chars filter and punctuation erasure.

TF*IDF (Term Frequency*Inverse Document Frequency) [8] method assigns non-binary weights related on a number od occurences of a word. Weighting exploits counts from a background corpus, which is a large collection of documents; the background corpus serves as indication of how often a word may be expected to appear in an arbitrary text. TF*IDF calculation determines how relevant a given word is in a particular document [9].

Besides term frequency $f_{w,d}$ which equals the number of times word w appears in document d, size of the corpus D is also needed. Given a document collection D, a word w and an individual document $d \epsilon D$, TF*IDF value can be calculated [9]:

$$TF * IDF_w = f_{w,d} * \log \frac{D}{f_{w,d}} \quad (1)$$

Total score for each word is given by multiplying TF*IDF value with attitude of a term. Attitude can have one of three values depending on the word polarity; -1 for word with negative polarity, +1 for word with positive polarity and 0 for neutral words. Final weigths, which now represent attitude of each document , are grouped on the level of document (tweet) and binned into three bins to give one of three final results for each tweet; positive, negative or neutral (Fig. 2).

Row ID	TwitterText	Total_grade
Row0	"51 users just unfollowed me. Via @FindUnfollower http://t.co/sA3	Negative
Row1	"@encef @officialccrp @OfficialCharice try na n Vha 2 restore it bu	Neutral
Row2	"Fantastic short film presents a day in the life of an iPhone http://	Positive
Row3	"Finally got Ubuntu running on my nexus 10 :D"	Positive
Row4	"Foursquare For iPhone Completely Redesigned For iOS 7, Get It F	Positive
Row5	"Fully Charged: Nexus 5 gets a camera boost, Ryan Seacrest give	Positive
Row6	"Great holiday gift! A ????? ghost tale that's NOT a haunted house	Negative
Row7	"Having fun with #ClumsyNinja for iPhone! Join me now for FREE!	Positive
Row8	"I don't think the Nexus is gonna change my mood anymore"	Positive
Row9	"I liked a @YouTube video http://t.co/jLF4wrp3z3 Google Nexus 7	Positive
Row10	"I'm selling dope, straight off the iPhone."	Negative
Row11	"Just when I've almost decided I no longer want a Nexus 5. A nev	Positive
Row12	"New Straight Input Lychee Glossy Cellphone Bag Cover For iPhon	Positive
Row13	"Obama banned from using iPhone 'for security reasons' http://t.c	Negative
Row14	"RT @Prizzy: So peak when your iPhone breaks :(Without a phone	Negative
Row15	"RT @net4tech: Android 4.4.1 arrives, the picture part of the #Ne	Positive
Row16	"Tbh i dont wanna have an iphone anymore"	Negative
Row17	"The Golden Phone | iPhone 5s Available In Gold, Platinum, and Ro	Positive
Row18	"Town tomorrow or tonight to get 2 iphone cases and a new scree	Positive
Row19	"i dont know if i should get that iPhone or shoes !"	Neutral
Row20	"iPad Mini with Retina Display + Nexus 7 (2013) Giveaway! @tldto	Positive
Row22	"so my iphone officially gave up on me. :(((((((((((((((((((USELESS	Negative

Figure 2. Results of analyzed tweets

VI. RESULTS

A. *Online reviews results*

Online reviews analysis counts all terms in one reviews, and gives mean grade on the review level, for every category. Results can be aggregated on the level of gadget (Fig. 3) or specific model and monitored over time time to analyze trendings.

PID	LgBl40NewChocolate	LgCookieLiteT300	LgKe600
ValueForMoney	4,25	3,75	4
Content_Composition	3,33	2,54	1
Quality	3,54	3,03	3,57
UserExperience	3,33	2,75	2,33
Service_Support	4	3	4
Accessibility_Usage	3,67	3,8	3,82
Look_Appearance	3,25	2,67	4,25

Figure 3. Analyzed online reviews results aggregated on the level of LG products

Simple tag cloud (Fig. 4) gives the overview of the most used phrases in online reviews:



Figure 4. Tag cloud for online reviews

B. *Twitter posts results*

Tweets analysis results are given in the form of positive, negative and neutral. Neutral grade can be avoided, which is easily accomplished by changing grade bins and removing a bin for neutral grade. Grade can be calculated on the level of specific post and then aggregated for all posts of one gadget product (Fig. 5). Sample of 40000 tweets were used for iPhone product analysis.

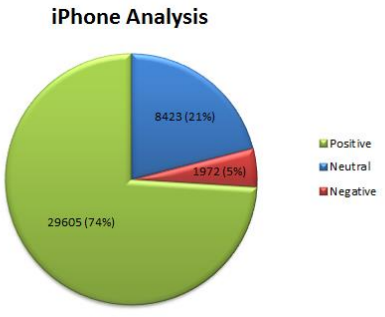


Figure 5. Analyzed tweets results aggregated on the level of gadget brand

Simple tag cloud (Fig. 6) gives the overview of the most used words in tweets:



Figure 6. Tag cloud for tweets

C. *Online reviews graded with Twitter dictionary*

When online reviews are graded with the dictionary made for analysis of tweets (Fig. 7), results are fairly good because each word is separately processed and graded. Dictionary developed for sentiment analysis of tweets is usable for the sentiment analysis of online reviews.

Row ID	ReviewText	Total_grade
Row531	"this mobile is very bad its hilly hang up .very falth what not think any oparate not	Negative
Row532	"nly thg s 3g s nt tr so sad"	Negative
Row533	"very bad phone 3g option no or ringtone option no so bad very very very very very	Negative
Row534	"it does show often that storage is low even if i have enough memory space..cant w	Positive
Row535	"> In reply to Anonymous @ 2013-08-30 18:34 from U$4N - click to read 4800 rupe	Positive
Row536	"> In reply to Vishal @ 2013-08-24 09:42 from bCds - click to readHai friend.. what	Positive
Row537	"which one is good celkona60 or LG Optimus L3 E400"	Positive
Row538	"hai frnds the phone is very nice in features design etc but a small disappointment in	Positive
Row539	"It is a Great phone for a spec perspective .....Only problem is that the internal mem	Neutral
Row540	"internal memory very lowwwwwwwwww...even battery back up very bad"	Negative
Row541	"it is very cheap mobile but its Ram very low"	Negative
Row542	"Hw is it clarity &exactly price?"	Positive
Row543	"Excellent smart phone with advanced features at economical price"	Positive
Row544	"> In reply to Ravi Varman ravi @ 2013-08-27 17:33 from Xu0X - click to readHello d	Positive
Row545	"I bought 63 it is good to use"	Positive
Row546	"oh this is a very good mobile.This is appropriate for my rough use.This is a fantastic	Positive

Figure 7. Results of online reviews graded with the Twitter dictionary

D. *Twitter posts graded with online reviews dictionary*

Tweets graded with dictionary developed for sentiment analysis of online reviews (Fig. 8) are not succesfuly analyzed because of absence of the more complicated phrases that tweets lack. Therefore, tweetscan't be graded on the detailed level, and positive-negative analysis or analysis with simple grade from 1 to 5 is most suitable for sentiment analysis of social media.

text	ValueForMoney	Content_Quality	Accessibility_Usage	Look_Appearance
"10 Best photo editing iP			5	
"I love the new design! ht				4
"Just when I've almost de			4	
"Missed it?: 'Echo Prime			4	
"My New PhoneiPhone 5		4,5		
"New Replacement Silver				
"RT @MisterBanatero: P	4			
"Town tomorrow or tonigh				4

Figure 8. Results of tweets analysis graded with the online reviews dictionary

VII. CONCLUSION

Sentiment analysis of online reviews is less complicated process and gives more detailed results, but developing a dictionary for these kind of text documents takes more time and resources than for social media.

On the other side, sentiment analysis of social media posts is more complicated process, but dictionary development is less complicated task. Also, dictionary developed for sentiment analysis of social media, on a single word level, can be successfuly applied on sentiment analysis of online reviews, but vice versa situation does not give results. Social media posts are hard to analyze on the phrase or sentence level because of theirs unique structure and grammar.

[1] B. Pang, L. Lee, "Opinion Mining and Sentiment Analysis, Foundations and Trends in Information Retrieval", vol. 2, No. 1–2, 2008, pp 1-135

[2] http://knime.org/

[3] http://nutch.apache.org/

[4] http://hadoop.apache.org/

[5] http://hbase.apache.org/

[6] T. Wilson, J. Wiebe, P. Hoffmann, Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis, University of Pittsburgh, Pennsylvania, 2005

[7] J. Kim, J. Li, J. Lee, "Discovering the Discriminative Views: Measuring Term Weights for Sentiment Analysis", Pohang University of Science and Technology, Republic of Korea, 2009, pp 253-261

[8] C. C. Aggarwal, C. Zhai, Mining Text Data, Springer Science+Business Media, New York, 2012, pp 48-52

[9] J. Ramos, "Using TF-IDF to Determine Word Relevance in Document Queries", Rutgers University, New Jersey, 2012

Implementing the additional knowledge in the Croatian Scientific Bibliography

Tomislav Jagušt*, Jadranka Stojanovski** and Mirta Baranović*
* University of Zagreb, Faculty of EE and computing / Department of Applied Computing, Zagreb, Croatia
** University of Zadar, Zadar / Ruđer Bošković Institute, Zagreb, Croatia
tomislav.jagust@fer.hr, jadranka.stojanovski@irb.hr, mirta.baranovic@fer.hr

Abstract - The Croatian Scientific Bibliography - CROSBI is a bibliography web site that stores information about scientific papers published in Croatia in the last 15 years. At the present, CROSBI consists of over 400,000 records and is daily accessed and updated by a large number of researchers. As the appearance and purpose of database (and hence its structure) changed over the years, some inconsistencies emerged, and existing data could not be used to its fullest potential, e.g. linking papers from the same journal or conference, or analysis of scientific activity in relation to the place of article publication. This paper explores the possibilities of improving and enriching the CROSBI database, using different techniques, from data cleaning and sanity checking, to creating links between CROSBI and similar online bibliographic databases. In order to improve data quality, book records from CROSBI were linked to the corresponding Google Books records, and using Semantic Web technologies, a number of links between CROSBI and online bibliographic databases was created. The resulting data set can be used to further improve the quality of CROSBI by creating a recommender system, or to improve and open up new possibilities in the analysis of existing data.

I. INTRODUCTION

The need to collect and organize data about scientific research in Croatia arose early after gaining independence as a country in 1992, when it was realized that data about scientific output are scattered across different index publication, subscription databases and local bibliographies. Small but regular support from the former Croatian Ministry of Science and Technology (today Ministry of Science, Education and Sports) enabled the start of the first phase of the Croatian Scientific Bibliography – CROSBI[1] development, and first release in 1997 [1]. CROSBI is designed, created and maintained at Rudjer Boskovic Institute Library in Zagreb. In the beginning, the primary goal of CROSBI was to collect the data on scientific output of the current research projects financed by the Ministry of Science and Technology and to make them publicly available [2]. The input was provided using a simple web interface, and monitoring and curation was organized by librarians. CROSBI was designed as a single entry point for various publications as books, book chapters, journal articles, conference papers, theses, reports, manuscripts, patents etc., but was also open for papers not yet published like submitted journal articles ("in press"), unpublished manuscripts, and other types of intellectual products related to the academic and research community. CROSBI served also as an Open Access (OA) repository offering free access to deposited publications long before OA movement became turning point in the scholarly publishing area. The data stored in CROSBI was used mostly for an ad-hoc insight into a comprehensive bibliography of an institution, research project or individual researcher [3].

Today, CROSBI consists of 410,000 records, about 130,000 of which are bibliographic data about journal articles, and more than 150,000 are about conference papers. Furthermore, CROSBI provides an access to 25,000 records with full-text available [4]. Authors/Researchers themselves enter the data into the database, so there is a high level of up to data.

The CROSBI system is based on open source solutions: Linux, PostgreSQL DBMS (replacing mSQL) and Apache. PostgreSQL is used for database management because it is unified database server with a single storage engine, with reliable and fast performance in complex operations, and the Perl interface used in cgi-scripts on the Web server. Apache web server is used together with SSL encryption mechanism to provide HTML content to end-users.

During the years, some inconsistencies in the database emerged, like data glitches, incomplete or missing data, missing lookup tables, etc. Since incorrect or inconsistent data could complicate database maintenance, prevent the exploitation of data to its fullest potential and lead to false conclusions, we decided to explore the possibilities of improving the quality and enriching CROSBI data, by implementing different approaches and techniques, from ETL-like procedures to the Semantic Web technologies. The goal is to enhance CROSBI from a stand-alone database of bibliographic records to a highly hyperlinked data set that can interact with other information resources on the web. Development of CROSBI will follow the principles of Linked Data, with resolvable URIs for all objects, ensuring structured RDF data. Using rich ontologies and semantic linkages data stored in CROSBI will not only be available on the web, but will really become an integral part of it.

In the future CROSBI is seen as an interoperable tool, exchanging data with others repositories, archives and databases, which will enable much easier and controlled

[1] http://bib.irb.hr/

This work was partially supported by the "Semantic integration of heterogeneous data sources" project, funded by the Croatian Ministry of Science, under grant number 036-0361983-2012.

input process and provide enriched data beyond bibliographic level. The tools and methods applied in this research and presented in this paper will enable such kind of transformation.

II. CURRENT DATABASE STRUCTURE

There are several inconsistencies and shortcomings of the current database structure, which are mainly the result of its revisions and adjustments to different needs during the years.

Some data which is a part of CROSBI is scattered over different databases. Most of bibliographic information about publications is stored in one database, but data about authors and research projects are in separate databases. Although located on the same server and within the same database management system, the fact that they are separated into different databases is an obstacle that prevents data integration and consumption of CROSBI as a whole, e.g. there are no foreign keys, and it is not possible to create advanced web services that expose data for public consumption, like RDF graph, since tools that provide these features cannot work with multiple databases at once.

The database is partly denormalized, e.g. table "Proceedings_article" contains information about the papers in the conference proceedings, together with information about proceedings itself, and also with information about the related conference. Similarly, table "Journal_article" contains information about the papers in journals combined with information about journals.

Many attributes (about 70%) in the database relations are of type *character varying*, and web form fields used to enter or edit data are, or at least were, in some point in the past, simple HTML text input fields. Consequently, database columns which are mandatory and can not be skipped during the entry process, contain a lot of inaccurate data, missing data, incomplete data or typographical errors (e.g. attribute "year" contains values like "-", "not yet published", "in press", "2011/12", "2012g.", "20111" etc.). Furthermore, there are same entries written in slightly different versions (all caps, camel case, with or without Croatian diacritics). Fields that were left empty by the author were not stored into the database as null, but rather as empty string, or contain some sort of "missing data" label, presented with "-" or "n/a".

Many of these discrepancies were corrected during the years, mostly through improved web interface, data-type or range constraints, and better control of user entries, but a certain amount of data entry errors are still present in CROSBI, mostly among the old entries.

There are columns that contain non atomic values, and are in conflict with the First normal form. Place of book, proceedings publication and conference location are two most prominent cases of these fields, where users often entered more than one value, e.g. "Rijeka - Opatija" or "Zagreb, Hrvatska".

III. DATA TRANSFORMATION AND CLEANING

To clean and 'sanity' check the unstructured free-text data, a set of custom ETL procedures was prepared and executed as described in **Error! Reference source not found.**.

A copy of current CROSBI database was made, and used as a test model to explore the possibilities of improving the database content, enabling advanced search and data analysis options, increase the user experience and ease the interconnection possibilities with other similar databases.

First, the database was normalized, which resulted in creation of several new relations ("Language", "Proceedings", "Event", "Theses type", etc., and their appropriate junction or linking tables). In total, 22 new relations were created.

The next step was to load data into newly created tables. While some lookup tables were loaded from other sources (e.g. a list of universities and faculties in Croatia was loaded from ISVU[2]), most of tables were filled with data from the old denormalized tables, mainly by selecting distinct values, and then manually cleaning and separating only the relevant records.

Some tables required extra transformation effort, e.g. a relation "Place" was first loaded with distinct values from all relations containing attribute "place", then the non-atomic values were split, filtering out country names and removing the unnecessary characters (blanks, commas, semicolons). The last step was to unify alternate city names (e.g. "Beč", "Vien" and "Wienna" with "Vienna"), and repair some frequent typos, like "Zabreg" or "Zabeg" with "Zagreb".

The data about proceedings and conferences was extracted using different text similarity approaches. Database table "Proceedings_article" was split into several tables, namely "Proceedings_paper", "Proceedings" and "Event". The original table contained 146,628 records, and newly created tables have 125,779, 68,834 and 53,227 records, respectively.

TABLE I. shows comparative overview of the number of records initially loaded into newly created tables and a number of records after the different data cleansing techniques were performed.

TABLE I. A NUMBER OF RECORDS IN NEWLY CREATED TABLES BEFORE AND AFTER DATA CLEANSING

Table	***Starting noumber of records:***	***Number of records after cleansing***
Language	223	36
Proceedings	79,897	68,834
Event	86,572	53,227
Place	3,160	1,159

[2] http://www.isvu.hr/

IV. LINKING CROSBI WITH EXISTING ONLINE SOURCES

In order to further improve the quality of data in the database, records from CROSBI were matched against a number of different available datasets. Three different techniques were applied: offline matching (for datasets which were available as a whole database dump or in similar form), writing a custom (SOAP or REST) web service client, or matching using the semantic web technologies and available SPARQL endpoints.

A. OFFLINE MATCHING

Offline matching was conducted with data from Croatian Higher Education Information System (Croatian acronym: ISVU) and GeoNames[3] geographical database. Database dumps of publicly available data were obtained, and loaded into the CROSBI database. Data matching was conducted in several iterations, each iteration decreasing the number of remaining unmatched records. On the other hand, every subsequent iteration was computationally more demanding and its results contained more errors (mostly false positives).

For instance, places from CROSBI were matched against GeoNames database in the following order: at first, matching was restricted only to Croatian towns with more than 15,000 inhabitants, and the whole string was matched. Some frequent places that do not meet those criteria (like Opatija and Cavtat) were added matched manually at the same time. The remaining data was then matched against a list of big European cities, then against a list of all Croatian cities, and against the list of alternate city names in the whole World. These steps connected about 90% of cities, with very good accuracy, from around 60,000 rows in the beginning; only about 5,000 were still unmatched. The last step in a row was fuzzy string matching (matching only part of the word, Trigram[6], Levenshtein and Soundex[7] techniques), but since the unmatched data contained a lot of erroneous data (like "Unknown city" or "Whole Europe") the number of hits was small, with a lot of false positives.

B. ONLINE MATCHING - CUSTOM WEB SERVICES

1) Google Books

Google Books[4] is an online service that allows its users to search full text of more than 30 million books and magazines. In addition to web search, Google Books exposes a Representational State Transfer (REST) Web Service - Google Books API, which allows integration of third party applications with Google Books database. The custom REST Web Service client was created, and CROSBI book data was matched against Google Books database. About 1,500 records (of 10,000 in CROSBI) were matched. Since only full title was matched, the number of hits could be further improved by implementing partial matching or some other sort of fuzzy search.

2) Google Scholar

One of the biggest online resources of scientific papers is Google Scholar[5], a free web search engine that indexes the full text of scholarly literature across many disciplines and sources. Although it is free for personal use, Google Scholar does not provide an API, and use of automated data crawlers is in conflict with their usage rules. Matching of a small sample of CROSBI data against Scholar search engine yielded some very good results (around 50% of articles were successfully matched). In the future Google will hopefully include Scholar into its API, and CROSBI could be interlinked and enriched with additional data from Scholar's huge database.

3) Web of Knowledge

Another big online resource of scientific works is Web of Science[6] (WoS), academic citation indexing and search service by Thomson Reuters. There is a SOAP Web Service available to all subscribers, which allows querying their database by different criteria, and retrieving a set of basic information about discovered articles. A part of CROSBI records that had "Current Content_index" and "Science Citation Index_index" flags set, was checked against WOK database, and about 65% of records were successfully matched.

C. ONLINE MATCHING - SEMANTIC WEB TECHNOLOGIES

Online libraries and bibliographic databases like CROSBI can gain much benefit from The Semantic Web technologies, like Ontologies (specifically Web Ontology Language - OWL) and SPARQL Protocol and RDF Query Language (SPARQL)[8]. Since some of the online bibliographic databases, search engines or Web sites have a semantic version, accompanied with a SPARQL endpoint (a REST Web Service that accepts SPARQL queries and returns results in a simple XML form), a custom Semantic tool was made to match CROSBI data against different SPARQL endpoints. In our research, PubMed (Medline), DBLP, ACM, CiteSeer and Jucs.org were queried with a subset of CROSBI data. Taking into account the restrictions of endpoints, response and computation time of text similarity functions in SPARQL, only simple text search was performed on most of the endpoints. The results are shown in TABLE II.

TABLE II. DATA MATCHING AGAINST DIFFERENT SPARQL ENDPOINTS

Endpoint	*Url:*	*Matches*
PubMed	http://pubmed.bio2rdf.org/sparql	5567
DBLP	http://dblp.rkbexplorer.com/sparql/	1513
ACM	http://acm.rkbexplorer.com/sparql/	168
CiteSeer	http://citeseer.rkbexplorer.com/sparql/	51
JUCS	http://jucs.org:8181/d2rq/sparql	5

[3] http://www.geonames.org/
[4] http://books.google.com/
[5] http://scholar.google.hr/
[6] http://www.webofknowledge.com/

V. RESULTS

Different database modifications, transformations, and data matching against data from various online bibliographic and library sources were performed. The CROSBI database was normalized, most of the inconsistencies and deficiencies were removed, and data accessibility and richness was improved. Some of the results are shown in TABLE I. and TABLE II.

Since the new database tables were created, the possibilities to use or analyze data in a different way emerged. Some examples of that would be: analysis and evaluation of Croatian universities by a number and type of papers published in different journals or conferences, analysis of scientific production by region or city, and similar.

VI. CONCLUSION AND FUTURE WORK

Conducted activities improved the content of the CROSBI database and opened up new possibilities to use and analyze its data.

Links with bibliographic data on different online databases stimulates the possibilities of further data discovery and enrichment. Since the used online databases, as well as CROSBI, are being constantly updated with new data, data matching should become a regular periodical activity, so newly added records would also get interconnected and enriched.

The existence of links with other databases allows cross checking the CROSBI data with a validated data sets. Also data enhancement, by adding missing or related information, was made possible.

Although there is still a rather small number of Semantic sources available and the number of matched records is, for now, relatively small, the Semantic web technologies showed some significant advantages over the other used technologies. The Web Service Client that consumes SPARQL endpoint needs minimal adaptation to work with different endpoints, data format is open and well defined, and interconnection with other sources or inclusion of additional knowledge into existing bibliographic database is simple by using the *owl:sameAs* property.

The future work should further explore the possibilities of more advanced text comparison methods, to improve the matching efficiency and even more increase the data quality. Also, a newly discovered and collected data could be utilized in a recommender system implemented inside a CROSBI web interface, which would help and guide users in correcting and improving the data about their papers and books.

ACKNOWLEDGMENT

We would like to express our special thanks to CROSBI developer prof. Ivo Batistić, from the University of Zagreb, Faculty of Science, for creating the database system and investing a lot of hard work in the maintenance and improvements of CROSBI in last 15 years. Special thanks go also to system's librarian Alen Vodopijevec from Ruđer Bošković Institute, for continuous technical support to the CROSBI bibliography and provision of data needed for this research.

This work was partially supported by the "Semantic integration of heterogeneous data sources" project, funded by the Croatian Ministry of Science, under grant number 036-0361983-2012. The authors would like to thank the funding agency for their support.

REFERENCES

[1] Stojanovski, J., & Slavic, A. (1999). Electronic bibliography - its reliability and its impact on the concept of bibliography in general. In T. Aparac, T. Saracevic, P. Ingwersen, & P. Vakkari (Eds.), Conference on Conceptions of Library and Information Sciences - COLIS (pp. 1–10). Lokve, Croatia: Benja Publishing.

[2] Stojanovski, J. (1999). Bibliography in the network environment: Croatian Scientific Bibliography (CROSBI). In Information Technologies Interfaces ITI 1999.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] Stojanovski, J. (2002). Croatian Scientific Bibliography (CROSBI) - four years experience. In A. L. Markham, James W. ; Hyett, David J. ; Duda (Ed.), Managing resources in a sea of change: Proceedings of the 27th Annual Conference of the International Association of Aquatic and Marine Science Libraries and Information Centres (IAMSLIC) and the 9th Conference of the European Association of Aquatic Scien (pp. 65–75).

[4] Croatian Scientific Bibliography - Total statistical data http://bib.irb.hr/skupni_podaci?lang=EN, Accessed: February 24, 2014.

[5] H. Muller, J. Freytag (2003), Problems, methods and challenges in comprehensive data cleansing, Humboldt-Universitt zu Berlin, Institut fr Informatik, Berlin

[6] PostgreSQL 9.3 Documentation: pg_trgm module, http://www.postgresql.org/docs/9.3/static/pgtrgm.html, Accessed: February 21, 2014.

[7] PostgreSQL 9.3 Documentation : The fuzzystrmatch module, http://www.postgresql.org/docs/9.3/static/fuzzystrmatch.html, Accessed: February 21, 2014.

[8] Svensson, L. G. (2013). Are Current Bibliographic Models Suitable for Integration with the Web?. Information Standards Quarterly, Winter, 25(4), 6-13.

The impact of training data tailoring on demand forecasting models in retail

M. Karan*, D. Pintar*, Z. Skočir*, M. Vranić*, A. Alajković**, J. Milojević**and M. Pleša**
*University of Zagreb
Faculty of Electrical Engineering and Computing
Zagreb, Croatia
**Konzum d.d., Zagreb, Croatia
E-mail address: {mladen.karan, damir.pintar, zoran.skocir, mihaela.vranic}@fer.hr,
{adrian.alajkovic, jelena.milojevic, marina.plesa}@konzum.hr

Abstract -Demand forecasting plays a very important role in retail business. Retail information systems commonly store large amounts of data which are subsequently used by sophisticated data mining tools for building forecasting models. Quality of these models is usually measured through their predictive accuracy as their most important property, followed by other measures which consider average underestimate and overestimate costs etc. Even though the choice of data mining algorithm is usually paramount, training set cleansing and preparation has a significant influence on final model performance. This article discusses and analyses the impact of training set preparation and tailoring on a final forecasting model performance used in a real world example from the retail industry.

I. INTRODUCTION

Today's retail industry is a highly challenging and competitive business area. In Croatia, retail industry dealing with fast-moving consumer goods (FMCG) is not yet consolidated and there is strong competition between most powerful players in this field. Thus, quick adoption of new technologies and adaptation is crucial for survival on the market.

In FMCG industry effective and efficient handling of data, information, people and goods is very important. Besides experienced employees in charge, good information systems support has a crucial role in keeping this retail giants adequately running. Forecasting user demand is a key part in supply chain management process. The impact of incorrect forecasting is two-fold: overestimating can result in large storage costs, potential loss of perishable commodities and frozen capital issues, while underestimating causes loss of potential profit and decreased customer satisfaction.

Due to the recent economic situation and decline of purchasing power promotional sales have increased significantly and became an important segment of overall sales traffic. This puts further emphasis on building specific predictive models which target promotional sales in particular. Additional complexities arise due to the nature of demand of promotional goods which usually behave less predictably than regular demand, especially since promotional items tend to be offered in stores which do not hold those items as their regular selection. Consequently, building predictive models for promotional sales is a challenging yet potentially high-yielding research task.

The approach to building predictive models usually requires two important decisions: the choice of data mining algorithm and the choice of dataset on which this algorithm will be used. Nowadays, retail industry produces and centrally stores large amounts of fine-grained data, usually containing information about each individual transaction in each individual store. If this data encompasses a large enough time period (such as few years), it can be assumed that certain regularities could be inferred from the data. These regularities can be uncovered if correct data mining algorithms are chosen coupled with properly prepared training datasets.

Training set preparation is an essential part of building successful predictive models. Usually it is not wise to use all available data. In real life scenarios historical data usually contains a lot of noise and atypical examples which result from unforeseen situations. E.g. a problem with goods delivery causes absence of a product in a certain store which – if only looking at the data - may lead to a conclusion that this product was present yet wasn't sold at all in that particular store. Alternatively, a highly seasonal store can have low regular sales yet achieve great promotional sale results for only a few days a year due to holiday traffic – if a single predictive model was built on this data it would generalize very poorly. These considerations imply that before using the data for model training all potential outliers should be identified and removed. This paper deals mostly with the following step of the analysis: the issue of determining *which* data and *in what amount* should be used as input for forecasting model training to achieve the best compromise between

the efficiency of the model building process, available resources and final model performance.

The article is organized as follows: Section II presents the related work on the subject. Section III describes our evaluation methodology, while in Section IV real-life available data is presented. In Section V, the solution framework for analysis is introduced. Section VI present the details of our experiments and discuss their results. Finally, we conclude in Section VII.

II. RELATED WORK

High quality forecasting of user demand can have significant impact on the success of FMCG business. Consequently, a large volume of research has been focused on this task. For a concise overview of various approaches to this task we refer to [2]. Furthermore, as explained in [3], forecasting methods can be divided into two major groups:

- statistical models relying on modelling regularities inferred from past data
- judgment-based models relying on informed judgments about future events.

Statistical models tend to outperform judgement-based models. However, studies from [4] indicate that a hybrid approach using judgment-based adjustment of statistical model output done by a domain expert works best. Regarding promotional forecasting, [5] suggests that there are three main factors that influence promotional sales. These are promotion style, item promotion history and store promotion history. These can be analysed in various ways. An approach using intervention analysis on time series is presented in [6]. Regression style modelling using 66 variables was used to produce a very successful system in [5]. We proceed along the same lines using regression analysis. Consequently, our approach belongs in the statistical models group.

In machine learning, experience shows that training on noisy data can often produce suboptimal models. There have been many approaches to tackle this problem. In [7,8] classifiers are trained to clean up the training set for the final statistical model. Alternatively some of many uni- and multivariate statistical methods for outlier detection can be used (e.g. [9,10]). In this work we use filtering rules devised by domain experts and experiment with sampling the data in various ways.

III. MEASURES FOR MODEL EVALUATION

Decision making and planning depends heavily on reliability of forecasting models. To properly evaluate our forecasting models, we use several different yet complementary measures which are widely accepted by domain specialists [2]. These measures consider different aspects of forecasting errors made by models.

The measures we employ in this paper are as follows:

- MAPE (*Mean Absolute Percentage Error*)
- WMAPE (*Weighted Mean Absolute Percentage Error*)
- WMPE (*Weighted* M*ean Percentage Error*)

To calculate these error metrics we define:

$\mathbf{y_p}$– amount predicted
$\mathbf{y}$ – amount achieved
$\mathbf{Y}$ – total amount achieved, $\sum y = Y$

Now we calculate MAPE as follows:

$$\mathrm{MAPE}=\frac{1}{N}\sum\frac{|y_p-y|}{y} \quad (1)$$

where the absolute percentage error (APE) of a single example is given by:

$$\frac{|y_p-y|}{y}$$

While MAPE is a good indicator of model performance it assigns equal importance to all examples. In practice this is often undesirable since from an economic point of view good predictions on certain examples are more important. Such cases include products with large traffic. E.g. the APE of a predicted demand of 4 instead of 5 is considered equal to the APE of a predicted demand of 400 instead of 500. To alleviate this issue a weighted measure might be more appropriate when evaluating models used in real world applications.

We calculate WMAPE as the weighted average of APE as follows:

$$\mathrm{WMAPE}=\frac{\sum\frac{|y_p-y|}{y}\cdot y}{\sum y}=\frac{\sum|y_p-y|}{Y} \quad (2)$$

Note that we use volume (achieved amounts) as weights, giving more importance to items that sell in larger volumes. This weighing scheme reflects the requirement of model optimization from a business perspective. Alternative weighting schemes are also possible, for example taking into account the value or price of the product. E.g. revenue from selling one diving suit is much higher than the one gained from selling one matchbox, which may lead to a conclusion that the more valuable items can be considered more important, even if sold in relatively small amounts. In contrast, one may also argue that the number of customers dissatisfied with not finding the requested item can be much higher in the latter case. Since in the FMCG industry WMAPE is commonly weighted by volumes [1, 2] this is the approach adopted in this paper.

While WMAPE is the most commonly used measure in FMCG industry, it does not reveal certain properties of the model which might be considered interesting or relevant from the business perspective. For example, business users are often interested in the overall behaviour of the model when it comes to underestimating or overestimating predicted values. This is commonly referred to as a downward or upward bias, which at store-level may mean that certain store is being systematically over or under-stocked with promotional items. WMPE measure provides a way of quantifying this bias through the following formula:

$$WMPE=\frac{\sum\frac{y_p-y}{y}\cdot y}{\sum y}=\frac{\sum(y_p-y)}{Y}=\frac{\sum y_p-\sum y}{Y} \quad (3)$$

While each of these measures offers a slightly different view on model performance, for the final evaluation of models we will rely on the WMAPE measure as the most informative with respect to business goals.

IV. AVAILABLE DATA

We use a real world database containing information about promotional sales spanning a period of two and a half years. The data set was collected by one of the largest Croatian FMCG retailers. The data set we use is derived by aggregating and filtering this database.

The products we focus on are part of a predefined hierarchy of products. The highest level of the hierarchy consists of 3 product classes which we will arbitrarily denote as "A", "B" and "C". Different product classes may exhibit drastically different behaviour when considering promotional sales volumes. These differences are due to different usage and buying patterns (e.g. perishable vs. non-perishable goods). Consequently, we build individual models for each of the product classes.

To acquire satisfactory models, focusing on each of the mentioned classes is not enough. Besides filtering out outliers, certain products which demonstrate different behaviour have to be removed also. Domain experts testify that certain periods of time (e.g. holidays) also have to be modelled separately. Within research described in this paper we focused on a purified subset of our data set derived as by applying the following filters:

1. holiday seasons – promotional sales during this periods are very atypical when compared to the rest of the year due to somewhat different customer behaviour.
2. promotions without price reduction – examples of such promotions would be "buy 2 get 1 free" which are atypical and considerably less frequent in our data set.
3. special items – items that domain experts considered as highly specific exhibiting irregular behaviour compared to other items in the data set
4. other outliers – items which are extremely atypical by some criterion. Such examples are very likely the result of an error in data input or an external problem (e.g. goods could not reach a store due to transport issues). We use the following criteria:
   a. extremely low increase in sales – items where data suggests the increase in sales was miniscule (e.g. less than 1%)
   b. extremely high increase in sales – items where the increase in sales is huge (e.g. 50 times)
   c. high differences in store availability – items which change their availability in stores during promotions very drastically, making proper generalization from past data very difficult
   d. extremely high discounts – items where the relative difference between the regular and promotional price is very large
   e. items with extremely low sales
   f. items with a large number of "sibling" items within the product hierarchy

After applying a combination of the above filters we derive a purified data set for our experiments.

V. SOLUTION FRAMEWORK

Our solution framework is presented in Figure 1. More detailed explanation follows.

A realistic scenario of building predictive models for promotional sales in the FMCG industry could involve having large amounts of detailed historical data. Once the outliers are eliminated, one issue that commonly arises is whether the rule of "more data is always better" does apply at all times. First, there is an issue of algorithm complexity and scalability; available resources and timeframe assigned for model building may not be sufficient to handle the size of dataset available for training. Furthermore, even if the quality of predictive model rises with more data used, a question may arise dealing with the issue of diminishing returns. If a small

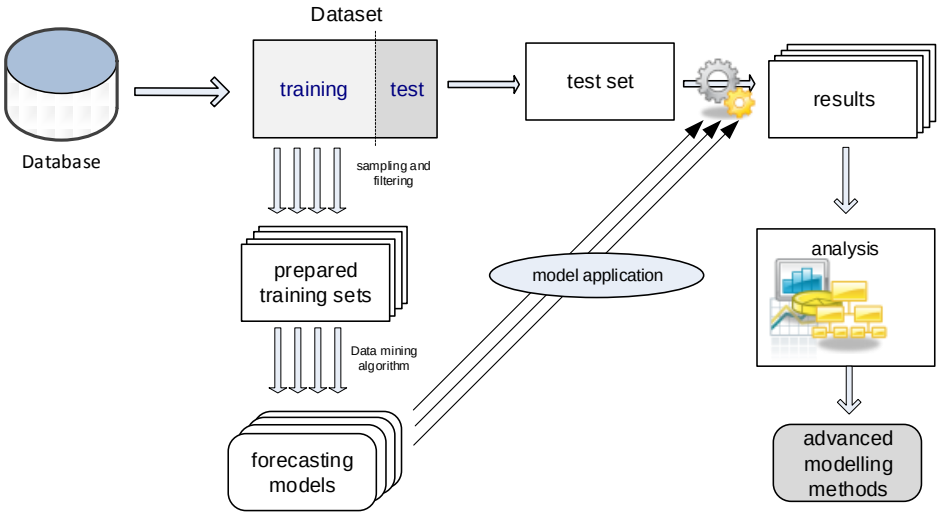


Figure 1.Graphical representation of analysis process flow

sample of available data may produce a predictive model only infinitesimally worse than the model built by using all available data, it is perhaps feasible to use less data for training to free up resources for trying out additional algorithms or approaches. In addition, customer behaviour changes over time so going further and further in the past might even start degrading the model despite increasing the amount of data used. This article deals with these issues and offers guidelines in handling similar problems.

We present a framework for efficiently determining the best method for tailoring the training set to our particular problem – forecasting promotional sale volumes.

The data set described in the previous chapter was used for building the forecasting model. The data is split into a training and test parts, statistics of which are given in Table I. We then apply various combinations of proposed filtering steps. In preliminary experiments we concluded that best results are obtained when applying all filters to both the training and test set. However, some of the filters (4a and 4b) require the value of the target variable which wouldn't be known in advance in a realistic scenario. However, analysis showed that examples which would be eliminated by these filters constitute a very small part of the whole test data set. In addition, domain experts agreed that these examples are erroneous and thus can be omitted when evaluating forecasting models. Additional optional steps are other modifications of the training set such as e.g. down sampling it to leave only the most recent data.

After preparing the training and test set we build a statistical forecasting model on the training set through the method of linear regression. The most important features used by the model include predicted baseline demand (demand in the case of no promotion), information about prices and discounts, position in the product hierarchy, and information about increased product availability during promotions. The final set of features used was determined by a "forward selection" procedure during preliminary experiments. Once the models are trained, they are used to generate predictions for examples in the test set. The performance is then evaluated on the held out dataset using the metrics described in section III

TABLE I. DATA SETS' STATISTICS

Data-set	Starting date	Ending date	No. of promotions	No. promotion/ product
training	30.8.2010.	31.12.2012	191	19186
test	1.1.2013.	20.5.2013.	20	5939

Note that instead of LR we could have used more complex models such as neural networks or support vector regression. These models, however, have considerably longer training times and are less convenient for experiments which include repeated testing with various combinations of parameters. Instead of finding optimal models the focus was placed on tailoring the training set to improve performance. It is expected that any improvements observed in these experiments will, to some extent, transfer to more complex models as well.

VI. DEPENDENCY OF MEASURES ON TAILORING OF TRAINING DATA

A. *Effects of set size*

To explore the effects of training set size on results we created subsets of the training set varying in size from 5% to 100% of the original set. The subsets were created by randomly sampling examples from a uniform distribution. Each sampled subset was used for training the linear regression model which was then tested on the prepared held out testing dataset. To account for the variance due to sampling, each experiment was repeated 30 times and the

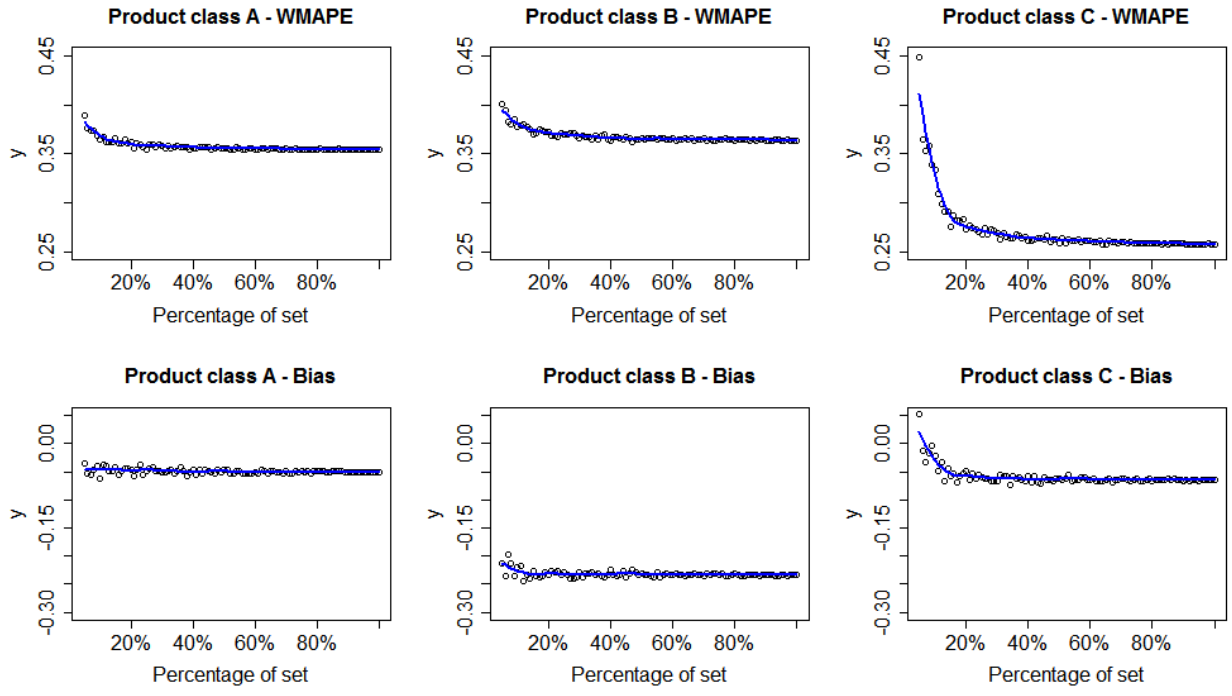


Figure 2. Evaluation measures dependency on random sampled portion of training set

average result was reported. Results are given in Figure 2.

The first row shows the WMAPE measure for product classes "A", "B", and "C" respectively. The second row provides the WMPE measure (called "Bias" here for the reasons of clarity) in the same manner. Axis ranges in each row are aligned for easier comparison.

Analysis of the graphs reveals that when considering the WMAPE measure, product class C is considerably more dependent on the amount of training data. Moreover, results on this product class are much better than on the other two. This is in line with the nature of class "C" which contains items which exhibit high regularity in customer demand. Results on other two classes also slightly improve with larger training datasets. The WMPE measure exhibits the same trends as WMAPE and remains stable for all product classes after using at least 30% of the training data.

A general observation is that all graphs tend to converge rather early. This indicates that adding more training data would not necessarily provide a significant performance boost. Furthermore, the analysis suggests that roughly 30% of the data is quite representative and could thus be used for the purposes of fine tuning more complex models which scale poorly.

B. *Effects of set size and recency*

Similarly to investigating effects of training set size we were interested in whether training set recency plays a significant role in the quality of generated models. In essence, this experiment is identical to the previous one, apart for the training set sampling method. Instead of sampling a certain number of examples *(N)* randomly, we simply take *N most recent* examples.

Note that here we investigate not only the effects of recency but the joint effects of both recency and size. *A priori* logic would dictate that an increase in either recency or size would be beneficial. However, there is a trade-off between these two dimensions. Taking more data increases the training set size but lowers recency and vice versa. To determine the optimal trade-off between these dimensions is one of the research goals in this paper.

Results are given in Figure 3. The method of display is identical as in the previous experiment. With respect to WMAPE, product class C again exhibits much better results when compared to the other classes. In contrast to the previous experiment, increasing the size does not necessarily increase the overall performance of the forecasting model. In fact, for product class B the optimal performance is achieved when using only 20% of the most recent data. Results on product class C, which previously had the most benefit from increasing size, are now quite good from the start (again roughly 20% of the most recent data), but still show modest improvement with taking more data.

Another interesting observation is that for product class A, bias improves with increasing the size of training data while for the other two product classes it slightly deteriorates.

In general, results suggest that filtering the training set chronologically can in certain cases not only reduce training times at no performance cost but can even increase the performance of the forecasting model.

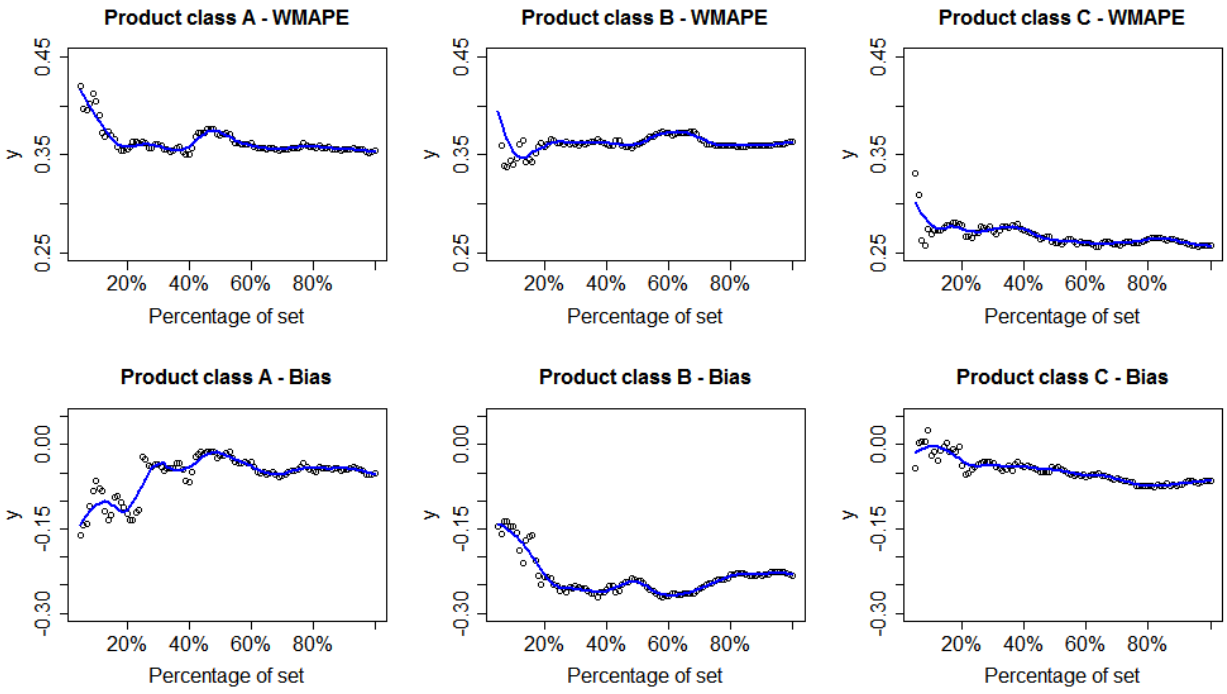


Figure 3. Evaluation measures dependency on chronologically filtered portion of training set

VII. CONCLUSION

We present a framework for optimizing promotion demand forecast model performance by filtering the training set by size or chronologically. This approach has several advantages. It allows us to reduce the amount of training data used to train the models, with little or no impact on performance. It also provides the analyst the opportunity to fine tune the more complex models before building the final forecasting model. We empirically determined good thresholds for each of the product classes present in our data set.

In future work we will focus on using more advanced yet computationally more expensive models such as neural networks or support vector regression. Another possible venue of research is using a more fine grained division of the data set, derived by a clustering procedure using measures defined in [11], [12], and training separate models for each cluster, each with an optimally tailored training set.

REFERENCES

[1] Sivillo, Joel K., and David P. Reilly. "FORECASTING CONSUMER PRODUCT DEMAND WITH WEATHER INFORMATION: A CASE STUDY." *Journal of Business Forecasting* 23.4 (2004).

[2] Koottatep, Pakawkul, and Jinqian Li. *Promotional forecasting in the grocery retail business*. Diss. Massachusetts Institute of Technology, 2006.

[3] Silver, Edward Allen, David F. Pyke, and Rein Peterson. Inventory management and production planning and scheduling. Vol. 3. New York: Wiley, 1998.

[4] Sanders, Nada R., and Larry P. Ritzman. "Judgmental adjustment of statistical forecasts." *Principles of forecasting*. Springer US, 2001. 405-416.

[5] Cooper, Lee G., et al. "PromoCast™: A new forecasting method for promotion planning." *Marketing Science* 18.3 (1999): 301-316.

[6] Leonard, Michael. "Promotional analysis and forecasting for demand planning: a practical time series approach." *with exhibits* 1 (2001).

[7] Brodley, Carla E., and Mark A. Friedl. "Identifying mislabeled training data." *arXiv preprint arXiv:1106.0219* (2011).

[8] Jiang, Yuan, and Zhi-Hua Zhou. "Editing training data for kNN classifiers with neural network ensemble." *Advances in Neural Networks–ISNN 2004*. Springer Berlin Heidelberg, 2004. 356-361.

[9] Ramaswamy, Sridhar, Rajeev Rastogi, and Kyuseok Shim. "Efficient algorithms for mining outliers from large data sets." *ACM SIGMOD Record*. Vol. 29. No. 2. ACM, 2000.

[10] Aggarwal, Charu C., and Philip S. Yu. "Outlier detection for high dimensional data." *ACM Sigmod Record*. Vol. 30. No. 2. ACM, 2001.

[11] Vranić, Mihaela, Damir Pintar, and Dragan Gamberger. "Adapting hierarchical clustering distance measures for improved presentation of relationships between transaction elements." *Journal of Information and Organizational Sciences* 36.1 (2012): 69-86.

[12] Vranić, Mihaela, Damir Pintar, and Zoran Skočir. "Integrating quantitative attributes in hierarchical clustering of transactional data." *Agent and Multi-Agent Systems. Technologies and Applications*. Springer Berlin Heidelberg, 2012. 94-103.

Data mining meets Network Analysis: Traffic Prediction Models

Teo Eterovic*, Sasa Mrdovic * , Dzenana Donko* and Zeljko Juric*
* Faculty of Electrical Engineering/ Department of CS and Informatics, Sarajevo, BiH
{teterovic, smrdovic, ddonko, zjuric} @etf.unsa.ba

Abstract - Most research on network traffic prediction has been done on small datasets based on statistical methodologies. This research analyzes an internet traffic dataset spanning multiple months using the data mining process. Each data mining phase was carefully fitted to the network analysis domain and systematized in context of data mining. The second part of the paper evaluates various seasonal time series prediction models (univariate), including ANN, ARIMA, Holt Winters etc., as a data mining phase on the given dataset. The experiments have shown that in most cases ANNs are superior to other algorithms for this purpose.

I. INTRODUCTION

Network traffic prediction is a crucial part of every serious network system. Network prediction enables network administrators to strategically plan and respond to network issues such as network infrastructure scaling, testing of new protocols/devices and to deal with network problems that already happened and could occur again in the future.

The construction of such a network traffic predictor could be the fundament of a network traffic generator that could be used for precise network traffic simulation and planning.

Lelend at al.(1994) has shown and proved [1] that the Ethernet traffic in LAN networks has an self-similarity nature. In 1998 *Feldman et. Al.* writes about the fractal nature of WAN traffic [2]. Their models have been verified in several experiments as in [3,4,5] on different datasets and network domains including the internet traffic[4] as a special class of network traffic. Other network traffic properties include multiscalarity and their nonlinear nature with long rage dependence as shown in [6]. This nature of network traffic enables us to do a complex and, with the available machine learning tools, a sufficiently accurate traffic prediction.

Although there are several papers that deal with the problem of network traffic prediction [7,8,9,10,11] most of them use simple - linear statistical models for time series prediction. Also most of the papers deal with only one dimension of the network traffic mostly the data they got from the SNMP logs. The problem with SNMP logs is that they miss the packet inter-arrival time parameter that is crucial for network generators. There are less papers that use machine learning models (in most cases simple ANN models or rarely SVM). Another problem is that most research has been done on relatively short time series that span usually over a few weeks focusing on short-term or medium-term time series prediction.

II. RELATED WORK

In the literature various models have been proposed in order to model the network traffic as a prediction problem. The methods vary from strictly statistical methods based on fitting statistical distributions to specific sources, models based on auto regression and moving average (AR, ARMA, ARIMA, SARIMA, GARCH)[8] or exponential smoothing methods such as Holt Winters that showed some superiority in daily predictions and computation time. Another approach are the machine learning models especially the non-linear ANN and kernel based SVM models[9][10], ensemble methods based on ANNs that give good results for short-term time series raging from five minutes or one hour[11]. There are also hybrid methods based on the combination of ANNs and ARIMA models[7].

III. DATA MINING - TIME SERIES ANALYSIS

For the purpose of this paper a new internet traffic dataset has been created on a web server. The dataset is recorded in *PCAP* format with *tcpdump* and spans from (10.10.2013. until 22.02.2014.) and it has been made publicly available at [12].

The data analyzed in this paper spans for several months but in order to simplify the graphs a focus has been set on several weeks to analyze the different time-series properties. The dataset includes several events like weekends, state and religious holydays and the daylight saving time switch on 25.10.2013. This switch can cause issues with some algorithms and the analysis. A data mining / KDD process methodology has been used to analyze the data as described by Fayyad et.al in [13].

A. Preprocessing

This part of the process reduces the dimensions of the dataset and extracts only the data that is useful to the prediction model. This is also the step where the training/ testing/validation datasets are created. Data preprocessing may include cleaning, normalization, transformation, feature extraction/selection [14]. Also a data warehouse could be constructed as shown in [15] that is used to analyze the web logs. A similar approach could be used for network traffic.

For the purpose of the analysis the pcap file has been preprocessed: features selected, cleaned, normalized and transformed to *CSV*. Than it was imported to a *RDBMS* in order to make fast aggregations and enable filters (groupby/where/partitions) for experiment dataset generation with different parameters.

The features chosen in the feature extraction/selection process where: absolute packet arrival time, packet size, delta time (packet interarrival time) and the protocol. The delta time and protocol features have been chosen because they are needed for a proper traffic generator implementation.

B. Outlier detection

According to [16] an outlier can be defined as an observation that is significantly distanced from all the other observations. It is very important to identify all the outliers and to filter them out because in some cases they can produce massive issues when using statistical learning algorithms.

Outliers can be easily identified on a scatterplot as show in figure 1 for a multiday time series between the 10.10.2013 and 03.11.2013.

Several outlier detection algorithms have been tested including Distance Based Outlier Detection, Density outlier detection and LOF.

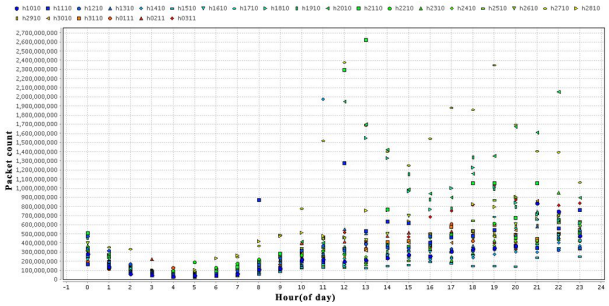


Figure 1 Outliers are visible on a time series scatterplot

Distance based outlier detection [17] searches for ***n*** outliers based on their distance from the ***k*** nearest neighbors (kNN)

Density based outlier detector DB(p,D) searches the global space for outliers. This algorithm takes an assumption that the outliers are those observations that have at least a ***p*** proportion from all other observations that are more distant than ***D***[18]. This algorithm differs from the previous one because it takes all observation into account and not just the k nearest.

The problem of this detector for this specific domain is that we have to specify the ***p*** and ***D*** parameters which is not very practical in context of the linear long time growth of network traffic and time series trends/seasonality. Although this algorithm has shown better results than the previous one by doing a global outlier detection, it is not the best choice in his plain form for this domain because of the parameter estimation problem.

Local outlier factor *(LOF)* ***algorithm*** [19] is similar to the previous ones. The LOF algorithm takes the distance from the k nearest neighbors and estimates the density.

The algorithm compares the local density of the observation with his all neighbor densities. Observations that have a much smaller density than they neighbors are considered outliers.

It's obvious that table 1 – left (distance based outlier detection) shows that there are anomalies at 11 o'clock because it's a big value(1098212) but it's hard to figure out what the anomalies are because we have to guess the number of outliers which is not very suitable for this domain.

hour	outlier	1410
0	false	304371
1	false	177879
2	true	97935
3	true	77836
4	true	63254
5	true	43206
6	true	51625
7	true	87473
8	true	115103
9	false	140210
10	false	184605
11	true	1098212
12	false	261378
13	false	234104
14	false	279164
15	false	275711
16	false	230653
17	false	248877
18	false	260348
19	false	300785
20	false	345683
21	false	326048
22	true	409741
23	true	398130

hour	outlier	1410
0	false	304371
1	false	177879
2	false	97935
3	false	77836
4	false	63254
5	false	43206
6	false	51625
7	false	87473
8	false	115103
9	false	140210
10	false	184605
11	true	1098212
12	false	261378
13	false	234104
14	false	279164
15	false	275711
16	false	230653
17	false	248877
18	false	260348
19	false	300785
20	false	345683
21	false	326048
22	false	409741
23	false	398130

hour	outlier	1410
11	14.972	1098212
22	2.259	409741
23	2.035	398130
20	1.430	345683
5	1.429	43206
21	1.339	326048
6	1.329	51625
1	1.321	177879
4	1.242	63254
10	1.235	184605
3	1.152	77836
9	1.141	140210
7	1.098	87473
8	1.060	115103
19	1.054	300785
2	1.051	97935
16	1.037	230653
0	1.035	304371
13	1.002	234104
14	0.958	279164
17	0.940	248877
15	0.937	275711
12	0.931	261378
18	0.931	260348

Table 1 Left-Distance based outlier detection(with euclide distance); Middle – Density outlier detctor; Right-Local outlier factor

If we set the outlier count parameter ***n*** to 1 the algorithm will successfully detect the big number but it will leave out the others.

Table 1 – middle shows the results of applying the Density based outlier detector. The parameters distance and proportion where set to 7000 and 0.9 respectively. The results are great but the problem is how to define the distance and proportion (the distance parameter is a constant but the problem is the network traffic is growing linear over time[20]).

Table 1 – right shows the results of the LOF algorithm ordered by the outlier factor. We can clearly see that it detected the 11 o'clock anomaly successfully but there is still the threshold problem. Nevertheless it's easier to figure it out than for the previous two methods.

A graphical method could also be used to detect the daily anomalies as seen in figure 2. Its obvious that there are anomalies in the days mentioned before and identified by the outlier detector.

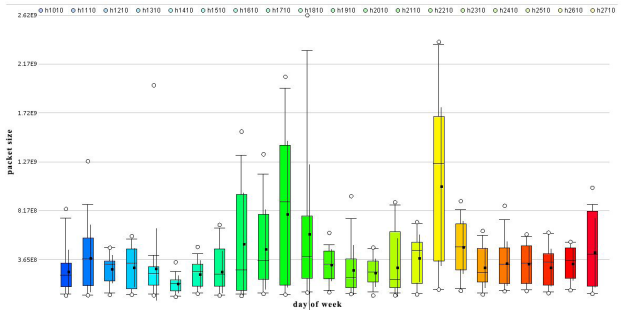


Figure 2 Quartile diagram used for anomaly detection

IV. TIME SERIES ANALYSIS OF THE DATASET

Figure 4(end of paper) shows the packet count rate of internet traffic for three and a half months on daily basis.

From figure 4 it is obvious that the traffic is similar between days and that the daily peek is from 7 till 11 PM while the off peek is from 3 to 5 AM which means that the traffic is mostly from the same time zone. It's also easy to detect the anomalies on the graph.

A. Time series components

To analyze the nature of the time series we have to decompose it to components: the trend component, seasonal component and the noise component.

In figure 5,6,7 we can see the specific components for a 10 days (240 hours) sample of the dataset. In figure 5 we see that the trend is fairly constant except the dropdown on the 100th hour (14.10.2013.) which was a day before a religious holyday (the students go home).

In figure 6 (the seasonal component) we see that the day is made of two periods. The first one from 0 to 50 has four waves, that fits to the previous observation that one day has two parts the (night) off peak and the peak period(day). In figure 7 we can clearly see the anomaly identified before.

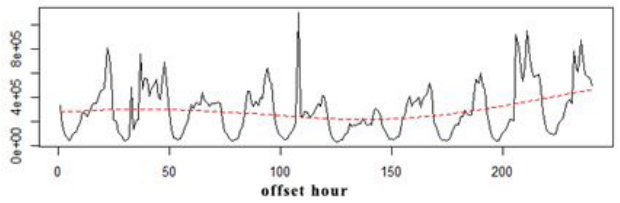


Figure 5 - The trend component

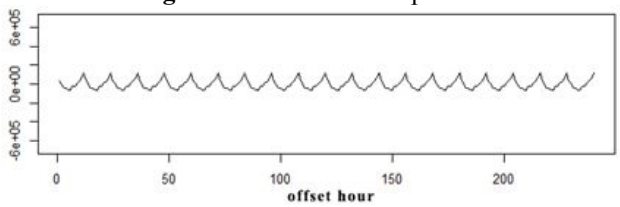


Figure 6 - The seasonal component

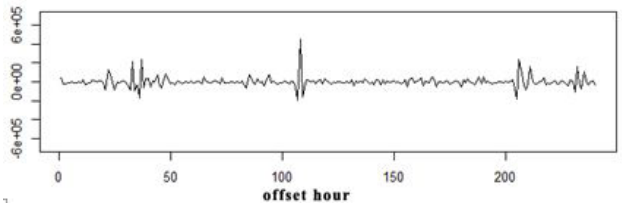


Figure 7 - The noise component

B. The Hurst parameter

By estimating the Hurst exponent we get a value that can indicate the long time memory of the time series [21] it also indicates if the time series is self-similar. An R/S analysis has been used on the given dataset to estimate the Hurst parameter. The results are given in table 2.

	Size	Delta	Count	ALL
Corrected R/ S Hurst exponent:	0.753247	0.6093524	0.7100859	0.9676726
Theoretical Hurst exponent:	0.5457227	0.5457227	0.5457227	0.5334291
Corrected empirical Hurst exponent:	0.7810718	0.4850182	0.6998232	0.9323203
Empirical Hurst exponent:	0.8152428	0.5140385	0.7312095	0.9528319

Table 2 - Hurst exponent estimation

If the Hurst parameter is smaller than 0.5(H<0.5) or greater than 0.5(H>0.5) then this is an indication that there is a trend. If the Hurst parameter is equal to 0.5 the time series is random. It's shown in [22] that time series with a greater value of the Hurst parameter give better results in ANN prediction. The results in table 2 show an H value greater than 0.5 which means there are signs of a trend.

C. Correlation between days

An interesting factor for the time series analysis is the correlation between days in a week as well as between weeks and years. A correlation matrix between days for a 10 days sample is shown in table 3 shown at the end of the paper.

It's easy to spot that the days 14, 21 and 11 (holiday) are anomalies because they have a very low correlation with all the other days.

We also see small anomalies on Friday the 25.10; 18.10 and the Saturday 12.10; 19.10 and 26.10. These are the days when most of the students go out (they are the targeted users of the websites hosted on this server).

day1	day2	corr	day1	day2	corr
h1010	h3010	0.970	h2410	h3010	0.916
h1710	h2510	0.967	h2210	h0311	0.912
h1010	h2310	0.956	h2610	h3010	0.912
h2910	h0311	0.949	h2510	h2910	0.909
h2910	h3010	0.946	h2510	h0311	0.906
h2210	h3110	0.940	h1310	h2610	0.899
h2410	h2910	0.939	h1710	h3010	0.899
h1610	h2610	0.934	h1010	h2910	0.896
h1710	h0311	0.929	h2610	h2910	0.896
h1310	h1610	0.926	h2410	h2610	0.896
h2210	h0111	0.926	h3110	h0111	0.895
h1710	h2910	0.926	h1310	h2010	0.895
h0111	h0311	0.926	h3110	h0311	0.894
h2310	h3010	0.917			

Table 4 Correlation between days sorted

An interesting thing occurred when we analyzed the most correlated days as shown in table 4. Although it's logical that the day of week should have the highest correlation with the same day next week the analysis has

shown different. The days that are near each other have the highest correlation during the week.

A correlation histogram for the days of week has been made as shown in figure 8. The histogram shows that the correlation for the sum of packet size per hour per day of week has a big value (around 0.8) and that the peaks are around the central value of 0.9. This fits a Beta(8,2) distribution.

The results have shown that the analysis for the delta parameter and packet size is very similar to the analysis made for the packet size.

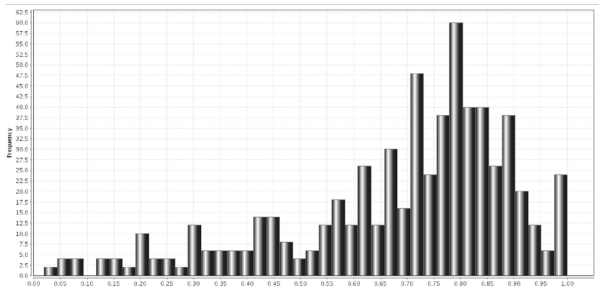


Figure 8 – Histogram - packet size correlation per hour per day per week(vertical axis frequency)

V. UNIVARIATE PREDICTION MODELS

The most popular time series prediction models are ARIMA and Holt Winer for linear prediction and ANNs for nonlinear predictions.

In this experiment we focused on a univariate time series predictions based on individual domain features.

A. Holt Winter

The Holt Winter model is a simple exponential smoothing model for TS prediction. The results of a 24 hour prediction of the time series is shown in figure 9. From the graph it's obvious that the prediction is just an average of the time series.

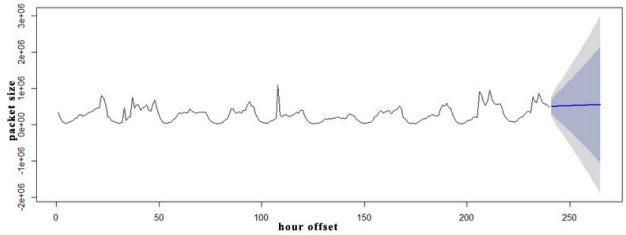


Figure 9 - 24h Time series prediction with Holt Winter

B. ARIMA

Autoregressive integrated moving average has shown good results for self-similarity internet traffic prediction[23] and with predicting time series that show a trend. During the experiment several ARIMA have been used and the model that best fitted the traffic was the ARIMA(2,0,3) model with p=2; d=0 and q=3. The 24 hour prediction on the given dataset is shown in figure 10.

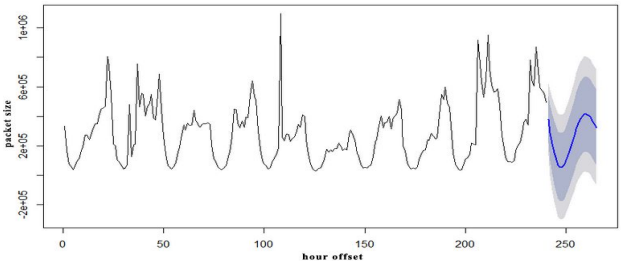


Figure 10 ARIMA(2,0,3) 24h prediction with non-zero mean

The result isn't bad but if we try to expand the prediction to five more days (120 hours) days we get the result shown in figure 11. So ARIMA in it's plain form is good for short term predictions but fails on long term predictions.

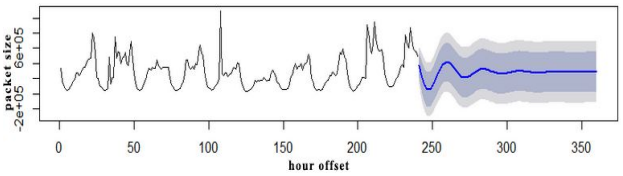


Figure 11 ARIMA(2,0,3) 5 days prediction with non-zero mean

C. ANN

Several experiments[7,11,13] have shown that the prediction with ANNs a nonlinear statistical learning model gives the best prediction results.

In order to analyze the time series a resolution of the time series sampling rate has to be chosen. The ITU rec-E.492 recommends to take a 15 min. or one hour aggregation frame [24]. Our analysis has shown that the best results for this particular dataset and the selected algorithms is a one hour frame.

An ANN with only 22 seasonal lags in the layers and with time series frequency 3 for the window has been used during the experiments. The 24h prediction can been seen in figure 12. As we can see the results are much better fitted than ARIMA so it fits better for the construction of a network traffic generator.

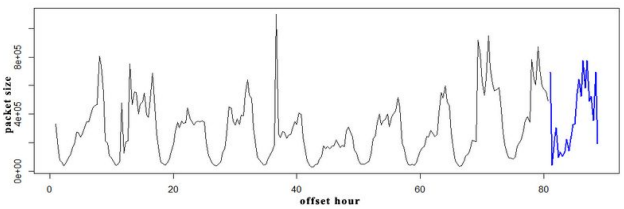


Figure 12 - ANN 24h prediction with 22 seasonal lags and frequency 3

The problem with this approach is that we use only one dimension the packet size and the neural network is adapting to that dimension. One solution is to use an ANN multivariate model in order to do a much longer prediction without the memory effect.

For the second experiment an 27 seasonal lag ANN configuration with 3 days prediction has been used as shown in figure 13.

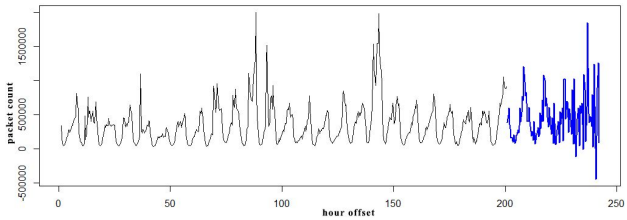


Figure 13 ANN(27,1) 3 days prediction with 27 seasonal lags and frequency 3

VI. RESULTS

A. *Fitting errors and performance*

Several error measurements have been used to evaluate the different models including ME, RMSE, MAE, MPE, MAPE, ACFI. Because of the univariate limitation of this work every dimension of the domain had to be evaluated separately. The fitting error results are shown in table 5,6,7.

	ANN	ARIMA	HoltWinter
ME	-345.6	-246.2	177387.5
RMSE	77451.7	144005.4	333465
MAE	45998.2	79901.6	217246.6
MPE	-9.8043	-17.6479	49.8455
MAPE	18.6512	31.0240	65.7479
MASE	0.25865	0.32807	1.21914
ACF1	0.059133	-0.00108	0.045426

Table 5 - Fitting errors for packet count per hour

	ANN	ARIMA	HoltWinter
ME	-205956	10955995	194601375
RMSE	125857101	226003847	417423245
MAE	65042165	11124690	24943242
MPE	-15.576	-26.010	50.054
MAPE	26.478	43.630	70.327
MASE	0.29317	0.50143	1.113607
ACF1	0.0489	0.0300	-0.0644

Table 6 - Fitting errors for packet size per hour

	ANN	ARIMA	HoltWinter
ME	1.8e-05	3.0e-05	0.00968482
RMSE	0.00403604	0.00656261	0.01967587
MAE	0.00253094	0.0045137	0.01317839
MPE	-4.489	-16.192	47.346
MAPE	16.14	33.22	73.21
MASE	0.21070	0.37578	1.096781
ACF1	-0.02278	0.006316	0.397162

Table 7 - Fitting errors for interarrival delta per hour

The results show that ANNs outperform the other models in model fitting performance but this could be also an indication for overfitting.

The indication that the model fits the data correctly is not the only factor that is important for the prediction.

Cross validation is a statistical method that could help to better evaluate the model but the method is a little bit more complex to be used on time series. A common name in the literature for time series cross validation is "forecast evaluation with rolling origin".

B. *Cross validation*

A cross validation for the time series could be constructed by splitting the dataset into several parts. One part could be the training set (in our case 80% of the dataset) and the other one the validation set (20%).

One could take 80% of the training set and then predict the next 20% and then compare the predicted ones with the real ones from the validation set that wasn't used for the training.

For the verification if the prediction was successful a correlation between the 20% predicted and the 20% validated could be done (better result would be with windowing).

The validation results cross-correlation coefficients are shown in table 8. It is obvious that ANNs outperform the ARIMA model

	Size	Count	Delta
ANN	0.6046795	0.6115911	0.7085994
ARIMA	-0.5489083	-0.3409841	0.7394064
HW	-0.6820613	0.7975048	0.6904266

Table 8 Cross validation 24 hours scope. The values are cross-correlation coefficients

It's interesting that Holt Winter outperforms the ANN and ARIMA, this is due the fact that we use cross-correlation without a time window and the result of the holt winters prediction is a straight line(avg of the time series). This shows the importance to use multiple parameters in order to evaluate the model.

VII. CONCLUSION

In this paper we show that the network traffic time series prediction fits the data mining process model. We have gone through different aspects of the analysis. This is important because most of the papers use traditional process models and small dataset while the network traffic time series prediction problem could be easily classified as a Big Data problem. It was shown that neural network outperform other linear models in the context of precision and that they could be a good model for a network traffic generator based on historical data on top of a data warehouse.

REFERENCES

[1] W. E. Leland, M. S. Taqqu, W. Willinger, D.V. Wilson. 1994. On the self-similar nature of Ethernet traffic (extended version). IEEE/ACM Trans. Netw. 2, 1 (February 1994), 1-15.

[2] A. Feldmann, A.C. Gilbert, and W. Willinger. Data networks as cascades: investigating the multi fractal nature of internet wan traffic. In P roc. ACM SIG COMM, pages 42-55, 1998.

[3] Crovella, M.E.; Bestavros, A., Self-similarity in World Wide Web traffic: evidence and possible causes, Networking, IEEE/ACM Transactions on , vol.5, no.6, pp.835,846, Dec 1997

[4] D. Chakraborty, A. Ashir, T. Suganuma, G. Mansfield Keeni, T. K. Roy, and N. Shiratori. 2004. Self-similar and fractal nature of internet traffic. Int. J. Netw. Manag. 14, 2 (March 2004), 119-129.

[5] Marie, R.R.; Blackledge, J.M.; Helmut E Bez, "On the fractal characteristics of Internet network traffic and its utilization in covert communications," Internet Technology and Secured

Transactions, 2009. ICITST 2009. International Conference for , vol., no., pp.1,6, 9-12 Nov. 2009

[6] Dharmadhikari, V. B.; Gavade, J. D., An NN approach for MPEG video traffic prediction, Software Technology and Engineering (ICSTE), 2010 2nd International Conference on , vol.1, no., pp.V1-57,V1-61, 3-5 Oct. 2010

[7] G.Peter Zhang, Time series forecasting using a hybrid ARIMA and neural network model, Neurocomputing, Volume 50, January 2003, Pages 159–175

[8] S Jung, C Kim, Y Chung, A Prediction Method of Network Traffic Using Time Series Models,Computational Science and Its Applications - ICCSA 2006

[9] X. Mu, N. Tang, W. Gao, L. Li, Y. Zhou, A One-Step Network Traffic Prediction, Advanced Intelligent Computing Theories and Applications. With Aspects of Artificial Intelligence, Lecture Notes in Computer Science Volume 5227, 2008, pp 616-621

[10] Weidong Luo, Xingwei Liu, Jian Zhang, SVM-based analysis and prediction on network traffic, ISKE-2007 Proceedings, 2007

[11] Cortez, P., Rio, M., Rocha, M. and Sousa, P. (2012), Multi-scale Internet traffic forecasting using neural networks and time series methods. Expert Systems, 29: 143–155. doi: 10.1111/j.1468-0394.2010.00568.x

[12] Dataset URL: http://etf.ba/netdataset

[13] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, From data mining to knowledge discovery: an overview., Advances in knowledge discovery and data mining., 1996

[14] Kotsiantis, D. Kanellopoulos, P. Pintelas, Data Preprocessing for Supervised Leaning, International Journal of Computer Science, 2006, Vol 1 N. 2, pp 111–117

[15] K. P. Joshi, A. Joshi, Y. Yesha, On Using a Warehouse to Analyze Web Log, Distributed and Parallel Databases March 2003, Volume 13, Issue 2, pp 161-180

[16] F.E. Grubbs, Procedures for detecting outlying observations in samples, Technometrics 11 (1): 1–21, February 1969

[17] S. Ramaswamy, R. Rastogi, and K.Shim, Efficient algorithms for mining outliers from large data sets. SIGMOD Rec. 29, May 2000

[18] E. Knorr, R. Ng: Algorithms for Mining Distancebased Outliers in Large Datasets , VLDB, 1998, pp392 - 403

[19] V.A., Local outlier factor, Wikipedia / Breunig, M. M.; Kriegel, H.-P.; Ng, R. T.; Sander, J. (2000). "LOF: Identifying Density-based Local Outliers". Proceedings of the 2000 ACM SIGMOD international conference on Management of data. SIGMOD '00: 93–104.

[20] Gardner, E. S., Jr. "Exponential smoothing: The state of the art." Journal of Forecasting 4, 1985, 1–28

[21] H.E. Hurst, 1951, Long-term storage of reservoirs: an experimental study, Transactions of the American Society of Civil Engineers, Vol. 116, pp. 770-799.

[22] B. Qian, K. Rasheed, 2004, “Hurst Exponent and financial market predictability,” IASTED conference on “Financial Engineering and Applications”(FEA 2004), pp. 203-209,

[23] R.F.Nau, “Introduction to ARIMA”, Duke University Lectures

[24] ITU, “Recommendation E.492 Traffic reference period”, ITU, 1996

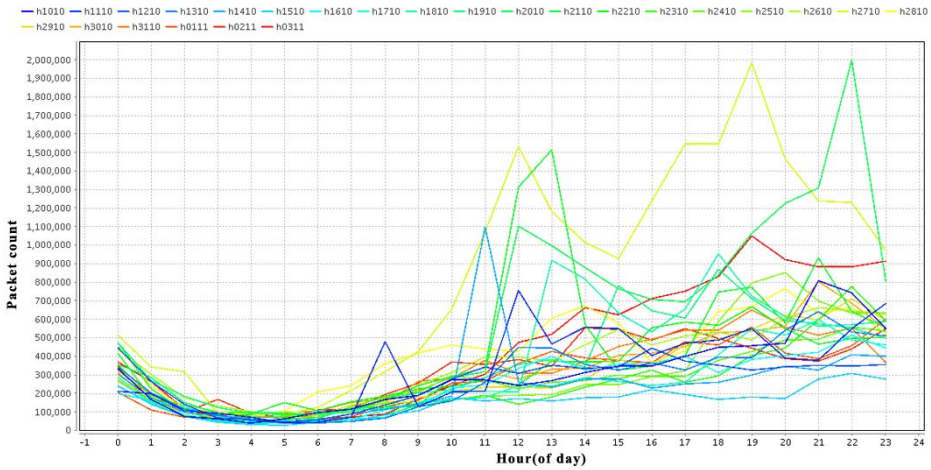


Figure 4 - Packet count aggregation based on hour basis the lines show the different days

Attributes	h1010	h1110	h1210	h1310	h1410	h1510	h1610	h1710	h1810	h1910	h2010	h2110	h2210	h2310	h2410	h2510	h2610
h1010	1	0.353	0.569	0.771	0.196	0.866	0.834	0.846	0.476	0.592	0.720	0.305	0.669	0.956	0.877	0.781	0.843
h1110	0.353	1	0.433	0.634	0.068	0.422	0.568	0.422	0.420	0.398	0.672	0.621	0.492	0.320	0.441	0.309	0.547
h1210	0.569	0.433	1	0.790	0.369	0.782	0.771	0.555	0.605	0.623	0.620	0.466	0.746	0.540	0.727	0.481	0.813
h1310	0.771	0.634	0.790	1	0.250	0.785	0.926	0.767	0.672	0.623	0.895	0.706	0.749	0.720	0.814	0.681	0.899
h1410	0.196	0.068	0.369	0.250	1	0.235	0.256	0.131	0.020	0.082	0.119	0.066	0.210	0.165	0.290	0.093	0.319
h1510	0.866	0.422	0.782	0.785	0.235	1	0.862	0.676	0.461	0.579	0.664	0.292	0.683	0.887	0.845	0.623	0.853
h1610	0.834	0.568	0.771	0.926	0.256	0.862	1	0.799	0.720	0.678	0.884	0.571	0.828	0.822	0.884	0.727	0.934
h1710	0.846	0.422	0.555	0.767	0.131	0.676	0.799	1	0.614	0.732	0.765	0.406	0.812	0.823	0.884	0.967	0.803
h1810	0.476	0.420	0.605	0.672	0.020	0.461	0.720	0.614	1	0.722	0.722	0.569	0.798	0.425	0.575	0.557	0.699
h1910	0.592	0.398	0.623	0.623	0.082	0.579	0.678	0.732	0.722	1	0.611	0.303	0.820	0.565	0.669	0.723	0.783
h2010	0.720	0.672	0.620	0.895	0.119	0.664	0.884	0.765	0.722	0.611	1	0.757	0.733	0.679	0.728	0.668	0.820
h2110	0.305	0.621	0.466	0.706	0.066	0.292	0.571	0.406	0.569	0.303	0.757	1	0.451	0.202	0.361	0.283	0.445
h2210	0.669	0.492	0.746	0.749	0.210	0.683	0.828	0.812	0.798	0.820	0.733	0.451	1	0.651	0.816	0.776	0.820
h2310	0.956	0.320	0.540	0.720	0.165	0.887	0.822	0.823	0.425	0.565	0.679	0.202	0.651	1	0.870	0.793	0.808
h2410	0.877	0.441	0.727	0.814	0.290	0.845	0.884	0.884	0.575	0.669	0.728	0.361	0.816	0.870	1	0.850	0.896
h2510	0.781	0.309	0.481	0.681	0.093	0.623	0.727	0.967	0.557	0.723	0.668	0.283	0.776	0.793	0.850	1	0.724
h2610	0.843	0.547	0.813	0.899	0.319	0.853	0.934	0.803	0.699	0.783	0.820	0.445	0.820	0.808	0.896	0.724	1
h2710	0.538	0.669	0.695	0.803	0.340	0.535	0.791	0.709	0.719	0.722	0.849	0.709	0.842	0.457	0.663	0.620	0.768
h2810	0.648	0.519	0.689	0.758	0.215	0.591	0.777	0.802	0.797	0.711	0.793	0.562	0.842	0.608	0.839	0.749	0.822
h2910	0.896	0.463	0.629	0.804	0.195	0.807	0.885	0.926	0.622	0.767	0.769	0.326	0.855	0.878	0.939	0.909	0.896
h3010	0.970	0.415	0.651	0.828	0.232	0.835	0.880	0.899	0.608	0.724	0.790	0.375	0.790	0.917	0.916	0.838	0.912
h3110	0.690	0.453	0.801	0.761	0.379	0.719	0.831	0.778	0.745	0.846	0.744	0.411	0.940	0.645	0.803	0.729	0.873
h0111	0.700	0.589	0.675	0.802	0.184	0.686	0.886	0.809	0.796	0.756	0.789	0.524	0.926	0.650	0.807	0.753	0.831
h0211	0.611	0.598	0.798	0.780	0.299	0.744	0.839	0.608	0.722	0.746	0.700	0.424	0.819	0.570	0.758	0.532	0.874
h0311	0.797	0.509	0.667	0.829	0.145	0.738	0.880	0.929	0.707	0.838	0.804	0.446	0.912	0.775	0.865	0.906	0.860

Table 3 - Aggregate packet size per hour: Correlation matrix

An approach to conceptual modelling of ETL processes

Saša Dupor*, Vladan Jovanović**
* sIT Solutions HR d.o.o., Rijeka, Croatia / Department of Informatics, University of Rijeka, Rijeka, Croatia
sasa.dupor@gmail.com
** Department of Computer Sciences, Georgia Southern University, Statesboro, Georgia, USA
vladan@georgiasouthern.edu

Abstract – In this paper, we propose a new conceptual model based on the visualization of data flow showing transformations of records accompanied with attribute transformations. A data warehouse development relies on the development of Extract, Transform, and Load (ETL) processes responsible for extracting data from transactional systems, data quality checks, transformation and conversion and loading data to target tables or cubes in a data warehouse. During the planning and design phases for data warehouse, the ETL conceptual model should be developed not only to show an overview of the whole process, but also to map sources, targets and necessary transformations, verify transformation logic, ensure that requirements are met and the needed structures and content exist. Once created, the conceptual model can be used for development of an ETL process, maintenance/optimization and redesign of a process due to new business requirements and database schema changes, all of which makes it a valuable addendum to data warehouse documentation.

Keywords: ETL process, data warehouse, conceptual model

I. INTRODUCTION

The need to collect, store and analyze data which can be generated through business or any other form of human activity have inspired and driven development of data warehouses. At first, data was analyzed by creating reports directly from transactional systems which were used to run business applications. But as the volume of data grew over time, so did the reporting and analysis requirements. If analytical reporting and a burst of business transactions (updating a DB) were run at the same time both tasks would suffer due to high workload. The solution was to build a special structure containing data for analysis – data warehouse.

Data warehouses (DW) are central places that contain historical data once generated on one or more transactional systems. Data can be used for analysis of past events, decision support, predicting future events and other purposes. To support specification and design of data warehouses, data models are created which differ from models used for transactional DB system.

Today, various DW modelling techniques exist, including the third normal form (3NF) model, the dimensional model [2] and the Data Vault model [4]. Different data models imply that source data must undergo a series of transformations to conform to a target structure. Also, data has to be cleansed and the quality insured so that analyses, which rely on that data, are accurate and relevant. Frequently only a portion of source data is used and modified in the process before being loaded into data warehouse.

The main goal of ETL processes is to facilitate moving and transformation of data. ETL is a collection of programs which adds significant value to data. It is far more than plumbing for getting data out of source systems and into the data warehouse. Specifically, the ETL system:

- Removes erroneous and/or corrects missing data
- Provides measures of confidence in data
- Captures the flow of transactional data for safekeeping
- Adjusts/reformat data from multiple sources to be used together
- Restructures data to fit end-user tools

The basic purpose of the ETL system is to get data out of the source and load it into the data warehouse. It is also used to clean and transform the data along the way. But the complexity of a particular implementation arises depending on data sources, business rules, existing software, and reporting applications. The challenge is to keep the focus on the simple overall mission of the ETL system [3].

Building the ETL system easily consumes up to 70 percent [1, 3] or even 80 percent of the resources needed for implementation and up to 55 percent of maintenance of a typical data warehouse [27]. Despite the fact that ETL takes up a great deal of resources and is the relevant component in the development and maintenance of a DW, there is limited literature that covers the architecture, design techniques, and methodology of data integration. Front-end business intelligence applications based on the DW provide executive dashboards with the multicolored pie and bar charts. But the processes that integrate, move, and populate the data have been largely ignored or misunderstood because it is simply hard, tedious, and highly disciplined work [1].

More challenges arise with the collection of more data, which can be structured or unstructured and comes from even more sources, often referred to as “big data.” Business users want to analyze more variables and get more insight about their business, demanding that all data

is stored and processed. ETL processes will have to handle these requests, find a way to combine sources using emerging technologies and combine structured and unstructured data in one integrated DW.

Planning and implementing the DW which fulfills the expected requirements requires extensive planning and development of ETL processes, which should start with a representation at the conceptual level as a starting point for understanding of developers and users [31]. So far, conceptual models have been developed for defining DW structure e.g. one described in [5], but the standardized method for ETL conceptual modelling has not been agreed upon yet, and researchers in the field agree that there is a relatively small amount of research literature and few approaches of ETL conceptual modelling technique exist, which may indicate more effort should be devoted to the development [12, 23, 32, 13].

The conceptual model proposed in this paper is inspired by previous work in the field and the need to present the ETL process on an abstract level. It is guided by the idea that process should be visualized to provide architects and developers with the „big picture“ of data transformations performed when moving data from source to target, thus enabling them to perform maintenance, optimization and further development of the process while keeping quality documentation.

We will concentrate on designing ETL processes which are created to run periodically (daily, weekly, monthly) to load data warehouse and data marts, with goals to ensure maximum productivity and minimum running time. Cases of special issues concerning initial batch loads and real-time and ad-hoc loads will not be covered.

Our contribution to the topic consists of:

- The proposal of a model which visualizes data flow of ETL process and separately lists attribute transformations
- The definition of a basic set of transformation elements which are used to visualize the data flow and reduce complexity of a process
- The definition of necessary information to be displayed in visualized data flows and in attribute transformation tables
- The concept of decomposing the visualization of the ETL process to a textual description using pseudo code derived from the visualization

The paper is organized as follows: In Section 2 we describe relevant work in this field. In Section 3 we present the proposition for conceptual modelling of ETL process. Section 4 presents a motivating example on which we demonstrate the model. Finally, in Section 5 we conclude the analysis and present some directions for future work.

II. RELATED WORK

There are different approaches present in the area of ETL conceptual modeling, published in research papers that propose different methods of model development. State of the art research in the area is systematically presented in [11] and [12].

In [11] a survey has been carried out covering issues concerning ETL process in DW, namely: conceptual and logical modeling of ETL processes; problems of each stage of the E-T-L; problems that pertain to the entirety of an ETL process; and review of some research prototypes of academic origin.

In [12] authors apply the method of systematic review to identify, extract and analyze the main proposals on modeling conceptual ETL processes. The main proposals are identified and compared based on the features, activities and notation of ETL processes. Questions that are considered include:

- At what level of design are ETL processes modeled?
- Does the approach define a methodology for defining the ETL process?
- Is the approach integrated with a comprehensive framework for the development of activity diagrams?
- Is the approach supported by a tool for modeling?
- What activities are described in the ETL approach?

One of the first proposals of the basis for ETL conceptual representation is described in [29]. Authors proposed a conceptual model customized for the tracing of inter-attribute relationships and the respective ETL activities. The proposed model is constructed in a customizable and extensible manner, and the authors developed a set of graphical elements for representing frequently used ETL activities.

Based on that proposal, the authors then presented a set of design steps that constitute the methodology for the design of the conceptual part of the overall ETL process. This design step includes: identification of the proper data stores; candidates and active candidates for the involved data stores; attribute mapping between the providers and consumers; and annotating the diagram with runtime constraints [28]. Further research led to the method of automatic transition from a conceptual model to a logical model [24] and responded to the need of automatic generation of program code and metadata directly from a model.

Further steps in that direction were made with the suggestion of the Entity Mapping Diagram (EMD) as a graphical model for representing ETL operations required to map data from sources to the target data warehouse or data mart [9, 32]. An improved design model and an optimization algorithm in mapping conceptual to logical models were proposed in [17]. Optimization techniques were described in [25], translation to logical models and proposal of architectural graph in [30].

Described modelling approach is based on the custom based visual representation which has been improved through time, but it does have limitations when describing

processes involving entities with large number of attributes.

Somewhat different approach is based on the Unified Modeling Language (UML). In [14] the authors define a set of UML stereotypes that represent the most common ETL tasks and use UML for the seamless integration of the design of the ETL processes with the DW conceptual schema. Further work resulted in a presentation of a disciplined framework for the modeling of relationships between sources and targets in different levels of granularity. To provide complementary views of the design artifacts in different levels of detail, the authors based their framework on a principled approach in the usage of UML packages, to allow zooming in and out the design of a scenario. The authors also extended UML to model attributes as first-class modeling elements. [13]. Additional set of modelling elements based on activity diagrams has been presented to capture dynamics of ETL processes [16]. While this approach tends to bring standards into modelling by using UML, it adds complexity to the representation. Additionally, activity diagrams visualizations would benefit from additional information about the process without adding complexity to the visualization.

An approach based on Semantic Web technologies was presented to facilitate the process of selecting relevant information from available data sources and appropriately transforming it to populate the data warehouse. The authors have used ontologies to formally and explicitly specify the semantics of the data stores' schemata, and presented possibility of automating, to a large degree, the process of the ETL workflow creation [26]. There is also another proposal of a tool called GEM [18] based on semi-automatic requirement driven design using OWL ontologies.

Authors in [7, 8] consider building ontologies a tedious task and they present model-driven framework, based on BPMN metamodel. They claim conceptual models can be straightforwardly mapped into logical models and implemented in any vendor tool, using Model-to-Text transformations. BPNM remains a choice to business-driven development of conceptual models, which can be improved by providing a new layer of abstraction that is based on a set of patterns [6]. However, BPNM is not completely suited to represent ETL, and additions are proposed also in [15] along with translation to logical and physical model.

There is another proposition of framework for modeling ETL processes named KANTARA [10] which offers a graphical working area and encapsulates the technical complexity of data.

Described proposition showcase the need and necessity to define common standardized conceptual ETL model, whether by defining new framework or adapting to existing standard. However, the most descriptive and informative model being simple enough to enter wider usage is still to be recognized.

We will also mention Data Vault modelling architecture and methodology [4] in a context of conceptual ETL modelling because it describes concept of standardization with only three types of DW tables defined and ETL code generated using repeatable patterns, which speeds up overall DW development. By loading raw data, which hasn't been transformed and/or cleansed and doing so when the data is ready and without constraining and conforming the data, it is possible to achieve very high throughput of data loading to the DW. Since there are no constraints and cleansing (business rules are enforced as far as possible downstream of the process, when building data marts from data warehouse), ETL processes don't break. This way the goals of speeding up ETL development and optimization of the processes are achieved.

III. PROPOSAL OF AN ETL CONCEPTUAL MODEL

The motivation for this proposal comes from a need to show an overview of the ETL process which was encountered through working on several data warehouse development projects. During the development, a conceptual DW model is created according to user requirements and with regard to available sources. At that time, a conceptual ETL model should also be created, which will guide developers into creating logical and physical models, respectively.

It is a realistic expectation that during the DW lifecycle the business users will request changes to the loading process, whether it were to be caused by business needs, compliance or regulatory requirements or even transaction systems changes by adding or removing functionalities. On the flip side, it is also realistic to expect that the developers who perform ETL maintenance or upgrades will change, because they will switch to other projects, change employers or there will be some other reason for workforce fluctuation.

In addition, the ETL processes are developed not only to retrieve information from various transactional system sources and load it to the DW, but also for loading specialized data marts that individual departments can use or are intended for different reporting applications. Recent developments tend to standardize the process of loading the DW by unifying types of tables and developing the method of data mapping, e.g. Data Vault model [4].

ETL can be developed by using 3GL/4GL (generation language) or visual tools. Different vendors developed different visual tools [19, 20, 21, 22] which considerably simplified the development of ETL by visualizing sources, targets and transformations. Each of these tools has a different set of predefined tools/transformations some of which are common (e.g. sorter, aggregator, joiner), but they also offer specific ones (e.g. deduplicator, rank, SQL qualifier). This situation can lead to different implementation of the same ETL concept when using different tools. Additionally, some of the tools require exact sequence or combination of transformations for some concepts to work, e.g. sorting records before joining or entering expressions to rank records before selecting the top n records. These specific rules and software requirements can add up to the complexity of visual mappings.

If the developer explores existing processes through visual data integration mappings, and tries to learn about

the concept of a process, all the necessary information cannot always be seen from the initial mapping representation. Information like join conditions in joiner or pivot setup for pivot transformation can be retrieved by opening individual transformations or additional dialog boxes. Showing details of a transformation often means showing all the fields involved in a transformation on the screen, and in cases of working with large datasets, the viewport can become filled quickly, causing the developer to lose focus. The exploration becomes even harder if the process is divided in phases through several mappings.

Consequently, we point out specific problems that may occur during the development and maintenance of the ETL:

- difficulties in communication between business users and ETL architects and developers caused by the lack of a standardized model
- time lost while maintaining the ETL process in which the developer must understand and explore the entire process in order to understand exactly what is happening with the data
- duplication of data processing and transformations caused by unawareness of all steps in the process, resulting in longer execution time and increasing complexity of data mappings, thus making maintenance more demanding - a phenomenon best described with the causal conditional "If you do not see the process, you will replicate that process" [1].
- missing the "big picture" of processing data from the transactional system source to the DW and then on to specialized data mart(s), regardless whether the process is constructed with visual mapping tools or with 3GL/4GL tools
- lack of adequate documentation which makes the ETL process harder to "read" and understand

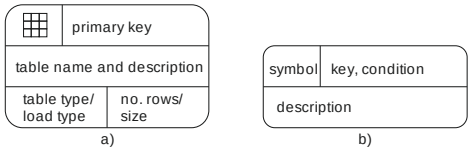


Figure 1. Visual representations of tables and record transformations

The model that we propose must fulfill the following criteria [1]:

1. Identify the major source data stores – What are the expected systems that the data will be extracted from and what is their size?
2. Document initial volumetric by source system
3. Review the data integration environment for reusable components
4. Define initial business rules and document them
5. Identify the major target data stores

The most important goal that we set up when creating the proposed conceptual model is that the model is as simple as possible and easy to read. It should be minimally affected by the complexity of the source and destination tables.

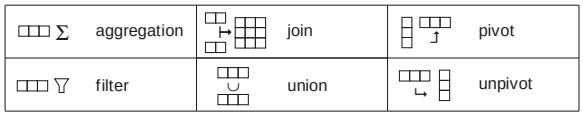


Figure 2. Transformation symbols

The ETL process is complex because there are multiple goals which have to be achieved. Currently, there are as many as 34 subsystems of ETL recognized [2]. In this paper, we propose simplification of the process in the conceptual model by separating the data flow containing records transformations only from attribute transformations. Data flow should be visualized and represent records by their keys, while the attributes will be described in accompanied table.

When creating the visual part of the model, we concentrate on the data record, as this is the main unit which is transformed. One data record is identified by the key and should have a described meaning in the model with a sentence which states the meaning of one record. We use the representation of a table (source or target) as shown on figure 1a. Elements of the visual block are:

- table type describes if it is database table or flat file
- load type denotes type of loading for the table, namely TI (truncate/insert), I (insert), U (update), M (merge) or SCDx (Slowly changing dimension where x denotes type) [2] if applicable
- no. rows/size states the projected or actual number of rows and size of the whole table in bytes.

Record based transformations are represented with visual element shown on figure 1b. Every transformation has its symbol which states what transformation takes place, key or condition by which it is performed and its description. The list of symbols is shown on figure 2.

There are different transformations defined by some vendors in ETL tools, but they can be reduced to one of mentioned transformations, e.g. splitting records can be

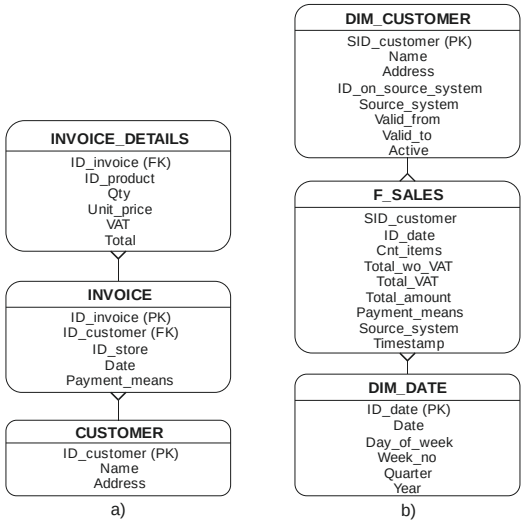


Figure 3. Model of source and target tables

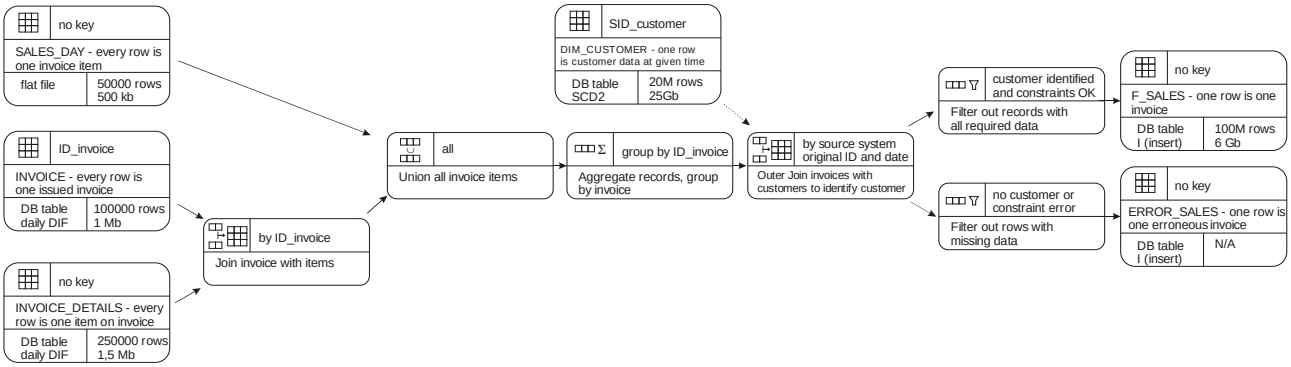


Figure 4. Visualizing the data flow of ETL process

done by series of filters and lookup can be done by outer join.

After visualizing data flow, it should be possible to decompose it to a series of sentences, going from one transformation to the other and combining the type of transformation, key and description, which combined tell a "story" of data record transformations. At that moment we have a description of the entire process which can then be communicated between ETL architects, developers, business users and other groups or individuals of interest.

During the ETL process, the transformation of attributes also takes place and should be included on the conceptual model. We propose that the attributes be shown in an additional table, so that the visualization remains simple and readable even if tables have a large number of columns. We create tables by starting from target requirements and requested data types, then map the attributes to source tables, try to describe the meaning of an attribute, describe how is it calculated and specify basic data quality rules.

IV. MOTIVATING EXAMPLE

We illustrate the usage of the proposed conceptual model through a motivating example.

Suppose we have a situation in the source system as shown in Figure 3a and we want to create a DW fact table of invoices which is connected to the customer and date dimension shown in Figure 3b. Let's assume that we get the data from two channels, sales in stores are recorded in the database system and loaded directly from it and online sales are in a flat file. Customer dimension tracks customer records and their historical values as separate rows (SCD type 2) and date dimension contains information about each date. We load the data daily, during the batch window at night.

Visualization of the process is shown on figure 4. We are able to "tell the story" about this process by constructing pseudocode sentences, for example: The process reads invoices and invoice items for the current day from the database and joins them. Those invoices with invoice items are unioned with invoices and invoice items loaded from the daily flat file and aggregated, grouped by invoice. We join every invoice to identify a customer on the invoice. Invoices which have all the required data are written to F_SALES table and invoices without a customer or missing data are written to error table.

We have described a process of record transformation with visualization, but we still need to describe attribute transformations. Steps to fill out the table include collecting requirements of the target and filling the rest of the table with data as shown in Table 1.

V. CONCLUSION

In this paper, we proposed a method for conceptual modeling of ETL processes which contributes to simplification of complex processes by showing a simple visual overview of the process. This proposal arose from recognition of diverse needs in planning, development and

TABLE I. ATTRIBUTES DESCRIPTION

Attribute	*Data type*	*Source(s)*	*Description/Meaning*	*Calculation*	*Constraint*
SID_customer	number	-	surrogate key of customer in DWH	-	not empty
ID_date	number	-	surrogate key of date in DWH	convert date to yyyymmdd format	not empty
Cnt_items	number	invoice_details, flat file	count items on invoice	count items on invoice	not empty
Total_wo_VAT	number	invoice, flat file	total amount without VAT	sum amounts without VAT	not empty
Total_VAT	number	invoice, flat file	total amount of VAT	sum VAT	not empty
Total_amount	number	invoice, flat file	total amount paid	sum total amounts	not empty
Payment_means	text	invoice, flat file	means of payment (e.g. cash, check)	-	not empty
Source_system	text	-	from what system the data arrived	-	not empty
Timestamp	date+time	-	date and time of entry in DWH	-	not empty

maintenance of ETL processes, all of which requires significant resources, time and money. It is estimated that the ETL processes take more than half of the budget for development and maintenance of data warehouses, but the standard for the ETL conceptual model has not yet been agreed upon.

The proposal meets the requirements of being able to easily present the process so it can easily be read, quickly developed, efficiently maintained and optimized according to the requirements of business users, overall ensuring minimal time is spent for understanding either new-to-be-developed or existing process. Simple and clear overview of the process supplies developers with the big picture and prevents creation of multiple processes that repeat data transformations and increase the total time required for ETL.

The key aspect of this model is that the data flow containing records transformations is separated from attribute transformations. Data flow is visualized and represents records by their keys, while the attributes are described in an additional table, so the visualization is not affected in the case of complex tables with 100 or even more attributes.

Further research on this topic may include addendums to the model in accordance with the conclusions that can arise after using the method on a higher number of ETL processes. Also, practitioners and researchers recognize a need for a way to simplify representation and exploration of the origin of data in the DW (data lineage). Conceptual model could offer a solution to that need. One of the challenges is the automatization of mapping conceptual models to logical models and maintaining the conceptual and logical models up to date and synchronized with the physical model. We believe it would be useful to develop software that would allow the design of models, as well as display of information that are of interest to all stakeholders – ETL architects, developers, business users and other stakeholders.

REFERENCES

[1] A. D. Giordano, "Data Integration Blueprint and Modeling: Techniques for a Scalable and Sustainable Architecture," IBM Press, 2011.

[2] R. Kimball, M. Ross, "Data warehouse toolkit," 3rd edition, Wiley Publishing, 2013.

[3] R. Kimball, J. Caserta, "The data warehouse ETL toolkit," Wiley Publishing, 2004.

[4] D. Linstedt, "Super charge your data warehouse," 2011.

[5] M. Golfarelli, S. Rizzi, "Data warehouse design: Modern principles and methodologies," Mc Graw Hill, 2009.

[6] B. Oliveira, V. Santos, and O. Belo, "Pattern-Based ETL Conceptual Modelling," MEDI 2013, pp. 237–248, 2013.

[7] Z. El Akkaoui, J. N. Mazón, A. Vaisman, and E. Zimányi, "BPMN-based conceptual modeling of ETL processes," DaWaK 2012, pp. 1–14, 2012.

[8] Z. El Akkaoui, J. N. Mazón, A. Vaisman, and E. Zimányi, "A model-driven framework for ETL process development," in Proceedings of the ACM 14th international workshop on Data Warehousing and OLAP (DOLAP'11), 2011, pp. 45–52.

[9] S. H. A. El-Sappagh, A. M. A. Hendawi, and A. H. El Bastawissy, "A proposed model for data warehouse ETL processes," J. King Saud Univ. - Comput. Inf. Sci., vol. 23, no. 2, pp. 91–104, Jul. 2011.

[10] A. Kabiri, F. Wadjinny, and D. Chiadmi, "Towards a framework for conceptual modeling of ETL processes," Communications in Computer and Information Science, vol. 241 CCIS. pp. 146–160, 2011.

[11] P. Vassiliadis, "A survey of Extract – Transform – Load technology," International Journal of Data Warehousing & Mining, vol. 5, July-September, pp. 1–27, 2009.

[12] L. Muñoz, J. Mazón, and J. Trujillo, "Systematic review and comparison of modeling ETL processes in data warehouse," Inf. Syst. Technol. (CISTI), 5th Iber. Conf., 2010.

[13] S. Luján-Mora, P. Vassiliadis, and J. Trujillo, "Data mapping diagrams for data warehouse design with UML," pp. 191–204, 2004.

[14] J. Trujillo and S. Luján-Mora, "A UML based approach for modeling ETL processes in data warehouses," Concept. Model. 2003, vol. 2813, pp. 307–320, 2003.

[15] K. Wilkinson, A. Simitsis, U. Dayal, and M. Castellanos, "Leveraging business process models for ETL design," In Proceedings of the 29th International Conference on Conceptual Modeling (ER'10), Vancouver, Canada, November 1-4, 2010.

[16] L. Muñoz, J.-N. Mazón, J. Pardillo, and J. Trujillo, "Modelling ETL processes of data warehouses with UML activity diagrams," Move to Meaningful Internet Syst. OTM 2008 Work, pp. 44–53, 2008.

[17] Q. Yao, M. Bai and M. Huang, "The research of methodology in models mapping for ETL processes based on model driven," *2008 Int. Symp. Inf. Sci. Eng.*, pp. 325–329, Dec. 2008.

[18] P. Jovanovic, O. Romero, A. Simitsis, and A. Abelló, "Requirement-driven creation and deployment of multidimensional and ETL designs," Adv. Concept. Model, pp. 391–395, 2012.

[19] Informatica Powercenter http://www.informatica.com/us/products/data-integration/enterprise/ on 01.02.2014.

[20] Oracle Warehouse Builder, http://www.oracle.com/technetwork/developer-tools/warehouse/overview/introduction/index.html on 01.02.2014.

[21] Microsoft SSIS http://www.microsoft.com/en-us/sqlserver/solutions-technologies/enterprise-information-management/integration-services.aspx on 01.02.2014.

[22] Talend Open Studio for Data Integration http://www.talend.com/products/data-integration on 01.02.2014.

[23] D. Skoutas, A. Simitsis, and T. Sellis, "Ontology-driven conceptual design of ETL processes using graph transformations," J. Data Semant. XIII, vol. pages, pp. 120–146, 2009.

[24] A. Simitsis and P. Vassiliadis, "A method for the mapping of conceptual designs to logical blueprints for ETL processes," Decision Support Systems, vol. 45, no. 1. pp. 22–40, 2008.

[25] A. Simitsis, P. Vassiliadis, T. Sellis, C. Of, and I. Of, "Modeling and optimization of extraction-transformation-Loading (ETL) processes in data warehouse environments," IEEE, 2004.

[26] D. Skoutas and A. Simitsis, "Designing ETL processes using semantic web technologies," Proc. 9th ACM Int. Work. Data Warehous. DOLAP '06, p. 67-74, 2006.

[27] D. Skoutas and A. Simitsis, "Flexible and customizable NL representation of requirements for ETL processes," 12th Int. Conf. Appl. Nat. Lang. to Inf. Syst. NLDB 2007, vol. 4592, pp. 433–439, 2007.

[28] A. Simitsis and P. Vassiliadis, "A Methodology for the conceptual modeling of ETL processes," in CAiSE workshops, 2003, pp. 305–316.

[29] P. Vassiliadis, A. Simitsis, and S. Skiadopoulos, "Conceptual modeling for ETL processes," Proc. 5th ACM Int. Work. Data Warehouses DOLAP '02, pp. 14–21, 2002.

[30] P. Vassiliadis, A. Simitsis, and S. Skiadopoulos, "Modeling ETL activities as graphs," DMDW, pp. 1–32, 2002.

[31] N. Roussopoulos and D. Karagiannis, "Conceptual modeling: past, present and the continuum of the future," pp. 139–152, 2009.

[32] A. Boshra and R. Hendawi, "Entity mapping diagram for modeling ETL processes," *researchgate.net*, 2005.

UNDERSTANDER
Business Intelligence Seeker – User Agent

Violeta Damjanovic*, Wernher Behrendt*
* Salzburg Research Forschungsgesellschaft, Salzburg, Austria
violeta.damjanovic@salzburgresearch.at
wernher.behrendt@salzburgresearch.at

Abstract – UNDERSTANDER is an exploratory R&D project investigating Business Intelligence (BI) as a knowledge-gathering tool for supporting innovation processes in specific industrial domains. We develop a knowledge-based model that is grounded on the following three concepts: (i) Business Intelligence Model (BIM) developed by a Canadian project for BI technologies, (ii) taxonomy of Competitive Intelligence (CI) developed by CI practitioners, and (iii) Conceptual Dependency (CD) theory dealing with verb-oriented organization of knowledge. We also investigate whether knowledge-based models can be made adequate for the problem of gathering BI from web resources. Such a knowledge model can be used to "prime" a software agent and the agent can then search autonomously for relevant information on the Web. To express the interaction between agents, we consider "mechanism design", which is a framework for designing interaction between self-interested agents to achieve specific outcomes. A BI Seeker is a multi- agent application developed in WADE. (Workflows and Agents Development Environment). WADE is an extension of JADE, a popular Open Source framework that brings workflow mechanisms to agents. We present the design of the BI Seeker as an application in the field of Industrial Internet.

I. INTRODUCTION

Knowledge representation and understanding is a prerequisite to any form of intelligent behavior. Business Intelligence (BI) applications are expected to gather new knowledge that can be used to innovate a firm's current manufacturing process. Hence, our motivation in UNDERSTANDER is to employ the BI Seeker as a knowledge-gathering tool with the focus on improving industrial innovation processes.

In this paper, we present our approach to the design and implementation of the BI Seeker's knowledge model and its internal logic. The knowledge model is based on a controlled vocabulary and taxonomy of terms for BI and Competitive Intelligence (CI), and is organized in the form of "scripts" as envisaged in Conceptual Dependency (CD) theory. The agent has to recognize such scripts when encountering a relevant web resource.

Our motivation for the approach taken in UNDERSTANDER goes back to classical AI models and theories. Hence, we explore the core of Schank & Abelson's CD theory that deals with the verb-oriented organization of knowledge [1]. We implement our knowledge model in SWI-Prolog, which provides a stable execution environment for knowledge representations with actionable semantics. SWI-Prolog is also capable of providing Web Services, and generating HTML pages dynamically, as well as providing data for web applications (c.f. http://www.swi-prolog.org/web/).

Section II discusses related work in BI and CI. Section III presents the basic concepts of the BI Seeker's knowledge model: (i) Business Intelligence Model (BIM), (ii) a taxonomy of CI developed by practitioners, and (iii) CD theory. Section IV describes the implementation of our knowledge model through integration of its basic concepts. Section V introduces the BI Seeker's internal logic, which utilizes mechanism design and preference aggregation for agents. Section VI shows our scenario, which employs the knowledge model in the field of manufacturing of home heating systems. Section VII presents the BI Seeker and its multi-agent environment developed in WADE. Section VIII concludes the paper.

II. RELATED WORK IN BI AND CI

The term BI was introduced by Hans Luhn, IBM researcher, in [2]: *"...[business] intelligence system… utilizes data-processing machines for auto-abstracting and auto-encoding of documents and for creating interest profiles for each of the "action points" in an organization. Both incoming and internally generated documents are automatically abstracted, characterized by a word pattern, and sent automatically to appropriate action points."* In a period between 1958 and 1989, only a small number of papers on BI were published. In 1989, Howard Dresner (Gartner analyst) proposed using BI as an umbrella term to describe "*concepts and methods to improve business decision making by using fact-based support systems"*. From 1989, Gartner's definition of BI became widely used as the primary definition [3][4][5][6].

Nowadays, BI is defined as data technology that gathers and stores data, analyses data and delivers information (and knowledge), facilitates reporting, querying, and allows organization to improve decision making [7][8][9][10][11][12][5]. BI can also be defined as the process that transforms data into information, then into knowledge [13]. BI is rooted in the Decision Support Systems (DSS), but is also associated with Knowledge Management (KM) and CI [7][14][11][5][15].

The authors in [16] differentiate between three BI approaches:

- **the traditional BI approach** - it is concerned with data aggregation, business analytics, and data visualization [8][17][5].
- **the business processes approach** - it is focused on the integration of BI into business processes [18][13][5][6].
- **the adaptive BI approach** - this is concerned with BI for self-adaptive systems [9].

Novel BI tools apply semantic search and analytics technology (e.g. Oracle Social Engagement and Monitoring Cloud Service for capturing consumer "considerations and preferences" as well as insights from consumer-generated content in social media automatically [19]. Some solutions use keywords to analyze information, while others rely on Natural Language Processing (NLP) to disambiguate content. Some efforts have already resulted in standards such as OMG's Business Process Modeling Notation (BPMN) [20].

BI is closely related to CI and Knowledge Management Systems (KMS) [7][11][12][5]. Negash in [11] presents CI as a branch of BI, and refers to it as *"a systematic and ethical program for gathering, analyzing and managing external information that can affect company's plans, decisions and operations"*. CI is concerned with gathering information and analysis of the behavior of competitors on the market. The authors in [21] summarize a set of "measures of effectiveness" (MOEs) that had been previously developed and used by professionals in the commercial IT field, and reapplied those measures for the purpose of CI.

Regarding CI technology tools used so far, the authors in [22] refer to the Global Intelligence Alliance (GIA) survey given in [23] as the basic reference on CI organizations, activity areas, tools, education, and challenges. To analyze information, most companies use competitor analysis and SWOT analysis. Other tools cover financial analysis, profit and loss analysis, industry analysis, customer segmentation, etc.

Finally, the tools and methods of CI, BI, KMS, and AI need to be integrated *"to provide value-added information and knowledge toward making organizational strategic decisions", as* the authors in [14] and [16] agreed. In UNDERSTANDER, we share that view and combine several technological approaches to assess their potential synergies and economies of scale.

III. UNDERSTANDER KNOWLEDGE MODEL: BASIC CONCEPTS

The UNDERSTANDER knowledge model relies on three technology pillars: (i) BIM model, (ii) CI taxonomy, and (iii) CD theory.

The authors in [24] stated: although BI systems are widely used in many companies, data still remains trapped when it comes to their analysis and interpretation. The main reasons for that are the following: *"(i) Systems are still very technical and data-oriented, (ii) It is hard to understand what the data means, (iii) It is hard to design queries or make new reports without technical knowledge or a knowledge of the underlying data structure, (iv) There is still gap between business and IT-supplied data."* The gap between "business" and "data" is the greatest factor in the cost of applying BI technology, which remains the greatest barrier to the overall adoption of BI technology [25]. This motivated the research of Mylopoulos and his team to raise the level of abstraction of BI systems through a modeling language that emphasizes concepts more familiar to business, i.e. strategic models, business models and strategies, business processes, markets, trends and risks. The result is called the **Business Intelligence Model (BIM)**, which became a part of the Canadian BI Network project (c.f. http://bin.cs.toronto.edu/home/index.php).

In addition to BIM, we adopted **taxonomy of CI** that is being jointly developed by CI analysts. The CI taxonomy is publicly accessible from: http://www.mindmeister.com/18753652/competitive-intelligence (password: editme).

Finally, **CD theory** is a theory about verb-oriented organization of knowledge proposed in [1]. It discusses representation of the meaning of natural language sentences in a way that facilitates the drawing of inferences from the sentences. In addition, CD introduces the idea of a canonical meaning representation without duplicating information; for example, (a) different words and structures may represent the same concept, (b) CD is meant to be a language-independent representation.

Lytinen et al. in [33] give a survey of the work of Schank & Abelson, and their successors. Here we briefly summarize the main aspects of the CD theory. Lytinen's survey points out that CD assumes a richly structured knowledge base and relatively simple processing rules. Intriguingly, Tom Gruber and his colleagues, scientists in the field of ontology engineering, systems for knowledge sharing and collective intelligence, took a similar approach when developing the SIRI speech-driven assistant that is used in Apple's iPhones. *"…Perhaps the most important common thread that connects the CD family of research is the assumption that* ***intelligence is knowledge-intensive****, rather than processing-intensive. If the representations used to perform a task are rich enough, then processing strategies traditionally associated with AI programs, such as search, become less important. The result is simpler processing theories which are highly dependent on a very rich, and highly organized, knowledge base."* [33]

IV. DOMAIN VOCABULARY BUILDING

As a result of integrating the above-described concepts in UNDERSTANDER, we build our domain vocabulary by combining approaches presented in BIM and CI taxonomies and we investigate the use of CD theory for representing knowledge and reasoning. Domain vocabulary building comprises the following steps:

- implementation of core BIM concepts and its relationships;
- implementation of core CI concepts and its relationships, plus their alignment to the BIM concepts and relationships;

- implementation of core CD concepts and its relationships, and their alignment to domain vocabulary built in previous two steps

A. Implementation of the BIM Taxonomy

BIM uses several existing languages and technologies for capturing business strategy [25]: Strategy Maps and Balanced Scorecards [26], OMG's Business Motivation Model [27], Dynamic SWOT (Strength, Weakness, Opportunity, Threat) Analysis by Dealtry [28], Goal models. In addition, BIM deals with the following reasoning techniques: Goal Model Reasoning, Probabilistic Decision Analysis, Reasoning with Indicators, and Hybrid Reasoning.

Abstract concepts for describing "things" in BIM are inspired by the DOLCE foundational ontology [29]. The intentional and social concepts are further adopted from the core concepts in Goal-Oriented Requirements Engineering [30][31]. In addition, the notion of influence is adopted from influence diagrams [32], which is a recognized decision analysis technique. Concepts related to SWOT analysis [28] and others have been adopted from OMG's Business Motivation Model standard [27].

BIM language semantics incorporates Description Logic (DL). It also allows for publishing generic BIM models as ontologies on the Semantic Web [24]. For example, the BIM formal definition is implemented in Web Ontology Language (OWL), and encoded in Protégé editor in the form of the **BIM Upper Level Ontology**. BIM is also intended to support the modeling and analysis of a business organization at both a strategic and a tactical level. It provides a set of constructs for modeling and analyzing a business context including intention, situation, processes, actors, influences, key performance indicators, and more [25].

In order to make use of the BIM Upper Level Ontology [24] in UNDERSTANDER, we further encode this ontology into SWI-prolog, by defining common components of domain vocabulary such as: classes (concepts), attributes (properties, features), relations, individuals (instances), function terms, restrictions, rules, axioms, and events. Relations are ways in which concepts and individuals relate to each other. Restrictions are descriptions of what must be true in order for some assertions to be accepted as input. Rules are statements describing the logical inferences that can be drawn from an assertion in a particular form. Axioms are assertions in a logical form that together comprise the overall theory that the ontology describes in its domain of application. Events represent the changing of attributes or relations.

Figure 1 shows the core BIM concepts represented as a knowledge hierarchy. For example, `Goal` (see Figure 1) is an objective of a business; `Situation` reflects internal or external factors influencing fulfillment of goals; `Indicator` shows performance measure of strategic activities, etc.

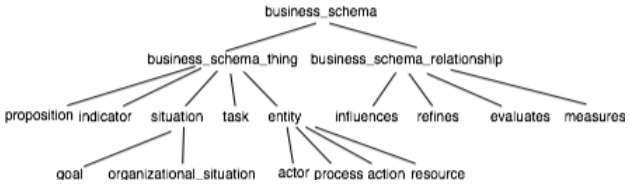


Figure 1. Knowledge hierarchy of the BIM

B. Implementation of the CI Taxonomy

The CI taxonomy consists of a "semantic cloud" of related concepts:

- **CI allied disciplines** – a list of relevant fields, e.g. market intelligence, knowledge management, etc.;
- **CI ecosystem** – considers the main elements of client/ vendor side, various organizations and associations;
- **CI decision support** – lists various decision support systems, i.e. strategy & portfolio, branding, etc.;
- **CI skills** - classifies existing approaches to applying CI e.g. researching CI or analytical CI.

Table 1 shows the alignment between the CI taxonomy and the BIM taxonomy. `CI ecosystems` classifies entities such as various clients (sales, marketing, library, etc.), vendors (small firms, major consultancies, etc.), organizations and associations. Hence, we align `CI ecosystems` (CI taxonomy) with the `Entity` class (BIM ontology). `CI allied disciplines` classifies common CI areas such as economic espionage, corporate security, market intelligence, etc. This information gives a more abstract view on previously defined entities, and can be expressed by adding to BIM, a new sub-concept of `Entity` called `EntityCategory`. In addition, `CI decision support` lists common ways of supporting CI, e.g. branding, go-to-market or customer analysis. Hence, we map this concept with the BIM's `Situation` concept. Finally, `CI skills` is the most important concept that lists various analytical tools and methods used to support enterprise, finance, evolutionary analysis, environmental analysis, etc. In order to encode this concept, we extend the BIM taxonomy by adding a new sub-concept of `BusinessSchemaRelationship`, namely `BusinessSchemaAnalyses`, which further includes sub-concepts such as `CIFinancialAnalysis`, `CIEvolutionaryAnalysis`, `CIEnvironmentalAnalysis`, `CIEnterpriseAnalysis`, `CICompetitiveAnalysis`.

TABLE I. CI AND BIM TAXONOMY ALLIGNMENT

CI Taxonomy	*BIM Taxonomy*	*Comment*
CI Ecosystem	`Entity`	
CI Allied disciplines	`BusinessSchemaThing/ Entity/ EntityCategroy`	`EntityCategory` is a newly added concept
Ci decision support	`BusinessSchemaThing/ Situation`	
CI skills	`Business Schema Relationship/ Business Schema Analyses`	`BusinessSchemaAnalyses` is a newly added concept

C. Implementation of Conceptual Dependency Theory

CD theory includes two core concepts: (i) CD primitives and (ii) CD conceptual categories. CD primitives represent semantics of possible actions, such as INGEST, EXPEL, GRASP, MOVE, PROPEL, PTRANS, ATRANS, MTRANS, MBUILD, ATTEND, and SPEAK. CD conceptual categories are *location*, *time*, *physical object*, etc. Encoding CD theory in SWI-Prolog requires defining its common components and their alignment to our domain vocabulary. The following table provides examples of the taxonomy alignment (at the schema level) between the core CD concepts and BIM taxonomy.

TABLE I. CD AND BIM TAXONOMY ALLIGNMENT

CD concepts	*BIM Taxonomy*	*Comment*
CD primitives	`BusinessSchemaThing/ Entity/ Task`	
CD conceptual categories/ PP	`BusinessSchemaThing/ Entity/ Actor`	PP stands for "picture producer", or "physical object"
CD conceptual categories/ ACT (action)	`BusinessSchemaThing/ Entity/ Action`	ACT (one of 11 primitive actions in CD)
CD conceptual categories/ LOC (location)	`BusinessSchemaThing/ Entity/ Location`	`Location` is a newly added concept
CD conceptual categories/ T (time)	`BusinessSchemaThing/ Entity/ Time`	`Time` is a newly added concept
CD conceptual categories/ AA (action aider)	`BusinessSchemaThing/ Entity/ Resource`	AA stands for "action aider"

Figure 2 shows the new BI/CI taxonomy, after merging the core concepts of BIM, CI and CD

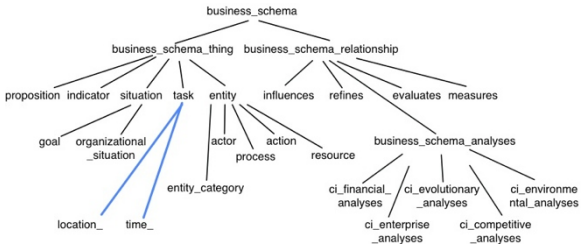


Figure 2. BI/CI concept hierarchy after merging BIM, CI and CD

The next task is to extend the generic BI/CI taxonomy by a specific domain vocabulary built e.g. for the home heating system scenario.

V. BI SEEKER – FUNCTONALITY DESIGN

Our knowledge model is to be used in an Industrial Internet application. At present, we consider two scenarios, one about 3D printing and the other about advanced home heating systems.

The BI Seeker's task is to find the most relevant business competitors and their commercial offerings. So the BI Seeker first searches for relevant web resources and then ranks them according to business criteria that can be set dynamically by making use of WADE's workflow mechanisms. In order to better assess competitors we employ an analysis technique known as mechanism design. Mechanism design looks at "*how preferences of several agents can be aggregated by a central designer"* [34]. *Note that mechanism design also addresses the issue of "strategic agents" (this means agents can even lie to others so as to maximize their individual payoffs), but we are not dealing with this case at present.* Thus, we are focused on designing preference aggregation mechanisms (ranking systems) to be employed within BI seeker, in the following real-world scenario.

VI. SCENARIO DESCRIPTION: CI FOR HOME HEATING TECHNOLOGY

We start with an introspective method to reach the following goal: *We want to know which firms are competing in the German speaking markets and what technologies they are selling. We are particularly interested in how they address the issue of integration into home automation systems. One possible approach is to first look for heater system manufacturers and then assembles a portfolio of technologies from their offerings. Another option would be to first look for technologies related to heating systems, and then look for manufacturers and their use of the technologies. The latter approach might yield more technologies than those currently on offer. Neither approach is likely to yield the specific issue of integration into home automation systems. On the contrary, we might be better off looking at pages about home automation in order to find those heating system manufacturers that are recognized as competent players in the home automation market.*

Our introspection will include the following steps:

- Step 1: Find competitors (manufacturer);
- Step 2: For each manufacturer, find types of heating systems (and relevant technology choices and developments);
- Step 3: For each manufacturer, search for offerings for home automation;
- Step 4: Find providers of home automation and what technologies and systems they are able to integrate into their solutions, and how.

VII. UNDERSTANDER MULTI AGENT ENVIRONMENT: THE APPLICATION

To perform the above steps, we develop a system of multi-agents that decide who are competitors in home heating systems technology and collect information about their technologies that are relevant for *home automation*. The question is how we can use our knowledge (domain vocabulary) to extract from the Web the information needed to compute the competences of manufacturers? How should we pool the preferences (findings) of several agents so as to best reflect the competences of competitors? To answer those questions, we employ our domain vocabulary, combining BIM, CI and CD concepts, to serve our system of multi-agents (BI seekers) as a knowledge reference.

BI seekers environment is implemented in WADE which supports the execution of tasks according to the workflow metaphor. One of the main advantages of WADE is in its workflow approach that brings the possibility of representing processes in a user-friendly graphical form. In addition, this approach utilizes the power of the Java programming language, and enables the usage of workflows for the internal logics definition that we use for the ranking of candidates.

Figure 3 shows the structure of our multi-agent system, which implements the scenario *BI Seeker for Home Heating Technology*. Here, we distinguish between two applications, namely the BI Home Heating Technology Application and the BI Ranking Application. The ***BI Home Heating Technology Application*** invokes a Web Service interface that collects the ranking results about specific companies, and/or their heating technologies, as input parameters. The ***BI Ranking Application*** finds both types of information by itself, and performs social ranking. A *Ranking Agent* is responsible for getting the user inputs and coordinating the ranking process, while a number of BI Seeker Agents correspond to the robots that retrieve the information about competitors in the field of interest (home heating systems, in our case).

All workflows that implement the relevant business logics to be executed by agents must be created in a separate module. Modules can be deployed at runtime, which allows hot-swapping of new business logics without the need to shut the system down and re-start it again.

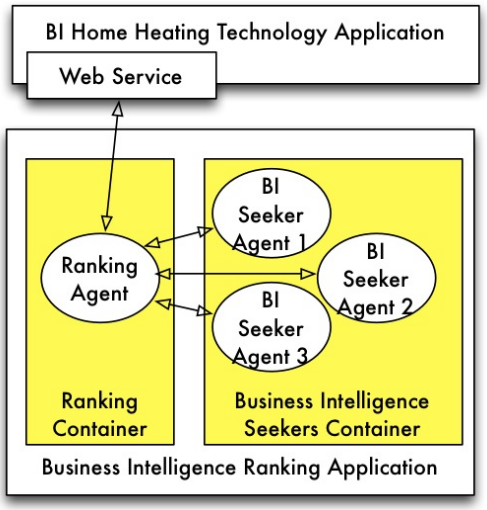


Figure 3. The structure of the BI Ranking Application

We now describe the social choice problem and its function in the context of our scenario. Let us consider a setting in which the set of agents is the same as the set of outcomes (see Figure 3 to figure out the relation between the agents and their strategies, and the outcomes).

Agents are designed as web crawlers with an internal ontology as described before (knowledge model presented in Section IV). Their task is to search for terms related to "*Heizungsanlagen*" (Heating Systems, eng.) (step 1 in the scenario). The agents are seeded with a number of web site URIs that either belongs to or point to, manufacturers of home heating systems. They gather publicly available information, which is then assessed against their ontology. In other words, we classify their technologies against an internal taxonomy of heating technologies (steps 2-4 in the scenario), and the agents then express their opinion about the importance of a particular heating technology (and manufacturer role), in order to determine a social ranking.

For the purpose of the first experimental setup, we investigate just three of the producers, one from Germany (*Buderus*), Austria (*KWB*) and Switzerland (*Elcotherm*). By firstly analyzing these three producers through humans we develop the agents' initial taxonomy of heating system technologies and components.

In order to perform ranking functionality, there exists a family of ranking algorithms. Ranking algorithms work by assigning agents numerical ranks that depend on the number of votes they have received, and breaking ties in favor of the agent who received a vote from the highest-ranked voter [13]. If this rule still yields a tie, it is applied recursively; when the recursion follows a cycle, the rank is a periodic rational number with period equal to the length of the cycle. At present, only the basic system is implemented and no performance testing has yet, taken place.

VIII. Conlusion and Future Work

In this paper we present our preliminary work on a multi-agent system that implements social choice and ranking functionality to perform BI/CI tasks in a specific manufacturing domain. More work is needed to really "understand" the nature of competitive assessment based on web resources.

Our work in UNDERSTANDER explores verb-oriented building of knowledge models as opposed to the prevalent, noun-oriented approaches. This approach has a potential to focus rather on the business service being provided, opposed to the business data being operated on. We are also aiming to provide an experimental context where our approach can be compared with and used by, Semantic Web and Linked Data technologies.

Acknowledgment

The research presented in this paper is based upon work supported by the UNDERSTANDER project (Austrian research grant 840230, awarded by FFG in the 2013 call for the "IKT der Zukunft" program).

References

[1] R.C. Schank and R.P. Abelson, "Scripts, Plans and Knowledge". Thinking: Readings in Cognitive Science. Proceedings of the Fourth International Joint Conference on Artificial Intelligence, Tbilisi, USSR, pp. 151-157, 1975.

[2] H.P. Luhn, "A Business Intelligence System". In IBM Journal, October 1958. Online available: http://altaplana.com/ibmrd0204H.pdf "", 1958

[3] D.J. Power. "A Brief History of Decision Support System." DSSResources.COM, Version 4.0, World Wide Web, 2008. Online available: http://dssresources.com/history/dsshistory.html.

[4] E. Turban, J.E. Aroson, T. Liang, R. Sharda. "Decision Support and Business Intelligence Systems." Pearson Prentice Hall, 2007

[5] E. Turban, R. Sharda, J.E. Aroson, D. King. "Business Intelligence: A Managerial Approach." Pearson Prentice Hall, 2008

[6] J. Zeller. "Business Intelligence: The Chicken or the Egg." BI Review Magazine, May 2007.

[7] T.D. Clark, M.C. Jones, C.P. Armstrong. "The Dynamic Structure of Management Support Systems: Theory Development, Research, Focus, and Direction." MIS Quarterly, 31, 2007, 579-615.

[8] S. Kudyba, and R. Hoptroff. "Data Mining and Business Intelligence: a Guide to Productivity." Idea Group Publishing, 2001.

[9] Z. Michalewicz, M. Schmidt, M. Michalewicz, C. Chiriac. "Adaptive Business Intelligence." Springer, 2007.

[10] L.T. Moss, and A. Shaku. "Business Intelligence Roadmap: The Complete Project Lifecycle for Decision-Support Applications." Pearson Education, 2003.

[11] S. Negash. "Business Intelligence." Communication of the Association for Information Systems, 13, 177-195, 2004.

[12] R.J. Thierauf. "Effective Business Intelligence Systems." Quorum Books, 2001.

[13] M. Golfarelli, S. Rizzi, I. Cella. "Beyond Data Warehousing: What`s Next in Business Intelligence." In DOLAP´04.1-6, 2004.

[14] J. Liebowitz. "Strategic Intelligence: Business Intelligence, Competitive Intelligence, and Knowledge Management." Auerbach Publications. 2006.

[15] J. Zeller. "Business Intelligence: the Road Trip." Information Management Special Reports, December 2, 2008.

[16] A. Azevedo and M.F. Santos, "Business Intelligence: State of the Art, Trends, and Open Issues". KMIS, pp. 296-300. INSTICC Press, 2009.

[17] M. Raisinghani. "Business Intelligence in the Digital Economy: Opportunities, Limitations and Risks." Idea Group Publishing, 2004.

[18] W.W. Eckerson. "Research Q&A: Performance Management Strategies." Business Intelligence Journal, 14, 24-27, 2009.

[19] "Social Media and Business Intelligence: Creating the Integrated Customer Hub." Oracle White Paper.

[20] Object Management Group. "Business process modeling notation (BPMN)." Online: http://www.omg.org/spec/BPMN/1.2/, 2009.

[21] P. Herring. "How Much is Your Competitive Intelligence Worth?" In SCIP2007, Vol. 10, No. 2, March-April 2007.

[22] M.L. Muller. "Global competitive intelligence practice." South African Journal of Information Management. Vol. 9(3), 2007.

[23] Competitive Intelligence in Large Companies - A Global Study. GIA White Paper 4/2005.

[24] J. Horkoff, D. Barone, L. Jiang, E. Yu, D. Amyot, A. Borgida, J. Mylopoulos, NSERC Report, 2012. Online available: http://www.cs.utoronto.ca/~jenhork/Presentations/BIM_TorontoSELab_July2012.pdf

[25] D. Barone, J. Mylopoulos, L. Jiang, and D. Amyot, "The Business Intelligence Model: Strategic Modelling", Technical report, 2010. Online available: ftp://ftp.csri.toronto.edu/csri-technical-reports/607/BIM-TechReport.pdf

[26] R.S. Kaplan and D.P. Norton, "The Balanced Scorecard: Transalting Strategy into Action", Aug. 1996, Harward Business Press Book.

[27] Object Management Group. Business Motivation Model. Ver. 1.1. Online available: http://www.omg.org/spec/BMM/1.1/

[28] T. R. Dealtry, "Dynamic Swot Analysis". Dynamic Swot Associates, 1994.

[29] A. Gangemi, N. Guarino, C. Masolo, A. Oltramari, and L. Schneider, "Sweetening ontologies with DOLCE". In Proc. of the 13th Int. Conf. on Knowl. Engineer. and Knowl. Manage. Ontol. and the Semant. Web, pp. 166–181, London, UK, 2002. Springer-Verlag.

[30] A. Dardenne, A. van Lamsweerde, and S. Fickas, "Goal-directed requirements acquisition", Sci. Comput. Program., 20(1-2):3–50, 1993.

[31] E. Yu, "Towards modelling and reasoning support for early-phase requirements engineering", In Proc. 3rd IEEE Int. Symp. on Requirements Engineering, IEEE CS, Washington, USA, 1997.

[32] R. Howard and J. Matheson. Influence diagrams. Readings on the Principles and Applications of Decision Analysis, Vol. II, 1984.

[33] S.L. Lytinen, "Conceptual Dependency and its Descendants". Computers, Mathematics, and Applications **23**(2-5):51-73, 1993.

[34] Y. Shoham and K. Leyton-Brown, "Multiganet Systems: Algorithmic, Game-Theoretic, and Logical Foundations", Cambridge University Press, 2009. Online available: http://www.masfoundations.org/

[35] T. Sandholm, "Automated Mechanism Design: A New Application Area for Search Algorithms", 2013. Online: http://www.cs.cmu.edu/~sandholm/amd_overview.cp03.pdf

Spend analysis systems

S. Maržić, P. Krneta and M. Pavlić
Ris d.o.o., Kastav, Croatia
{sanjin.marzic, perino.krneta, mile.pavlic}@ris.hr

Abstract – This paper introduces spend analysis, spend management and supplier management field. It describes the need for a spend analysis system, its' benefits and savings which such system can help to achieve. We present flowchart of spend analysis process in form of a dataflow diagram. We present main issues in development of a spend analysis software, and give an overview of current spend analysis systems.

I. INTRODUCTION

Usage of computers in business, development of information systems, development of programming languages and databases made businesses more efficient. Reporting is faster, richer in content, more accurate, many tasks are now automated. All of these technologies participated and are still participating in collecting large amounts of various (structured) data. There is data of company's core business, procurement, retail, human resources, revenues, spend. This data contains useful information and knowledge but to extract it one has to make various *ad-hoc* analyses over large volumes of data. Such analyses aren't feasible on transactional systems because these systems are optimized for daily work which consists of simple, predefined operations and predefined reporting. This problem was solved with data warehouses – data was periodically exported from transactional system into another database, optimized for complex analyses. In data warehouse data was additionally enriched (fulfilled) by data from other (external) sources to achieve better analyses.

During the 1980s in the USA a new need emerged – spend analysis [1]. This analysis had different requirements which couldn't be fully satisfied by data warehousing. In this analysis commodity and supplier data had to be treated differently. Spend analysis' goal is to reduce the company's expense by current expense analysis, supplier analysis, commodity analysis. Expenses are also reduced by centralizing procurement and supplier rationalization.

Spend management is a way for a company to control and optimize its expenses; it includes cutting operational costs as well as other expenses [2]. Spend management gives a company control over the money so that all business needs (manufacturing, delivering of goods and services) are satisfied in the most cost efficient way. Profit of a company is the difference between total incomes and total expenses, so profit can be augmented either by raising the income side or by reduction of expenses. In (current) period of financial crisis reduction of expenses is a common approach because the reduction of 1$ on expenses equals to raising incomes for 3$ [2].

II. SPEND ANALYSIS

Spend analysis is organization of expenses data (mostly procurement) and usage of supplier hierarchy and hierarchy of goods and services combined with amounts spent to:

(1) Find categories of expenses

(2) Find opportunities for strategic sourcing by gathering procurement requirements and supplier rationalization

(3) Reduce expenses with higher discounts (with contracts, volume discounts).

Finding categories of expenses implies categorizing goods and services (expenses) so that one can answer questions such as what was total amount spent on computer monitors. Transactional systems usually contain only data on specific purchase (e.g. specific model of computer monitor), but to answer mentioned question one should collect data of all purchases. For this kind of analysis a hierarchy of goods and services is required. This hierarchy will, for specific model of computer monitor, define that it is a monitor of specific size (e.g. 17' monitor), an LCD monitor, hardware element etc. Using such hierarchy analysis is possible inside and between categories, e.g. which category contains most expenses, how large is price range of products or services of same category. Such analysis can be used to detect corruption which can be manifested as big difference in price for products in same category. One of the existing hierarchies for categorization of products and services is UNSPSC (United Nations Standard Products and Services Code) [3]. This hierarchy was developed by United Nations and Duns & Bradstreet. In this hierarchy every product and service has a unique 8-digit code in format SS-FF-CC-DD where SS is 2-digit representation of segment, FF is family, CC class and DD commodity. Important features of UNSPSC are (1) simple drill-down and drill-up by using parts of code, (2) consistency which ensures that every element is defined in exactly one point in hierarchy, (3) completeness because all products and services that can be traded are included in hierarchy and (4) market tracking meaning that hierarchy is maintained as new products and services appear in the market.

Gathering procurement requirements enables enlargement of single procurement orders and reduction of total number of orders (by connecting smaller orders to the same supplier into larger ones or connecting orders for

similar products or services). With larger orders company can get better procurement conditions and lower prices. **Supplier rationalization** implies creating supplier hierarchy. With this hierarchy expenses and procurement data can be categorized and company can define which suppliers are preferred and which are not. Supplier hierarchy defines which suppliers (that appear in transactional system) are parts of the same company (e.g. offices of the same company or multiple records for the same company) or are connected in some other way. With supplier hierarchy the real expenses towards some supplier can be determined (in contrast to expense towards an office of a supplier). By combining supplier hierarchy with products and services hierarchy questions like "how many suppliers are supplying some product category" and "how is the expense in some product category distributed over suppliers" could be answered. With this data, available **expenses can be reduced with higher discounts** by reducing the number of suppliers for a product category. Thereby the expense towards remaining suppliers would enlarge which creates better negotiating position for getting higher discounts. By detecting undesirable (not preferred) suppliers costs with them can be lowered and transferred to other suppliers (which can lower risks and raise quality).

A. Who needs spend analysis?

Not all companies have the need for spend analysis and there are several reasons for that: spend analysis software is usually very expensive and therefore not applicable for small companies and such companies don't even have the need for spend analysis since their procurement is often centralized and not very large in volume, so it can be analyzed without software. Companies that can benefit from spend analysis are the ones that are physically dislocated, without central procurement, with multiple procurement centers. The fact that some company is physically dislocated is not enough for it to need spend analysis: e.g. a retailing company with multiple supermarkets won't be interested in spend management software because, although it has multiple supermarkets, it has central procurement and all procurement data at one place. On the other hand city administration could be interested in spend analysis because a city usually has several communal companies which are financed from the same city budget, but have independent procurement. Since each company is autonomous in procurement substantial savings can be done by gathering the procurement requests. Also, corruption can be detected by comparing the prices for similar products.

B. Savings

A question that can be asked is: what kind of savings can be expected after implementing spend analysis and in which areas of business?

Table 1 displays possible savings (in percentage) by categories. Such savings are possible with commitment to advanced purchasing methods (e.g. e-sourcing, e-auctions) [4]. All savings are achieved in one of three ways [1]:

- Strategic sourcing – finding the strategic suppliers and reduction of supplier number to achieve discounts and raise quality (quality is raised by choosing the right supplier)
- Reduction of maverick spend by choosing (contracting) suppliers and improvement of business processes (e.g. internal guidelines for procurement)
- Avoiding of spend leakage by controlling suppliers' compliance to contract terms.

TABLE I. POSSIBLE SAVINGS WITH COMMITMENT TO ADVANCED PURCHASING [4]

Category	Saving
Raw materials	2-5%
Packaging	10-20%
Indirect materials and services	10-20%
Information technology	15-30%
Professional services (including consulting, legal human resources)	8-15%
Capital projects	7-15%
Other indirects (nonproduction costs)	5-15%
Media/marketing/promotional items	10-20%
Logistics (rail, truck, barge, ocean, air)	7-15%

Third method is especially important because a part of savings that were negotiated and agreed to can be lost without control of compliance. All three points should be continuously repeated to achieve continuous savings.

Data in table 1 is relative to full amount spent per category, specific amounts can be estimated with amounts of total expenses for each category. Since savings are expressed as percentage, saving of 2% for raw materials can seem like an insignificant saving, but actually it is the largest saving for production business systems since raw materials in those systems are carrying the largest part of total expense. Larger percentage for other categories can now seem like a much smaller saving but, as figure 1 depicts, it is not always so.

Total raw materials cost	1,000,000,000
Saving 5%	50,000,000
Indirect expenses, services	500,000,000
Saving 15%	75,000,000

Figure 1. Example of possible savings [4]

C. Specific advangates of spend analysis

Spend analysis software enables its users an overview of expense data for the entire company and reduces the need for additional data processing regarding reports and trend analysis. Specific advantages:

- Visibility of all expenses of the company
- Substantial improvement of data accuracy and consistency
- Improvement of analyses quality over time
- Reduction of time required for creating ad-hoc reports from up to two weeks to almost momentarily
- Reduction of maverick spend for additional savings
- Ability to perform data cleaning and create ad-hoc reports without need for permanently employed data extraction experts

- Reduction of administrational expenses for processing and controlling of procurement orders
- Fast identification of savings opportunities
- Additional savings by selecting suppliers and larger orders
- Savings by identification of contracting opportunities
- Elimination of overpaying the supplier and guarantee of rebate realization
- Continuous improvement of compliance to contract terms and savings by tracking contracted and real prices
- Reduction of procurements' depending on IT experts for supplying the expense data
- Removal of disagreements about data quality by introducing the "single truth" for spend data.

III. SPEND ANALYSIS PROCES

Spend analysis, as a process, is similar to the data warehousing process – both processes are based on ETLA process (Extraction, Transformation, Loading and Analysis).

Data extraction is periodical fetching of data from the transactional system, as well as from other relevant data sources, to make analyses up to date. Data can be fetched from sources other than databases, like user files, internet, etc.

Data transformation is the most important part of the spend analysis process (besides the analysis itself). Quality of analysis step depends on quality of data transformation. In this step transactional data is being normalized (conversion of currencies, units of measure) and brought to the common format (e.g. putting all address data in one field with specific ordering of data elements or dividing data to multiple fields). Another important segment of data transformation step, which distinguishes it from standard transformation in data warehousing, is building the products and services hierarchy and supplier hierarchy. Supplier hierarchy means connecting naturally connected suppliers (e.g. branch offices with head office or duplicate entries) so that it is easy to determine total spend with some supplier and not with e.g. its branch offices. Products and services hierarchy is development of multilevel hierarchy which will allow analyzing spend on multiple levels. E.g. management will be interested in general spend distribution and IT department will only be interested in subcategory of IT equipment expenses and will analyze more closely its sub-categories. Duplicate entries are detected and merged, business data is being cleaned. Part of this step is also data enrichment: missing data is being filled and external sources are being used (e.g. stock market data, supplier financial data). Since there is no perfect classification algorithm (for classification of suppliers and products) there must be an option for user to correct the results of data transformation. Data of users' interference to transformation results should be used to make future transformations more precise.

Data loading is loading the transformed data into a database optimized for spend analysis.

Data analysis is step where reports are created using the transformed data. Analysis, as with data warehouse, must enable the user to perform slicing & dicing, allow multidimensional and hierarchical view on data. Visualization is very important element of analysis; it should be able to work with large set of data efficiently.

Figure 2 depicts a model of the process described. Diagram was drawn according to DFC (data flow chart) method [5, 6]. Figure 3 depicts data transformation sub process.

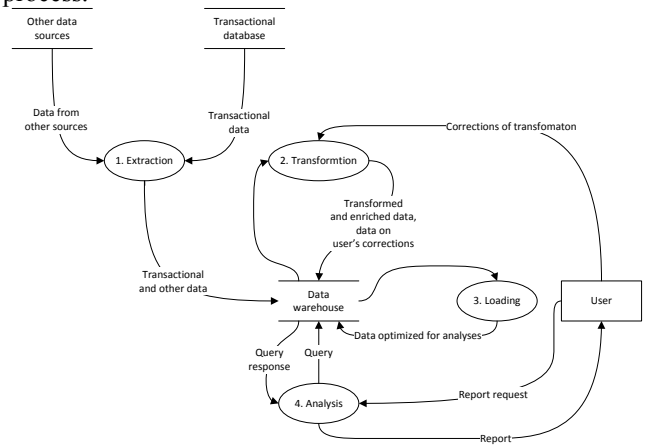


Figure 1. Spend analysis proces

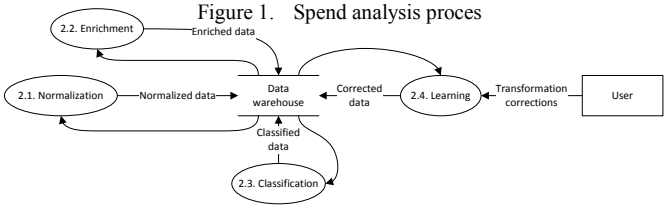


Figure 2. Data transformation process diagram

IV. SUPPLIER MANAGEMENT

Supplier management or supplier performance management is a process that consists of measuring, evaluating and controlling over suppliers' work, his business processes and habits with objective of reducing costs, lowering risk level and continuous progress [7].

Supplier management can be viewed as a business relation management for improvement of duplex understanding flow between a company and its suppliers, which especially concerns communication and negotiation about requirements and supply chain performance expectations [7]. Supplier management can include quality of delivered products and services, well-timed delivery, reliability of delivery (delivery of correct product in right amounts), adaptation to request changes, availability of required amounts of products or services when needed etc.

Supplier driven costs, or costs of poor quality (COPQ), can be divided into four groups [7]:

- Preventive expense – cost of actions that have to be performed to prevent appearance of bad products or services (work procedures, rules, planning, training)
- Evaluation expense – testing cost, evaluation or quality estimation cost
- Internal error expense – costs of low-grade products repairment (i.e. re doing the job)
- External expense – user support costs, dealing with customer complaints, loss of contracts (and reputation)

Low quality products and services costs are estimated to hold 10-25% of total sales, in that amount there is 25-75% cost produced by low supplier quality (2.5-18.75% of total sales) [7].

Low supplier quality costs are also affecting users by increasing price of company's products and services. Cost of low supplier quality rises with time needed to detect it. If problem is detected before the product is being used it can be rejected or returned to supplier which is the lowest cost; if, on the other hand, problem gets detected by the customer there are additional costs of customer service, complaint processing, returning the product, customer compensation and dealing with the returned product.

A good example of costs generated by using low quality product is cost of unusable X-ray image (table 2) [7]. Costs generated by one unusable X-ray image are paid time of technician and nurse who have to repeat the procedure, material expenses (new image, development materials, machine usage) and there is also a delay (and possible rescheduling) of the medical care that can't be given before usable X-ray image is taken. If such mistake happens five times per day and the cost of one error is evaluated to 33$ we get annual expense of 66,225$.

Information about total expense of low supplier quality is important because it helps prioritize such problems.

TABLE II. LOW SUPPLIER QUALITY EXAMPLE [7]

Element	Price
Technicians' time (30 minutes)	4$
Nurses' time (30 minutes)	9$
Film	7$
Development material	2$
Machine mainteanance	1$
Medical care delay	10$
Total price of one unusable image	**33$**
Annual expense (5 unusable imager per day, 365 days)	**66,225$**

V. SPEND MANAGEMENT SOFTWARE

First steps in spend analysis happened in 1980s, first spend management software appeared in early 2000s.

A. Properties of ideal spend analysis software [1]

In an ideal case spend analysis software would have the following elements:

1. Data selection and loading
2. Data cleaning, structuring and enrichment
3. Spend analysis
4. Knowledgebase management

Data selection and loading includes making a decision which data is going to be analyzed – what data is needed for spend analysis. With definition of necessary data, all data sources are defined [1]. Since data is fetched in a common database schema it should be brought to a compatible form, as data is arriving from heterogeneous sources. Data can be imported from different platforms, different databases, different file formats (Excel files, delimited files, etc.) [1].

Data cleaning, structuring and enrichment should have features:

1. Web oriented application with administrative and limited access – in cases where enrichment team is distributed each user should have access to specific products and services to prevent accidental (or unauthorized) change of data outside his/hers area of expertise
2. Data cleaning and enrichment – defined procedures for supplier data enrichment, classification of expenses (products, services), mechanism for conflict resolution. Expenses are classified by using rules (e.g. for supplier IBM expense is categorized as IT), these rules can be defined by an expert or by using computer learning algorithm. There can be a conflict set within set of rules (several rules with different outcome which can be activated in the same moment) so conflict resolution mechanism must be defined.
3. Defining and maintenance of hierarchies and dimensions – hierarchical schemes of products and services, supplier hierarchy, cost centers database are under continuous development so there must be one or more administrators for all these schemes, users should have limited authority over schemes and their parts to assure proper development of schemes.
4. Job scheduling – in large databases jobs can take a long time to complete, jobs should be well scheduled to assure the best system performance.
5. Import of end-user feedback – to improve system over time all feedback information should be collected to help in development of new rules based on collected information.

Spend analysis should have properties and functions:

1. Web-oriented application with administrative and limited privileges
2. Specialized data visualization tools
3. Reporting and dashboards

4. Sending feedback for changes in dimensions and hierarchies
5. Sending feedback for changes in classification
6. Immediate update of new data
7. *What-if* analyses

Knowledge base management means collecting data of user activity and to have an ability to learn upon collected data. System learns based on every action end user performs, so it should have ability to record every click of every user. Based on this information system can substantially improve its classification accuracy because it collects the data specific to user's industry. Knowledge database should have a lexicon to store data on three types of entities – concepts, instances and relations. Concepts are abstract terms (nouns, attributes, units of measure). Instances are descriptions of actual objects (e.g. "Dell 17' monitor"). Relations are structural and semantic links between concepts. Relations contain information on how concepts can be connected, which concepts are e.g. synonyms. Knowledge database is used in expense classification – the larger and more complete the lexicon is classification gets more accurate. Knowledge database management is an important task in spend analysis, there has to be a possibility to track changes in the database and to define a process for making changes, users who will have privilege of changing knowledge database.

In following sections some of the existing spend analysis applications are described.

B. Ariba

Ariba [8] is German software manufacturers founded in 1996, specialized for development of spend analysis software, contract management software, supplier management software. Their spend management application is "Ariba Spend Analysis". This application is used in software as a service model (SaaS). In 2012 Ariba was acquired by SAP.

Data imported into application is enriched with data from Dun & Bradstreet database (D&B). It is an American company that collects data on other companies to improve decision making, supplier chain management. Data is also enriched with data regarding ecology of business, company relations, and risks.

Properties of application that manufacturer points out is that it enables comparison with similar companies in the same industry; it enables predictive analyses and *what-if* analyses, classification (of products and services, suppliers) in predefined hierarchies or user defined hierarchies. Project manager is assigned by Ariba to every client.

C. Bravo Solution

Bravo Solution [9] is Italian spend management software manufacturer. The application "Bravo Solution Spend Management" is their application for spend management; as with Ariba, it is offered in SaaS model.

Properties of application that manufacturer points out is possibility to use multiple hierarchies to classify expense data, user can define their own hierarchies. Expenses are classified based on rules and classifiers trained by machine learning algorithm based on previous decisions. In supplier classification D&B database is being used as well as comparison to other available supplier attributes.

D. CVM Solutions

CVM Solutions [10] is supplier analysis software manufacturer. They are collecting supplier data so that suppliers themselves are supplying specific documents and reports, CVM Solutions are checking this data with public and closed sources and are continuously monitoring changes in supplier data. In this way they've collected a large volume of supplier data, so their service is supplier data cleaning (removing of multiple entries for same supplier, linking suppliers that are naturally connected), supplier data enrichment, analysis and control of company's suppliers and warnings in risky situations.

E. Applicability to Croatian (regional) circumstances

In Croatia and surrounding states there is not a great number of large companies which have the need for spend analysis, so spend analysis and management is not a developed field in the region. Above mentioned software is partially applicable to local companies because of several reasons: (1) D&B database and CVM Solutions database contains none or very small volume of data on regional companies so this data can't be used in supplier data cleaning; cleaning can only be done by record comparison, (2) mentioned solutions don't mention if their application interface is translated to Croatian language (or any of regional languages). In product classification some applications state that stop-words are used (words that don't contribute to the meaning of the statement and are ignored in text analysis, e.g. conjunctions, prepositions, articles), English stop-words aren't applicable to other languages so for Croatian and regional languages a research should be conducted to retrieve a list of such words. Another aspect that manufacturers aren't stating publicly is their software price. Since their customers are large corporations (they reference AT&T, Dell, Cisco, McDonalds, Mitsubishi), software price can be high because for these companies it will pay off with savings they will achieve. In region there is not very large number of large multinational companies, so their savings won't be so large; this is why software price will be a significant factor.

Existing software solutions are applicable to regional companies, but it can't be expected to give as good results as they achieve in "English speaking" companies. In Croatia a project MaxyTransparency [11] is started to develop a software solution that will be adjusted to language(s) and specific need of regional market.

VI. SPEND ANALYSIS ISSUES

In previous sections we've introduced spend analysis, requirements and causes of appearance, properties of an ideal software application and examples of existing

solutions. In development of a generic software application for spend analysis (or introduction of spend analysis to a business system) there are several issues (question) that have to be solved in different phases of the process. These issues and questions are displayed in table 3.

TABLE III. SPEND ANALYSIS ISSUES PER PHASES

Data extraction	
Refresh period	How often data extraction from transactional system has to be performed (weekly, monthly, once every quartal, yearly)? How recent analysis have to be? Should some data be imported more often?
Size of extraction	What data will be extracted/imported? Only new data or entire database will be refreshed? How to distinguish old and new data? How to recognize what data is already imported, but was changed meanwhile? Are some changes on transactional system required to optimize extractions? What data is allowed to be extracted? (important issue for demonstration of spend analysis power/ROI because real possibilities cannot be shown with missing data)
External sources	Which external sources will be used? How often? What is their price? How reliable/relevant are that sources?
Data transformation	
Product and services classification	How will products and services be classified? Which (how many) schemes will be used? How is the classification going to be performed? Which algorithms will be used?
Supplier hierarchy	How is the supplier hierarchy going to be built? How to detect/remove multiple records for the same supplier?
Data enrichment	What data will be used for data enrichment? (reliability of sources, relevance of sources, cost of this data)
Learning	How to improve future classification based on users interaction? Which algorithms will be used? How to train initial model (on what data)?
Errors	If an error is detected in data should source be corrected as well?
Data loading	
Schema	What schema (model) will be used? OLAP cubes are precalculating "answers to all questions", but for price of long refresh time and cubes are unable to give the *what-if* scenario analysis. Relational models, such as dimensional model, are faster in refresh, but queries take longer to execute.

Additional issues are present for spend analysis applications in regional circumstances. Some of reasons are language specificities (e.g. multiple word forms), writing specificities (special characters, often misuse of letters *č* and *ć*, replacements of *č* and *ć* with *c* or *đ* with *dj*). Spend analysis for companies that primary use English language can use results of numerous researches in natural language processing while in Croatian language and regional languages this research it is not possible because lack of interest for development of this area. There are no researches or tagged corpora for supplier (or products) matching model training.

Special problem is also to persuade companies that, if they are cohesive units, it is good for them to connect in unified procurement system to gather suppliers and products into a single supplier management system and usage e-supply (e.g. mutual e-auctions).

VII. CONCLUSION

In this paper we've introduced spend analysis. We've displayed expected results of spend analysis application usage and also given examples of losses, which happen without noticing, that occur when spend analysis is not applied. Process of spend analysis and supplier management is described. We've given list of features that software application for spend management should have. Also, we've analyzed usability of existing spend analysis applications on regional circumstances and identified issues that should be resolved in development of a spend analysis system.

Implementation of spend analysis system is recommended in every larger organization, public institution or private corporation, especially with large competition, lowering of income and market crisis. Spend management system can produce better results in combination with e-procurement system and supplier management system. E-procurement system helps the procurement sector; in combination with results from spend management system enables the procurement to focus on products and suppliers where most of savings can be achieved. Supplier management system holds an up-to-date supplier database, their products and services which are being used by both spend management system and e-procurement system. Such database can be created by joining several consumers and suppliers in single network: consumers can use it as an e-commerce system and suppliers as a platform for placement of their products and services.

One of suitable areas for implementation of these three systems is public sector (e.g. Republic of Croatia) and we hope that such system will be developed in near future.

ACKNOWLEDGMENT

The research has been conducted under the project "Extending the information system development methodology with artificial intelligence methods" (reference number 13.13.1.2.01.) supported by University of Rijeka (Croatia).

LITERATURE

[1] Pandit, K., Marmanis, H. (2008). *Spend analysis – the window into strategic sourcing*. J. Ross Publishing

[2] Wikipedia (2013). *Spend Management*, http://en.wikipedia.org/wiki/Spend_management

[3] Granada Research (2001). *Using the UNSPSC – United Nations standard products and services code*. Granada Research

[4] Rudzki, R. A., Smock, D. A., Katzorke, M., Stewart, S. J. (2006). *Straight to the bottom line*. J. Ross Publishing

[5] Pavlić, M. (2011). *Informacijski sustavi*. Školska knjiga, Zagreb

[6] Pavlić, M. (2011). *Oblikovanje baza podataka*, Odjel za informatiku Sveučilišta u Rijeci

[7] Gordon, S. R. (2008). *Supplier evaluation and performance excellence*. J. Ross Publishing

[8] Ariba http://www.ariba.com/

[9] Bravo Solution https://www.bravosolution.com/

[10] CVM Solutions http://www.cvmsolutions.com/

[11] Ris (2012) *Maxy Transparency: project documentation*. Ris d.o.o.

Poslovna inteligencija u sustavu informacijskog posrednika za elektronički račun

Danijel Matić*, Damir Kalpić**, Alan Čičak *
* InfoDom d.o.o., Zagreb, Hrvatska
** Sveučilište u Zagrebu Fakultet elektrotehnike i računarstva, Zagreb, Hrvatska
danijel.matic@infodom.hr, damir.kalpic@fer.hr, alan.cicak@infodom.hr

Sažetak - Elektroničko poslovanje je izvršavanje svih elemenata poslovnih aktivnosti ekonomskog procesa elektroničkim putem, odnosno stvaranje dodatnih vrijednosti koristeći se ICT tehnologijama. Informacijski posrednik je fizička ili pravna osoba koja u ime drugih obavlja otpremu, prijem i prijenos elektroničkih dokumenata.

U radu će biti prikazana upotreba poslovne inteligencije unutar sustava informacijskog posrednika za elektronički račun. Sustav služi za razmjenu prvenstveno elektroničkih računa, ali i za razmjenu ostalih dokumenata u elektroničkom obliku. Poslovna inteligencija je dio sustava informacijskog posrednika koji omogućava praćenje i upravljanje poslovnim procesima informacijskog posrednika. Primatelji elektroničkog računa mogu pratiti iznos troškova iz primljenih e-Računa te imaju uvid u strukturu troškova. Izdavatelji elektroničkog računa imaju uvid u uspješnost odvijanja poslovnih procesa isporuke e-Računa. Potporne službe informacijskog posrednika kroz sustav poslovne inteligencije nadziru odvijanje svih poslovnih procesa unutar sustava.

I. UVOD

Elektroničko poslovanje je oblik poslovanja koji prilikom transakcija i razmjena informacija koristi informacijske i komunikacijske tehnologije između sudionika u poslovnom procesu. Cilj je elektroničkog poslovanja automatizirati poslovne procese čime se smanjuju troškovi i povećava efikasnost. Elektroničko poslovanje pokriva poslovne procese cijelog dobavnog lanca.

Najraširenija elektronička isprava u svijetu je elektronički račun. Račun sadrži informacije koje su potrebne za fakturiranje. Podaci o računu mogu se koristiti za analizu poslovanja.

Poslovni partneri mogu u međusobnom poslovanju ugovoriti elektronički razmjenu poslovnih dokumenata. Fizička ili pravna osoba koja u ime drugih obavlja otpremu, prijem i prijenos elektroničkih dokumenata naziva se informacijski posrednik.

Informacijskim posrednicima neophodno je posjedovati pravovremene i relevantne informacije o svim aspektima poslovanja i odvijanju poslovnih procesa. Informacije je moguće brzo i jednostavno dobiti korištenjem poslovne inteligencije u sustavu informacijskog posrednika za elektronički račun.

Poslovna inteligencija (engl. Business Intelligence – BI) temelji se na podacima, koji se transformiraju u informacije potrebne za upravljanje i odlučivanje.

Sustav poslovne inteligencije u sustavu informacijskog posrednika treba osigurati da se podaci prikupljaju, transformiraju, skladište, kombiniraju, strukturiraju i čiste te treba osigurati kreiranje ključnih pokazatelja i izvještaja. Sustav poslovne inteligencije temelji se na skladištu podatka. Skladište podataka baza je podataka koja sadrži povijesne nepromjenjive podatke koji se prikupljaju i obrađuju radi potpore poslovnom odlučivanju [1].

Prema Kimballu [2] skladište je kopija transakcijskih podataka specijalno strukturirana za upite i analize. ETL (engl. *Extract, Transform, Load*) sustav je temelj skladišta podataka. Za unošenje podataka u skladište provode se ETL procesi. Dobro dizajnirani ETL sustav dohvaća podatke iz izvorišnog sustava, provjerava kvalitetu i dosljednost podataka, usklađuje podatke tako da se različiti izvori koriste zajedno i na kraju donosi podatke u prezentacijskom formatu koje pomoću aplikacija koriste krajnji korisnici za donošenje odluka [3].

Sustav informacijskog posrednika za elektronički račun sadrži važne podatke o poslovanju. Podatke je potrebno analizirati i pretvoriti u poslovne informacije koje se mogu koristiti za unapređenje i nadzor odvijanja poslovnih procesa unutar sustava informacijskog posrednika za elektronički račun.

Aplikacije koje pristupaju podacima BI sustava trebaju biti jednostavne i lake za uporabu, a informacije trebaju biti prikazane korisniku u najkraćem mogućem vremenu.

U ovom radu prikazana je upotreba poslovne inteligencije unutar sustava informacijskog posrednika za elektronički račun. U poglavlju II. opisana je generička arhitektura, funkcionalnosti sustava informacijskog posrednika i uloga sustava poslovne inteligencije unutar sustava informacijskog posrednika. U poglavlju III. izneseni su zahtjevi za implementaciju sustava poslovne inteligencije. U poglavlju IV predložen je dimenzijski model koji omogućava praćenje i upravljanje poslovnim procesima informacijskog posrednika. Izneseni su primjeri upotrebe BI sustava prilikom kreiranje izvještaja za različite korisnike sustava informacijskog posrednika.

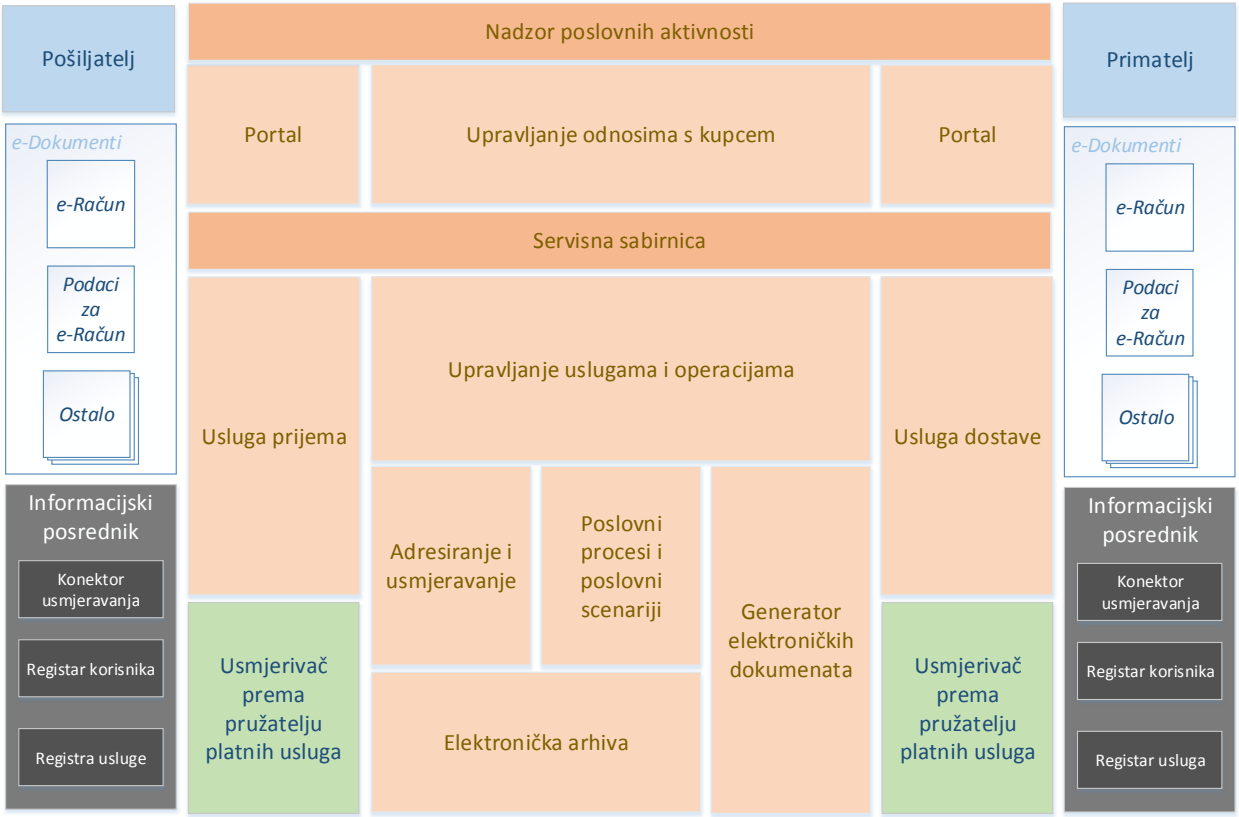
II. GENERIČKA ARHITEKTURA SUSTAVA INFORMACIJSKOG POSREDNIKA

U uvodu je spomenuto kako poslovni partneri mogu u međusobnom poslovanju ugovoriti elektroničku razmjenu poslovnih dokumenata i prilikom razmjene elektroničkih dokumenata mogu koristiti informacijske posrednike. Informacijski posrednik u ime drugih obavlja otpremu, prijem i prijenos elektroničkih dokumenata, odnosno realizira razmjenu elektroničkih poslovnih dokumenata. Informacijski posrednici osim razmjene elektroničkih dokumenta, obavljaju i funkcije elektroničke arhive, konverzije formata elektroničkih dokumenata, digitalnog potpisivanja i označavanje dokumenta elektroničkim žigom, izvještavanje korisnika, kreiranje elektroničkih dokumenata itd.

Generička funkcionalna arhitektura (slika 1.). implementira minimalni skup potrebnih funkcija za realizaciju poslovnih procesa informacijskog posrednika. Glavne funkcijske komponente sustava informacijskog posrednika za elektronički račun su:

- Nadzor poslovnih aktivnosti (engl. *Business activity monitoring* BAM) – funkcija nadzora uključuje sučelje putem kojeg se nadziru svi parametri funkcioniranja sustava od tehničke razine do razine poslovnih transakcija.
- Portal – dio sustava informacijskog posrednika čija je uloga pošiljateljima i primateljima omogućiti pristup informacijama vezanim uz proces isporuke, statusu predanih elektroničkih dokumenata, pregledu ugovorenih usluga i servisa. Funkcionalnost portala mora implementirati različite funkcije zavisno o profilu korisnika. Npr. privatni i poslovni korisnici, izdavatelji i primatelji i sl.
- Servisna sabirnica (engl. *Enterprise Service bus*) – predstavlja okosnicu sustava. Sve funkcionalne komponente koriste servisnu sabirnicu za međusobnu komunikaciju.
- Upravljanje uslugama i operacijama (engl. *Service Management & Operations*)
- Funkcija prijema elektroničkih dokumenata (engl. *Receiver Services*)
- Funkcija dostave elektroničkih dokumenata (engl. *Delivery Services*)
- Adresiranje i usmjeravanje (engl. *Addressing & Routing*)
- Poslovni procesi i poslovni scenariji (engl. *Business Processes & Scenarios*)
- Generator elektroničkih dokumenata (engl. *Electronic Document Generator*)
- Elektronička arhiva (engl. *e-Archive*)
- Usmjerivač prema pružatelju platnih usluga (engl. *Banking Gateway*)

U nastavku rada naglasak će biti na dijelove sustava koji koriste poslovnu inteligenciju.



Slika 1. Generička arhitektura informacijskog posrednika

A. *Nadzor poslovnih aktivnosti i Portal*

Potporne službe informacijskog posrednika kroz Nadzor poslovnih aktivnosti nadziru odvijanje svih poslovnih procesa unutar sustava. Nadzor poslovne aktivnosti je softver koji se temelji na poslovnoj inteligenciji i uključuje agregaciju, analizu i prezentaciju informacija u stvarnom vremenu o aktivnostima unutar sustava. Nadzor poslovnih aktivnosti pruža uvid u poslovne procese, točne informacije o stanju i rezultatima raznih operacija, procesa i transakcija. Omogućuje identifikaciju problema i rješavanja istih. Osim praćenja poslovnih procesa, Nadzor poslovnih aktivnosti može slati obavijesti koje se mogu koristiti kako bi se upozorili donositelji odluka o promjenama u poslovanju za koje je potrebno nešto poduzeti.

Na portalu se korisnicima omogućuje pregledavanje različitih unaprijed definiranih izvještaja. Portali koriste sustav poslovne inteligencije za izvještavanje tako što rade upite nad skladištem podataka.

III. ZAHTJEVI NA IMPLEMENTACIJU SUSTAVA POSLOVNE INTELIGENCIJE U SUSTAVU INFORMACIJSKOG POSREDNIKA

Sustav informacijskog posrednika za elektronički račun sadrži važne podatke o poslovanju i odvijanju poslovnih procesa. Iz tih podatka sustava može se oblikovati skladište podataka, koje čini temelj sustava poslovne inteligencije.

Prema Kimballu [2] neka od obilježja sustava poslovne inteligencije su:

- BI sustav mora omogućiti lak i brz pristup informacijama.
- BI sustav mora prezentirati informacije dosljedno.
- BI sustav mora biti prilagodljiv promjenama.
- BI sustav treba biti sigurna „utvrda“ za podatke i mora učinkovito kontrolirati pristup povjerljivim podacima.
- BI sustav mora služiti kao autoritativan i pouzdan temelj za poboljšanje odlučivanja. Skladište podataka mora imati prave podatke koji podupiru donošenje odluka.

Prije implementacije sustava poslovne inteligencije potrebno je identificirati koje informacije su potrebne korisnicima takvog sustava za analizu i donošenje odluka.

Sustav poslovne inteligencije u sustavu informacijskog posrednika mora omogućiti analiziranje osnovnih poslovnih procesa:

- proces dostave elektroničkog računa / elektroničkog dokumenta izdavatelja u sustav informacijskog posrednika
- proces isporuke elektroničkog računa krajnjem korisniku kroz različite kanale isporuke
- proces registracije i deregistracije krajnjeg korisnika za uslugu zaprimanja elektroničkog računa od izdavatelja

Iz perspektive informacijskog posrednika sustav BI treba sadržavati točne informacije o stanju i rezultatima razmjene elektroničkih dokumenata, podatke o transakcijama i dr. Potpornim službama informacijskog posrednika je potrebno osigurati informacije za praćenje rada sustava informacijskog posrednika i ključne pokazatelje uspješnosti rada sustava.

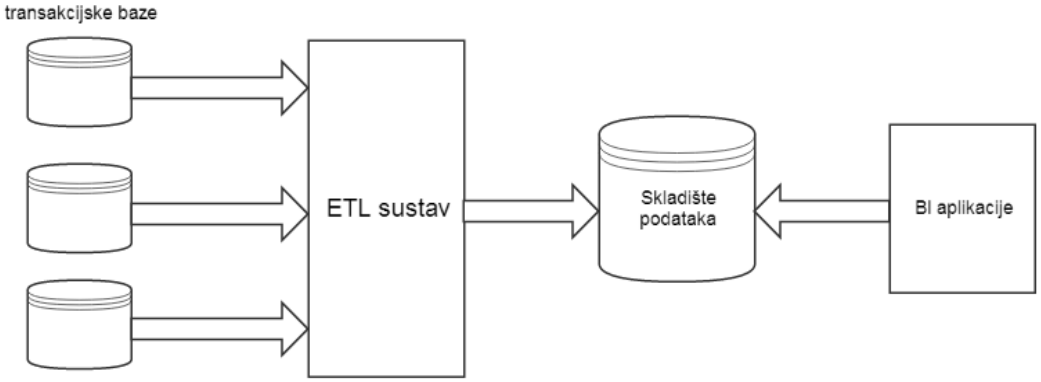
Krajnjem korisniku sustava trebale bi biti dostupne informacije o potrošnji po uslugama zaprimanja elektroničkog računa od izdavatelja. Također trebala bi mu se omogućiti statistika broja i iznosa računa po određenom izdavatelju i usluzi u određenom razdoblju.

Izdavatelji računa trebali bi imati informaciju o statusu isporuke računa i ostalih elektroničkih dokumenata krajnjim korisnicima. Izdavatelji bi htjeli imati informaciju o broju svojih korisnika koji koriste elektronički račun.

Sukladno iznesenim zahtjevima oblikovan je sustav poslovne inteligencije prikazan na Slika 2.

IV. SUSTAV POSLOVNE INTELIGENCIJA U SUSTAVU INFORMACIJSKOG POSREDNIKA

U prethodnom poglavlju izneseni su zahtjevi na implementaciju sustava poslovne inteligencije. Sustav



Slika 2. Sustav poslovne inteligencije [2]

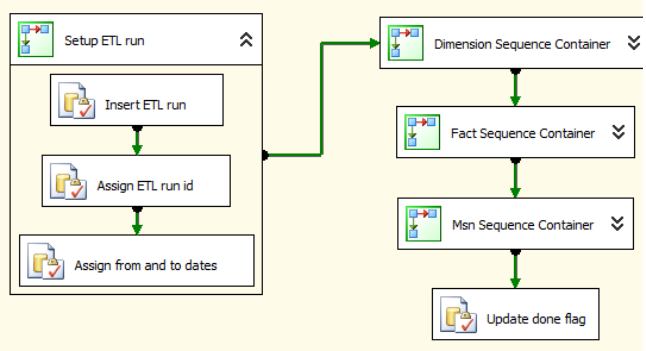
poslovne inteligencije u sustavu informacijskog posrednika za elektronički račun sastoji se od sljedećih komponenti:

- Transakcijske baze podataka – sadrže podatke iz transakcija poslovnih procesa informacijskog posrednika.
- ETL sustav – sastoji se od skupa procesa kojima se podaci dohvaćaju i čitaju iz izvorišnih transakcijskih baza podataka, transformiraju i spremaju u dimenzijski model skladišta podataka.
- Podatkovni dio BI sustava realiziran u skladu s konceptualnim (multi)dimenzijskim modelom podataka koji se implementira kao zvjezdasta shema u relacijskom modelu podataka i/ili kao nerelacijske multidimenzijske OLAP kocke, koje se kreiraju radi brže izvedbe upita.
- Aplikacije poslovne inteligencije za poslovno odlučivanje - rade upite nad skladištem podataka i/ili multidimenzijskim modelom OLAP kocki.

A. *ETL sustav*

ETL sustav uključuje sve procese između izvorišnih transakcijskih baza i skladišta podataka. Dohvaćanje podataka iz izvorišnih transakcijskih baza je prvi korak u ETL-u. Uključuje čitanje podataka iz izvorišnih sustava i kopiranje podataka u ETL sustav za daljnje manipulacije. Nakon toga slijedi niz transformacija. Ispravljaju se greške u podacima i podaci se raščlanjuju odredišne formate. Pri tome se kombiniraju podaci iz više izvora te se uklanjaju duplikati. Posljednji korak u ETL procesima je spremanje podataka u skladište podataka. Prilikom unosa podataka u skladište, ključno je zadovoljiti integritet dimenzijskog modela skladišta podataka vezan uz strukturu činjeničnih tablica i dimenzijskih tablica. Prilikom ETL-a prvo se pune dimenzijske tablice, a tek onda činjenične tablice. Budući da činjenične tablice sadrže strane ključeve dimenzijskih tablica [2], ključno je funkcijama za traženje (eng. lookup function) odijeliti ispravne vrijednosti stranih ključeva.

Slika 3 prikazuje kreirani proces ETL-a. Proces ETL-a je podijeljen u nekoliko koraka. Na početku se instancira proces i definiraju osnovne postavke procesa. Za potrebe praćenja i kontrole procesa ETL-a definirana je dodatna podatkovna struktura koju koristi ETL. Prilikom pokretanja procesa, ETL zapisuje koji paket ETL-a je pokrenut i vrijeme početka ETL-a. Zatim slijedi punjenje dimenzijskih tablica, činjeničnih tablica i mosnih tablica. Dodatno se mogu procesirati OLAP kocke. Nakon što je ETL završio, zapisuje se vrijeme završetka ETL-a i rezultat odvijanja procesa (uspješno ili neuspješno).



Slika 3. Koraci ETL procesa

Prilikom punjenja činjeničnih tablica može doći do grešaka. Primjerice funkcije za traženje ne mogu dodijeliti ispravne vrijednosti stranih ključeva iz dimenzija. Stoga je definiran mehanizam obrade grešaka. Za svaku činjeničnu tablicu u skladištu podataka definirana je njezina tablica grešaka u koju se spremaju greške prilikom ETL punjenja činjenične tablice. Jedan zapis u tablici grešaka sadrži podatke o vremenskoj oznaci greške, opisu greške, kôd greške, vrijednost pogrešnog podatka iz izvorišnog sustava i identifikator zapisa u izvorišnom sustavu.

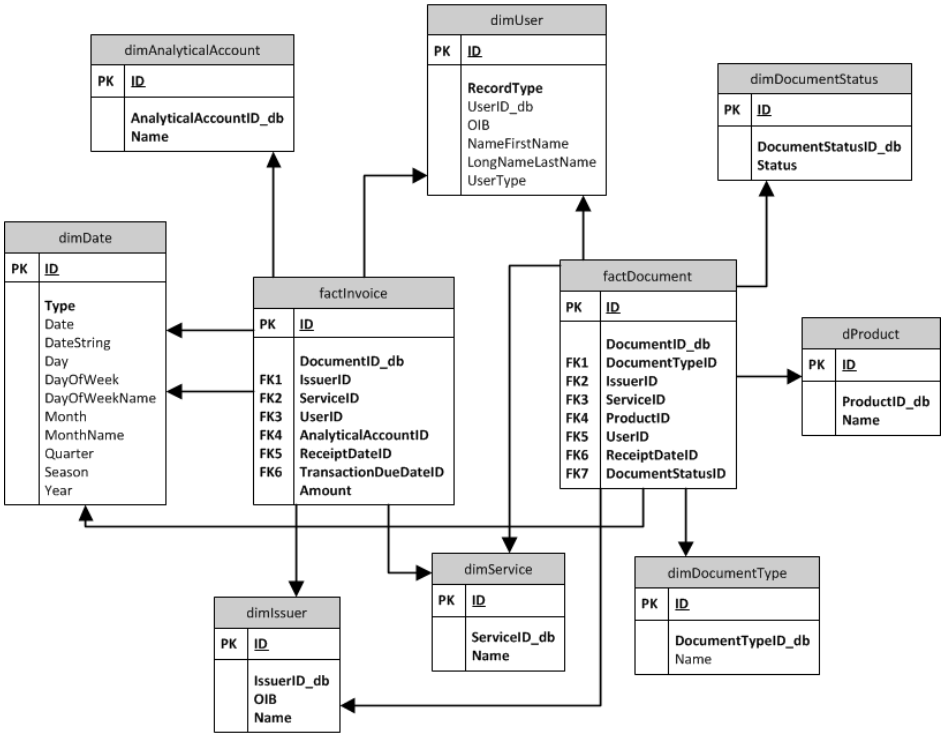
Sustav informacijskog posrednika sadrži mnogo podataka vezanih uz poslovne procese. Velika količina podataka može bitno utjecati na performansne ETL procesa. Stoga je korištenje inkrementalnog punjenja efikasnije.

B. *Model skladišta podataka*

Prilikom modeliranja skladišta analizirani su poslovni procesi navedeni u prethodnom poglavlju. U oblikovanom skladištu podataka svaki poslovni proces je predstavljen dimenzijskim modelom koji se sastoji od činjenične tablice, koja sadrži mjere događaja, i dimenzijskih tablica koje sadrže kontekst događaja. Više činjeničnih tablica dijele iste dimenzijske tablice. Činjenične tablice i njihove povezane dimenzije čine zvjezdasti model [2]. Predloženo skladište podataka informacijskog posrednika ima zvjezdasti model, a Slika 4. prikazuje dio predloženog skladišta podataka.

Nakon identifikacije koji poslovni procesi će se pratiti u skladištu podataka, potrebno je odrediti razinu granularnosti (razina detalja) podataka u skladištu podatka. Pri tome potrebno je odgovoriti na pitanje: Što znači jedan redak činjenične tablice? Za proces dostave elektroničkog računa / elektroničkog dokumenta izdavatelja u sustav informacijskog posrednika nastat će po jedan zapis u činjeničnoj tablici za svaki dokument krajnjeg korisnika koji je izdavatelj dostavio informacijskom posredniku. Prilikom procesa isporuke elektroničkog računa, činjenična tablica sadrži jedan zapis za svaki elektronički račun za uslugu izdavatelja koji je namijenjen krajnjem korisniku. U procesima registracije i deregistracije krajnjeg korisnika za uslugu zaprimanja elektroničkog računa od izdavatelja, činjenična tablica sadrži jedan zapis za svaki korisnički zahtjev za uslugom izdavatelja.

Granulacija definira primarnu dimenzionalnost činjeničnih tablica. Dimenzijske tablice sadrže tekstualni kontekst povezan s mjerljivim događajima poslovnih procesa. One opisuju "tko, što, gdje, kada, kako i zašto" je



Slika 4. Dio dimenzijskog modela skladišta

povezan s događajem. Dimenzije proizlaze iz analize poslovnog procesa i određivanje granulacije. Za procese informacijskog posrednika identificirane su sljedeće dimenzije: izdavatelj (*dimIssuer*), usluga izdavatelja (*dimService*), krajnji korisnik ili primatelj računa (*dimUser*), analitički konto korisnika u sustavu izdavatelja (*dimAnalyticalAccount*), produkt koji izdavatelj ima ugovoren u sustavu informacijskog posrednika (*dimProduct*), tip dokumenta (*dimDocumentType*), status dokumenta (*dimDocumentStatus*) itd. U procesima registracije i deregistracije krajnjeg korisnika za uslugu zaprimanja identificirane su još i dimenzije izvor zahtjeva (*dimRequestSource*), tip zahtjeva (*dimRequestType*) te status zahtjeva (*dimRequestStatus*). Nakon određivanja dimenzija potrebno je odrediti atribute dimenzija. Dimenzije imaju jednostavni surogatni ključ koji zamjenjuje ključeve izvorišnih podataka. Dimenzije dodatno sadrže i ključeve izvorišnih podataka kako bi se zadržala povratna veza na zapise u izvorišnim sustavima. Dimenzija izdavatelja računa (dimIssuer) i dimenzija krajnjeg korisnika (dimUser) sadrže osnovne matične podatke: naziv / ime i prezime i OIB. Dodatno se može proširiti s podacima o adresi. U izvorišnim sustavima podaci vremenskog tima su precizniji od razine jednog dana. Stoga oblikovano skladište sadrži vremensku dimenziju datuma (dimDate) i dimenziju koja predstavlja sekunde unutar dana (dimTime). Datumska dimenzija ima granulaciju do jednog dana. Izdavatelji računa ne izdaju račune samo na mjesečnoj bazi stoga je potreba uvesti i ostale atribute u dimenziji datuma čime se omogućava izvještavanja po danima u tjednu, mjesecima, kvartalima, godišnjim dobima i godinama. Vremenske dimenzije *dimDate* i *dimTime* inicijalno se napune vrijednostima i one se neće puniti tijekom procesa ETL.

Činjenične tablice sadrže mjere. Mjere su numerički atributi kojima se mjere procesi [2]. Zadnji korak prilikom dizajniranja skladišta podataka je određivanje mjera. U procesu dostave računa mjera je iznos računa.

Prilikom dostave elektroničkog računa od strane izdavatelja u sustav informacijskog posrednika, izdavatelj može greškom izdati račun za korisnika koji se ne nalazi u sustavu informacijskog posrednika. Informacijski posrednik može zaprimiti taj račun, ali ga ne može isporučiti korisniku. Stoga je u dimenziju korisnika dodan tip zapisa koji ima vrijednost 'Unknown' za nepoznatog korisnika.

Osim zvjezdastog modela sustav BI uključuje i multidimenzijske OLAP kocke. Online analitička obrada podataka (engl. *On-Line Analytical Processing* – OLAP) predstavlja grupu programa koji omogućavaju podršku odlučivanju i izvještavanje na temelju podataka iz skladišta podataka [4]. Pohrana podataka u multidimenzijske strukture zvane OLAP kocke [5] pruža veću brzinu izvedbe upita, mogućnosti bogatije analize i definiranje složenijih hijerarhija neodređene dubine nego pohrana u zvjezdaste sheme u relacijskom modelu. S druge strane, zvjezdasta shema pruža veći kapacitet.

C. Aplikacije poslovne inteligencije

U poglavlju II. ovog rada prikazana je generička arhitektura informacijskog posrednika. Portal je dio sustava informacijskog posrednika čija je uloga pošiljateljima i primateljima omogućiti pristup informacijama vezanim uz proces isporuke, statusu predanih elektroničkih dokumenata itd. Portali rade upite nad skladištem podataka i/ili multidimenzijskim modelom OLAP kocki. Funkcionalnost portala mora implementirati različite izvještaje zavisno o profilu korisnika: privatni i poslovni

korisnici, izdavatelji, primatelji itd. Prema zahtjevima na implementaciju sustava poslovne inteligencije u sustavu informacijskog posrednika na portalima bi trebali biti kreirani izvještaji. Krajnjem korisniku sustava trebale bi biti dostupne informacije o potrošnji, broju računa i iznos računa određenog izdavatelja u određenom razdoblju. Izdavatelji računa trebali bi imati informaciju o statusu isporuke računa i ostalih elektroničkih dokumenata krajnjim korisnicima. Potporne službe informacijskog posrednika trebale bi imati izvještaje kojima mogu pratiti odvijanje svih poslovnih procesa unutar sustava informacijskog posrednika.

V. ZAKLJUČAK

Sustav informacijskog posrednika za elektronički račun sadrži važne podatke koje je potrebno analizirati i pretvoriti u poslovne informacije. Dobivene informacije mogu se koristiti za unapređenje i nadzor odvijanja poslovnih procesa unutar sustava informacijskog posrednika. Kako bi se to omogućilo, potrebno je implementirati sustav poslovne inteligencije.

Predloženim sustavom poslovne inteligencije moguće je analizirati osnovne poslovne procese informacijskog posrednika. Sustav omogućuje lak i brz pristup informacijama. U skladištu podataka su podaci detaljno predstavljeni (niska razina granulacije) što osigurava prilagodljivost skladišta naknadnim promjenama. Predloženo skladište sadrži podatke koji su razumljivi krajnjem korisnicima, omogućuje brzo izvođenje upita i može se lako proširiti novim mjerama, dimenzijama i dimenzijskim atributima.

LITERATURA

[1] Z. Skočir, I. Matasić, B. Vrdoljak, "Organizacija obrade podataka", Merkur A.B.D., Zagreb, 2007.

[2] R. Kimball, M. Ross, "The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling", Third Edition, 2013.

[3] R. Kimball, J. Caserta, "The Data Warehouse ETL Toolkit. Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data", 2004

[4] Codd F, Codd S.B, Salley C.T. Technical Report. San Jose, Calif: Codd EF & Associates "Providing OLAP (Online Analytical Processing) to User-Analysts: An IT Mandate", 1993.

[5] Gray J, Bosworth A, Layman A, Pirahesh H. Data cube: A Relational Aggregation Operator Generalizing Group-by, Cross-Tabs and Sub-totals", 95-22. MSR Technical Report. Washington, DC: Microsoft Corporation; 1995.

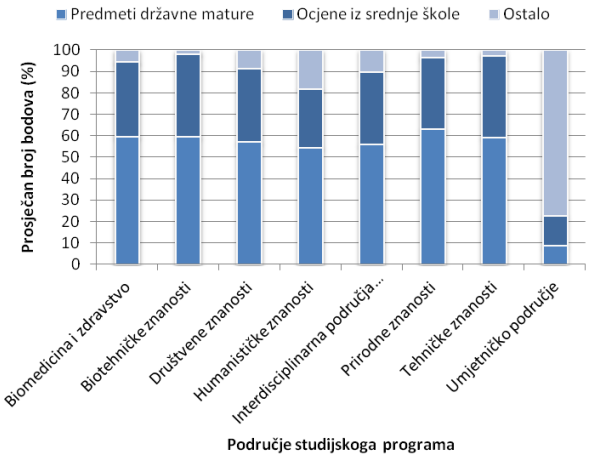
Analiza podataka u svrhu predlaganja studijskih programa u Nacionalnom informacijskom sustavu prijava na visoka učilišta

I. Budišćak*, V. Mornar* i D. Mlinarić*
* Fakultet elektrotehnike i računarstva, Sveučilište u Zagrebu, Hrvatska
ivan.budiscak@fer.hr, vedran.mornar@fer.hr, danijel.mlinaric@fer.hr

Sažetak – Nacionalni informacijski sustav prijava na visoka učilišta učenicima četvrtih razreda srednjih škola pruža mogućnost kreiranja svoje liste prioriteta najviše deset studijskih programa koje žele studirati. Neki od njih su neodlučni i ne znaju koje studijske programe odabrati. Učenicima se u sustavu prikazuje ostvareni broj bodova i plasman samo za studijske programe koji se nalaze na njihovoj listi prioriteta pa nisu u mogućnosti znati mogu li ostvariti pravo upisa na studijske programe koji nisu na njihovoj listi prioriteta. Budući da se rangiranje učenika obavlja periodički, svakih sat vremena, dodavanje i brisanje studijskih programa s liste prioriteta nije učinkovito rješenje za pronalazak studijskog programa koji odgovara učenikovim interesima i mogućnostima. U sklopu rada izradit će se model skladišta podataka za analizu ostvarenih rezultata učenika. Predložit će se postupak kojim će se neodlučnim učenicima predložiti studijski programi na temelju ocjena iz prva tri razreda srednje škole i postupak kojim će se odabrati studijski programi, koji se ne nalaze na listi prioriteta, za koje se učenicima dodatno može prikazati ostvareni broj bodova i mogući plasman.

I. UVOD

Za potrebe projekta državne mature razvili smo Nacionalni informacijski sustav prijava na visoka učilišta (NISpVU) koji objedinjuje cijeli postupak od prijava ispita državne mature i studijskih programa do rangiranja učenika za upis na studijske programe. Učenici mogu odabrati najviše deset studijskih programa koji čine njihovu listu prioriteta, a pravo upisa ostvarit će na onom studijskom programu koji se nalazi najviše na njihovoj listi prioriteta, a na čijoj se rang listi nalaze unutar upisne kvote. Slika 1. prikazuje prosječan udio bodova prema komponentama vrednovanja za sva područja studijskih programa. Najveći udio u bodovima, koji se u postupku rangiranja dodjeljuju učenicima, imaju rezultati ostvareni na ispitima državne mature i ocjene iz srednje škole. Kako bismo neodlučnim učenicima predložili studijske programe, uz ocjene koje su nam poznate za prva tri razreda srednje škole, potrebna nam je procjena rezultata ispita državne mature. Postupak za procjenu rezultata ispita državne mature opisat će se u III. poglavlju a u IV. poglavlju opisat će se postupak za predlaganje studijskih programa koji se temelji na prosječnoj ocjeni predmeta u prva tri razreda srednje škole i procijenjenom rezultatu ispita državne mature. U VI. poglavlju opisat će se model skladišta podataka iz kojega se mogu dohvatiti podatci potrebni za provedbu postupka predlaganja.



Slika 1. Prosječan udio bodova prema komponentama vrednovanja za sva područja studijskih programa

II. HRVATSKI SREDNJOŠKOLSKI SUSTAV [1]

Srednje škole, ovisno o vrsti nastavnog plana i programa koji provode, dijele se na:

- **gimnazije** - opće i specijalizirane.
- **strukovne škole** - tehničke, industrijske, obrtničke i druge, što se određuje prema vrsti nastavnog plana i programa.
- **umjetničke škole** - glazbene, plesne, likovne i druge, što se određuje prema vrsti nastavnog plana i programa.

Nastavni programi podijeljeni su u strukovna područja, a strukovna područja u obrazovne sektore.

Srednje obrazovanje učenika u gimnazijskim programima obrazovanja završava polaganjem državne mature (DM). Srednje obrazovanje učenika u strukovnim i umjetničkim programima obrazovanja, koji traju najmanje četiri godine, završava izradom i obranom završnoga rada u organizaciji i provedbi škole. Učenici četverogodišnjih strukovnih i umjetničkih srednjih škola mogu polagati i ispite državne mature, ako to žele.

Državna matura je skup ispita iz određenih nastavnih predmeta koje je učenik učio tijekom svoga najmanje četverogodišnjega srednjoškolskog obrazovanja. Cilj je državne mature provjera i vrednovanje postignutoga

znanja i sposobnosti učenika, stečenih obrazovanjem prema propisanim općeobrazovnim nastavnim planovima i programima. Na državnoj maturi učenici prijavljuju tri obvezatna predmeta (Hrvatski jezik, Matematiku i strani jezik), a od ponuđena 22 izborna predmeta učenici mogu odabrati najviše šest [2].

Sva visoka učilišta u Republici Hrvatskoj prihvatila su rezultate ispita državne mature kao jedan od uvjeta za rangiranje učenika za upis na studijske programe. Osim vrednovanja rezultata ispita državne mature, visoka učilišta vrednuju i ocjene iz srednje škole, a za pojedine studijske programe uvjetuju se i dodatne provjere posebnih znanja, vještina i sposobnosti potrebnih za uspješno studiranje. Osim toga, kod upisa na mnoge studijske programe na sveučilištima u Republici Hrvatskoj vrednuju se i neka dodatna postignuća učenika. Zbroj bodova prema svim navedenim komponentama vrednovanja je broj bodova s kojima se na kraju postupka kandidati rangiraju za upis na visoka učilišta.

III. PROCJENA REZULTATA NA ISPITU DRŽAVNE MATURE

Procjenu rezultata na ispitu državne mature moguće je učiniti na temelju ostvarenih rezultata učenika iz prethodnih generacija. Možemo reći da, s određenom vjerojatnošću, učenici s boljom prosječnom ocjenom predmeta u prva tri razreda srednje škole ostvaruju bolji rezultat na ispitu državne mature [3]. Vjerojatnost je tim veća što je veći koeficijent korelacije između prosječne ocjene predmeta u prva tri razreda srednje škole i rezultata ostvarenoga na ispitu državne mature.

Za procjenu rezultata na ispitu državne mature, na temelju prosječne ocjene predmeta u prva tri razreda srednje škole, predlaže se korištenje postupka regresije [4], [5], [6]. U postupku regresije ciljna vrijednost $y \in \mathbf{R}$ je kontinuirana, a na temelju primjera $\boldsymbol{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\}$ potrebno je naučiti nepoznatu funkciju $f : \boldsymbol{X} \rightarrow \mathbf{R}$ tako da, idealno, $y^{(i)} = f(\mathbf{x}^{(i)})$, gdje $\boldsymbol{X}$ predstavlja ulazni prostor primjera. Skup svih primjera $\boldsymbol{D}$ za postupak regresije, sastoji se od prosječne ocjene predmeta u prva tri razreda srednje škole i ostvarenoga rezultata (postotka riješenosti) na ispitu državne mature, za određeni skup učenika. Skup primjera može sadržavati podatke za učenike jednoga odjeljenja, jedne škole ili učenike koji pohađaju određeni program, strukovno područje ili obrazovni sektor.

Vrijednost koeficijenta korelacije između prosječne ocjene predmeta u prva tri razreda srednje škole i rezultata ostvarenoga na ispitu državne mature razlikuje se od škole do škole. Zbog toga je potrebno razviti više regresijskih modela, primjerice za svaku školu, umjesto jednog univerzalnog regresijskog modela za sve učenike.

IV. ODABIR STUDIJSKIH PROGRAMA ZA PREDLAGANJE

Odabir studijskih programa koji će se predložili učeniku obavlja se na temelju procijenjenih rezultata ispita državne mature. Ovdje je bitno naglasiti da u izbornom dijelu učenici mogu prijaviti predmet koji nisu učili u srednjoj školi. To znači da za predmete koje učenik nije učio u srednjoj školi ne možemo napraviti procjenu rezultata na ispitu državne mature.

Postupak kojim će se učeniku predložiti studijski programi sastoji se od ovih sedam koraka:

1) Definira se skup svih studijskih programa $\boldsymbol{Q}$.
2) Odabire se skup učenika $\boldsymbol{U}$ čiji će se rezultati koristiti kao skup ulaznih primjera za učenje (50 %), provjeru (25 %) i testiranje (25 %) postupka regresije. To mogu biti učenici istoga odjeljenja, škole, programa, strukovnoga područja ili obrazovnoga sektora, koji su prethodne školske godine ostvarili upis na neki studijski program u ljetnome upisnom roku.
3) Odabire se skup predmeta $\boldsymbol{P}$. To su svi predmeti koje je učenik učio u srednjoj školi i za koje postoje rezultati ispita državne mature u prethodnoj školskoj godini za učenike iz skupa $\boldsymbol{U}$.
4) Za svaki predmet $p_i \in \boldsymbol{P}$, postupkom regresije opisanom u III. poglavlju, procjenjuje se učenikov rezultat na ispitu državne mature r_i.
5) Za svaki predmet $p_i \in \boldsymbol{P}$, odabire se skup učenika $\boldsymbol{K_i} \subseteq \boldsymbol{U}$ koji su na ispitu državne mature ostvarili rezultat u intervalu $[r_i - \varepsilon, r_i + \varepsilon]$. Na ovaj način obuhvatiti ćemo učenike koji su na ispitu ostvarili sličan rezultat, tj. rezultat koji je bolji ili lošiji za vrijednost ε. Vrijednost ε može se odrediti empirijski na podatcima prethodnih generacija učenika.
6) Za svaki predmet $p_i \in \boldsymbol{P}$, svim studijskim programima $s_j \in \boldsymbol{Q}$ određuje se postotak učenika n_i, iz skupa $\boldsymbol{K_i}$, koji su ostvarili pravo upisa.
7) Svakom studijskom programu $s_j \in \boldsymbol{Q}$ pridjeljuje se težina t_j određena medijanom vrijednosti iz skupa $\boldsymbol{M_k} = \{n_i\}$, $i \in [1, |\boldsymbol{P}|]$. Konačno rangiranje studijskih programa provodi se na temelju pridijeljene težine t_j.
8) Od tako rangiranih studijskih programa učeniku se predlaže prvih deset.

Umjesto promatranja studijskoga programa na kojemu su učenici ostvarili pravo upisa, u šestome koraku predloženoga postupka, možemo promatrati znanstveno polje ili područje studijskoga programa. Tako možemo dobiti rang-listu znanstvenih polja ili područja studijskih programa i dodatno prilagoditi predložene studijske programe. Za svako od najboljih deset znanstvenih polja ili područja studijskih programa, može se uzeti studijski program koji pripada tom znanstvenom polju ili području, a koji se nalazi najviše na rang-listi studijskih programa dobivenoj prethodno opisanim postupkom. Dodatnu prilagodbu predloženih studijskih programa može obaviti i učenik odabirom znanstvenih polja ili područja studijskih programa koji odgovaraju njegovim željama ili interesima ako ih ima.

Na temelju procijenjenih rezultata ispita državne mature i ostvarenih ocjena u prva tri razreda srednje škole, moguće je procijeniti i broj bodova za rangiranje koje će učenik ostvariti na ponuđenim studijskim programima. Usporedbom s rang-listama iz prethodnih godina, moguće je, s određenom vjerojatnošću, procijeniti učenikov plasman na ponuđenim studijskim programima a time i

hoće li ostvariti pravo upisa. Ako učenik želi studirati na nekom od ponuđenih studijskih programa na kojem s procijenjenim plasmanom ne ostvaruje pravo upisa, prikazana procjena može ga potaknuti na dodatno učenje predmeta koji se vrednuju za upis na tom studijskom programu. Dodatnim učenjem učenik bi mogao ostvariti bolje rezultate na ispitima državne mature a time i bolji plasman od procijenjenog.

V. ODABIR STUDIJSKIH PROGRAMA ZA DODATNI PRIKAZ OSTVARENIH BODOVA

Nakon što su poznati rezultati ispita državne mature u NISpVU postoje svi potrebni podatci za rangiranje učenika na svim studijskim programima. U tom trenutku promjene na rang-listama uzrokovane su isključivo izmjenama koje učenici rade na svojim listama prioriteta. Učenici koji nemaju pravo upisa na nekome od željenih studijskih programa nemaju informaciju o plasmanu na srodnim studijskim programima ako se oni ne nalaze na njihovoj listi prioriteta. Ako učenik na svoju listu prioriteta doda studijski program, mora čekati izračun bodova koji se obavlja periodički, svakih sat vremena. Brisanje prijave za studijski program na kojemu je učenik ostvario pravo upisa, ima za posljedicu da učenik koji je na tom studijskom programu bio rangiran kao prvi koji se ne nalazi unutar upisne kvote ostvaruje pravo upisa a gubi ga na studijskom programu na kojemu je imao pravo upisa prije promjene. Promjene se nastavljaju lančano, dok se ne dođe do studijskoga programa kojemu je upisna kvota veća od broja prijavljenih učenika. Dodavanje studijskoga programa na kojem učenik po broju bodova ostvaruje pravo upisa ima za posljedicu gubitak prava upisa za učenika koji je bio rangiran kao posljednji unutar upisne kvote. Slično kao i kod brisanja prijave, promjene se nastavljaju lančano. Upravo zbog toga što prijava odnosno odjava može izazvati lančanu reakciju, izračun bodova obavlja se periodički.

Nakon svakoga izračuna bodova i generiranja rang-lista može se obaviti i izračun bodova za studijske programe koji se ne nalaze na učenikovoj listi prioriteta. Izračun se može obaviti samo za one studijske programe koji pripadaju istom znanstvenom polju kao i studijski programi koji se na učenikovoj listi prioriteta nalaze na višem prioritetu nego studijski program na kojemu je ostvareno pravo upisa. Dobiveni skup studijskih programa može se smanjiti ograničavanjem na studijske programe koji se izvode u županiji u kojoj učenik ima prebivalište ili prema nekomu drugom geografskom kriteriju. Alternativno se izračun može obaviti za studijske programe koje bi odredili postupkom opisanim u IV. poglavlju ali sada sa stvarnim rezultatima koje je učenik ostvario na ispitima državne mature.

VI. SKLADIŠTE PODATAKA

Za potrebe analize nužno je izraditi skladište podataka iz kojega se mogu iščitati potrebni podatci u prethodnim školskim godinama. Za izradu modela skladišta podataka korišten je Kimballov model [7]. Skladište podataka sastoji se od triju zvjezdastih shema, a sheme su povezane uporabom zajedničkih dimenzijskih tablica. Dimenzije i tablice činjenica opisat ćemo u nastavku ovoga poglavlja.

A. *Dimenzije*

Kreirane su ove dimenzije:

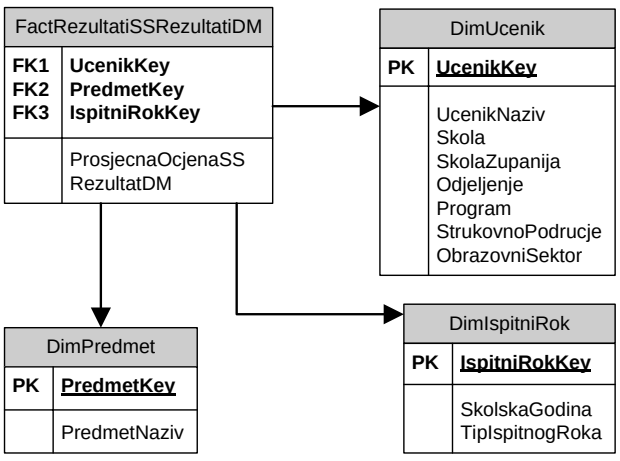
- *Učenik* – podatci o odjeljenju, školi, nastavnom programu, strukovnom području i obrazovnom sektoru
- *Predmet* – predmeti koji se polažu u sklopu državne mature
- *Studijski program* – podatci o nositelju i izvođaču, znanstvenom polju i području studijskoga programa
- *Razina ispita* – može biti viša ili osnovna
- *Ispitni rok* – školska godina i oznaka ispitnoga roka

U predloženim dimenzijama, dimenzija *Ispitni rok* može se smatrati vremenskom dimenzijom pri čemu je koristan pregled podataka prema školskim godinama ili ispitnim rokovima u jednoj školskoj godini.

B. *Tablice činjenica*

Slika 2. prikazuje tablicu činjenica koja se koristi kao izvor podataka za skup primjera za učenje, provjeru i testiranje postupka regresije opisanim u III. poglavlju. Uz tablicu činjenica prikazane su i pripadajuće joj dimenzijske tablice. Tablica činjenica sadrži ove atribute:

- *Učenik* – učenik koji polaže ispit državne mature
- *Predmet* – predmet koji se polaže ispitom državne mature
- *Ispitni rok* – ispitni rok na kojemu je proveden ispit državne mature
- *Prosječna ocjena* – prosječna ocjena u prva tri razreda srednje škole
- *Rezultat DM* – ostvareni rezultat na ispitu državne mature



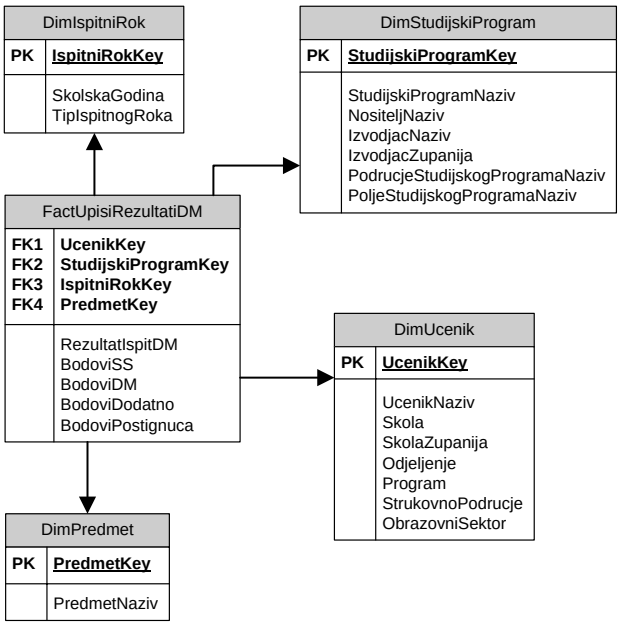
Slika 2. Tablica činjenica s pripadajućim dimenzijskim tablicama koja sadrži prosječnu ocjenu u prva tri razreda srednje škole i rezultat na ispitu državne mature za predmete koje su učenici polagali u sklopu državne mature

Podatci u tablici činjenica prikazanoj na slici 2. mogu se upotrijebiti i za određivanje korelacije između prosječnih ocjena predmeta u prva tri razreda srednje škole i rezultata ostvarenih na ispitima državne mature.

Pri izračunu rezultata ostvarenoga na ispitu državne mature valja naglasiti da se ispiti u obvezatnom dijelu državne mature polažu na dvjema razinama, osnovnoj i višoj [8]. Kako bi se rezultati ispita na osnovnoj i višoj razini mogli međusobno uspoređivati rezultat ostvaren na višoj razini množi se s faktorom 1,6 [8].

Slika 3. prikazuje tablicu činjenica koja se koristi u predloženom postupku za odabir studijskih programa za predlaganje, opisan u IV. poglavlju. Uz tablicu činjenica prikazane su i pripadajuće joj dimenzijske tablice. Tablica činjenica sadrži podatke o studijskim programima na kojima su učenici ostvarili pravo upisa i rezultate koje su učenici ostvarili na ispitima državne mature, a sadrži ove atribute:

- *Učenik* – učenik koji polaže ispit državne mature
- *Studijski program* – studijski program na kojemu je učenik ostvario pravo upisa
- *Ispitni rok* – ispitni rok na kojemu je proveden ispit državne mature i ostvareno pravo upisa na studijski program
- *Predmet* – predmet koji se polaže ispitom državne mature
- *Rezultat ispita DM* – rezultat ostvaren na ispitu državne mature
- *Bodovi DM* – broj bodova ostvaren na svim ispitima državne mature koji su vrednovani za upis na studijski program



Slika 3. Tablica činjenica s pripadajućim dimenzijskim tablicama koja sadrži podatke o studijskim programima na kojima su učenici ostvarili pravo upisa i rezultate na isptima državna mature koje su ti učenici polagali

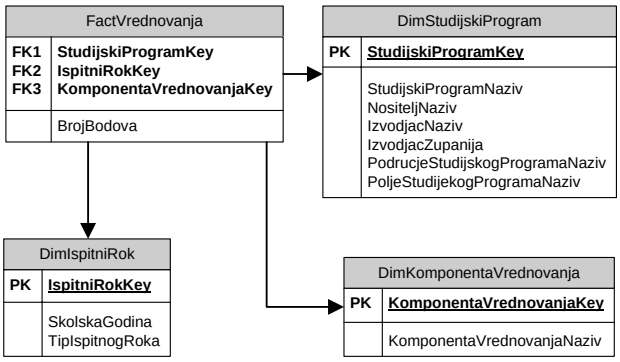
- *Bodovi SS* – broj bodova ostvaren temeljem ocjena iz srednje škole
- *Bodovi dodatno* – broj bodova ostvaren na provjerama posebnih znanja, vještina i sposobnosti
- *Bodovi postignuća* – bodovi ostvareni na temelju učenikovih dodatnih postignuća

Ovdje valja napomenuti da se prijave dvopredmetnih studija u NISpVU vode kao prijave dvaju studijskih programa pa je moguće da jedan učenik ostvari pravo upisa na dvama studijskim programima.

Iako se u predloženom postupku za odabir studijskih programa za predlaganje koristi samo rezultat ostvaren na pojedinačnim ispitima državne mature, tablica činjenica sadrži i sumarne bodove za studijski program na kojemu je učenik ostvario pravo upisa. Ti se bodovi mogu iskoristiti kako bi se dodatno ograničio skup studijskih programa koji će se predložiti učeniku. Primjerice možemo isključiti studijske programe na kojima su učenici posebnim postignućima i dodatnim provjerama ostvarili znatan broj bodova jer u tom slučaju bodovi ostvareni na ispitima državne mature nemaju jednaku težinu kao kod učenika koji te bodove nisu ostvarili.

Model podataka za opis vrednovanja studijskih programa je kompleksan i izvan okvira ovoga rada te ga ovdje nećemo opisivati. Bitno je naglasiti da se za svaku komponentu vrednovanja može odrediti maksimalan broj bodova koje učenici mogu ostvariti te da ukupan zbroj bodova iznosi 100. Slika 4. prikazuje tablicu činjenica korištenu za određivanje prosječnoga udjela bodova prema komponentama vrednovanja za sva područja studijskih programa. Uz tablicu činjenica prikazane su i pripadajuće joj dimenzijske tablice. Tablica činjenica sadrži podatke o maksimalnom broju bodova prema komponentama vrednovanja za sve studijske programe, a sadrži ove atribute:

- *Studijski program* – studijski program na kojemu je učenik ostvario pravo upisa
- *Ispitni rok* – ispitni rok za koji su definirana vrednovanja



Slika 4. Tablica činjenica s pripadajućim dimenzijskim tablicama koja sadrži maksimalan broj bodova prema komponentama vrednovanja za sve studijske programe.

- *Komponenta vrednovanja* – oznaka komponente vrednovanja
- *Broj bodova* – maksimalan broj bodova koji učenici mogu ostvariti za komponentu vrednovanja

VII. ZAKLJUČAK I PLANOVI ZA DALJNJE ISTRAŽIVANJE

Kako bismo pomogli neodlučnim učenicima koji ne znaju koje studijske programe odabrati, u III. i IV. poglavlju predložen je jednostavan postupak za odabir studijskih programa koji im se mogu predložiti. Postupak se zasniva na procjeni rezultata ispita državne mature na temelju ocjena iz prva tri razreda srednje škole. Uspješnost predloženoga postupka može se provjeriti testiranjem na podatcima u aktivnoj školskoj godini s time da se procjena temelji na podatcima iz prethodne školske godine.

Predloženi postupak uzima u obzir samo jesu li ili nisu učenici upisali studijski program no ne i jesu li ga završili i s kojim uspjehom. Kada bismo imali podatke o uspjehu na upisanom studijskom programu, imali bismo i bolju informaciju o tome jesmo li predloženim postupkom odabrali studijske programe na kojima učenik može ostvariti najbolji rezultat, tj. ostvariti svoj potencijal. Analogno sustavima za predlaganje proizvoda za kupnju kao u [9], [10], gdje je potrebno uzeti u obzir i zadovoljstvo kupaca nekim proizvodom, ovdje se na sličan način može koristiti podatak o uspjehu na upisanom studijskom programu. Smatramo da bi se korištenjem tih podataka moglo pridonijeti profesionalnoj orijentaciji učenika koji upisuju studij. Analogno, postupak bi se mogao upotrijebiti i za profesionalnu orijentaciju učenika koji upisuju srednju školu. Za takvu analizu uspjeh učenika trebalo bi se pratiti tijekom cijeloga obrazovanja, od osnovne škole do visokoga obrazovanja.

LITERATURA

[1] Mrežne stranice Ministarstva znanosti, obrazovanja i sporta [Web]. http://public.mzos.hr/Default.aspx?sec=2239

[2] Pravilnik o polaganju državne mature [Web]. http://public.mzos.hr/fgs.axd?id=19419

[3] Statističko izvješće o rezultatima državne mature u školskoj godini 2009./2010. [Web]. http://dokumenti.ncvvo.hr/Dokumenti_centra/DM2010/Statistika/izvjestaj_2010.pdf

[4] E. Alpaydm, Introduction to Machine Learning. The MIT Press, 2009, pp. 34-37

[5] C.M. Bishop, Pattern Recognition and Machine Learning. Springer, 2006, pp. 138-177

[6] S. Russell, P. Norvig, Artificial Intelligence: A Modern Approach (3rd Edition), Prentice Hall, 2009, pp. 717-727

[7] R. Kimball, M. Ross, The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling. Wiley Computer Publishing, 2002.

[8] Brošura za učenike u redovnom sustavu srednjeg obrazovanja u Republici Hrvatskoj [Web]. http://public.mzos.hr/fgs.axd?id=19569

[9] J. Bobadilla, F. Ortega, A. Hernando, and A. GutiéRrez. 2013. „Recommender systems survey“. Knowledge-Based Sysems, vol. 46, pp. 109-132, July 2013.

[10] Soltani, A.; Akbarzadeh-T, M.-R., "Confabulation based recomender system", Computer and Knowledge Engineering (ICCKE), 2013 3th International eConference on , vol., no., pp.82,86, Oct. 31 2013-Nov. 1 2013

Aplikacija za uvoz podataka iz vanjskih izvora u skladište podataka

Krešo Matejčić*, Zrinko Mršo**
*IN2 d.o.o., Zagreb, Croatia, kreso.matejcic@in2.hr
**IN2 d.o.o., Zagreb, Croatia, zrinko.mrso@in2.hr

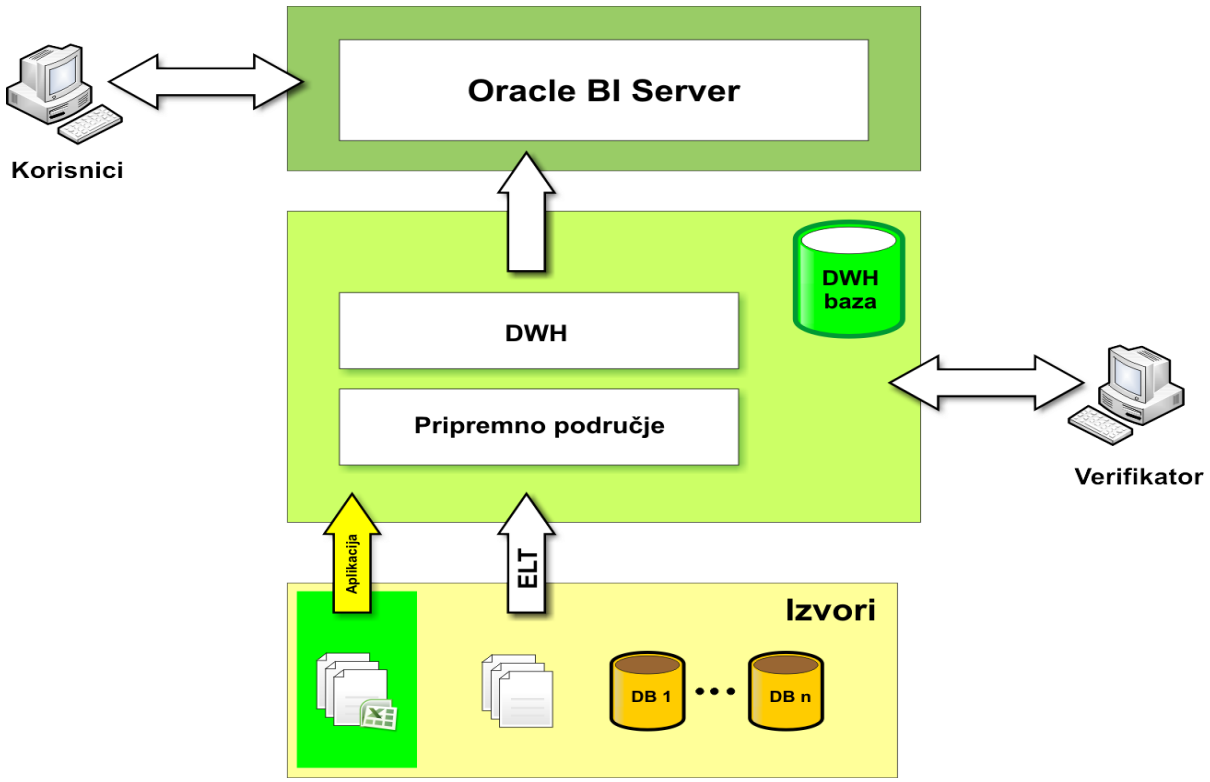
Sažetak - Aplikacija za uvoz podataka iz vanjskih izvora u skladište podataka (DWH) omogućuje korisniku da samostalno unosi vanjske podatke u sustav za potporu poslovnom odlučivanju (BI sustav). Korisnici BI sustava često imaju potrebu za podacima koji nisu obuhvaćeni redovitom ekstrakcijom podataka iz standardnih izvora (transakcijske baze podataka, ERP sustavi i sl.). Ti podaci nisu obuhvaćeni ekstrakcijom iz raznih razloga: dolaze sporadično, dolaze svaki put u drugačijem obliku, potječu iz izvora koji nisu povezani s centralnim informacijskim sustavom poduzeća. Pomoću aplikacije za uvoz podataka korisnik ima mogućnost da sam preslikava podatke iz excela (obične ili pivot tablice) na vrijednosti mjera ili dimenzija u skladištu podataka. Aplikacija se brine za konzistentnost skladišta podataka, tj. ne dozvoljava unos na nepoznate dimenzijske vrijednosti, te traži od korisnika da dimenzijske vrijednosti koje nisu u sustavu preslikava na neku dimenzijsku vrijednost. Podržana je verifikacija podataka, u smislu da uneseni podaci neće biti vidljivi korisnicima (iako su unijeti u sustav) dok verifikator ne označi podatke vidljivima. Aplikacija je bazirana na Oracle tehnologiji i integrirana je sa OBIEE izvještajnim alatom.

I. UVOD

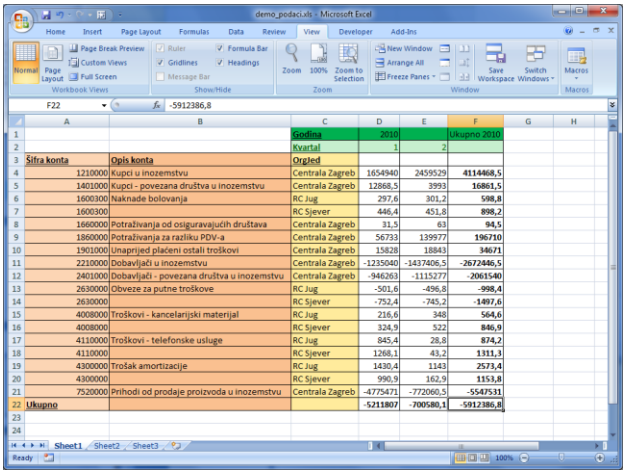
Glavni izvori za skladište podataka (DWH) su obično razne baze podataka koje postoje u poduzećima. No, vrlo često postoji potreba za uvozom podataka iz različitih vanjskih izvora. Tipičan vanjski izvor su Microsoft Excel datoteke, a može se raditi i o drugim izvorima poput raznih internih aplikacija na starijim platformama, koje nisu integrirane u novije sustave, ali se i dalje koriste za određene svrhe.

Uobičajeno je da podatke iz vanjskih izvora uvozi IT poduzeća ili implementator skladišta podataka. Samim tim je više ljudi uključeno u proces, a poslovni korisnici su ovisni o IT-ju. Posljedica te ovisnosti je vrlo često i zakašnjela dostupnost podataka iz vanjskih izvora u sustavu za podršku poslovnom odlučivanju.

IN2 aplikacija za uvoz podataka iz vanjskih izvora omogućava poslovnim korisnicima da samostalno uvoze podatke, ocjenjuju njihovu ispravnost, odnosno kontroliraju cijeli proces. Ne postoji ovisnost o IT sektoru ili implementatoru.



Slika 1. Arhitektura sustava



Slika 2. Podaci za unos

II. ARHITEKTURA SUSTAVA

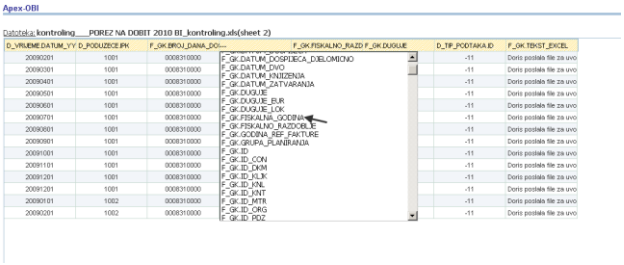
Arhitektura sustava prikazana je na slici 1. Radi se o sustavu za podršku poslovnom odlučivanju (BI) u potpunosti baziranom na Oracle tehnologiji. Front-end dio je Oracle Business Intelligence Enterprise Edition (OBIEE), koji služi kao BI server [3]. Izvor podataka za OBIEE je skladište podataka implementirano u Oracle 11gR2 bazi.

Skladište podataka puni se ELT (extract-load-transform) procesom [2], što znači da se svi podaci prvo dovedu u pripremno područje u bazi te se potom transformiraju i pune u skladišno područje (DWH), u zvjezdastu strukturu mjera i dimenzija [1].

Sama aplikacija za uvoz podataka implementirana je Oracle APEX tehnologijom [4]. Pomoću aplikacije podaci se pune u DWH, tj. aplikacija unutar sebe ima implementiran jednostavni ELT proces.

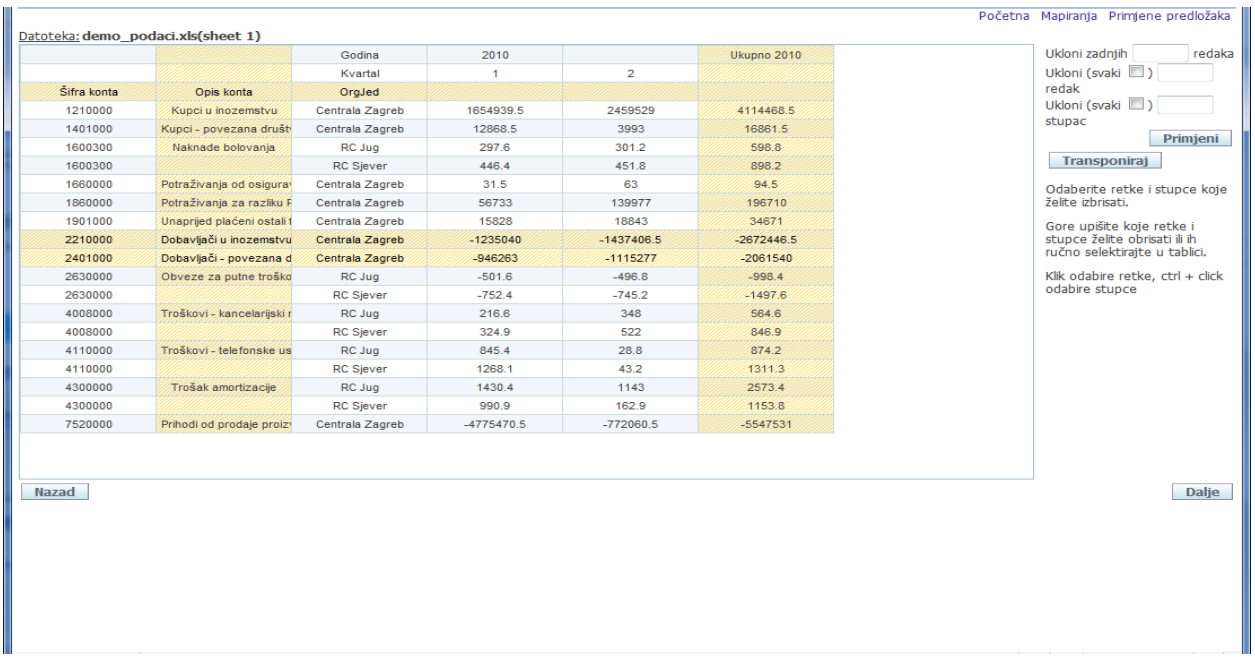
III. OPIS MOGUĆNOSTI APLIKACIJE

Svrha aplikacije je da korisniku omogući samostalan unos vanjskih podataka direktno u skladište podataka. Podaci se uvoze u tablicu mjera skladišta podataka. To znači da se barem jedan stupac iz vanjskog izvora treba moći preslikati na neku mjeru iz skladišta podataka, a ostali stupci na dimenzije koje opisuju tu mjeru. Naravno, vanjski izvor može sadržavati i više mjera, raspoređenih u više stupaca.

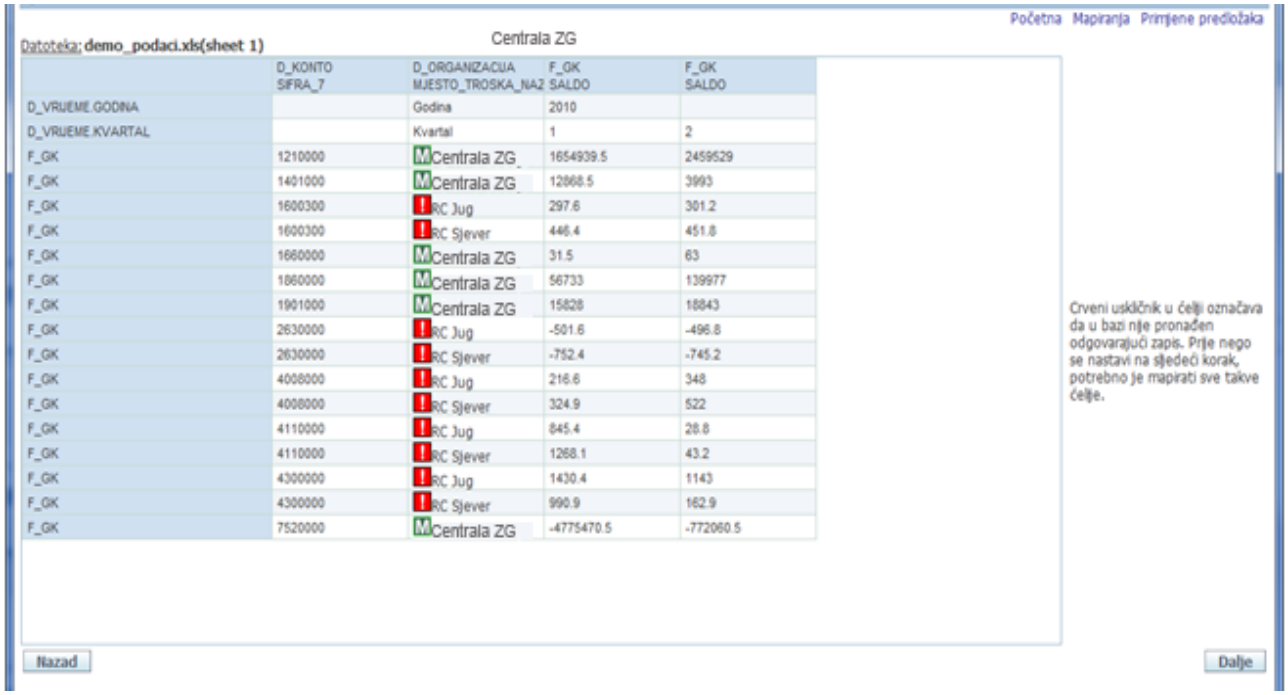


Slika 3. Preslikavanje dimenzija i mjera

Tipična je situacija da se dodatni korisnički podaci nalaze u excelu. Korisnik učitava excel kroz aplikaciju te označava stupce koji predstavljaju mjere i koje vrijednosti mjera uvozi (ne mora uvesti kompletan vanjski izvor, Excel). Zatim određuje koji stupci predstavljaju pojedine dimenzije i preslikava vrijednosti iz dimenzijskih stupaca excela na vrijednosti iz dimenzija skladišta podataka.

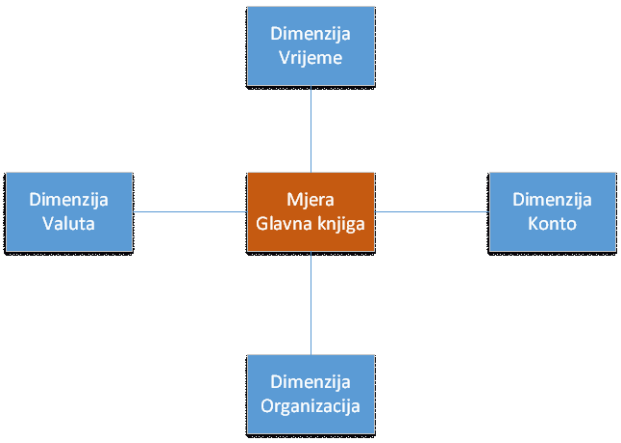


Slika 4. Odabir podataka za unos



Slika 5. Preslikavanje vrijednosti na dimenzijske vrijednosti

Za potrebe ponovnog učitavanja excel datoteke iste strukture, aplikacija podržava spremanje definiranih preslikavanja putem predložaka. Predlošci omogućavaju znatno brži uvoz svaki sljedeći put.



Slika 6. Model podataka

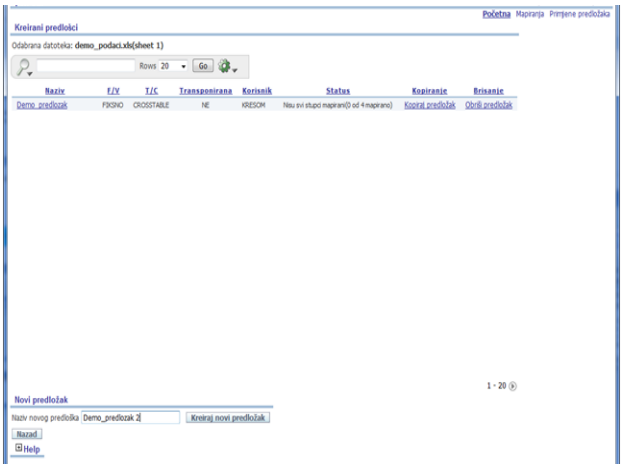
Ono što je dodatna vrijednost ove aplikacije je da omogućava uvoz excela *bilo kakve* strukture. To znači da stupci mogu svaki put biti dostavljeni u drugačijem rasporedu, pa čak i s drugim mjerama i dimenzijama. Korisnik putem aplikacije gore opisanim postupkom ponovno preslikava stupce excela na model skladišta podataka i cijeli proces se ponavlja.
Dakle, aplikacija ne uvjetuje striktnu strukturu dostavljenih podataka iz vanjskog izvora. Upotreba predložaka omogućava znatno ubrzavanje uvoza u slučaju repetitivnih dostava u istoj strukturi.

IV. PROCES UNOSA PODATAKA

Podaci koje korisnik učitava nalaze se u excel datoteci (Slika 2.), u obliku pivot ili obične tablice.
Izvorišni podaci mogu biti dostavljeni kao jednostavna tablica s vrijednostima po stupcima ili kao zaokretna (pivot) tablica. Prvi korak procesa uvoza podataka je učitavanje excel datoteke (ili drugog vanjskog izvora) u aplikaciju. Aplikacija datoteku sprema u bazu, kako bi original datoteke bio spremljen za naknadnu analizu, kontrolu ili za izvanredne okolnosti. U aplikaciji se odabiru stupci i redovi excel tablice koji će se uvesti u skladište podataka. Ako je potrebno isključiti neke stupce ili redove iz odabranog područja, treba ih samo označiti – ne treba ih brisati iz dostavljene datoteke. Aplikacija će prepoznati da ti redovi i/ili stupci nisu za prijenos u skladište podataka.
Slika 4. prikazuje proces odabira podataka za prijenos.

Nakon toga potrebno je preslikati stupce iz excela na stupce dimenzija i/ili mjera (Slika 3.).
Korisnik prvo odabire tablicu mjera na koju će se preslikavati vanjski izvor ili datoteka. Sljedeći korak je preslikavanje stupaca mjera iz datoteke na stupce (mjere) iz DWH tablice. Budući da aplikacija radi unos u zvjezdastu strukturu skladišta podataka, odabirom tablice mjere korisniku se na odabir za preslikavanje nude samo one dimenzije koje su povezane s tom tablicom mjera (Slika 6.).
Na osnovu preslikanih stupaca excel datoteke na određene dimenzijske stupce, aplikacija vrši provjeru da

li vrijednosti sadržane u tim stupcima iz excela, postoje u skladištu podataka. Ako ne postoje, aplikacija neće dozvoliti unos takvih podataka, jer bi se oni vezali na nepoznatu dimenzijsku vrijednost, već od korisnika traži da vrijednost dimenzijskog stupca preslika na neku vrijednost tog stupca iz skladišta podataka. Ako željena vrijednost ne postoji u bazi, aplikacija nudi korisniku mogućnost da unese novu vrijednost u dimenziju skladišta podataka. U praksi se pokazalo da se unos dimenzijskih vrijednosti na takav način iznimno rijetko koristi.
Slika 5. prikazuje preslikavanje vrijednosti.
Korak preslikavanja vrijednosti ključan je korak u cijeloj aplikaciji. Budući da je upravo to najkompleksniji korak prilikom unosa podataka u skladište podataka mimo ELT (ili ETL) procesa.



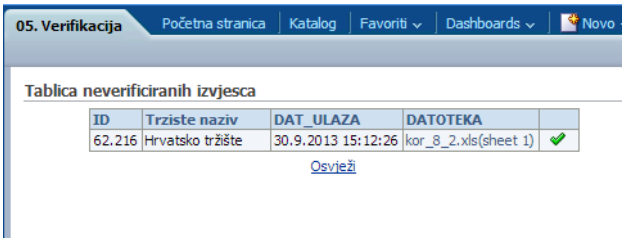
Slika 7. Rad s predlošcima

Preslikane vrijednosti pamte se u repozitoriju aplikacije tako da je dovoljno da korisnik preslikava vrijednost samo prvi kad se ona pojavi. Prilikom naknadnih uvoza, ta vrijednost će biti prepoznata kao preslikana.
Sva preslikavanja stupaca excela na stupce dimenzija i mjera bilježe se u aplikaciji u obliku predloška. Na taj način, idući puta kada korisnik želi učitati excel datoteku iste strukture može jednostavno preskočiti sva preslikavanja stupaca i aplicirati već napravljeni predložak na datoteku (Slika 7.).
Ove dvije funkcionalnosti, spremanje preslikanih vrijednosti u repozitorij i automatsko preslikavanje pri naknadnim uvozima te spremanje predložaka, značajno olakšavaju upotrebu aplikacije i ubrzavaju proces naknadnih uvoza. Jednom kad su sve vrijednosti preslikane, korisnici mogu u vrlo kratkom vremenu osvježiti skladište podataka, a sve to bez potrebe za IT stručnjacima i poznavanjem strukture baze skladišta podataka.

V. KONTROLA I POTVRDA UVOZA

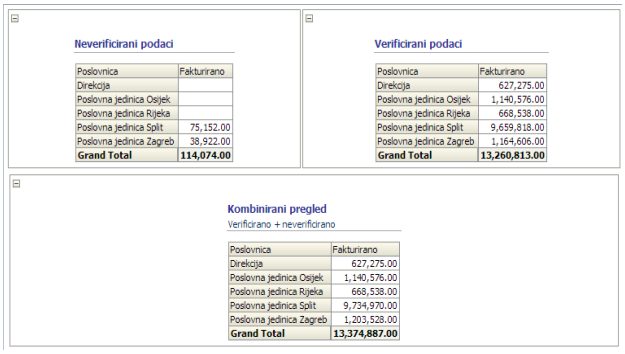
Sve učitane datoteke spremaju se u repozitorij aplikacije koji se nalazi u Oracle bazi. Na taj način uvijek je moguća kontrola i uvid u originalnu datoteku.

U skladištu podataka, na svakoj mjeri, postoji servisna dimenzija koja sadrži podatak od kuda je podatak stigao. Na taj način može se identificirati gdje su završili podaci iz pojedine excel datoteke ili drugog vanjskog izvora



Slika 8. Verifikacija podataka

Podaci učitani iz excela, neće biti vidljivi korisnicima izvještaja dok se ne verificiraju, ali će biti vidljivi verifikatoru. Na taj način verifikator ima mogućnost pregleda kompletnog sustava sa učitanim podacima, a da ih ostali korisnici ne vide (Slika 8.). Kada ih verifikator označi kao ispravne, tj. verificirane, oni će postati vidljivi svima. Aplikacija je u potpunosti integrirana sa OBI (Oracle Business Intelligence) alatom i proces verifikacije odvija se kroz OBI (Slika 9.).



Slika 9. Pregled za verifikaciju

VI. ZAKLJUČAK

Uvođenjem aplikacije za uvoz vanjskih podataka u skladište podataka korisnicima je otvorena mogućnost da sami imaju kontrolu nad uvozom takvih podataka. Povećana je učinkovitost jer više nije potrebna osoba iz informatičke podrške za taj posao. Smanjena je mogućnost greške, jer korisnik sam uvozi podatke i u načelu ih bolje razumije te prije može uočiti grešku pri unosu, ako se dogodi. Upotrebom predložaka omogućen je jednostavan i brz periodički unos datoteka iste ili slične strukture. Verifikacija podataka osigurava ispravnost izvještaja.
Aplikacija je dobro prihvaćena od strane korisnika BI sustava i koristi se u praksi.

LITERATURA

[1] Ralph Kimball, Margy Ross, „The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling", Wiley, 2013

[2] Ralph Kimball, Joe Caserta, „The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data", Wiley, 2004

[3] Oracle Business Intelligence Enterprise Edition 11*g* documentation is part of Oracle® Fusion Middleware Online Documentation Library 11*g* Release 1 (11.1.1.7.0), Oracle, 2014

[4] Oracle Application Express 4.2 Documentation, Oracle, 2014

DE/GLGPS

International Conference on

DIGITAL ECONOMY AND GOVERNMENT, LOCAL GOVERNMENT, PUBLIC SERVICES

Steering Committee

Chairs: Mladen Mauher, Polytechnic of Zagreb, Croatia
Edvard Tijan, University of Rijeka, Croatia

Members: Andras Gabor, Corvinus University of Budapest, Hungary
Jože Gričar, University of Maribor, Kranj, Slovenia
Walter Ukovich, University of Trieste, Italy
Christian Kittl, Evolaris Research Lab, Graz, Austria
Anton Lavrin, Technical University of Košice, Slovakia
Bernhard Katzy, Universität der Bundeswehr München, Germany

PAPERS

Information SecurityAnd The Right To Privacy in Digital Economy– The Case Of Republic Of Croatia

Marija Boban, PhD*,
*Assistant Professor at University of Split Faculty of Law, Split, Croatia
E-mail: marija.boban@pravst.hr

Abstract - The complexity of the changes that have affected digital economy, as well as modern society in whole, is primarily reflected in the frequent use of terms such as information, data (and information) security, personal information, the right to privacy and the information society. As a result, numerous articles, studies and research are dealing with this topic. The starting point of the paper presents the concept of information, personal data, privacy and information security. Defining privacy or infringe the rights of persons (individuals) and personality (individuality) sets the basis for the protection of personal data and set up an adequate framework for information security in the Republic of Croatia. Also, in addition to the protection of data in this paper a special part presents the debate on the right to correction of information as an important part of data protection and at the same time as the way of protecting the rights of citizens in the information society - primarily in the Digital Economy and Government/Local Government/Public Services.

Key words: information, data security, information security, privacy, correction of information, Croatian law.

I. Introduction

The complexity of the social changes that have affected modern society is primarily reflected in the frequent use of terms such as information, information security, personal information, right to privacy, information society and digital economy. Numerous articles, studies and research dealing with this topic. The fundamental driver of these changes was the Internet. Of course, the primary goal of the Internet as such was ensuring that easier and faster access to information in order to be available to everyone in the network without any restrictions.[1] Although it seemed to suit the needs of the time, today it appears as one of the greatest weaknesses of the Internet , especially when it comes to the subject of data privacy and right to acces information in the public administration which especially in the area of public and local governments that include minimum investments in information security (if any!).

At this point it is important to keep in mind that first developed the technology transfer of various types of messages as well as digital form of communication , as used on the Internet now, but not the equivalents protective mechanisms which in practice lead to unwanted and quite frequent cases of abuse , which leads to the question of the status of the Internet as a global communications infrastructure. Overall, the business of the 21st century is known as the digital economy due to the fact that new knowledge and information has become the most important commodity in the market , and the Internet as a medium of exchange has imposed the same as the basic infrastructure of society that enables the exchange of data spatially very distant participants , and also those who are willing to abuse everything they available through a global network.[2]

II. Relevance of the concepts of information security and the right to privacy in the digital ECONOMY

The concept of information we encounter in the most varied situations of use in everyday life to those in specialized scientific fields. It is a basic feature of the information age, information science, technology and society itself. From the variety of meanings which owns, in this part of the processed information is one aspect of which are associated with the concept of the message as the information carrier. By definition, *information* is the result of processing, analyzing and organizing data in a way that adds a note to the recipient. In other words, it is primarily the context in which the data was obtained. Information as a concept has a variety of meanings, from the mundane to the technical use.[3]

[1] Also KOGUT, B., (ed.) „The global internet economy“, MIT Press, Massachusetts Insitute of Technology, 2004., p. 1 - 41.

[2] For more information on digital economy see MALECKI, E. J., MORISET, B., „*The digital economy: business organization, production processes, and regional developments“,* Taylor & Francis, 2008., p. 13 – 17.

[3] On the one hand, the analysis of the conceptual nature of the term of information often leads to a kind of relative historicism which reflects the constant change of meaning, and a confusing number of arbitrary definitions that do not provide the foundations of a new understanding of the term. On the other hand, the analysis by setting nature of the term information points to two problems: a large number of equivalent evidence fail to show the available information and certain personality when choosing. This presents a serious difficulty of each researcher's history one clue who wants genealogically and critical to clarify the concept of information with etymological methods. See HORIĆ, A., "Information - history of a term on Capurro understanding of the term information," Croatian Library Journal 50, 1/2 (2007), p. 98. See also

Since the data and information are often used interchangeably, it is important to distinguish these terms. Specifically, the definition of information is that this is data put into the semantic context , while the data outside of the context. According to the Law on Data Secrecy, Official Gazette . 79/2007 (hereinafter LDS) " data presents any document , or any written , amplified , drawn image , printed, filmed , photographed , magnetic , optical, electronic , or any other record of information , knowledge , measure , process , object, oral communication or information, which, given its content, has the importance of confidentiality and integrity of their owners. " (LDS Art. 1 Paragraph 1). Or in other words , the data is useless until it conveys some information . According to the definition of information is a set of characters that mean something to the recipient , or discovering something new . Information is a term with many meanings depending on the context , but as a rule closely related to such concepts as meaning, knowledge , perception, instruction , communication , and various mental processes . Simply put , the information is received and understood as the message . But above all , it is the result of processing , manipulating and organizing data in ways that they build on the knowledge the person who receives this information.

On the other hand, the message is materialized form of information. The main feature of the message is that it is informative, and that the person to whom it is intended (the recipient) provides a (new) information. Information is always about something, and in this sense, it may be true or false. To be as useful or appropriate it is important that all participants in the communication channel have a certain knowledge about the subject matter. Although theories and models that consider the recipient necessarily imply the existence of the sender and therefore do not attach importance to the idea that information can come from the environment, it should be noted that the information and the message we must look at is always strictly linguistic context, because even us and the noise that interferes with the flow communication gives specific information.

Exhaustive legal definition of personal data on the Personal Data Protection Act (Official Gazette no. 103/0, 118/06 - hereinafter PDPA), according to which personal information is any information relating to an identified natural person or a natural person who is can identify that a person whose identity can be determined directly or indirectly, in particular on the basis of one or more factors specific to his physical, physiological, mental, economic, cultural or social identity (PDPA Article 2.) However, in an attempt to define a thorough personal data the question whether the concept of distinguishes personalities from the concept of privacy was left open? If we try to primarily determine the meaning of personality in Croatian language should provide some preliminary remarks. If we were to literally translate the words used in the western countries, then we are talking about the same word or translation would be the same as the English word privacy[4], similar to the personality of the German -Croatian-English dictionary, the word privacy is translated as secrecy of trust, privacy, solitude, peace, and so on until the word Personklichkeit, or in English, personality translates as personality, identity, personality, etc.[5] Thus, the concept of personality[6] covers an area that explicitly represents a person and her personal life, and the concept of privacy encompasses a wider range of features people taken individually, do not have to have some meaning, but linked to a stable structure resulting portrait of the person, which usually only the person needs to know . Therefore, the origin of the narrower the definition of personal data as well as information specifically related to an identified natural person or a natural person or a person whose identity can be determined directly or indirectly.

Also, privacy is one of the supporting values of Western legal culture. It is based, on the one hand, the belief that every human being has value in itself, and on the other to the primal human need for the existence of a particular protected area from which everyone else was off psychologically and materially. [7] Using the phrase "right

CAPURRO, R., "Information: ein Beitrag zur etymologischen und ideengeschichtlichen Begründung des Informationsbegriffs" Saur,. Munich, in 1978., p. 116.

[4] According to the dictionary of foreign words the word " private " (derived from the word " PRIVATUS " - personal , Your) means something personal and has no official character (unofficial). Further defined as confidential, secret , private, confidential , pets , domestic , exclusive, limited, nejavan , closed. See KLAIĆ , B., Dictionary of foreign words , foreign words and loanwords , Croatian Heritage Publishing House, Zagreb, 1990. , p. 1090. The term of privacy itself has been derived from the Latin word, which originally just means those places that are not allowed access to the public. Indicated the need for the inviolability of the private sphere is sublimated in the rich tradition of human thought, especially in the writings of luminaries like Voltariea , Montesquieu , Diderot , Rousseau, Bentham , Gibbon , Hume , Adam Smith , Kant, Locke, B. Franklin , Th. Jefferson and many others. But before it was recognized as an individual right provided with adequate protection, privacy has long figured only as a moral- philosophical ideal. On this see RADIĆ , Ž. , " The genesis of the right to privacy " , Proceedings of the symposium 'Right to privacy , Novalja - Split, 2008., p. 46 See also Guardster , The Right to Privacy , "p. First available at http://www.guardster.com/?Tutorials-The_Right_to_Privacy (25th 06th 2011).

[5] The United States Supreme Court has stated that American citizens have the protection of the Fourth Amendment (freedom from search and seizure absent warrant) when there is a reasonable expectation of privacy. Without a reasonable expectation of privacy, however, there is no privacy right to protect. "Various guarantees create zones of privacy. The right of association contained in the penumbra of the First Amendment. The Third Amendment in its prohibition against the quartering of soldiers. The Fourth Amendment explicitly affirms 'the right of the people to be secure in their persons, houses, papers, effects, against unreasonable searches and seizures'. The Fifth Amendment in its Self Incrimination Clause. The Ninth Amendment provides: 'The Enumeration in the Constitution, of certain rights, shall not be construed to deny or disparage others retained by the people." Griswold v. Connecticut, 381 U.S. 479 (1965) See GUARDSTER , The Right to Privacy , "p. First available at http://www.guardster.com/?Tutorials-The_Right_to_Privacy (10. 09. 2013.).

[6] Instead of the notion of privacy as synonymous in the Croatian language uses the term "personality" since according to its content, which is determined by factors outside the person, in connection with it, is a component of "persona". Cf. Brezak, M. "The right to privacy", Croatian Heritage Publishing House,Zagreb, 1998, p. 12 by Tomislav Ladan, editor of Croatian Lexicographic Institute in Zagreb (conversation led with him on January 11, 1995)

[7] Ibid., BREZAK, M,, p.12-13.

to privacy" when determining primarily legal and safety and protection of the methodology, as well as a set of rules that should be included in the protection of all the features of the situation, the conditions and circumstances in which the individual realizes his need for inner peace, what I want and which, taking into account all social conditions, have the right, it should be clear when it comes to normative regulation.[8]

In today's busy world threats to privacy rights have the effect of violation of information personality that in the strictest sense of the word is defined as a claim of individuals, groups or institutions to independently decide when, how and what information about yourself to cede to others.[9] In a broad, generally accepted meaning of the term information personality actually defines the term information security under which implies that the individual, in terms of the information society, decide when, where, how and how much will disclose personal information, taking into account their rights and needs, but and the rights and needs of the community where the person lives.[10] Circumstances and conditions of personal rights are defined depending on different variables: depending on the culture, structure or sociological tradition of a society.[11]

In the complex environment of the digital economy modern security management information systems acknowledge that most risks cannot be completely eliminated and that they need to be managed in a cost effective manner. In information strategies, especially created for the purpose of information security is important to concentrate on the development of a methodology for the assessment and analysis of threat and vulnerabilities within the context of a security risk management.[12] It refers mostly to the threat and vulnerability assessment method developed with the needs of mobile computer systems in mind. Most of the information strategies consist of four stages: a) Assessment Scope, b) Scenario Construction & Modeling, c) Threat Agent & Vulnerability Analysis, and d) Stakeholder Evaluation. These kinds of strategies actively involve stakeholders and focuses upon a technical, socio-technical and business aspect of the system.[13]

III. The Legal Legislative In Data Privacy, inofrmation security and the right to privacy in the Republic Of Croatia

Apart from regulatory framework, the legal framework of eGovernment model in Croatia if regulated by the following legislative:

- The Constitution of Republic of Croatia; Official Gazette of Republic of Croatia, NN 56/90, 135/97, 8/98, 113/00, 124/00, 28/01, 41/01, 55/01, 76/10, 85/10
- Law on Personal Identification Number,
- Official Gazette of Republic of Croatia, No. 60/08
- Law on Protection of Personal Data, Official Gazette of Republic of Croatia 103/03
- Law on Right to Access Information, Official Gazette no. 172/03
- Law on Information Security, Official Gazette of Republic of Croatia 79/07

In this area, as one of the most questions that the information society faces in today's digital econoy is confidentiality of information, the second is availability and the third is integrity of the information. Together these three security principles form the basics of information security model which is namely intented for protection of access to information.[14], the conflicts of

[8] Also, it is important to compare the perception of personality through property relations that even possible violation of personal data is seen as a violation of people (causing mental anguish, etc.) and is not a property in the classical sense, nor of material values. See KRNETA, S., Civil law protection of the private life in European law, Yearbook of Law Faculty in Sarajevo,1970., p. 81-97. See also, Dika, M., "Protection of personal rights in litigation and enforcement", in Klaric, P. (ed.) "Responsibility for non-pecuniary damages for the infringement of personal rights, , Zagreb, 2006.,, p. 267. Also see RADOLOVIĆ, A., "The right of personality in the new Law on Obligations," Journal of Faculty of Law in Rijeka, Vol 27 no.1, p.133-134.

[9] According to WESTIN, A., Privacy and Freedom, Atheneum, New York, 1970., p.7

[10] s a basis for the definition of personal data protection is primarily defined by the concept of personal data. In accordance with the legal point of view, personal rights are defined as subjective immaterial rights legal system recognizes a person in respect of its personal, compensates personal resources and opportunities as their protection of private rights. On this definition, it is necessary to interpret the definition of personal rights as one of the key foundations of the protection of persons and personal data. With respect to protecting people and individual rights is important to distinguish the concept of protecting the physical security of the person and the protection of privacy and individual personality. Personality is , however, necessarily related to privacy, which is based on deep personal definition of each individual. Interference on the personality of violating the fundamental rights guaranteed by the legislation and the question is can you ever recover damages in this case učinjena.Cf. Brezak, M. p. 22

[11] For example, privacy in certain cultural communities is given by hiding women's faces on foundation in religion, but this is a good example of how the protection of personal rights is functioning. In the same way, in modern terms, the "chat-room-in" or talking through aliases or nicknames of people who access this website in order to protect the personality not to sign their data on the first and last name in order to prevent misuse of the same. Every person on earth is born with the need for personality, for its own framework in which to develop and prosper. On this see beyond TUCAK, I., "Rule of Law and Human Rights," Legal Life, no. 14/2007., p. 483

[12] See BÖHLEN, M. H., E-government: towards electronic democracy : international conference, TCGOV 2005 Bolzano, Italy, March 2-4, 2005 proceedings, Springer, 2005., p 122

[13] More on this approach at BOBAN, M., Building eGovernment model on the principles of new economy trends and international standards considering protection of citizen privacy and personal data, E-Society Journal Research and Applications, Volume 1, Number 2, Zrenjanin, December 2010., str. 1 – 15

[14] Refering to presentation available on the Internet KLAIĆ, A.: „EU's information security expectations", Conference Infos-week, Zagreb, 14-18. svibanj 2007., p.5. (unpublished): the first domain's property is confidentiality as the main property; the second and third domain's is privacy while in the 4th domain disclosure is not desireable but it is not harmful.

rights of access to information and protecton of personal data.[15] (as shown on Fig. 1.).

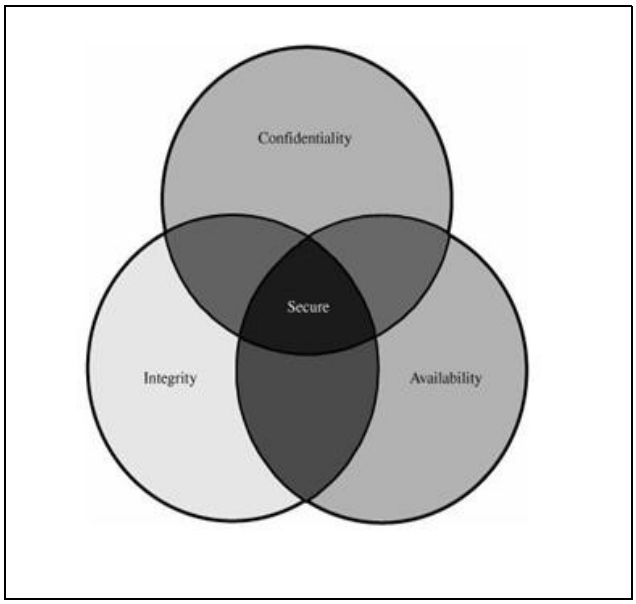


Figure 1. Three main principles of information security in digital economy

The Republic of Croatia, as the entire member states of the European Union, is also signatory of the European Convention on Human Rights and the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data, the European Commission was concerned that diverging data protection legislation would emerge and impede the free flow of data within the EU zone. Therefore the European Commission decided to harmonize data protection regulation and proposed the Directive on the protection of personal data, which member states had to transpose into law by the end of 1998. In full name, Directive 95/46/EC on the protection of individuals with regard to the processing of personal data (Official Journal L 281 , 23/11/1995 P. 0031 – 0050) and on the free movement of such data contains a number of key principles which must be complied with. Anyone processing personal data must comply with the eight enforceable principles of good practice.

In contemporary information society, information is made pursuant to the process of public information, an integral part of the democratic and political process. Information becomes an important element of freedom and the right to disseminate information to a large extent depends on the legitimacy and management capabilities of data collections. It is obvious that the process of disseminating information, its printing, publication or broadcast in the media is subject to the erroneous interpretation. Logically, one of the constitutional and distinctive means of limiting the freedom of public information for the protection of personal rights is the right of correction of information. The Constitution of Republic of Croatia ("Official Gazette" no. 56/90, 135/97, 8/98-consolidated text, 113/2000 124/2000 - consolidated text, 28/2001 and 41/2001-consolidated text, 85 / 2010-revised text - hereinafter Constitution) guarantees the right to a correction which has been determined by the Constitution and the law (Constitution Art. 38). The process of disseminating information, their printing, posting or displaying is regulated by the legal regulations governing public media. In Croatia, the information is accomplished through media that are subject to national legislation and Media Law[16] ("Official Gazette", no. 59 / 04, still-ML).

Legally, it is interesting to view the correction information is one of the logical, constitutional and differential means limiting the freedom of public information to protect the rights of personality. ML entitles legal or natural person that within 30 days of the publication of false or incomplete information from the editor requires disclosure of the correction of the notice of violating her rights and interests.[17] Correction must be published in the first, and if it arrived late, in the second program content upon receipt, without changes and additions, in the same or equivalent in the programming area, of the same or equivalent manner in which the information was published, in radio and television programs published reading in the same program and the period of time or in the same kind of identical ratings.[18] The ML also regulates rules of court proceedings if the editor refused to publish a correction or do not publish on time and in the manner prescribed by law. Considering the essence of the freedom of public information in the legal freedom of reporting, advocates of logical restrictions on the freedom of the press in the correct information, see the means to support the principle of duty of truth, i.e., transmitting only the truth in reporting. For supporters of the constitutional limitations on freedom of media boundaries of public information derived from the protection of personal honor, and the principle of equality of citizens. For the theory differing

[15] See REDDICK, C. G., Politics, democracy, and e-government: participation and service delivery, Idea Group Inc (IGI), 30. 3. 2010., p.373

[16] Media are: newspapers and other print, radio and television programs, news agencies, electronic publications, teletext and other forms of daily or periodical publishing editorial program content transmission media, voice, sound or image. The media are not books, textbooks, newsletters, catalogs or other holders disclosure intended exclusively for educational, scientific and cultural process, advertising, business communication, operation of companies, institutes and institutions, associations, political parties, religious and other organizations, school newsletters , "Official Gazette" of the Republic of Croatia, the official journal of local and territorial (regional) governments and other official announcements, posters, flyers, brochures and banners, and video sites without live images and other free information, unless this Act provides otherwise. (ML, Article 2. § 1)

[17] According to ML everyone has the right to require that the editor at no charge to publish a correction of information that they were violating the rights or interests. Have the right to correct and legal persons and other organizations and bodies, if the information was violated their rights and interests. The purpose of corrections is to correct inaccurate or incomplete informacije.Objava correction may be requested within 30 days of the publication of information. (ML Art. 40, paragraphs 1 and 2)

[18] On this see Zlatović, D., "The right to correction and response to the new Media Law," Law and Taxation, no. 3, March 2007., p. 38 - 42

demarcations of the correction of the information, its intent is to defend certain constitutional and preferential rights established by the Constitution and by the law regulations. The Croatian Constitution guarantees "the right to correction to anyone who has been violated by public news as his/hers constitutional and legal right." (Constitution Art. 38).
Through the mass media it is also exercised the constitutional right to freedom of expression. The exercise of this right requires freedom of the press and other mass media, freedom to establish any public information, the ban on censorship and the right of journalists to freedom of reporting and access to information. The Republic of Croatia guarantees the freedom of expression and freedom of the media , which includes the freedom of expression , collection , research, publication and dissemination of information, publication and dissemination of the press, the flow of information , media openness , pluralism and diversity of media , the availability of public information , production and broadcasting of radio and television broadcasting , freedom of establishment of legal persons for the performance of public information (ML Art. 3 Para 1 and 2). However, freedom of information can lead to the publication of false and derogatory information, and cause a breach of privacy. (ML Art. 7 Para 1) is therefore legally entitled to a correction and reply and the right to compensation for the damage caused.[19] On the other hand, prevent printing and distributing printed items and broadcasting is a criminal offense.

Also, the right of journalists, editors, printers and authors of published reports that refuse to give information on the source of published information or information that is intended to publish, is also protected by ML in a specific way. Specifically, the „journalistic secret" , a term that is conventional for such retention is different from the professional secret in that it is only the right , but not the obligation . So, the reporter may refuse to reveal the source of information, but there is no legal obligation to maintain the confidentiality of the data. (ML Art. 30 Paragraph 1) The court shall request for paying sources to adopt if it is necessary to protect the public interest in a particularly important and serious circumstances, and no reasonable alternative measures of disclosure of such information, and if the public interest in disclosure of the source clearly prevails over the interest of its protection.[20] According to the interpretation of the Code of Criminal Procedure (Official Gazette 62/03- still CPC), the journalist is released duty to testify about sources of information obtained in carrying out their profession , except in criminal proceedings for offenses against honor and reputation committed mass communication (CPC art. 244 § 1).[21]

In Croatia, the right to access the information is explicitly regulated by the Law on Right to Access Information – (further LRAI - Official Gazette no. 172/03). which stipulates that public authorities are public bodies, local and territorial (regional) governments, legal persons with public authorities and other persons to whom are delegated public authorities? (Article 3 paragraph 2 LRAI). The goal of LRAI is to enable and provide access to information to physical and legal persons by public action by public authorities. (Article 2 , LRAI) Also, all public authorities were required to ensure organizational, financial, technical and other requirements to implement the provisions of the Act. According to this, “the right of access to information" is defined "as the right of beneficiary to seek and receive information as well as the obligation of public authorities to provide access to the requested information, or to disclose information even when there is no special requirement for publication is their obligation to a particular law or other general regulations. "(Article 3 Paragraph 5, LRAI).
On the side of public authorities in the LRAI model, there are following assumptions: the non-existence of exceptions to the right of access; consistent application of the principle of access to information; legal action concerning a request and consistent application of measures for the realization of the right of access as well as the organization of collecting, processing and providing information. So for the availability of information it is essential to meet the specific requirements that the applicant's side, and on the side of public authorities.[22] On the information seekers: it is necessary to first know the legislation on right to access information; such content making demands that will result in the requested information; it is necessary to know the matter, within which one can find the required information and also it is an essential and status information seeker and the financial capability.

[19] See KOVAČIĆ, T., "Liability for breach of personal rights of the legal person publishing the information in the media (I)," Law and Taxation, no. 8, August 2008., p. 40-45

[20] According to art. 233 Paragraph 3 of the CPC in the criminal proceedings are not subject to seizure records of the facts that have been done by journalists and their editors in the media about the sources of information and data that they learn in the course of their occupation and who were used while editing the media, which are in their possession or the editor in which they are employed.

[21] According to Article 244 CPC obligation of testimony exempts: 1) the spouse or common-law partner accused, 2) relatives of the accused in a vertical line, relatives of the lateral line to the hird degree and relatives by marriage to the second degree, 3) an adoptee and adoptive defendant, 4) lawyers, notaries, tax consultants, doctors, dentists, pharmacists, midwives and social workers about what they are doing their jobs out of the defendant, and in Section 5) journalists and their editors in the media about the sources of information and data that they learn in the course of their occupation and who were used while editing the media, except in criminal proceedings for offenses against honor and reputation committed through the media.

[22] The law stipulates the obligation of submitting reports by public authorities on the implementation of the previous year to 31 January, and the Office submits a report to the Croatian Government to 28th February for the previous year. The Croatian Government is obliged to submit the report on the implementation until the 31st of March to the Croatian Parliament for adoption. By experience, the Central Administration Office regularly reports on the right to access information and enter it in a timely manner in the regular procedure in the Croatian Parliament. For example see report available at www.vlada.hr/hr/content/download/7052/59557/file/6%20-145.pdf (12. 03. 2014.)

Also, the main stipulation of the LRAI is public interest connected with legal, political, economic and financial interests of the Republic of Croatia – in this case. But sometimes it is not clear what the term "public interest" actually means. Often it is defined as a public benefit that may result from publication of certain information. But it is very difficult to explain what would be the benefit, because of course it depends from case to case. In this question, we should not ignore the fact that the exemption of this right to access the information as a public interest from the obligation to keep information that contain personal data of citizen's – beginning even with only a name – a secret. In this case it shouldn't no matter how the public services are motivated by public (government) interest or the right to access information of importance to the entire legal, political, economic and financial system of a country.

In this cases in the legal practice is used the test of public interest and proportionality which requires the public authority to assess whether, in each case in the public interest to allow access to the information or whether you would benefit from providing access to information was greater than any harm (proportionality between constraints and objectives to be protected). The theory emphasizes that the public interest test should be used very carefully and conscientiously as it requires a higher level of decision making between the two conflicting rights or interests. The test involves balancing the public interest and assessment in order to determine which is better for the public interest: to disclose or not disclose information. The public interest may change over time and depends on many circumstances, these changes and depends on the current situation. Public interest test requires an assessment on a case by case basis, considering the different variable factors.

IV. Conclusion

Since the information, in nowdays so called „information era" has began to rule the world, the question of the data security has becoming increasingly relevant issue in the experts discussions. Also, the question of information, data security and the personal privacy take a fundamental part of the discussion related to the personality and information security connected to the new tecnololgies. In this paper, afterwords the thorough debate, still remains the question whether indeed litigation, and the Law itself, can really undo the damage of the revelation and correction of inaccurate information? Whether the civil lawsuit can put a "price" on someone's "privacy", and whether mechanisms, so clearly laid, in the Constitution and connected laws actually function in practice? Especially in the field of Government, Local Government and Public Services? Although the litigation related to the disclosure and correction of information is laid on expedited basis - remains an open problem of compensation and protection of personal data and personal information security in order to prevent encroachment into the foundations of human rights and dignity. Furthermore, there remains the question of "irreparable" harm violating the right of personality, personal's honor, reputation, dignity, one's private and intimate sphere, the name, character or work which in Croatia in the field of public public administration – still has no legal practice in general.

References

[1] Böhlen, M. H., E-government: towards electronic democracy : international conference, TCGOV 2005 Bolzano, Italy, March 2-4, 2005., Proceedings, Springer, 2005., p 122

[2] Boban, M., Building eGovernment model on the principles of new economy trends and international standards considering protection of citizen privacy and personal data, E-Society Journal Research and Applications, Volume 1, Number 2, Zrenjanin, December 2010., str. 1 – 15

[3] Brezak, M. "The right to privacy", Croatian Heritage Publishing House,Zagreb, 1998.

[4] Capurro, R., "Information: ein Beitrag zur etymologischen und ideengeschichtlichen Begründung des Informationsbegriffs" Saur,. Munich, in 1978., p. 116.

[5] Dika, M., "Protection of personal rights in litigation and enforcement", in Klaric, P. (ed.) "Responsibility for non-pecuniary damages for the infringement of personal rights, , Zagreb, 2006.,, p. 267.

[6] Ehmann, H., "Das Allgemeine Persönlichkeitsrecht-zur Transformation unmoralischer in unerlaubte Handlungen" Canaris / Heldrich / Hopt / Roxin / Schmidt / Widmair (Ed..), 50 Jahre Bundesgerichtshof Band I. Bürgerliches Recht, Munich, in 2000., p. 613

[7] Horić, A., "Information - history of a term on Capurro understanding of the term information," Croatian Library Journal 50, 1/2 (2007), p. 98

[8] Klaić, B., "Dictionary of foreign words , foreign words and loanwords" , Croatian Heritage Publishing House, Zagreb, 1990.

[9] Klaić, A.: „EU's information security expectations", Conference Infos-week, Zagreb, 14-18. svibanj 2007., p.5. (unpublished)

[10] Kogut, B., (ed.) „The global internet economy", MIT Press, Massachusetts Insitute of Technology, 2004., p. 1 - 41

[11] Kovačić, T., "Liability for breach of personal rights of the legal person publishing the information in the media (I)," Law and Taxation, no. 8, August 2008., p. 40-45

[12] Krneta, S., "Civil law protection of the private life in European law", Yearbook of Law Faculty in Sarajevo,1970., p. 81-97

[13] Radić , Ž. , " The genesis of the right to privacy " , Proceedings of the symposium 'Right to privacy , Novalja - Split, 2008., p. 46

[14] Radolović, A., "The right of personality in the new Law on Obligations," Journal of Faculty of Law in Rijeka, Vol 27 no.1, p.133-134

[15] Saračević, T., "An Essay on the Past and Future (?) Of Information Science Education and Historical Overwiev", Information Processing and Management (15) 1-15 (1979), p. 3.

[16] Webster, F.,"Theories of the Information Society", Cambridge: Routledge, 2002.

[17] Westin, A., Privacy and Freedom, Atheneum, New York, 1970.

[18] Tucak, I., "Rule of Law and Human Rights," Legal Life, no. 14/2007., p. 483

[19] Tuđman, M., "Theory of Information Science", Informator, Zagreb, 1990.

[20] . Zlatović, D., "The right to correction and response to the new Media Law," Law and Taxation, no. 3, March 2007., p. 38 - 42

E-business models for web based social networks in the Republic of Macedonia

Dimitar Jovevski*, Edvard Tijan **
* Faculty of Economics, University ss. Cyril and Methodius-Skopje, Macedonia
** University of Rijeka, Faculty of Maritime Studies, Rijeka, Croatia
djovevski@eccf.ukim.edu.mk, etijan@pfri.hr

Abstract - Since their introduction, social media sites (SMSs) such as Facebook, LinkedIn and Twitter have attracted millions of users, many of whom have integrated these sites in their daily practices. As of this writing, there are hundreds of SMSs, with various technological affordances, supporting a wide range of interests and practices.

Social media are gaining popularity and are increasingly used in regular operations of many companies, including start–ups, small, medium-sized, and large organizations. The purpose of this paper is to explore the different e-business models and organizational factors that impact the successful use of social media in companies in the Republic of Macedonia. This research will focus on the extent of impact that social media have on organizational capabilities and business performance.

From the research it could be found that successful deployment of set of external and internal factors can lead to the improvement of business performance (financial and nonfinancial) of the companies. According to the results, the successful use of web-based social networks can improve (1) CRM integration, (2) open communication within a company and (3) ICT integration.

Keywords: social media, social networks, critical success factors, e-business models

I. INTRODUCTION

The world in which we are living and the balance between companies and consumers are constantly changing. Companies need to implement new information and communication technologies (ICT) to meet the needs of consumers. Web-based social networks are included in the lifestyle of many people in the world. This fact enables the companies to have another channel to communicate with their customers through these networks. Social networks like web - based services incorporate various technologies (blog, forum, and social games).

In today's digital environment web-based social networks are important in the everyday operations of the companies. Because of its unique ability to share information between users, the web-based networking slowly but surely transformed into a powerful tool with which companies can communicate with their audience. As a result, companies increasingly seek to create fun and informative pages on social networks and to maintain their presence. Their presence directly affects their business performance.

Emerging technologies gain popularity as enabling tools for cooperation among businesses in business networks [1], [2] whereas the applications market is flourishing [3]. Companies that take advantage of the latest social media technologies seem to outperform their competitors and report benefits like lower costs and improved efficiencies [4]. In this context it is important to understand the specific impact that social media have on business performance[5]. The identification of a direct connection between the two will support the shift towards Enterprise 2.0 – a new business environment in which the companies will maximize the benefits they can obtain by integrating social media suites into their daily operations.

The subject of research in this paper is the analysis of e-business models that include the impact of various factors for successful web-based social network by organizations. This analysis involves a group of independent factors that directly or indirectly affect the success of enterprises. These groups of independent factors are: the use of ICT in organizations, organizational structure and architecture of business processes, strategy, affecting dependent variables such as increased sales and brand awareness through web-based social networking.

Empirical research includes exploratory and conclusive research with three parts. The first part is to create a conceptual model for successful use of web-based social networks, the second includes testing the relationship between various internal and external dimensions of successful use of web-based social networks and the third include testing of the relationship between individual dimensions and business performance.

Exploratory research used quantitative methods such as analysis of secondary data and interview techniques with experts. The primary data used statistical methods: structural equation modeling and confirmatory factor analysis and exploratory factor analysis.

II. E-BUSINESS MODELS FOR SOCIAL NETWORKS AND BUSINESS IMPACT

In this section we define social media as a part of Web 2.0 technologies (A), refer to the resource-based view of the firm and factors for successful use of social media and their impact on organizations (B)

A. Defining social media

The term Web 2.0 was coined in 2001 by O'Reilly [6] in a conference brainstorming session to reflect the transition from the manager generated content era to the user-generated era. O'Reilly identifies seven differences between Web 1.0 and Web 2.0: the web as a platform; the harnessing of collective intelligence; the data as the next Intel Inside; the end of the software release cycle; the lightweight programming models; the software above the level of a single device; the rich users' experience [6]. The term was widely adopted and definitions have been formulated for Web 2.0, all emphasizing collaboration and enhanced communication, as well as user involvement. Harris and Rea [7] define Web 2.0 as "a perceived second generation of Web development and design that facilitates communications and secures information sharing, interoperability, and collaboration on the Web".

Bell and Loane [8] define Web 2.0 as "a set of economic, social, and technology trends that collectively form the basis for the next generation of the Internet – a more mature, distinctive medium characterized by participation, openness and network effects".

Web 2.0 technologies share common characteristics that distinguish them from previous generations of Web development. Firstly, Web 2.0 brings about an emphasis on collaborative learning as well as on user engagement through participation. Secondly, Web 2.0 is regarded as user friendly, as it enables immediate publication and wide distribution of user generated content. The driving force behind the new wave of applications stands in their content and data management systems, as well as in their architecture of participation that encourages user contributions. Further on, the new generation of applications uses the web as a development platform. Most Web 2.0 tools are based on the Software as a Service technology [8].

Web 2.0 and social media have different meanings, but they are directly connected, and it can be said that social media derive from Web 2.0. Kaplan and Haenlein [9] define social media as "a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of user-generated content". In this paper we use Web 2.0 when we address the technology platform, but most of the time we will talk about applications that derive and use technological foundation of Web 2.0 that is social media.

Social media is a phrase that describes the platforms and other tools that connect people into social networks (of their choice) online. Some of the household social media names worldwide (other than Facebook) are MySpace, LinkedIn, Twitter, YouTube, Foursquare Google+ [10].

The individuals that make up these online networks use social media to organize themselves according to (1) their interests and (2) their preference for the way they choose to share, store and deliver information within their community (or "network") [11].

Individuals have made it clear they wish to connect with other individuals that they choose to, in the way that they choose to, and around topics and for reasons that they choose to. The individual is at the centre of the social media revolution. Companies must be present there, for sure it can be said that this is changing the way business is done.

B. E-business models for successful use of web based social networks in companies.

In the paper, the resource-based view will be pursued as a perspective on organizations to explain the impact of using web based social networks on company performance. The resource-based view states that organizations obtain a set of certain resources (like human resources or ICT infrastructure) that could be company specific. The particular combination of resources forms the basis for company's competitiveness and performance. A distinction can be made between resources and capabilities. While resources serve as basic units of analyses, capabilities are repeatable patterns of action in the use of resources to create, produce, or offer value to a market [12]. Note that resources (like Web 2.0 tools) may be obtained easily, but that it is not easy to develop business-wide capabilities to use the resources to enhance business performance. So, while any company can reach resources easily, capabilities embedded in business practice cannot be reached. We focus on Web 2.0 based capabilities, more on capabilities related to web based social networks, when a company is using them.

In the literature different models exist, describing different factors related to the use of web-based networks in companies. These different models are regarding the factors from different perspectives and views. Each of the models is defining a set of critical success factors (CSFs) that are directly connected to web based networks and responsible for improving business performance. Daniel Crain Smith [13] defines several CSFs related to social networks that are crucial for a company when using Web 2.0 technologies such as social networks. These CSFs are: age of the employees, support from the top management, organizational structure and social pressure from the environment. The main disadvantage of his model is the lack of precise and clear definition of factors that are directly connected to business performance. Colin Smith's [14] model, compared to the previous one,

succeeded in distinguishing several factors that can measure successful use of social networks in a company. Smith suggests income from sales of new products as one of the most important factors that can measure successful use of social networks. Many authors mention other CSFs that are important when one company is using social networks like: human resources, organizational culture and organizational structure. Reagans and B. McEvily [15] and Burt [16] are also defining a set of CSFs responsible for successful use of social networks in one company. According to this group of authors, the most important factors are the type and structure of the networks, demographics of the network and social capital of the organization. Multidimensionality of this model is making it complicated, besides the fact that there are factors which the company cannot influence. Another model that describes different CSFs is Philip Chen's model [17]. Is his model, Chen has defined four groups with two factors that are important for successful use of social networks. These groups are: strategy, infrastructure, openness, and collaboration. Chen clearly defines and confirms that this set of CSFs is responsible for successful use of Web 2.0 technologies to improve business performance from financial and nonfinancial perspective.

From the reviewed literature, most of the authors suggest more or less the same set of the CSFs. For our purposes, we decided to use the two groups' model (or the multistage model). The first group consists of internal factors and the second group consists of external factors. Internal factors are directly connected to every day work of a company and its capability to successfully use social networks. These factors are subsequently divided into the following subgroups: functional factors, strategy factors and technical factors. In the group of functional factors are internal integration of social networks and openness in using. Strategic factors are: alignment of business and Internet marketing strategies and alignment of Internet marketing strategy and CRM (Customer Relationship Management). The third group are technical factors like IT infrastructure and organizational structure. Second subgroup of factors are external ones related to the social pressure that one organization could have from its competitors and users in their environment. These two groups of CSFs are influencing a group of independent financial and nonfinancial factors that are directly related to business performance of a company. These independent factors are income from sales, decreasing costs, and brand recognition.

Based on the literature and the different models mentioned above, a conceptual model was created, with eight groups of factors in total.

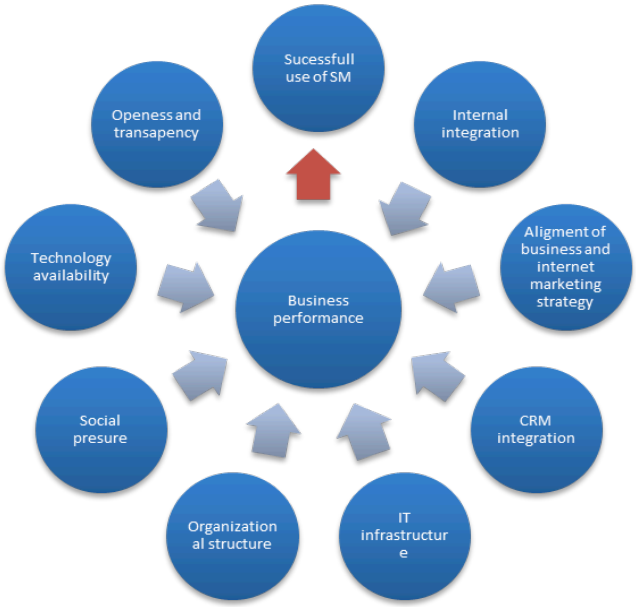


Figure 1. Conceptual model of CSFs for the successful use of social media

III. RESEARCH

For this study, a survey instrument was used. The collected data was analyzed using Structural Equation Modeling (SEM) techniques to validate, evaluate, and analyze the relationships between the specific factors (independent variables) and business performance (dependent variable).

The sample for testing consisted of 110 Macedonian companies, which actively used web-based social networks in the past year. The descriptive statistics show that 91% of the companies in the survey were from Skopje, 20% were in trading industry, 13% in software industry, 12% in telecommunication industry and 14% in media and newspapers. According to the size 40% were small companies, 26% were medium companies, 20% were micro companies and 13% were large companies.

To test the reliability, an internal consistency of this study was analyzed and reviewed using Cronbach's alpha. Construct validity is the extent to which a set of measured items actually reflects the theoretical latent construct those items are designed to measure [18], [19]. In this study, all measures were analyzed for reliability and validity using a Structural Equation Modeling (SEM) confirmatory technique. Confirmatory factor analysis (CFA) was used to construct a measurement model composed of ten constructs of this research model. The measurement model estimates each construct between latent variables (independent variables) and dependent variables using measuring items. To assess convergent

validity for construct validity, composite reliability (CR) and average variance extracted (AVE) from measures were examined [18].

Cronbach's alphas assessing the internal consistency of the study's measures are in range from 0.71 to 0.91. They were above the acceptable threshold of 0.70, suggesting adequate reliability [19].

A confirmatory factor analysis was conducted to determine the convergent and discriminant validity of the 10 constructs. As suggested by Hair [18], convergent validity was assessed by examining the factor loadings, computing the composite reliability of, and average variance extracted for the model constructs. Constructs have convergent validity when the factor loadings are statistically significant, the composite reliability exceeds the criterion of 0.70, and the average variance extracted is above 0.50. As suggested by Hair [18] and Fornell and Larcker [20], discriminant validity was assessed by comparing the squared correlations (between the constructs) and the average variance extracted for a construct. Constructs have discriminant validity when the squared correlations are lower than the average variance extracted for a construct. The fit of the measurement model was assessed using the following statistics and indices: Chi-square, the ratio of the Chi-square to the degrees of freedom (df), Comparative Fit Index (CFI), and Root Mean Squared Error (RMSEA). Chi-square/df values less than 3 indicate good model fit and between 2.0 and 5.0 is acceptable level [21]. CFI values in 0.80 to 0.90 indicate a confidence level [22]. RMSEA values less than 0.90 indicate good fit [20].

The factor loadings, composite reliabilities, and average variance extracted for the model constructs are shown in Table I. From the table it can be concluded that some of the factor loading do not fit the accepted ratios. The index GFI and AGFI are not fitting the acceptance level bellow 0.90. Also RMSEA is over the recommended level of 0.90 [19]. The index CFI is below 0.8 which is also below the recommended level. Only relative index CMIN which represent Chi-square is between 1 and 3 [23].

TABLE I. CFA TESTING WITH 10 CONSTRUCTS

χ^2	1019.429
Level of freedom	620
Level of probability	0,000
GFI	0,631
AGFI	0,56
RMSEA	0,91
CFI	0,735
CMIN	1,664

Next step in CFA is to determine the level of adequacy, by using the squared multiple standardized regressive coefficients. From the analysis, it can be concluded that 13 factors which are explaining the latent variables are below the accepted level of 0.40 [23]. Because the CFA shows that the some of the constructs in the model do not fit the factor loadings, adjustment of the model was required.

After refining and adjusting the model, a confirmatory factor analysis was conducted for the second time. The factor loadings were good, but one of the 10 constructs was left out. The factor loadings are shown in a table 2.

TABLE II. CFA TESTING WITH 9 CONSTRUCTS

χ^2	324,230
Level of freedom	216
Level of probability	0,000
GFI	0,774
AGFI	0,687
RMSEA	0,80
CFI	0,895
CMIN	1,501
PNFI	0,589

From the Table II. it can be concluded that all indexes for model fit are better then the previous model with 10 constructs. RMSEA is within the recommended value of 0.9, CFI is above recommended minimal value of 0.8. Also PNFI (Parsimony-Adjusted Measures), wich show the level of adjustment of the model is better then the first tested model. Also CMIN is in the recommended value between 1 and 3.

To evaluate the fit of the proposed model, the fit of the whole model was tested and assessed. The magnitude (i.e., statistical significance) and direction (i.e., positive or negative) of the individual parameters (i.e., the path coefficients) were assessed. Overall, the goodness-of-fit indices show that the proposed model has a good degree of fit with the data; the ratio of the Chi-square to the degrees of freedom was 3.80, a number smaller than 4.0 is considered very good, and between 2.0 and 5.0 is acceptable [18], the CFI 0.89 was within 0.8 to 0.9 confidence level [19] and the RMSEA 0.08, less than 0.1, which is considered a good fit [22].

The path coefficients and corresponding standard errors and t-values are presented in Table III

TABLE III. STANDARDIZED REGRESSION COEFFICIENTS OF SEM

			t-value	P - value
BusFin	<---	VnatInt	.070	.908
BusFin	<---	PolOtv	1.954	.008
BusFin	<---	BusIMS	.177	.775

			t-value	P - value
BusFin	<---	CRMInt	1.965	.060
BusFin	<---	ITInf	.475	.088
BusFin	<---	OrgPos	1.910	.059
BusFin	<---	NedTeh	1.935	.036
BusNFin	<---	NedTeh	1.935	.022
BusNFin	<---	OrgPos	-.574	.188
BusNFin	<---	ITInf	.335	.137
BusNFin	<---	CRMInt	1.633	.248
BusNFin	<---	BusIMS	.270	.587
BusNFin	<---	PolOtv	1.964	.018
BusNFin	<---	VnatInt	-.134	.779

*Note:*** p-value <0,01; :** p-value <0,05; :* p-value <0,10*

From the values above it can be concluded that 50% of the regression coefficients are statistically significant. Statistical significance is expected to be in the range of p-value <0.10 because of the size of the example [18]. All regressive coefficients are positive except the last one (Business non-financial and internal integration), but there is no statistical significance.

Statistical significance can be found between the next regressive coefficients: culture of open communication - financial and non-financial performance; CRM integration and financial performance; IT infrastructure and financial performance; organization structure and financial performance; availability of technology - financial performance and non-financial performance. There are 5 statistical significant relations related to financial and 2 with non-financial performance. Of all relations, the strongest is CRM integration and financial performance (t-value = 1,965), and the weakest is the availability of technology - financial performance and non-financial performance (t-value = 1.935). From here it can be concluded that seven of fourteen initial hypotheses were confirmed.

IV. CONLUSION

The subject of research in this paper was the analysis of e-business models that include the impact of various factors for successful web-based social network by organizations. This analysis involves a group of independent factors that directly or indirectly affect the success of enterprises.

The results of the conducted SEM research shows that statistical significance exists between the proposed seven critical factors and financial or non-financial performance of the companies. The results show that there is significance in the relationship with financial performance (5 links) and non-financial performance (2 links). These links are: the culture of open communication and financial performance; CRM-integration and financial performance, IT infrastructure and financial performance, organizational structure and financial performance, unavailability of technology and financial performance, unavailability of technology and non-financial performance; policy open communication and non-financial performance.

In terms of financial performance, it can be concluded that the link between CRM- integration and financial performance is the strongest. This relationship is the strongest in the overall model. Namely, if companies invest in social CRM data integration in their system, it will significantly affect and increase their sales and profits. In terms of non-financial performance the strongest link is between the policy of open communication and non-financial performance.

REFERENCES

[1] A. Ansari, O. Koenigsberg, and F. Stahl, "Modeling Multiple Relationships in Social Networks," *J. Mark. Res. JMR*, vol. 48, no. 4, pp. 713–728, Aug. 2011.

[2] S. Belleghem, "Social media integration," *Steven Van Belleghem*, 2013. [Online]. Available: http://stevenvanbelleghem.com/. [Accessed: 13-Aug-2013].

[3] L. Eduardo, "Web-based business models in 2013," 27-Feb-2013.

[4] R. Agnihotri, P. Kothandaraman, R. Kashyap, and R. Singh, "Bringing 'Social' into Sales: The Impact of Salespeople's Social Media Use on Service Behaviors and Value Creation," *J. Pers. Sell. Sales Manag.*, vol. 32, no. 3, pp. 333–348, Summer 2012.

[5] E. Fischer and A. R. Reuber, "Social interaction via new social media: (How) can interactions on Twitter affect effectual thinking and behavior?," *J. Bus. Ventur.*, vol. 26, no. 1, pp. 1–18, Jan. 2011.

[6] T. O'Reilly, *What is Web 2.0*, 1 edition. O'Reilly Media, 2009.

[7] A. Harris and A. Rea, "Web 2.0 and Virtual World Technologies: A Growing Impact on IS Education," *J. Infomation Syst. Educ.*, vol. 20, no. 2, pp. 137–144, 2009.

[8] J. Bell and S. Loane, "'New-wave' global firms: Web 2.0 and SME internationalisation," *J. Mark. Manag.*, vol. 26, no. 3–4, pp. 213–229, May 2010.

[9] A. M. Kaplan and M. Haenlein, "Users of the world, unite! The challenges and opportunities of Social Media," *Bus. Horiz.*, vol. 53, no. 1, pp. 59–68, Jan. 2010.

[10] S. Godin, *Tribes: we need you to lead us*. New York: Portfolio, 2008.

[11] D. Chaffey, *Internet marketing*. Harlow: Financial Times Prentice Hall, 2008.

[12] J. Barney, "Firm Resources and Sustained Competitive Advantage," *J. Manag.*, vol. 17, no. 1, pp. 99–120, 1991.

[13] D. C. Smith, "Social Media Correlates of Organizational Climate," ProQuest LLC. 789 East Eisenhower Parkway, P.O. Box 1346, Ann Arbor, MI 48106. Tel: 800-521-0600; Web site: http://www.proquest.com/en-US/products/dissertations/individuals.shtml, 2009.

[14] C. J. Collins and K. G. Smith, "Knowledge Exchange and Combination: The Role of Human Resource Practices in the Performance of High-Technology Firms," *Acad. Manage. J.*, 2006.

[15] R. Reagans and B. McEvily, "Network Structure and Knowledge Transfer: The Effects of Cohesion and Range," *Adm. Sci. Q.*, vol. 48, no. 2, pp. 240–267, Jun. 2003.

[16] R. S. Burt, *Structural holes: the social structure of competition*. Cambridge u.a.: Harvard Univ. Press, 2005.

[17] P. Chen, “Factors Affecting Business-to-Business Electronic Commerce Success: An Empirical Investigation.” 2010.
[18] J. F. Hair, W. C. Black, B. J. Babin, and R. E. Anderson, *Multivariate data analysis (7th ed.)*. Upper Saddle River, NJ: Prentice Hall, 2009.
[19] R. P. Bagozzi and Y. Yi, “Specification, evaluation, and interpretation of structural equation models,” *J. Acad. Mark. Sci.*, vol. 40, no. 1, pp. 8–34, 2012.
[20] C. Fornell and D. F. Larker, “Evaluating Structural Equation Models with Unobservable Variables and Measurement Error,” *J. Mark. Res.*, vol. 18, pp. 39–50, 1981.
[21] E. Groenland and J. Stalpers, “Structural equation modeling: A verbal approach,” *Nyenrode Res. Pap.*, vol. 12, no. 2, pp. 1–39, 2012.
[22] J. B. Schreiber, F. K. Stege, J. King, N. Amaury, and E. A. Barlow, “Reporting Structural Equation Modeling and Confirmatory Factor Analysis Results: A Review,” *J. Educ. Res.*, vol. 99, no. 6, pp. 323–338, 2006.
[23] J. F. Hair, *Multivariate data analysis*. Upper Saddle River, NJ: Prentice Hall, 2010.

E-preduvjeti fiskalizacije računa u Hrvatskoj

Ivana Arbanas

Veleučilište "Nikola Tesla" u Gospiću/Poslovni i Prometni odjel, Gospić, Hrvatska
ivana.arbanas@velegs-nikolatesla.hr

Sažetak – Obveznici fiskalizacije računa moraju uz ostale prilagodbe, sukladno Zakonu o fiskalizaciji u prometu gotovinom (dalje Zakon), osigurati i određene e-preduvjete za njenu provedbu. Cilj rada je istražiti i prikazati koje e-prilagodbe za provedbu fiskalizacije trebaju izvršiti obveznici provedbe fiskalizacije računa te koliki su troškovi njihove nabave i uspostave. Zaključak istraživanja je da fiskalizacija računa utječe na povećanje troškova poslovanja svih obveznika njene provedbe, i to najviše na povećanje troškova poslovanja malih poduzeća i obrta koji nisu koristili, ili su koristili u manjoj mjeri, informacijske tehnologije i softver u poslovanju te je potrebno ukazivati i poticati na iskorištavanje mogućnosti povećanog korištenja suvremenih tehnologija uzrokovanog provedbom fiskalizacije za razvoj e-poslovanja, novih proizvoda, marketing i širenje tržišta, odnosno povećanje prihoda.

Ključne riječi: e-preduvjeti fiskalizacije računa, obveznici fiskalizacije, troškovi nabave e-preduvjeta za fiskalizaciju računa, razvoj e-poslovanja.

I. UVOD

Radi osiguravanja nadzora nad plaćanjima u gotovini, tj. smanjenja poreznih utaja pri plaćanjima gotovinom, donesen je i u primjeni je od 1. siječnja 2013. Zakon o fiskalizaciji u prometu gotovinom (NN 133/12). Ovim Zakonom propisano je tko su obveznici fiskalizacije u prometu gotovinom (dalje fiskalizacija), koji su rokovi njene provedbe, postupak i način provođenja, evidencija obveznika fiskalizacije, gotovinska plaćanja između obveznika fiskalizacije te nadzor nad provođenjem Zakona. Sam postupak provođenja fiskalizacije detaljnije je reguliran Pravilnikom o fiskalizaciji u prometu gotovinom (NN 146/12), dalje Pravilnik. Obveznici fiskalizacije koji obavljaju plaćanja u gotovini osim Zakonom propisanih obveza za sve obveznike fiskalizacije (izrade internog akta s popisom poslovnih prostora i načinom njihova označavanja, prilagođavanja elemenata računa sukladno Zakonu i ostalim poreznim propisima te postavljanja upozoravajućih naljepnica o obvezi izdavanja i uzimanja računa) moraju osigurati i određene e-preduvjete za komunikaciju s PU i provedbu fiskalizacije računa.

II. OBVEZE I E-PREDUVJETI OBVEZNIKA FISKALIZACIJE ZA PROVEDBU FISKALIZACIJE IZDAVANJA RAČUNA

Oni obveznici fiskalizacije koji obavljaju plaćanja u gotovini obveznici su i fiskalizacije računa. Npr. poduzeće koje je registrirano kao veleprodaja obveznik je fiskalizacije, ali ne nužno i fiskalizacije računa. Naime, sukladno Zakonu između ostalih obveznici fiskalizacije su i pravne i fizičke osobe obveznici poreza na dobit pa sukladno tome ukoliko poduzeće posluje kao veleprodaja, ali ako nema promete u gotovini, nego samo plaćanja općom uplatnicom na transakcijski račun ili s jednog transakcijskog računa na drugi, nije obveznik fiskalizacije računa [1], ali ako ima plaćanja u gotovini, tj. ostvaruje plaćanja novčanicama, kovanicama, čekovima, karticama ili drugim sličnim plaćanjima, obveznik je i fiskalizacije računa, osim za djelatnosti Zakonom izuzete ili oslobođene od fiskalizacije. Dok mali obveznici fiskalizacije mogu odlučiti žele li račune ispostavljati iz unaprijed u Poreznoj upravi (PU) ovjerene knjige računa ili kao svi ostali obveznici fiskalizacije računa. Obveznici fiskalizacije računa, uz sve obveze koje imaju svi obveznici fiskalizacije sukladno Zakonu, dužni su prije slanja prvog računa koji se odnosi na taj poslovni prostor, potpisati digitalnim certifikatom i putem uspostavljene elektroničke veze s PU dostaviti i poruku o poslovnim prostorima u kojima obavljaju plaćanja gotovinom, s naznakom svih podataka iz čl. 19. Zakona te račune izdavati putem elektroničkih naplatnih uređaja, softverski prilagođenih za potrebe fiskalizacije, koji im omogućavaju digitalno potpisivanje poruka i internetsku vezu s PU, tj. ovjeravanje računa jedinstvenim identifikatorom računa (JIR) od strane PU [2], [3], [4] te osigurati knjigu uvezanih računa ovjerenu u PU iz koje će ispostavljati račune u slučaju prekida elektroničke veze ili kvara elektroničkog uređaja.

Za provođenje samog postupka fiskalizacije izdavanja računa obveznici iste trebaju:

- "prilagoditi ili nabaviti naplatni uređaj sa softverom koji podržava provođenje fiskalizacije izdavanja računa
- od Financijske agencije (FINA-e) nabaviti digitalni certifikat koji se u postupku fiskalizacije koristi za elektroničko potpisivanje elemenata računa te za identifikaciju obveznika fiskalizacije prilikom elektroničke razmjene podataka
- nabaviti aplikaciju za elektroničko potpisivanje te
- povezati naplatni uređaj na Internet radi razmjene podataka s PU"[3].

U Republici Hrvatskoj (RH) primjenjuje se softverski model fiskalizacije s certifikacijskim sustavom, tj. obveznici fiskalizacije sami, sukaladno svojim poslovnim potrebama, odlučuju hoće li, tj. mogu li već postojeći naplatni uređaj (POS ili PC blagajnu) softverski prilagoditi, sukladno Zakonu [2] i Pravilniku [4] i tehničkim specifikacijama PU [5], [6], tako da im omogućava elektroničko potpisivanje podataka, internetsko povezivanje i razmjenu podataka s PU ili trebaju li nabaviti neki novi elektronički uređaj i odgovarajući softver koji im to sve omogućavaju te trebaju osigurati internetsku vezu za razmjenu podataka s PU.

III. DIGITALNI CERTIFIKATI

Za potrebe fiskalizacije FINA-a, kao jedini registrirani ovlašteni izdavatelj digitalnih certifikata (CA) u RH, izdaje dvije vrste cetifikata:

- Produkcijske aplikativne certifikate za obveznike fiskalizacije
- DEMO aplikativne certifikate za informatičke tvrtke koje će obveznicima fiskalizacije vršiti uspostavu sustava fiskalizacije.

A. Digitalni aplikativni certifikati - obveznici fiskalizacije

Osim nabave ili prilagodbe postojećeg elektroničkog naplatnog uređaja, koji mora omogućiti uporabu programske podrške za elektroničko potpisivanje elemenata računa i elektroničku vezu (Internet) za razmjenu podataka s PU, obveznik fiskalizacije računa za provedbu fiskalizacije mora nabaviti i produkcijski aplikativni digitalni certifikat FINA RDC: Fiskalx (x - redni broj certifikata izdanog za određenog obveznika fiskalizacije) koji izdaje Financijska agencija (FINA). Produkcijski aplikativni certifikat se koristi za elektroničko potpisivanje elemenata računa te za identifikaciju obveznika fiskalizacije prilikom elektroničke razmjene podataka. Ovaj certifikat nabavlja sam obveznik fiskalizacije ili njegov opunomoćenik, tj. informatička tvrtka koja obvezniku fiskalizacije obavlja uspostavu sustava fiskalizacije ne može ga nabaviti u njegovo ime, osim ako osoba ovlaštena za zastupanje obveznika fiskalizacije ne dodijeli pravo skrbnika djelatniku informatičke tvrtke, sukladno čl. 4. Pravilnika. Preduvjet za izdavanje produkcijskog aplikativnog certifikata je da obveznik fiskalizacije bude evidentiran u Registar digitalnih certifikata koji vodi FINA (Sustav PKI-Public Key Infrastructure). Ukoliko obveznik fiskalizacije već koristi neki od certifikata FINE (za neku drugu elektroničku uslugu FINA-e - Regos, PDV, ID...) već je registriran u Sustav PKI, tj. nije potrebna ponovna registracija. Registracija u sustav PKI vrši se predajom potrebne dokumentacije u nekom od Registracijskih ureda FINA-e ili elektronički putem aplikacije RDC Obrasci on-line kojoj se može pristupiti putem internetskih stranica FINA-e.

Za registraciju u sustav PKI potrebno je dostaviti sljedeću dokumentaciju:

- Rješenje o upisu u nadležni registar (tijelo koje ga je osnovalo).
- Obavijest državnog zavoda za statistiku o razvstvanju prema Nacionalnoj klasifikaciji djelatnosti (NKD-u) i matičnom broju (samo za pravne subjekte).
- Presliku osobne iskaznice osobe ovlaštene za zastupanje pravnog subjekta [7].

Za preuzimanje digitalnog aplikativnog certifikata potrebno je predati sljedeću dokumentaciju (ili presliku iste): "1. Zahtjev za izdavanje aplikativnog certifikata za fiskalizaciju – 1 primjerak, koji se može preuzeti na internetskoj stranici Financijske agencije,

2. Ugovor o obavljanju usluga certificiranja – 2 primjerka, koji se može preuzeti na internetskoj stranici Financijske agencije,

3. presliku identifikacijskog dokumenta skrbnika certifikata, te

4. dokaz o uplati naknade za izdavanje produkcijskih aplikativnih certifikata za fiskalizaciju" [7].

Nakon obrade predane dokumentacije, skrbniku certifikata, na način naznačen u Zahtjevu za izdavanje produkcijskih aplikativnih certifikata, dalje u tekstu Zahtjev, (sms-om, e-mail-om ili u Registracijskom uredu FINA-e), izdaju se referentni broj i autorizacijski kod, čijom kombinacijom može certifikat preuzeti on-line pomoću CMS Fina portala za preuzimanje certifikata, na Internet adresi https://mojcert.fina.hr/finacms/. Ako je skrbnik certifikata fizički identificiran te ako je u Zahtjevu za izdavanje produkcijskih aplikativnih certifikata naveden podatak o njegovom broju mobitela i adresi e-pošte, aktivacijske kodove skrbnik certifikata dobiva na ta dva kanala i to; referentni broj na e-mail, a autorizacijski broj putem SMS-a. Ako u Zahtjevu nije naveden broj mobitela skrbnika certifikata, aktivacijske kodove može preuzeti u Registracijskom uredu FINA-e. Rok za preuzimanje aktivacijskih podataka/kodova za preuzimanje produkcijskih aplikativnih certifikata nije određen. "Skrbnik je fizička osoba koja je zadužena za preuzimanje, uporabu, brigu i čuvanje privatnog ključa, odnosno certifikata koji je izdan za aplikaciju. Skrbnik može biti i osoba zadužena za zastupanje poslovnog subjekta" [7].

"Da bi obveznik fiskalizacije izdavanje aplikativnih certifikata za fiskalizaciju u potpunosti obavio elektronički, mora zadovoljiti sljedeće preduvjete:

- da osoba ovlaštena za zastupanje posjeduje kvalificirani digitalni certifikat izdan na FINA e-kartici/tokenu ili CoBranding kartici/tokenu banke,
- da je osoba ovlaštena za zastupanje poslovnog subjekta ujedno i skrbnik certifikata,
- da se na Zahtjevu za izdavanje certifikata naznači podatak o adresi e-pošte i broju mobitela na koje će biti isporučeni aktivacijski podaci/kodovi.

Postupak izdavanja provodi se na sljedeći način:

- osoba ovlaštena za zastupanje putem aplikacije RDC obrasci on-line popunjava elektronički obrazac Zahtjeva u kojem sebe ujedno imenuje i skrbnikom certifikata,
- na temelju podataka navedenih u certifikatu, automatski se generira Ugovor o obavljanju usluga certificiranja,
- osoba ovlaštena za zastupanje Zahtjev i Ugovor potpisuje naprednim elektroničkim potpisom,

- osoba ovlaštena za zastupanje/skrbnik zaprima par aktivacijskih podataka/kodova: referentni broj i autorizacijski kod, odvojeno; referentni broj na e-mail, a autorizacijski kod na SMS,

- kombinacijom dobivenih podataka, skrbnik vrši on-line preuzimanje certifikata" [8], pomoću CMS FINA portala za preuzimanje certifikata.

Nakon toga FINA dostavlja ovjeren i potpisan ugovor o certificiranju na adresu e-pošte navedene u Zahtjevu.

Ukoliko osoba ovlaštena za zastupanje ne posjeduje odgovarajući kvalificirani digitalni certifikat ili obveznik fiskalizacije nije evidentiran u sustav PKI, tada se Zahtjev za izdavanje certifikata može elektronički popuniti i poslati putem aplikacije RDC obrasci on-line, čime će biti automatski predbilježen u Fininoj bazi, što ubrzava proces. Istodobno je Zahtjev potrebno ispisati, potpisati i dostaviti u Finu s ostalom dokumentacijom potrebnom za izdavanje aplikativnog certifikata. Nakon obrade Zahtjeva, skrbnik certifikata dobiva aktivacijske kodove i može on-line preuzeti digitalni certifikat, na prethodno opisan način.

Pri preuzimanju produkcijskih aplikativnih certifikata, obveznici fiskalizacije računa (poslovni subjekti) trebaju se pridržavati FINA-ine Procedure za preuzimanje produkcijskih aplikativnih certifikata za fiskalizaciju. Sve detaljnije informacije, kao i sva potrebna dokumentaciju i procedure za zahtijevanje i preuzimanje produkcijskog aplikativnog certifikata mogu se pronaći na internetskoj stranici FINA-e, http://www.fina.hr/Default.aspx?sec=1543.

U Fininoj Radnoj uputi za preuzimanje digitalnog certifikata detaljno je opisan postupak rješavanja problema kad Internet Explorer i Mozilla Firefox upozoravaju na mogući sigurnosni problem prilikom pristupa stranici https://mojcert.fina.hr/finacms/, tj. prikazan je postupak prihvaćanja stranice kao sigurne u Windows ili Linux okruženju ili se za rješavanje ovog problema savjetuje, i opisuje kako, prvo obaviti postupak preuzimanja RDC CA root certifikata i njegove pohrane unutar Windows Certificate Store-a.

"Certifikati Fine su usklađeni s X.509 v3 standardom i preporukom RFC 5280 koja regulira uporabu X.509 v3 certifikata na internetu" [9]. Aplikacijski certifikati koje Fina izdaje za potrebe fiskalizacije su certifikati u soft obliku i njihovo korištenje je moguće na raznim platformama, tj. usklađeni su s međunarodnim standardima i nisu ograničeni ni na jednu specifičnu računalnu platformu i standardne su razine sigurnosti. Svrha digitalnog certifikata je autentikacija poreznog obveznika čime se on pouzdano i sigurno predstavlja PU i garantira se da je izdani račun upravo došao od tog poreznog obveznika te se njime omogućuje primjena elektroničkog potpisa, sukladno Zakonu o elektroničkom potpisu. Elektronički potpis jamči integritet i nepromjenjivost podataka na računu (potpis bi bio narušen ukoliko bi se naknadno mijenjali podaci nakon potpisa). Fina kao ovlašteni izdavatelj certifikata (CA), utvrđuje identitet i provodi registraciju svih poslovnih subjekata prije izdavanja certifikata, kako bi se ti podaci kasnije koristili unutar servisa kao jamstvo PU da je račun potpisan upravo certifikatom određenog poreznog obveznika. Aplikativni digitalni certifikati izdaju se u soft obliku pa se mogu koristiti (kopirati) na više prodajnih mjesta i više blagajni, unutar jednog poslovnog subjekta (OIB-a). Moguće je koristiti jedan certifikat na više računala kao i koristiti više certifikata na jednom računalu. Pri korištenju više certifikata na jednom računalu (vođenje poslovanja više obveznika fiskalizacije) potrebno je pri preuzimanju/pohranjivanju pojedinog certifikata voditi računa o imenovanju p12 datoteka, s obzirom da je certifikat vezan uz OIB poslovnog subjekta. [9].

Postavlja se pitanje zašto se za potpisivanje računa ne koristi napredni elektronički potpis koji ima veću razinu sigurnosti. Budući da je napredni elektronički potpis vezan uz fizičku osobu potpisnika i omogućen je primjenom kvalificiranog certifikata, koji se mora izdati na smart kartici ili USB key sa chipom (mediju-kripto uređaju koji zadovoljava određene propisane standarde) to bi značilo da bi uz nabavku uređaja i svaki djelatnik na svakoj blagajni morao posjedovati svoju karticu/key s certifikatom, dok korištenje softverskog certifikata bez potrebe za kripto uređajem pojednostavljuje izradu aplikativnih rješenja, s obzirom da se onda ona ne moraju prilagođavati tehničkim preduvjetima za primjenu smart kartice/keya, a uz to se mogu koristiti i na više blagajni, što smanjuje troškove i jedino je u aplikativnom certifikatu moguće definirati naziv aplikacije (fiskal) iz čega je vidljivo da se on koristi za potrebe fiskalizacije. Upotreba elektroničkog potpisa nema utjecaj na povećanje podatkovnog prometa, a samim time i povećanje troškova izazvanih time. Naime, obveznik fiskalizacije, sukladno Zakonu, dužan je osigurati 24 satnu internetsku vezu, koja omogućuje komunikaciju s PU, a paketi prometa koje nude operateri kreću se uglavnom od 1GB do FLAT rate (neograničenog prometa), a poruka, zajedno s potpisom, je veličine nekoliko KB (engl. Kilobyte), a potpis povećava poruku za 2 – 3 KB, odnosno potpisana poruka je oko 300 000 puta manja od najmanjeg paketa prometa iz ponude, što znači da je moguće poslati 300 000 računa unutar jednog mjeseca uz korištenje najmanjeg paketa prometa. Digitalni aplikativni certifikat vrijedi pet godina, a cijena mu je 300,00 kn + PDV [9].

B. Demo aplikativni certifikati - informatičke tvrtke

Informatičkim tvrtkama koje obveznicima vrše uspostavu sustava fiskalizacije na njihov zahtjev, isključivo za potrebe testiranja rješenja na naplatnim uređajima obveznika fiskalizacije, Fina izdaje DEMO aplikativne certifikate. Ovi certifikati su iste konstrukcije kao i produkcijski aplikativni certifikati, ne naplaćuju se, a izdaju se na razdoblje od dvije godine. Također je potrebno da poslovni subjekt bude registriran u sustav PKI. Ukoliko je već registriran tada treba samo predati dokumentaciju potrebnu za izdavanje Demo certifikata za fiskalizaciju i to:

- Zahjev za izdavanje DEMO aplikativnog cerfikata za fiskalizaciju – 1 primjerak.
- Presliku osobne iskaznice koja će biti skrbnik certifikata.

Upute "Procedura za izdavanje DEMO certifikata za fiskalizaciju", cijene licenci i ostali potrebni obrasci za izdavanje ovog certifikata mogu se pronaći na internetskim stranicama FINA-e, http://www.fina.hr/Default.aspx?sec=1542 .

Za digitalno potpisivanje poruka koje sadrže propisane elemente računa (XML-a) Fina je razvila programsku komponentu modul XmlSinger, koji je moguće integrirati u aplikativna rješenja. Poruke se potpisuju aplikativnim certifikatom za fiskalizaciju. Unutar modula nalazi se dll koji se može povezati s .NET aplikacijom, odnosno jar, za ugradnju u java aplikacije. Modul XmlSinger izdaje se za određeni poslovni subjekt, a njegov rezultat je ispravno potpisan HML računa, odnosno račun s potpisom koji će prihvatiti PU, nakon slanja putem web servisa. Računi se mogu potpisati isključivo aplikativnim certifikatom poslovnog subjekta za kojeg je i izdan modul XmlSinger. Informatička tvrtka koja će uspostavljati sustav fiskalizacije kod obveznika fiskalizacije mora za korištenje XmlSinger modula nabaviti za svakog poslovnog subjekta jednu licencu, tj. jednu ili više njih ovisno o tome kod koliko obveznika fiskalizacije uspostavlja sustav fiskalizacije. Cijenu licence/i plaća tvrtka koja je licence naručila [10].

Ako je obavljena fizička identifikacija skrbnika certifikata, tada on aktivacijske kodove može dobiti na e-mail, a ukoliko nije, tada ih može preuzeti u zaštićenoj omotnici u Fini. Aktivacijske kodove za preuzimanje ovog certifikata potrebno je aktivirati u roku od 30 dana. Ukoliko nisu aktivirani u tom roku potrebno je zatražiti izdavanje novih. Nakon preuzimanja DEMO certifikata isti je moguće koristiti/kopirati na onoliko mjesta/računala koliko je potrebno.

Korisnik je odgovoran za zaštitu lozinke kojom se pristupa privatnom ključu pripadajućeg Demo aplikacijskog certifikata, tj. ne smije zaporku zapisivati i pohranjivati u nezaštićenom obliku na računalu za fiskalizaciju.

U programsko rješenje može se ugraditi mogućnost da kod potpisivanja elemenata računa softver koristi zaštićenu pohranjenu zaporku i tako može pristupiti privatnom ključu i certifikatu za potpisivanje. Ukoliko je potrebno pronaći gdje je certifikat pohranjen na računalu treba provjeriti certificate store Windowsa (Internet Explorer->Tools->InternetOptions->Content>Certificates ili preko mmc snap in-a) [9].

IV. ANALIZA PONUDE UREĐAJA I INTERNETA ZA FISKALIZACIJU RAČUNA

Potrebno je naglasiti da vlada zabluda da je obavezno nabaviti novi hardver za potrebe fiskalizacije. Ukoliko poduzetnik posjeduje računalo ili neki drugi elektronički naplatni uređaj prosječne konfiguracije i isti omogućuje spajanje na internet i komunikaciju s PU, nije potrebno kupovati novi uređaj, već postojeći softverski prilagoditi za potrebe fiskalizacije.

U nastavku rada dat je prikaz ponuda elektroničkih uređaja, softvera i paketa interneta za potrebe fiskalizacije, do kojih se došlo slučajnim izborom, istraživanjem za stolom, tj. na temelju podataka dostupnih na internetu u trenutku pisanja rada, s tim da je dat samo prikaz tih podataka, a ne i ocjena ili izbor najboljeg rješenja, s ciljem da se korisnicima uštedi vrijeme pretraživanja ponuda i prikaže što je moguće veći broj istih, kako bi oni sami, tj. uz savjet informatičkih stručnjaka procijenili što je najbolji i istodobno najučinkovitiji izbor za njih, s obzirom na vrstu njihova poslovanja, dosadašnje korištenje ili nekorištenje elektroničkih uređaja i interneta u poslovanju.

A. Ponuda elektroničkih uređaja i softvera za fiskalizaciju

Pri izboru elektroničkih uređaja i/ili softvera za potrebe provođenja fiskalizacije, uz specifičnosti posla svakog poduzetnika pojedinačno, potrebno je skrenuti pozornost na to što se nudi u ugovoru o održavanju i prilagodljivosti programa za fiskalizaciju tipu poslovanja. Pri izboru softverskog rješenja treba obratiti pažnju na sljedeće: na koje uređaje se može implementirati, njegovu jednostavnost, mogućnost preinaka i podrške u slučaju zakonskih ili potrebnih periodičnih promjena u poslovanju. Najjeftinije niti najskuplje rješenje za fiskalizaciju ne mora nužno biti najbolji izbor, te je preporučljivo sve prethodno testirati, tj. provjeriti cjelokupnu funkcionalnost uređaja i softvera u praksi.

Kako piše Poslovni dnevnik [11] poduzetnicima se na tržištu nudi više od 300 različitih fiskalnih blagajni, tj. rješenja za fiskalizaciju, različitih „mogućnosti" i cijena. Tako se fiskalne blagajne tvrtke Fiskal 1 mogu naći i u prodavaonicama Konzuma po cijeni do 2 000 kuna, a mogu se odmah početi koristiti, podrška korisnicima se pruža kroz automatska osvježenja softvera, a hardver, odnosno Android Tablet i printer za ispis računa reklamiraju se kao i svaka druga kupljena roba. Zagrebačka tvrtka Infokom razvila je cloud rješenje koje se može koristiti i na Android tabletima, a koje se naplaćuje kroz mjesečni najam od 150 kuna na više. Tvrtka Lunacon iz Mostara razvila je vlastito rješenje softvera "Fiskalne Blagajne POS Sector", koje je usklađeno s hrvatskim zakonima, a prodaju ga kroz najam za 215 kuna mjesečno ili jednokratno za 3 550 kuna u što su uračunate eventualne zakonske izmjene.

Prema pisanju Jutarnjeg lista [12] na tržištu vlada raznolikost ponude, nesrazmjer cijena, nuđenje nepotpunih i dvojbenih usluga što dodatno zbunjuje korisnike. Prema njihovim saznanjima, do kojih su, kako pišu, došli uz savjet informatičkih stručnjaka, za manje poduzetnike koji se prvi put susreću s fiskalnim blagajnama najkvalitetnije, a istodobno financijski najpovoljnije rješenje je nabava tableta s Android sustavom, koji se može kupiti za manje od 1 500 kuna (bez PDV-a), a sadržava hardverski i softverki dio i pisač i ne zahtijeva nikakve dodatne troškove (licence za operativni sustav, antivirusni program-nema virusa i slično), jer je djelomično zasnovan na Linuxu, a može obavljati sve zadatke za potrebe fiskalizacije. Prema istom navodu, Hrvatski telekom (HT) nudi sva tri dijela za fiskalizaciju (hardver, softver i internet za komunikaciju s PU), a nudi ga putem najma. Uz HT na-

jam računala i programa za fiskalizaciju nude i drugi ponuđači, i to programski dio po cijenama od 50 do 250 kuna, a računala od 200 kuna na više. Za prodaju se programi osposobljeni za fiskalizaciju nude po cijenama od 1 200 kuna na više, a računala od 2 500 kuna na više. Također postoje i brojne ponude polovnih računala. Navodi se da većina ponuda uglavnom uključuje operativni sustav Windows, što zahtijeva dodatnu investiciju od oko 900 kuna za licencu, te još oko 200 kuna za antivirusni program, budući da se radi o sučelju sklonom prihvaćanju virusa. Također se navodi da nekoliko ponuđača nudi svoja rješenja na Linux platformi koja je besplatna i na kojoj virusi gotovo ne postoje, a čija je primjena moguća i na računalima skromnije konfiguracije što doprinosi uštedi pri nabavi računala, budući da je većina računala uglavnom presnažna tj. nepotrebno skuplja za funkciju maloprodajne blagajne. Linux rješenje je cijenama bliže klasičnim PC blagajnama srednje klase, a njihovo se programsko rješenje nudi samo kroz najam.

Kako piše na internetskim stranicama Hrvatske udruge MIPRO [13], HT je svoju ponudu obogatio Mini fiskalnom blagajnom koja korisnicima nudi najjednostavniju prilagodbu obvezama fiskalizacije, lako je prenosiva, zauzima malo prostora, jednostavna je za korištenje i instalaciju i nije potrebno ispunjavati nikakve infrastrukturne preduvjete te se zbog toga nameće kao idealno rješenje za sve obveznike fiskalizacije, a posebice za obveznike fiskalizacije koji rade na terenu. HT nudi dvije varijante ovog uređaja/blagajni: Mini prijenosnu fiskalnu blagajnu koja vezu s PU može ostvarivati putem mobilnog i/ili fiksnog interneta i Mini fiskalnu blagajnu koja se na Internet spaja putem fiksne linije.

Mini prijenosna fiskalna blagajna nameće se kao idealan izbor za obveznike fiskalizacije kao što su trgovci na tržnicama i svi ostali koji veći dio posla obavljaju na otvorenom i terenu, a Mini fiskalna blagajna svima koji pretežit dio posla obavljaju u uredima, tj. imaju mogućnost pristupa fiksnoj linii, npr. za odvjetnike, turističke djelatnike, frizere i sl. Navodi se da se izborom nekog od ovih uređaja dobiva potpuno rješenje za fiskalizaciju i da se uređaji mogu koristiti neovisno o operateru, rad na njima je jednostavan, a sadrže blagajnu (uređaj), pisač, mogućnost spajanja na internet i aplikaciju, tj. blagajnički softver, a sadrže snažnu bateriju za rad do dva dana prije nego ju je potrebno opet puniti. Korisniku su, na blagajni ili u sklopu web aplikacije, u sklopu usluge, na raspolaganju i izvještaji o izdanim računima, unošenje i ažuriranje cijena i sličnih podataka te ova rješenja za fiskalizaciju omogućuju izdavanje računa u slučaju prekida internetske veze s PU, koji će onda naknadno biti automatski poslani u trenutku uspostave veze.

Mini fiskalne blagajne moguće je nabaviti samostalno ili u paketu s internetom HT-a. Navodi se da je za poslovanje u pokretu najbolji izbor "Mini paket koji uključuje usluge Mini prijenosnu fiskalnu blagajnu i pristup internetu putem M2M (Machine to Machine) veze za 199 kuna mjesečno (cijena bez PDV-a" [13], pri čemu je sav internetski promet u mobilnoj mreži prema PU uključen u mjesečnu naknadu za mobilni Internet, tj. cijena je uvijek ista bez obzira na broj izdanih računa i korisniku je na raspolaganju 24 satna korisnička i tehnička podrška. Više o ponudi HT-a za prilagodbu fiskalizaciji dostupno je na internetskim stranicama HT-a, https://www.hrvatskitelekom.hr/poslovni/ict/poslovna-rjesenja/fiskalizacija.

Za korisnike koji posjeduju računalo s pristupom internetu i printer, kako piše na internetskim stranicama hrvatske udruge MIPRO [13], za svođenje troškova na minimum kao najbolje rješenje za fiskalizaciju nameće se Vip aplikacija za fiskalizaciju, jer u tom slučaju nisu potrebna nikakva dodatna ulaganja i aplikacija se može koristiti odmah po pokretanju, izdavanje računa je jednostavno, "a cijena kompletne usluge, koja uključuje i održavanje aplikacije i telefonsku podršku, iznosi 60 kuna + PDV mjesečno, bez ugovorne obveze"[13]. Korisnici koji instaliraju mobilni printer mogu ovu uslugu koristiti i na tabletu ili pametnom telefonu, a također se može koristiti uz vlastite uređaje i postojeći Internet. Korisnici mobilnog interneta imaju besplatan podatkovni promet za potrebe fiskalizacije. Navodi se da Vipnet uz Vip blagajnu nudi i mobilne uređaje, tablet i printere po povoljnim cijenama. Više informacije o ponudi Vipa za fiskalizaciju na http://www.vipnet.hr/poslovni/fiskalizacija.

Poslovni dnevnik, također, piše da tvtka Fiskal 1 nudi i Fiskal1 GO koje je "mobilno cjelovito rješenje koje uključuje 7" tablet, softvere Fiskal1 te Bluetooth baterijski printer. Njegova je cijena 1992 kune plus PDV". Prema istom navodu, tvrtka GRC iz Dugog sela za postojeće korisnike koji već imaju instalirana njihova softverska rješenja, nudi nadogradnju postojećeg softvera za potrebe fisklaizacije. Također se navodi da tvrtka Datalab, nudi softversko rješenje Pantheon. "Rješenje za fiskalizaciju nude gotovo svim djelatnostima, a početna cijena iznosi 999 kuna"[14].

U nastavku su navedene internetske stranice različitih ponuđača uređaja i softvera za fiskalizaciju, za potrebe različitih djelatnosti, na kojma se mogu proučiti njihove ponude, tj. obveznici fiskalizacije proučavajući iste mogu okvirno izračunati i predvidjeti troškove vezane uz potrebne e-prilagodbe za fiskalizacju. RI. POS d.o.o., Računalni inžinjering, Rijeka također nudi hardver i rješenja za fiskalizaciju. Njihova ponuda dostupna je na http://www.ri-pos.hr/software.html. Ponuda rješenja za fiskalizaciju tvrtke Fiskal1, http://www.fiskal1.hr/. Poduzeće Napredni informatički sistemi d.o.o. također nudi softversko rješenje za fiskalizaciju-Fiskalux, koje se sastoji od tableta, printera i softvera za fiskalizaciju, više o njihovoj ponudi na http://fiskalux.hr/#home. Remaris-Integralni poslovni sustavi d.o.o. također nudi rješenje za fiskalizaciju RemarisPOS fiskalne blagajne, a njihova ponuda dostupna je na http://www.remaris.com/.

Jedan od ponuđača opreme i softvera za fiskalizaciju je i tvrtka Digitron d.o.o. koja ima dugogodišnje iskustvo u proizvodnji popularnih digitrona. Paletu njihovih proiz-

voda i usluga moguće je detaljnije proučiti na internetskoj stranici http://www.digitron.hr/. Za više djelatnosti i Kigo Grupa d.o.o. u suradnji sa Symmetria d.o.o. nudi različitu opremu i programe za fiskalizaciju, a njihova ponuda se može pogledati na KigoKasa.com, http://kigokasa.com/. Neoinfo d.o.o. za informatiku, trgovinu i usluge-Moji računi, na http://moji-racuni.hr/, također nudi softver i hardver za fiskalizaciju.

Prema pisanju Poslovnog dnevnika vlada je napravila dobar potez što nije naručila licencu za fiskalne blagajne jer bi one onda bile dosta skuplje te da bi bilo upitno kako uniformirani program uklopiti u sve vrste poslovanja i koliko bi jednostavno bilo njegovo održavanje, a ovako je omogućena modernizacija na način da je iskorišten obvezan pristup internetu za smještanje programa za fiskalizaciju u "Cloud". Ovakvoj kasi se s bilo kojeg uređaja može pristupiti u bilo kojem trenutku s bilo kojeg mjesta na svijetu, s bilo kojeg uređaja koji ima pristup internetu. "Takva kasa radi na iPhone-u, iPadu, Androidu, te svim verzijama Windowsa i Linuxa" [15].

I većina ranije spomenutih tvrtki nudi cloud rješenja za fiskalizaciju, ali uz njih i sve više ostalih ponuđača nudi različita tzv. cloud rješenja za fiskalizaciju, koja se ne instaliraju na računalo korisnika nego se aplikacija za fiskalizaciju smješta u tzv. oblak, tj. premješta se na infrastrukturu drugih tvrtrki koja se iznajmljuje, tako npr. funkcioniraju Gmail ili Facebok, i često su takva rješenja besplatna i omogućavaju održavanje bez dolaska na lokaciju korisnika, tj. potrebno je imati još samo Internet preglednik novije generacije i osigurati vezu na Internet [16]. Na Fiskalna. hr (Inter-Net d.o.o) [17] također se nude slična rješenja za fiskalizaciju, njihova ponuda dostupna je na http://www.fiskalna.hr/.

Još jedna od aplikacija koja radi na sličnom principu je i internetska aplikacija KlikUred koju je razvila varaždinska informatičaka tvrtka Orion informatika. S ovom aplikacijom mali poduzetnici i obrtnici, obveznici fiskalizacije, mogu "izdavati i fiskalizairati račune samo uz notebook ili tablet koji ima pristup internetu i pisač te mjesečni izdatak niži od pretplate za mobilni telefon!"[18]. Kako piše t/portal aplikacija KlikUred je jedna od sedam startup projekata koje je Grad Varaždin nagradio iz Fonda svježih ideja kojemu je cilj potpomagati ostvarenje cjelovitih i održivih poduzetničkih zamisli mladih poduzetnika, u ovom slučaju studenata suradnika ORION informatike i djelatnika iste tvrtke. Njihov glavni cilj je bio omogućiti cjelokupno uredsko poslovanje u "oblacima", a za početni paket usluga treba mjesečno izdvojiti 49 kuna. Osim izdavanja računa, ponuda, radnih naloga i sl. tu će se nalaziti i super jeftino telefoniranje, e-mali besplatne videokonferencije te mnoge druge informatičke usluge, a imaju i projekt Mobilni KlikUred, tj. rješenje za sve one koji rade na terenu, dostavljače, terenske prodavače i sl. Jedino što je potrebno za rad je njihov Android 3G Tablet i bluetooth pisač [18], https://www.klikured.hr/. Pozitivno je i to što piše da imaju u planu otvaranje novih radnih mjesta, jer planiraju ovaj proizvod plasirati i na međunarodna tržišta, na kojima je tvrtka već prisutna. Iz tvrtke Raverus d.o.o. nude Desktop i Web rješenje za fiskalizaciju, https://www.fisoffice.hr/Default.aspx. Rep.hr. (Računalni poslovi i Internet vijesti) [19] piše da je obrt RGB-toneri ponudio besplatnu online aplikaciju za fiskalizaciju koja je namjenjena malim obrtima. O sigurnosti certifkata na serveru i mogućnostima ovog rješenja, http://www.rep.hr/vijesti/internet/ponudili-besplatnu-aplikaciju-za-fiskalizaciju/5398/ i besplatna fiskalna, online fiskalna blagajna za fiskalizaciju, http://www.besplatna-fiskalna.com. Postoji još mnogo ponuda hardvera i softverskih rješenja za fiskalizaciju i vjerojatno se s većinom ponuđača mogu dogovoriti različite kombinacije rješenja i cijena, sukladno specifičnostima i potrebama poslovanja obveznika fiskalizacije, a većina obveznika fiskalizacije i u svom susjedstvu ima malu informatičku tvrtku koja može ponuditi dobra rješenja za fiskalizaciju ili dati informatičke savjete.

Pri pisanju rada svi ponuđači hardvera i softverskih rješenja za fiskalizaciju pronađeni su slučajnim izborom, pretraživanjem njihovih ponuda dostupnih na Internetu u trenutku pisanja rada, niti jedno rješenje nije preporučeno kao najbolje i najpovoljnije. Pri eventualnoj odluci o kupnji nekog od u radu navedenih rješenja za fiskalizaciju potrebno je, najbolje uz pomoć informatičkih stručnjaka, proučiti detaljno svaku ponudu, kontaktirati detaljnije svakog izabranog ponuđača, ispitati funkcionalnost uređaja i rješenja za fiskalizaciju u praktičnom poslovanju, ispitati sigurnost podataka, pouzdanost i dostupnost usluge i podrške u budućnosti, koje se dodatne usluge nude i ocijeniti koje rješenje je najisplativije i istodobno najpraktičnije za nesmetano i efikasno poslovanje, uzimajući u obzir specifičnost svake pojedine djelatnosti.

B. Ponuda paketa interneta za fiskalizaciju

Za provođenje fiskalizacije osim potrebnog hardvera i softvera obveznici fiskalizacije trebaju osigurati i internetsku vezu putem koje će se računi slati na ovjeru u PU. Prema pisanju Poslovnog dnevnika [11] HT ima "kompletnu ponudu koja uključuje mjesečni najam PC blagajne, LCD monitora, POS printera, blagajničke aplikacije Luceed Kasa usluge i 3G mobilni Internet već od 300 kuna mjesečno." Također piše da nude pet usluga pristupa internetu u skladu s uvjetima fiskalizacije. Više o uslugama pristupa internetu na Internetskim stranicama https://www.hrvatskitelekom.hr/poslovni/internet-poslovni, https://www.hrvatskitelekom.hr/poslovni/ict.

Poslovni dnevnik [11] također piše, da Vipnet nudi tri tarife „čije se cijene kreću od 35,06 kn mjesečno", http://www.vipnet.hr/poslovni/fiskalizacija?icn=menu.Pišu i da je Tele2 ponudu za fiskalizaciju razradio u suradnji s Metronetom i da nude usluge i na optičkoj i bakrenoj mreži kao i mobilni pristup, a jedna od usluga je i SLA koja garantira da neće biti nedostupnosti više od 15 minuta mjesečno. Također navode da Optima telekom ima niz usluga za fiskalizciju, a među njim i ADSL, stalnu vezu i optičku vezu, a pišu da u H1Telekomu kažu da u svojoj ponudi imaju niz paketa pristupa Internetu koji odgovara-

ju svim poslovnim namjenama pa tako i potrebama fiskalizacije, a cijene se kreću od 49 kuna mjesečno". Detaljniji prikaz spomenutih tarifa pristupa Internetu dostupan je na linkovima u nastavku: Tele2 d.o.o., http://www.tele2.hr/poslovnikorisnici/usluge/fiskalizacija/vb60/#vise, Metronet, http://www.metronet.hr/gotovipaketi/promo-paketi/fiskalizacija/, Optima telekom d.d., http://www.optima.hr/hr/poslovni-korisnici/optimalna-rjesenja/halo-surf/, H1 Telekom d.d., http://www.h1telekom.hr/Naslovnica/tabid/54/Default.aspx., Iskon Internet d.d. također nudi različite pakete usluga za poslovne korisnike, više na http://www.iskon.hr/Za-tvrtke.

V. ZAKLJUČAK

Iz u radu prikazanih ponuda ponuđača softvera, uređaja i interneta za fiskalizaciju može se izračunati da osiguravanje e-preduvjeta za fiskalizaciju računa u prvoj godini primjene fiskalizacije povećava godišnje troškove poslovanja poduzetnika okvirno za oko 1 100 do 5 000 kuna pa i više, ovisno o prethodnom korištenju ili nekorištenju suvremenih tehnologija te veličini i specifičnosti djelatnosti poduzetnika. Fiskalizacija povećava troškove poslovanja svih njenih obveznika, a pogotovo malih obrtnika i poduzeća koji nisu ili su u manjoj mjeri koristili suvremene tehnologije u poslovanju, jer oni moraju nabaviti i uređaje i ili softver, osigurati vezu za Internet, a velika većina i prilagoditi se radu na tim uređajima i digitalni certifikat, pa imaju i znatno povećanje troškova i utrošak vremena u odnosu na njihove ionako male prihode, ali PU smatra da su ti troškovi mali i neće imati značajniji utjecaj na poslovanje obveznika fiskalizacije računa. Iako je fiskalizacija povećala troškove poslovanja pridonijet će i povećanom korištenju suvremenih tehnologija u poslovanju obrta i malih poduzeća, koji su najviše pogođeni gospodarskom krizom (pad kupovne moći, nemogućnost financiranja i sl.), a najmanje su koristili suvremene tehnologije u poslovanju. Naime, prema istraživanju koje je 2012. za Google Hrvatska provela agencija za istraživanje tržišta GfK [20], na uzorku od 700 malih i srednjih tvrtki iz cijele Hrvatske, iako 90% malih i srednjih tvrtki u RH ima pristup Internetu, a čak 99% tvrtki koje imaju više od 10 zaposlenih, samo 20% njih koristi Internet kao medij za poboljšanje prodaje svojih proizvoda, iako su svi ispitanici svjesni važnosti Interneta za poslovanje i povećanje izvoza. Tek 16% korisnika Interneta obavlja kupnju preko Interneta, a samo 13% ispitanika je imalo trening za promociju proizvoda na internetu, što ukazuje na mogućnosti razvoja e-trgovine, ali i potrebe razvoja obrazovanja i potpora poduzetnicima u ovom području kako bi povećali prihode, a pri čemu postoje znatne mogućnosti za financiranje istog iz EU fondova.

LITERATURA

[1] M. Vuraić Kudeljan, D. Dodig, "Primjena Zakona o fiskalizaciji u prometu gotovinom", Računovodstvo i financije, Hrvatska zajednica računovođa i financijskih djelatnika (Rif), Godina LIX, siječanj 2013., str. 126, 127.

[2] Zakon o fiskalizaciji u prometu gotovinom, Zagreb, Narodne novine d.d., br. 133/12, [Online], str. 2822 i dalje, http://narodne-novine.nn.hr/clanci/sluzbeni/2012_12_133_2822.html, 20. 12. 2013.

[3] M. Vuraić Kudeljan, D. Dodig, "Prikaz Zakona o fiskalizaciji u prometu gotovinom", Računovodstvo i financije, Hrvatska zajednica računovođa i financijskih djelatnika (Rif), Godina LVIII, prosinac 2012., 62 – 70., str. 63, 66.

[4] Pravilnik o fiskalizaciji u prometu gotovinom, NN 146/12

[5] Ministarstvo financija, Porezna uprava, Apis IT d.o.o., Fiskalizacija u prometu gotovinom, Tehnička specifikacija za korisnike,Verzija 1.1, http://www.porezna-uprava.hr/HR_Fiskalizacija/Aktualnosti%20dokumenti/Fiskalizacija%20-%20Tehnicka%20specifikacija%20za%20korisnike_v1.1.pdf, 20. 12. 2013.

[6] Ministarstvo financija, Porezna uprava, Apis IT d.o.o. Fiskalizacija u prometu gotovinom, Tehnička specifikacija za korisnike,Verzija 1.2, http://www.porezna-uprava.hr/HR_Fiskalizacija/Aktualnosti%20dokumenti/Fiskalizacija%20-%20Tehnicka%20specifikacija%20za%20korisnike_v1.2.pdf, 20. 12. 2013.

[7] Financijska agencija (FINA), Fiskalizacija, Obveznici fiskalizacije-produkcijski certifikati, Procedura (Radna uputa) za zahtijevanje aplikativnog digitalnog certifikata, http://www.fina.hr/Default.aspx?sec=1543, 22. 2. 2013.

[8] Financijska agencija (FINA), Fiskalizacija, Elektronička predaja obrazaca, http://www.fina.hr/Default.aspx?sec=1559, 25. 2. 2013.

[9] Financijska agencija (FINA), Fiskalizacija, Najčešća pitanja i odgovori, O certifikatima za provedbu zakona o fiskalizaciji, http://www.fina.hr/Default.aspx?art=10121, 22. 12. 2013.

[10] Financijska agencija, Fiskalizacija, Informatičke tvrtke-DEMO certifikati, http://www.fina.hr/Default.aspx?sec=1542, 22.2. 2013.

[11] B. Ivezić, "Na tržištu se poduzetnicima nudi više od 300 različitih fiskalnih blagajni", Poslovni dnevnik, 24. 1. 2013., http://www.poslovni.hr/hrvatska/na-trzistu-se-poduzetnicima-nudi-vise-od-300-razlicitih-fiskalnih-blagajni-228128, 15. 2. 2013.

[12] V. Panić, "KAKO JEDNOSTAVNO DO POREZNE BLABAJNE Fiskalni tablet: Najbolje i najjeftinije rješenje za sve obrtnike" Jutarnji list, 6. 12. 2012., http://www.jutarnji.hr/fiskalni-tablet--najbolje-i-najjeftinije-rjesenje-za-sve-obrtnike/1070668/, 22. 2. 2013.

[13] Hrvatska udruga MIPRO, "Mini fiskalna blagajna HT-a – jednostavno, mobilno i povoljno rješenje za nove obveznike fiskalizacije", 1. ožujka 2013., http://www.mipro.hr/Default.aspx?TabId=36&newspage=3&newspro=0&language=hr-HR&newsid=85798ea2-5003-4f93-82fb-e6f6730d9cb0 , 25. 3. 2013. , Vip blagajna-jednostavno i povoljno rješenje za fiskalizaciju, 22. veljače 2013. http://www.mipro.hr/Default.aspx?TabId=36&language=hr-HR&newspage=2&newspro=0&newsid=7b22b021-c6dc-47d1-a588-1ba7553b2a8f , 25. 3. 2013.

[14] J. Ban,/ VLM, " 'Kači' Vas fiskalizacija? Evo gdje povoljno kupiti mini blagajne", Poslovni dnevnik, Tehnologija, 12. 4. 2013. 13:58, http://www.poslovni.hr/tehnologija/mini-blagajna-za-samo-199-kuna-237361 , 30. 4. 2013.

[15] A. Prtenjača, "Razvoj prema poslovanju u oblacima'", Poslovni dnevnik, 28. 1. 2013. http://www.poslovni.hr/tips-and-tricks/razvoj-prema-poslovanju-u-oblacima-228578, 30. 4. 2013.

[16] Portal Danas.hr., Odlična kvaliteta, besplatnog programa fiskalizacije!, http://danas.net.hr/promo/odlicna-kvalitetabesplatnog-programa-fiskalizacije, 30. 5. 2013.

[17] Fiskalna.hr, http://www.fiskalna.hr/, 30. 5. 2013.

[18] I. Kruhoberec/tportal.hr, VARAŽDIN PREPOZNAO ODLIČNU IDEJU, S ovakvim mladima nema straha za budućnost, 10. 2. 2013. 9:06, http://www.tportal.hr/vijesti/hrvatska/242215/S-ovakvim-mladima-nema-straha-za-buducnost.html, 30. 4. 2013.

[19] Portal Računalni poslovi i Internet vijesti/rep.hr, "Ponudili besplatnu aplikaciju za fiskalizaciju", 15. 2. 2013. 11:42, http://www.rep.hr/vijesti/internet/ponudili-besplatnu-aplikaciju-za-fiskalizaciju/5398/, 30. 5. 2013.

[20] Lider, Google: "Svega 20 posto malih tvrtki koristi internet za unaprjeđenje prodaje", Lider, 15. siječnja 2013. 12:33, http://liderpress.hr/tehnopolis/svega-20-posto-malih-i-srednjih-tvrtki-koristi-internet-kao-alat-za-unapredenje-prodaje-/, 7. 1. 2014.

Legal regulation of public information infrastructure in Republic of Croatia

doc. dr. sc. Goran Vojković
dr. sc. Tihomir Katulić

University of Zagreb, Faculty of Transport and Traffic Sciences
Vukelićeva 4, Zagreb, Croatia
E-mail: goran.vojkovic@fpz.hr

Faculty of Law of the University of Zagreb,
Trg maršala Tita 14, Zagreb, Croatia
E-mail: tihomir.katulic@pravo.hr

Abstract – Draft proposal of the Croatian State Information Infrastructure Act was made public in December 2013. This Act, due to be entered into legislative procedure by summer 2014, aims to regulate rights, obligations and responsibilities of the state regulatory bodies concerning establishment, development and management of the state information infrastructure system. The Act adopts a unified method of establishing and managing public registries and conditions the state information infrastructure has to implement with regard to public registries as well as using common data interchange platform within the system of state information infrastructure. The draft proposal regulates government to citizen (G2C) and government to businesses (G2B) services. The aim of the article is to analyse the current draft proposal, consider its compatibility with the European legal framework as well as with national legislative and, where current provisions fail to meet the legislative standards, to propose modifications de lege ferenda.

Keywords – information infrastructure, public sector, public registers, master register, metaregister, regulation, Croatia

I. INTRODUCTION

On the 12th of December 2013, the Public Sector Informatization Coordination Commission of the Government of the Republic of Croatia adopted the draft proposal of the State Information Infrastructure Act ("Draft Proposal") [1]. The government ministry in charge of the develepoment of the mentioned Draft Proposal is the Ministry of Public Administration.

The Draft Proposal was made available to the public on the 19th of December which marked the start of the public discussion [2]. As of the beginning of February, the official documents regarding the public discussion, received comments and criticism from experts and the general public are yet to be made available to the public.

The proposed Act, due to be entered into legislative procedure by summer 2014, aims to regulate rights, obligations and responsibilities of the government institutions and other bodies concerning establishment, development and management of the state information infrastructure system.

The Act adopts a unified method of establishing and managing public registries and conditions the state information infrastructure has to implement with regard to public registries as well as using common data interchange platform within the system of state information infrastructure. The draft proposal regulates government to citizen (G2C) and government to businesses (G2B) services as well as other users and further development of information society in the Republic of Croatia. The Act is an important step toward further integration of the Republic of Croatia into established European as well as global information networks and services.

Establishment and development of state information services is undoubtedly a challenging task considering the rapid development of information technology, information services, electronic media, privacy and other data protection issues etc. Furthermore, both from the organizational as well as financial, legal and security perspectives, establishing and managing public registries while at the same time insuring their interoperability, authenticity and information security is a daunting task even for societies far ahead of the Croatian one regarding transition into information society.

It is obvious that this complicated issue needs to be properly addressed through legislative effort - even more so in light of the facts concerning public spending on various informatization projects in the recent years. According to the publicly available state budget reports, the State and various other public sector bodies, institutions and other budget funded entities spend around billion Croatian Kuna a year on various IT projects. While budget informatization spending has been on this level for a number of years, the real question is has it had the much needed impact on the quality and speed of state administration and on the total costs of government.[3]

We feel that the Draft Proposal of the State Information Infrastructure Act suffers from numerous inconsistencies and nomotechnic defects that, proposers good intentions notwithstanding, bring the success of the legislative measures in question.

II. REASONS FOR NEW LEGISLATIVE ACTION

The published draft proposal contains an additional document with the main motives regarding the proposition to adopt measures to regulate state information infrastructure.

While correctly stating that Croatian legal framework at this time does not include a set of rules to regulate establishment, development and management of the state information infrastructure, it does describe the main problem of the current state of state information infrastructure, as well as perhaps public administration in general: "*Just like the development of the real world institutions which are developed autonomously and function autonomously, like islands, so is the development of the information systems. Connecting them into an integral sytem that can serve citizens and business entities is something delegated to the very same citizens and business entities who are burdened to carry information and documents from one institution to another, regardless of their state of informatization*".[4]

This *modus operandi* of public administration, both central as well as the one within local and regional self-government has often been underlined as one of the main antagonists in attracting significant foreign investments and businesses to the Republic of Croatia.

As a result, the citizens and businesses alike are expected to know:

1. the competences and area of jurisdiction of state institutions

2. their address and office hours

3. documentation they need to bring, along with the fact that they need to set aside significant time for each interaction with state authorities.

Both the citizens as well as business entities go through these steps each time they interact with public administration. Administration institutions relentlessly demand documents and data already available to other institutions within the administration, although the State Administration System Act [5], as well as the General Administrative Procedure Act [6] clearly state that civil servants have a duty to obtain data contained within public records established by government bodies. The document concludes that *"...currently there is no technical system in place to enact the said provisions*."

At this point, a remark needs to be made - the Draft Proposal does not only imply technical measures regarding infrastructure services, it also steps into the matter regulated by both the State Administration System Act as well as the General Administrative Procedure Act.

In order to solve the above mentioned problems the Draft Proposal:

- Establishes the central state internet portal as a unique administrative location in the virtual domain
- The National Autentification and Identification System is mandatory for all institutions
- Provides for the Public Sector institutions (see infra for clarification on "public sector") communicating with citizens through a personal electronic mailbox

These steps create legal conditions and prerequisites to enable the e-Citizens ("*e-Građani*") project.

Additionaly, the Draft Proposal describes and regulates public and master registries, authentic data, and interconnection of registries stipulating the obligation of the public sector bodies to access data from master registries without demanding the same data from citizens or business entities.

Article 47. p.2 of the Croatian General Administrative Procedure Act stipulates that "*Civil servant is required to acquire data regarding information in registers run by other administrative institutions or courts*". Its predecessor, the former federal General Administrative Procedure Act adopted 1991, that was in force until 2009. contained a very similar provision.

In general, a question arises why does a new legislative measure strive to regulate matter regulated by general principles of administrative procedure well over half a century ago instead of ensuring the application of existing provisions, just to improve what has correctly been detected as a problem - the absence of a systemic technical solution required to apply the existing provisions in the first place.

III. DRAFT PROPOSAL TERMINOLOGY

The terminology of the provisions in this Draft Proposal requires special scrutiny. Since the proposed Act, the State Information Infrastructure Act represents a general normative text, the terminology contained therein defines a number of extremely important concepts on a national level with repercussions throughout the Croatian legal system.

It has long been known that speed and efficiency of the public administration (or more precisely, lack of) present significant challanges to both the citizens and business entities in the Republic of Croatia. The main goal of this new legislative proposal is to regulate the legal framework of electronic governance in the Croatian legal system and to allow anyone with legal interest to access data harvested by the public administration and other entities performing public service.

While we can agree with the proposition that legal framework of e-government as it is in Croatian positive legal system needs additional normative effort in order to be fully effective as well as compliant to the existing normative mechanisms of (personal) data protection, right to access information, management of identity in digital environment and other aspects of what is increasingly being recognized as information law, there are certain inconsistencies and nomotechnical problems with the current draft proposal that merrit analysis.

In this chapter we shall present provisions of the draft proposal that we find inconsistent with existing legislative framework and try to offer solutions.

Article 1 of the draft proposal contains an unclear reference to "*public sector bodies responsible for establishing, developing and managing the system of the national infromation infrastructure*". The colloquial expression "*public sector bodies*" is not a legal standard. While it seems obvious that this provision actually refers to the bodies listed in the provisions of the Articles 2 and 3 of the State Administration System Act [5], the expression "public sector bodies" is referenced throughout the draft proposal including provisions in Articles 2.1.1, 2.1.4, 2.1.6., 2.1.9 etc. Furthermore, Article 2.1.20 defines "*public sector bodies as public administration bodies, state bodies and public services*".

While Croatian legal system does occasionally within a few other Acts, government Strategies and Decisions

contain the expression "*public sector bodies*", most notably the Budget Act [7]. However, this expression is not broadly accepted as a legal term that describes the range of public administration entities (central, regional and local bodies) and other legal persons with public/official authority within the Croatian legal system. Furthermore, the Budget Act in this regard provides a different definition of public sector bodies (Article 3.21 of the Budget Act defines public sector as ..." *state budget, units of local and regional selfgoverment, budget and extrabudget users of state budget and the budgets of units of local and regional selfgovernment"*). The draft proposal also contains another unclear expression - *"public services"* (2.10.) - defined as "*...public institutions and other legal persons whose means of operation are provided by the state budget"*.

It would appear more prudent to avoid usage of non-standard expressions in legal documents, especially general norms like Acts in order to avoid confusion regarding the scope and applicability of law. The proposal should instead refer to the provisions of the State Administration System Act and the Organisation and Scope of Ministries and Other Central State Administration Bodies Act (OG 150/11 and 22/12)

The Article 2 of the draft proposal contains basic expressions and concepts used throughout the proposal. Several of these warrant analysis, most notably *autentification* (2.1.2.), *electronic identity* (2.1.4.), *electronic certificate* (*"vjerodajnica"*, 2.1.5.*)*, *identification* (2.1.7), data (2.1.12), and aforementioned provision 2.1.20.

Concerning *autentification* (2.1.2) and *identification* (2.1.7), there seems to be some duplicity in terms of defining these procedures. The proposal states that "*Autentification is a formalized process of verification of users electronic public services certificates resulting in affirming/denying identity"*, whereas "*Identification is a process in which it is necessary to ascertain or recognize electronic public services users identity"*. It would be more appropriate to merge these provisions into a single provision, perhaps defining both identification and autentification as a process of establishing users identity and level of access to information system or service (public register or any other information system or service) by verifying his advanced electronic signature and qualified certificate, as defined by the Croatian Electronic Signature Act [8].

With regard to *electronic identity* (2.1.4.) and *electronic certificate* (*"vjerodajnica"*, 2.1.5.*)*, the draft proposal should be inline with the provisions of the existing Electronic Signature Act, which defines electronic signature, advanced electronic signature, certificate and qualified certificate. The draft proposal defines electronic identity as "*...a unique set of identification data regarding a certain subject (natural person, public sector body, information system) being kept in electronic form allowing for identification of related subject*". Meanwhile, the electronic certificate according to the draft proposal is "*...a set of data identifying the user of an electronic service which serves to prove electronic identity in order to allow access to electronic services"*.

At the same time, the existing Electronic Signature Act defines electronic signature (as well as advanced electronic signature, Articles 2.1 and 2.2) as "*...a set of data in electronic form associated or logically connected to other data in electronic form providing identification of user (signee) and authenticity of the signed electronic document*", and certificate (and qualified certificate) "*...as an electronic certificate that associates electronic signature verification data to a certain person thereby verifying that person's identity.*"

Furthermore, the definition of the master register ("*temeljni registar"*) is also questionable (2.1.18.). The master register is one of the fundamental notions of this Act. According to the Draft Proposal, "*a master register is a public register that collects and contains at least one authentic data item*". Evidently, almost every state register contains authentic data making this provision too broad without a clear purpose. In our opinion, master registries are in fact state registries ("*državne matice"*) and a limited number of other important registries organized and run by the state.

The question here is the relationship between electronic identity, electronic signature and electronic certificate. Seperate definitions relating to the same matter (in this context identifying users identity in an electronic environment via electronic signatures and certificates) present a challenge regarding legal interpretation, even more so when there is no clear distinction which of the conflicting norms is the more specialized one (*Lex specialis derogat legi generali*).

The solution to this situation is to adapt the draft proposal inline with the provisions of the Electronic Signature Act or add a provision to the draft proposal concerning the conflict between Electronic Signature Act and the draft proposal effectively derogating the prior law (*lex posterior derogat legi priori*).

IV. PUBLIC REGISTRIES

One of the most important parts of the Draft Proposal is the Chapter III titled "*Public Registers*". This Chapter, along with the rest of the Draft, contains several serious flaws and inconsistencies. Analyzing all of them in detail is not possible within limits of this short paper, so we shall concentrate on those we deem most important.

The Draft Proposal states that the purpose of master registries is to, through facilities of the state information infrastructure, to ensure availability of collected authentic data required to perform duties stemming from their jurisdiction to all bodies of the public sector (8.1).

Regarding this provision several questions come to mind. Why should data be made available only through facilities of the state information infrastructure, and even then why only to the bodies of public sector? Does this imply that all registers need to be based on a state controlled infrastructure, and what does this mean for the registers maintained by companies with public competences? Also, the provisions of the Right to Access Information Act allow interested parties to access data collected by the public sector bodies, under certain conditions.

Furthermore, where purpose of the proposed Act is concerned, there is no point in regulating the exact technical procedure by which the said purpose is to be achieved, especially considering the very dynamic of technological change information technology is renowned for. Since much data stored in master registries is already available to the general public, there is no point in limiting access to it by a specific norm.

Regarding the collection of authentic data, the Draft Proposal states "*authentic data is collected only once, and*

entered into the master register" (8.2). This is a classic provision stipulating that data should be entered into an electronic system only once, and henceforth should this data be required, it should be requested from said electronic system, not from citizens or other subjects.

The provision that "*the master register is established by decision of the Government of the Republic of Croatia following a proposal by the central administration body responsible for E-Croatia matters*" (8.3) also raises doubt. A decision by the Government of the Republic of Croatia by which a register is declared a master register is in essence a constitutive decison. We find that it is not necessary that Government should proclaim a register to be a master register by an individual decision (there are literally hundreds of registers!) when the definition of terminology used in the Draft Proposal already defines which register is, in fact, a master register. To illustrate, let us consider registers connected to maritime affaires - implementation regulation concerning maritime affaires contains several registers, and this is for a field of activity that is just one of more than hundred and fifty administrative areas regulated by Croatian law. [9]

Furtheremore, the following Article (9.2.) stipulates: "*Public registers are established by law or a general act of a legal person with public authority or by an international treaty, that clearly specify data to be collected, what data is authentic and who is the manager of the public register*." If a register is established by law or an international treaty then no separate decision by the Government of the Republic of Croatia is necessary. Even when a register is established by the means of a general act by a legal person with public authority, the public administration has authority and administrative mechanisms to analyze the establishment and organization of a public register (i.e. through mechanisms of administrative oversight - Article 21.1 of the State Administration System Act which grants the competent ministry adequate administrative oversight authority)

With this in mind, it is perhaps interesting to note that according to the Ordinance on Internal Organization of the Ministry of Public Administration there is no administrative oversight provided for activities administered by Directorate for e-Croatia [10]. Since the Ministry of Public Administration regulates that master registries are established following a proposal by the Directorate for E-Croatia, the Ordinance should be amended and the Directorate for E-Croatia should be provided with powers of administrative oversight.

According to the Draft Proposal, official records of the public sector bodies are established exclusively as public registers (9.1). Issue we detect here is the following - is the term "*official record*" comparable to the the term "*public register*"? Croatian nomotechnic practice has so far failed to produce a standard concerning this issue. Legislative practice itself has seen usage of distinctly different terms - registers, lists, accounts, record files etc. (registri, evidencije, upisnici, očevidnici itd.). With this in mind, even if we lack legislative standards, it may be prudent to consult theoretical division as it is currently understood in Croatian academia.

Registers or record files are statutory regulated data collections containing records on important data regarding excercising certain rights, performing duties on the grounds of administrative authority defined by law or protection of legaly defined public interest in a manner that contained records are made by a competent authority (i.e. boat register, road vehicle register, but also a land register etc.)

Inquest register or records of evidence are data collections of events and procedures established and maintained by competent authorities without regard or consent of the subject of inquiry (i.e. during a criminal investigation or a misdemeanor record).

Catalogues, inventories and other listings are data collections governed by internal rules and regulations concerning a certain activity without a formal legal grounds.

Of course, suggested division is not systemic nor formally, de lege lata, acknowledged. Current legislative practice contains various division, some rather peculiar (i.e. the one suggested by the Regulation on archive evidence records [11] which divides the archive evidence records into: books, general inventory, dossiers and evidence records (!)).

State information infrastructure Act should take advantage of this existing division, especially considering unequal and diverging legal status of the registers, records of evidence and data catalogues already in existence.

There is further evidence of terminological confusion. The Draft Proposal states: "*Public sector bodies maintaining master registers are obliged to, without delay and without asking for additional permits, to deliver (dostaviti) authentic data to a public register of another public sector body...*" (9.3). as well as: "*Establishment of public records for the purpose of collecting authentic data already being collected by other master registries is not allowed, instead data has to be taken up (preuzeti) from master registers*." Here a difference in terminology is present, however the actual meaning of these two different terms (*dostaviti, preuzeti)* is evidently the same.

Draft Proposal also iterates already present statutory obligations of the public bodies. Article 11 of the states: "*Public sector bodies maintaining master registries shall deliver without asking for additional permits data to another public sector body that uses said data within the scope of performing prescribed duties*". This provision is an exact copy of the Article 47.p.2. provision of the General Administrative Procedure Act that has the same objective. Perhaps this provision of the Draft Proposal is a reaction to the abysmal lack of compliance to this norm observable in the practice of the Croatian public administration and public sector bodies, however including the same provision into two diferent laws is not going to miraculously enable its enforcement but rather trigger further nomotechnic disorder. A proper way to ensure enforcement of this provision would be to apply existing oversight mechanisms.

Similar duplication of provisions can be also observed in the Article 15.p.2 of the Draft Proposal that states: "*Law or a general act abolishing a public register shall establish a method of continuity of further accesibility and processing of data from the abolished public register, especially an abolished master register.*" In and of itself, within the confines of this Draft Proposal, this provision seems in order. However, data from an abolished public register constitutes archive matter, as defined by the Archive and Archive Matter Act [12]. This Act defines archive matter as :" *...records or documents originated by action of legal or physical persons in performing their activities, and have lasting significance on culture, history and other sciences without regard to place and time of their origination, form or carrier medium stored on*".

When abolishing a register there may be some special circumstances that need to be defined by a competent authority (such law needs no special provisons in this regard apart from the fact that a public register is being abolished), however the lex specialis in this matter obviously exists - the Archives and Archive Matter Act - and should be applied.

Additionally, this Draft Proposal should contain specifications of longterm electronic data storage that have not been included into the Archives and Archive Matter Act, especially since Republic of Croatia has already adopted the relevant ISO 19005 standard. [13]

V. METAREGISTER

Another interesting feature of the Draft Proposal is the establishment of the Metaregister in the Article 16.

Metaregister is a public register established as a part of the state information infrastructure containing information on all public registers, specifically:

- legal grounds on establishment of the public register
- indication if the register is a master register
- indication of the kind of data collected by the public register and whether that data is authentic
- delivery of data and possible recepients of that delivery
- legal grounds of data delivery
- description of the data collection process
- overview of the manner and conditions of data utilization

Article 16. p.1 states that the purpose of establishment of the Metaregister is to ensure interoperability of all public registers, reminiscent of the Central Register implemented by the provisions of the Article 14 of the Personal Data Protection Act concerning the establishment and management of personal data collections.

In Article 16.p.2.sp.4 the Draft Proposal states "*…delivery of data and possible recepients of that delivery*" and continuing in the sp.5 "*…legal grounds of data delivery*". Meanwhile, Article 11 of the Personal Data Protection Act regulates the act of allowing a third party to receive collected and processed personal data as handing over rather than delivery (*davanje* instead of *dostava*). While both of these expressions stand for essentially the same thing, keeping them both in equivalent legal norms might needlessly incure legal uncertainty.

VI. MANAGEMENT OF THE STATE INFORMATION INFRASTRUCTURE

Finally, Articles 18 through 20 of the Draft Proposal regulate the management of the state information infrastructure.

While Article 18.p.1.sp.1 through 10 regulate the mandate of the body competent to oversee application of the State Information Infrastructure Act, Article 19 regulate the duty of that body to report to the Government of the Republic of Croatia on the progress and development of the state information infrastructure and Article 20 stipulates that the head of the competent body has duty to implement measures aimed at raising awareness regarding the activities and importance of the state information infrastructure.

Regarding this chapter of the Draft Proposal, our main concern lies within the provisions of the Article 18. More precisely, provision 18.1.10 is unclear on how this new Act relates to the provisions of Personal Data Protection Act as well as the Right to Access Information Act [14].

It is unclear what the expression "*verifies the compliance of the legal grounds and the request for access to information in public registers and authorizes the same*" actually means. However, provisions of Article 15. of the Right to Access Information Act, especially p.3 of said Article, regulate the terms and conditions how public authorities can limit access to information - these regulations do not provide for an additional level of control the Draft Proposal seems to impose.

Furthermore, this provision influences the Personal Data Protection Act as well. Article 32. and 33. of the Personal Data Protection Act regulate the competence of the Personal Data Protection Agency. The current provision in the Draft Proposal might obstruct several of the currently prescribed duties of the Agency, such as those regulated in Articles 32.1 and 32.5.

VII. CONCLUSION

The Draft Proposal, as presented to the public, is a goodhearted attempt motivated by positive and constructive intent, but falls short of the goal of facilitating effective legal framework for development of the state information infrastructure and new generation of services for citizens and businesses.

The main problem of the proposal is unnecessary and superflous mixing of technical concepts and measures within what is supposed to be a general normative text. As it is, this draft simply does not connect with the provisions of many applicable and existing laws and regulations, including but not limited to General Administrative Procedure Act, Electronic Signature Act, State Administration System Act, Archival Matter and Archives Act, Personal Data Protection Act, Right to Access Information Act as well as Office Transactions Regulation [15].

Ensuring better and more effective public administration and public sector performance is a question of well designed and implemented procedures, not merely a question of technology and technological infrastructure. Recent experiences show that major investments in information and communication technology divorced from rethinking and implementing modern business procedures in the public administration environment fails to produce an improvement in quality, availability and speed of public administration services.

In this regard, the Draft Proposal offers very little - management of state information infrastracture is again limited on management of the physical and software infrastructure, and not on the modernization and improvement of administrative procedures which should, in our opinion, represent a focus of activity for the Ministry of Public Administration.

REFERENCES

[1] Povjerenstvo za koordinaciju informatizacije javnog sektora, Zaključak o prihvaćanju Nacrta prijedloga Zakona o državnoj informacijskoj infrastrukturi i Prijedloga Uredbe o ustrojavanju javnog Registra o projektima državne informacijske infrastrukture (Registar ProDII) – KLASA: 650/01/12-01/37, URBROJ: 505-05-01-02-01/1-13-90, Zagreb, 12. prosinca 2013.

[2] Nacrt prijedloga Zakona o državnoj informacijskoj infrastrukturi, http://www.uprava.hr/default.aspx?id=13984 (visited: 6.2.2014)

[3] Business.hr, „Država rasipa milijardu kuna za IT bez nadzora i strategije," http://www.business.hr/ekonomija/ drzava-rasipa-milijardu-kuna-za-it-bez-nadzora-i-strategije-108863 (visited: 20.1.2014)

[4] Dokument za savjetovanje o Nacrtu prijedloga Zakona o državnoj informacijskoj infrastrukturi s obrazloženjem zakona http://www.uprava.hr/default.aspx?id=13984 (6.2.2014)

[5] Zakon o sustavu državne uprave (Narodne novine, br. 150/11, 12/13)

[6] Zakon o općem upravnom postupku (Narodne novine, br. 47/09)

[7] Zakon o proračunu (Narodne novine, br. 87/08, 136/12)

[8] Zakon o elektroničkom potpisu (Narodne novine, br. 10/02, 80/08)

[9] http://www.mppi.hr/UserDocsImages/more-popis%20propisa_10.PDF (15.2.2014)

[10] Uredba o unutarnjem ustrojstvu Ministarstva uprave (Narodne novine, br. 28/12)

[11] Pravilnik o evidencijama u arhivima [Narodne novine, br. 90/02).

[12] Zakon o arhivskom gradivu i arhivima (Narodne novine, br. 105/97, 64/00, 65/09, 125/11) i Pravilnik o evidencijama u arhivima [Narodne novine, br5. 90/02).

[13] Vojković, Goran: Dugoročna pohrana elektroničkih dokumenata, Pravo i porezi. XIX. (2010), 3; 30-34

[14] Zakon o pravu na pristup informacijama (Narodne novine, br. 25/13)

[15] Uredba o uredskom poslovanju (Narodne novine, br. 7/09)

* "Narodne novine" is "Official Gazette" of the Republic of Croatia.)

Content Management Systems for mobile, context-dependent Augmented Reality applications

M. Lengheimer*, G. Binder ** and T. Rößler **
* AVL List GmbH, Graz, Austria
** evolaris next level GmbH, Graz, Austria
mario.lengheimer@avl.com

Abstract - In the last few years, the proliferation of mobile devices with internet access, especially smartphones and tablets, has grown strongly. This has also positively influenced the mobile augmented reality (AR) market and allowed AR applications to achieve a high number of users. The necessary tools and frameworks for developing such applications are numerously available. However, there is still the question how to store and manage the displayed contents of the applications in an effective way. This is true, especially if the management of user individual content is considered. This paper describes how content management systems (CMS) for mobile AR applications can be designed, in order to allow an effective and efficient management of context-dependent and personalized content. Because of the possibility to create individual AR contents for each user, with context-dependent and personalized content management an improved effect on advertising and marketing activities can be achieved. The architecture of the described system additionally considers important system characteristics, which lead to high user acceptance and high customer value.

I. INTRODUCTION

Augmented reality (AR) applications combine digital elements with the view of the reality and displays them together e.g. on a smartphone display. The basis for this combination is the captured camera video stream of the real environment, which is displayed on a screen. The digital elements are added to this view as overlay elements. First mobile AR applications focused on displaying digital content based on the location of the device. In the last few years the development more and more focused on creating frameworks and tools in order to use computer vision technologies for creating AR applications. With these frameworks and tools it is possible to create a wide range of different applications. Especially in the field of marketing these developments allow to create applications to improve the advertising effect of marketing activities. As an example, such applications can be used at the point of sale to virtually try on items of clothing or to show additional information of a product. Other applications can provide the possibility to virtually visualize the product in different contexts or to enhance 2D print media or advertising spaces with 3D or live data overlays.

To provide the possibility to achieve a higher customer value with such AR applications, this paper deals with an approach how the content, which is shown as overlay, can be created and managed individually for different users.

At the beginning this paper describes the motivation of developing and designing such an approach. Afterwards the developed approach is described in detail.

II. MOTIVATION

The motivation of this work was to further increase the marketing effect of AR applications by adapting the content of the applications. The chosen approach to achieve this is to use a form of personalized communication with the consumers. Personalized communication or personalization of content means that offers, activities and information are adapted to the individual needs of the consumer [1]. With personalization several benefits for the user can be achieved. First, since only the content which fits to the user's needs is displayed, the recognition of the benefits can be increased. If promotions are found useful, a higher acceptance of these promotions can be achieved. [2] Second, by reducing the available information to the relevant information, an information overload can be prevented [3]. Third, a higher attention for the contents of the promotions can be achieved [2].

Before designing the solution how AR content can be managed in a personalized and context dependent way, analysis of currently available system on the market were conducted. These analysis mainly covered the AR authoring tools of Wikitude, Layar, Junaio, BuildAR, Point At Me and the Feedgeorge AR Plug-In for WordPress. The research included the analysis of the provided features and where possible the technical implementation. The result showed, that the existing AR authoring tools all provide the possibility to simply create and maintain AR content without the need of having deep technical knowledge. But it also showed, that these systems do not provide the possibility to define individual contents for different users. This result reinforces our main motivation, to design an overall system architecture to combine customer relationship management (CRM)

with content management (CM) systems for content authoring for AR applications and to show which benefits can be gained out of it.

An additional motivation was the wide range of possible use cases. Especially in the field of mobile marketing several use cases can be considered. An example could be to show additional information to a scanned product and to filter these information based on the user's interests. E.g. more details if users are interested in such kind of products or showing related products which fits to the interests of the user. Additionally specific offers can be provided to a selected group of users. Another use case could be to virtually enhance magazines or print media and adapt the shown information according to the age or gender of the user resp. the current date/time/place. In the context of the described solution the use case to enhance advertisements on billboards for tourists was addressed. This use case addresses tourists or foreigners who often do not know what events or exhibitions are present, which sights have special offers or which specific activities can be done and where. All the various offers can be advertised on billboards, posters or magazines but with only rare information and not filtered according to the individual needs and interests of different viewers. This can cause an information overload at the user's side. With the described system, the user (e.g. tourist) is able to scan the advertising space with a mobile device. As a result, the mobile device displays information of events and activities which fits to the user's interests. These information could include e.g. a music event for one user and for another user an exhibition matching his interests. Videos and 3D representations of the venue or event can be used to support the advertising effect. By interacting with the displayed content, the user can e.g. request further information or book tickets. The filtering of the displayed contents according to the user's needs can be seen as the main advantage. To use various multimedia elements for visualizing the information in a more appropriate way can be seen as an additional advantages.

III. AUGMENTED REALITY

Figure 1 shows a Reality-Virtuality Continuum which defines the three areas real environment, mixed reality and virtual environment. The areas at the endpoints only use real resp. virtual elements, whereas the mixed reality area combines real and virtual elements to visualize the user's environment. Dependent on the position on the axis, more real or more virtual elements are used. [4] AR uses more real elements and can therefore be defined as the extension of the reality by digitally adding virtual information to the view of the reality [5]. The combination of real and virtual elements can be imagined as follows. The user scans the surroundings with the camera of a mobile device. This captured video stream is shown on the display of this device. Additionally, digital elements are displayed, which fully or partly overlays the real objects from the video stream. A technological independent definition of AR describes AR as the combination of real and virtual information, which provides interactions in real time and registers the elements in 3D [6][7].

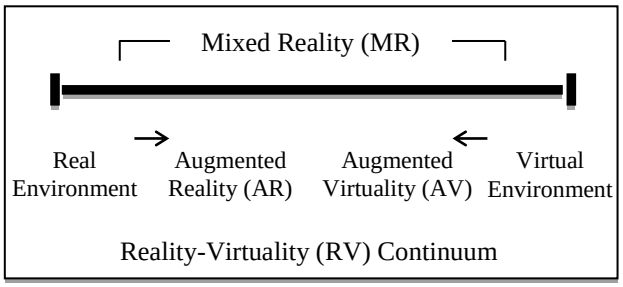


Figure 1. Cp. Simplified representation of a RV Continuum [4]

To determine which virtual information shall be shown as overlay elements, data from the reality, the real environment is used. To capture the reality either the location sensor of the device or computer vision technologies are used. With computer vision technologies specific markers on an image or the image itself can be recognized. [5][8] Based on the capured location, marker or image the virtual information can be filtered. To show the combined view, the display of mobile devices like smartphones or tablets, or head mounted displays can be used. Alternatively stationary devices with a camera and a display or devices projecting the virtual content directly onto the real object can be used. [7]

IV. CONTENT MANAGEMENT AND CUSTOMER RELATIONSHIP MANAGEMENT SYSTEMS

Content Management (CM) in general covers all necessary steps, which support the life cycle management of digital information. These steps include the creating, editing, publishing, storing, archiving and deleting of information. A CM system (CMS) assists the user when doing these steps. [9] Additionally to this requirements, a CMS must allow users with little or no programming knowledge to create and manage the content [10].

Customer Relationship Management (CRM) is the business strategy to combine the internal processes and functions with the external networks, in order to create and deliver value to the customer. To do this, high quality data about the customers, together with the information technology, is used. [11] CRM covers all steps to analyze, plan, do and check relationships with customers [12]. The purpose of CRM is the management of durable and profitable customer relationships [13].

In the area of marketing, the functionalities of CRM systems typically cover analysis and forecasting features. These are used to segment and categorize customers resp. to create forecasts for the market, purchase behavior and customer development. Additionally, this area includes features to support the campaign, territory and lead management. The campaign management deals with the planning, creation and evaluation of marketing campaign, whereas the territory management covers all features to organize marketing activities and campaigns according to specific regions. [14] In the context of this work these functionalities are combined with the features of a CMS in order to provide individual AR content to the users.

V. EXISTING SYSTEMS

To design and develop our system architecture, the architectures of existing systems were analyzed first. This

analysis mainly focused on their implementation and architectural view and mainly covered the content authoring systems from Wikitude, Layar, Junaio and BuildAR. Additionally the Vuforia platform were analyzed, which, compared to the other systems, only provides a framework for building applications and does not provide a proprietary solution to manage AR content. The analysis showed, that all systems provide the possibility to simply create AR content using a web-based user interface. Using these web applications, the user is able to upload reference images (view which shall be virtually enhanced) and to create and arrange digital overlay elements. To access this created AR content via a mobile device, proprietary mobile applications are provided by each of the manufacturers except BuildAR. BuildAR is also using the mobile application of Junaio. To create AR content and to access them with the mobile application the user doesn't need to have technical knowledge of the backend system, e.g. hosting a web service as data source for the mobile application. So this represents an easy solution for a wide range of users. Except these system characteristics, all analyzed systems are lacking of the possibility to created different AR views for the same reference image and presenting the most appropriate AR view to the user, based on the interests of the currently active user.

VI. OUR APPROACH

Focusing the addressed use case as described in section II, two different users need to be considered. The first one is the user of the backend system, which is further named as content author and represented by a marketing person who wants to create AR based commercials/advertisements. The second user is the user of the mobile application and in that use case represented by a tourist. Considering the whole scenario, this need to be split into two sub use cases, which are:

- Creating AR content for tourists with different interests or for different contexts individually (content author).
- Scanning billboards and getting individual AR content shown on a mobile device (tourist).

To provide individual AR content, the content author uploads the image of the view which shall be augmented to the backend system. This could be e.g. an image of a poster which is e.g. placed on a public billboard, in a magazine or in metro stations. If this image is uploaded, the content author creates several AR views for this single image. E.g. if there are currently several exhibition in the city the content author can create one AR view for tourists interested in arts and one for tourists more interested in technical areas. If the content author has created the different AR views, these views must be assigned to the different users/groups retrieved from the CRM system. With these steps the creation of the AR content is completed. In a second step, the tourist can scan the billboard, magazine page or poster with the mobile application and the mobile application presents the most appropriate AR view to the tourist. To do this, the tourist first needs to authenticate himself at the mobile application in order that the application knows him as the active user. Second, the mobile application requests the most appropriate AR view of the scanned billboard from the backend system in order to be able to presents this view to the user. The most appropriate AR view is one of the previously created AR views in the CMS and is selected based on matching user and contextual information. So different users, with different interests get different AR views presented. The criteria, which are used to assign AR views to users and to request the most appropriate AR view from the backend system covers the location of the device, the language which is selected on the mobile device, the gender, age and interests of the user as well as the current date, time and weekday. With these criteria a flexible assignment of AR views to different users in different contextual situations is enabled.

A. *Architecture*

Based on the analysis of the existing systems and the requirements of the use cases, the architecture of the system is designed as shown in the following figure.

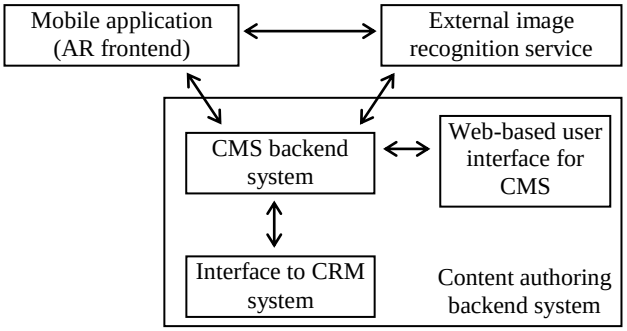


Figure 2. Overall system architecture

The overall system is basically created of three subsystems. These subsystems are the mobile AR frontend, the content authoring backend system and the backend system for image recognition. All three subsystems can be executed independently and are connected using standard web technologies. The mobile AR frontend is installed and executed on a mobile device, such as a smartphone or tablet. It is used for scanning billboards, requesting contents from the backend systems and displaying the contents for the user. Additionally this application can be used to register and authenticate users and to collect specific data about the user. The backend system for content authoring is separated into two parts. The first part is the CMS including the user interface which is used to create the AR views/content. An AR view is a collection of various virtual elements which belong together and are shown as one view to the user. Several AR views can be defined for the same billboard/image, in order to be able to create different views for different users. The second part of this backend system is the interface to the CRM system. This part is used to fetch all customer and campaign related data from the CRM system.

B. *Content authoring and image recognition backend systems*

This section describes the backend systems from a deeper technical point of view. As already mentioned before the backend systems consist of two parts, the

content authoring system and the image recognition service. The content authoring system is again divided into two parts, the CMS including the user interface and the interface to the CRM system. The CRM system itself is considered as a separate system, which is connected to the CMS by using a public interface. Therefore only the interface is drawn in the architectural overview. With this separation, the dependency between the systems can be kept as low as possible and the usage of a wide range of existing CRM systems can be enabled. Reusing existing CRM systems is important to avoid the necessity to manage customer related data multiple times (once within the existing systems and once within an AR CMS) and also because CRM systems with a huge amount of customer related data already exists in many companies. To additionally keep the mobile frontend independent from the CRM backend system, the mobile frontend only communicates with the CMS, which forwards requests to the CRM system if necessary. Requesting context-dependent and personalized content is entirely handled by the CMS.

The backend system for image recognition is used for two actions. Once to register reference images which shall be recognized by the mobile application and once to perform the recognition of the scanned billboards for the mobile application. To register reference images, the user need to create a new AR content within the CMS. The uploaded billboard image is automatically registered as a reference image at the image recognition service. To perform the image recognition for the mobile application, also the external system is used because the reference images belong to the created and active marketing campaigns and these images are changed continuously. If the image recognition would be done locally on the device, the user would have to update the mobile frontend always if the reference images are changed. To avoid this, the external service is used. Additionally the maximum amount of concurrently used reference images is limited to 100 images on several existing platforms, if the recognition is done locally in the application [15][16].

The backend systems and interfaces are implemented using standard technologies to ensure high accessibility and compatibility. The CMS is provided as web service with a browser based user interface. The interfaces between the mobile frontend and the CMS resp. the image recognition service and the CMS are implemented using the Representational State Transfer (REST) protocol.

C. Connection of CMS and CRM backend systems

Required information about users and active marketing campaigns are stored within the CRM backend system. To connect AR contents with users or user attributes, the active campaigns are used. With campaign activations, different marketing activities for different users, ages, genders, dates, times and regions can be defined. In the context of AR, the AR content/AR view represents an activity, which means that this AR content shall be shown on the mobile device if the campaign is considered as active. Since existing CRM systems shall be connectable to the CMS, the information of this assignment is stored within the CMS. The CRM system provides an interface, where the active campaigns and user data can be retrieved. If the CRM system does not provide an appropriate campaign management this can be handled by an intermediate layer between the CMS and the CRM interface. Then only the customer related data is requested from the CRM system and the campaign management is provided by the CMS user interface.

Figure 3 illustrates how the CRM and CM systems are connected using campaigns and campaign activations. Each reference image can have a list of AR views assigned. Each view can be linked with a campaign activation in order to be activated for e.g. various users or times. If the sample use case, as described in section II, is considered, this means that one reference image (the uploaded image of the poster) and two AR views (one for art-interested and one for technical-interested users) are created. Within the CRM system two marketing campaign activations are created (one for each AR view). As activation criteria once the user attributes related to arts and once the user attributes related to technical areas are set.

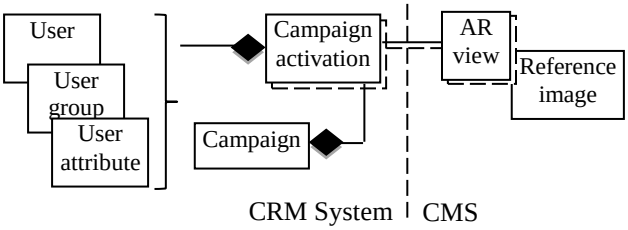


Figure 3. Connection CMS and CRM backend system

If the mobile application requests an AR view for a specific reference image, the most appropriate AR view is searched by comparing the attribute values which are set within the linked campaign activations and the attribute values which are set within the request. The following sections detailed describes how the search is done.

D. Requesting AR content

The following figure shows the communication sequence to request AR content from the backend system.

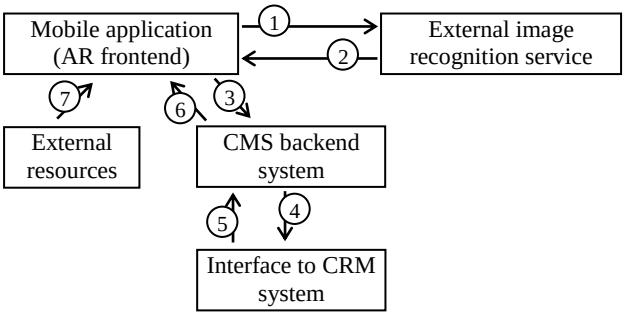


Figure 4. Communication for requesting AR content

To request AR content the mobile application first establishes a connection to the image recognition service and sends data from the captured video stream to this service (1). If the service recognizes a registered reference image within the received video stream data, the unique id of this reference image is returned to the mobile application (2). Afterwards this returned id is used, together with the user id and optionally the position and language of the mobile application, to create a HTTP GET

request to fetch the AR view from the web service of the CMS backend system (3). This request must be structured as shown in the following figure.

```
http://<base uri>/{ir_ref}/{usr_id}?lat={latitude}
                                   &lng={longitude}
                                   &alt={altitude}
                                   &lang={language}
```

Figure 5. Structure of a HTTP GET request to query AR content

Arguments which need to be set by the mobile frontend are encapsulated in brackets. Mandatory arguments are part of the URI. Optional arguments are defined as URI parameters. With the arguments `ir_ref` and `usr_id` the unique identifier returned from the image recognition service and the identifier of the currently authenticated user need to be passed. With the URI parameters `lat`, `lng` and `alt` the location of the device and with the URI parameter `lang` the present language on the mobile device can be passed. Considering the previously described sample use case only the mandatory arguments are required because the distinction between the two AR views only uses user related data which can be fetched using the user id. If the backend system receives a request, the most appropriate AR view is searched and returned. The search algorithm compares the values from the request, with the values which are set within the campaign activations. To do this, the user and campaign related information are retrieved from the CRM system (4), (5). The AR view which is assigned to the found campaign activation is returned to the mobile application (6). The values which are compared during the search cover a date range, time range and weekday selection to provide different content based on the time of day or the current weekday. Additionally the age and the gender of the user, the language and the location of the device are compared. Moreover, assigned user information are compared. These user information cover directly assigned users, user groups and user attributes. If one of these user settings match and all other compared attributes match, the campaign activation is considered as matching. If an optional URI parameter is not set, no filtering for this attribute is done and the comparison for this attribute is considered as matching. If several matching campaign activations are found, the list of matching campaign activations is ordered by a list of prioritized criteria. These criteria are:

- User is directly assigned
- At least one user attribute matches
- The user group, which is assigned to the user, is also assigned to the campaign activation
- The defined gender matches the gender of the user
- The age of the user is within the defined range

Considering the previously described sample use case the second criteria (user attributes) would be the matching one because user attributes are assigned to the campaign activations. If the content author creates a third AR view (e.g. special offer for users/tourists who often use the system) this AR view can be directly assigned to these users. Also if the user attributes match for e.g. the first AR view (interested in arts), the third AR view (special offer) is returned because directly assigned users have a higher priority. If no matching campaign activation is found, the first AR view for this reference image is returned in order to reduce the amount of returning empty views to a minimum. This approach is used, because returning a not perfectly matching content is still better, than returning an empty view or an error. Only if no contents are defined for this reference image, an error is returned. If a matching AR view is found, this is returned to the mobile frontend as a JSON formatted AR view description. If this description refers to any further external resources, these are separately downloaded by the mobile frontend (7).

VII. RESULTS

The present work shows an approach how AR content authoring tools can be designed in order to support the possibility to created different AR content for different users for the same reference image. The approach shows a flexible and extensible solution which can be either entirely implemented as described and also allows to exchange parts of the system with low effort because standard web technologies are used for inter-system communication. Additionally to the solution-design a prototype of the described system architecture was implemented in order to be able to evaluate the approach. The backend systems of the prototype is based on a Java web application running on a Tomcat web server combined with the Vuforia image recognition service. The mobile application is developed for Android and also uses Vuforia as AR engine. To further evaluate the designed system architecture, the prototype was used to carry out tests concerning response times and extensibility of the system. The following sections describes the results of the tests.

A. Response times

In order to be able to continuously update the content in the backend system without the necessity to always update the mobile application, the content as well as the image recognition service are provided by a web service. Therefore the response time of the application is related to the available speed of the mobile data connection. To evaluate that relationship, the time between the user starts scanning the environment and the AR view is displayed was measured using two different data connections. Once with a 3G connection with around 5 Mbit/s download and 3.9 Mbit/s upload. The second connection was a 2G connection with around 166 kbit/s download and 67 kbit/s upload. The tests showed, that the average response time with the fast connection was approximately 6 seconds, which can be seen as acceptable. Especially because the user interface of the mobile application changes from the image recognition view to the downloading view after about half of the time, which shows progress within the application. With the lower speed connection, the response time only for the image recognition was around 15 to 26 seconds, which can be seen as not acceptable. As a result it can be mentioned, that a high speed data connection for such a system architecture is required.

B. Extensibility

As part of this evaluation the extension of the contextual and personal criteria which are used to assign AR views to different users was considered. This showed, that criteria which addresses personal information can be easily added and removed, because these criteria are only known and handled by the content authoring backend system. The mobile application only sends the user id which is independent of the used personal criteria. The extension of contextual criteria which need to be collected by the mobile device, like the current position, is also possible but the deployment of such extension is more complex because the mobile application need to be updated as well. As a second part of this evaluation, the usage of the system with different mobile platforms were considered. Because standard web technologies are used for communication each mobile platform which supports HTTP communication and JSON or XML as data format and which is supported by the used AR engine framework can be used. Only the mobile application need to be implemented for the various platforms.

VIII. CONCLUSIONS

The described system provides the ability to create and distribute AR based commercials for different users individually. With this approach the customer needs can be fulfilled more accurately by presenting only relevant information to the user. Moreover, an information overload can be avoided. Both aspects results in a higher customer value and a higher acceptance of the system. The possibility to request further information by executing interactions with the displayed contents also results in a higher memorization and an enhanced customer loyalty [2]. On the backend system side, the complexity of the overall system can be reduced, because the linking between the AR content and campaign activations can be done by the same user interface. Low complexity is one key attribute to reach a high acceptance [17]. Additionally, personalized commercials can increase the click-through rate of an advertisement [18], which can be used as a rating for the effectiveness of the advertisement.

Because the image recognition is done remotely and the content is downloaded from a web service, the system always requires an active internet connection. This could probably prevent users from using the application, if they have no flat rate for the mobile data connection at the holiday destination. Additionally, a fast mobile data connection is required. This could prevent users from using the application in areas where the mobile data network is not well developed. Beside these two challenges there exists also another non-technical challenge for AR systems in general. This challenge focuses on the awareness for AR products in our society, which is still relatively low [19]. As a second non-technical challenge it will be necessary to convey the advantages and benefits to the user of the mobile AR applications in order to increase the usage rate.

REFERENCES

[1] K. Riemer, Personalisierung am Beispiel des Internet-Handels. In D. Ahlert, J. Becker, R. Knackstedt, & M. Wunderlich (Eds.), *Customer Relationship Management im Handel* (pp. 103-127). Berlin Heidelberg: Springer-Verlag, 2002.

[2] A. Kielholz, Online-Kommunikation: Die Psychologie der neuen Medien für die Berufspraxis. Heidelberg: Springer Medizin Verlag, 2008.

[3] M. Pfeffer, Erfolgspotenziale einer personalisierten Massenkommunikation über das Internet. Diplomarbeit. Norderstedt: Grin, 2002.

[4] P. Milgram, H. Takemura, A. Utsumi, and F. Kishino, Augmented reality: a class of displays on the reality-virtuality continuum. *Proc. SPIE 2351, Telemanipulator and Telepresence Technologies*, pp. 282-292, December 1995.

[5] C. Miller, *Augmented Reality ~ Enhancing Mobile. Whitepaper*. Hampshire: Juniper Research Ltd, 2012.

[6] R. T. Azuma, A Survey of Augmented Reality. *Presence: Teleoperators and Virtual Environments*(6), pp. 355-385, August 1997.

[7] G. Kipper, and J. Rampolla, Augmented Reality - An Emerging Technologies Guide to AR. USA: Elsevier, 2013.

[8] L. Madden, Professional Augmented Reality Browsers for Smartphones - Programming for Junaio, Layar, and Wikitude. Chichester: John Wiley & Sons, 2011.

[9] M. Eckert, Basiswissen SOA - BI - CRM - ECM: Grundlagen, Methoden, Praxis. München: IDG Business Media GmbH, 2011.

[10] N. Mehta, *Choosing an Open Source CMS.* Birmingham - Mumbai: Packt Publishing, 2009.

[11] F. Buttle, *Customer Relationship Management: Concepts and Technologies* (2. ed.). Oxford: Butterworth-Heinemann, 2009.

[12] G. Greve, Erfolgsfaktoren von Customer-Relationship-Management-Implementierungen. Wiesbaden: Deutscher Universitätsverlag, 2006.

[13] G. Raab, R. A. Ajami, V. B. Gargeya, and J. Goddard, *Customer Relationship Management: A Global Perspective.* UK: Gower Publishing Ltd, 2008.

[14] J. Schumacher, and M. Meyer, Customer Relationship Management Strukturiert Dargestellt: Prozesse, Systeme, Technologien. Berlin-Heidelberg: Springer-Verlag, 2004.

[15] Metaio, *Visual search | metaio Developer Portal*. Retrieved Juli 14, 2013, from http://dev.metaio.com/sdk/tutorials/visual-search, 2013

[16] Qualcomm. *Image Targets Workflow | Vuforia Developer Portal*. Retrieved Juli 24, 2013, from https://developer.vuforia.com/resources/dev-guide/developer-workflow-image-targets, 2013

[17] D. Sauerland, Zur Beziehung von Akzeptanz, Gerechtigkeit und Leistungsfähigkeit der sozialen Marktwirtschaft, Diskussionspapiere, Wirtschaftwissenschaftliche Fakultät, Universität Witten, Herdecke, No. 23/2011. Herdecke: Wirtschaftwissenschaftliche Fakultät, Universität Witten, 2012.

[18] M. Sonntag, *Untersuchungen zur Personalisierung*. Retrieved September 8, 2013, from: http://www.fim.uni-linz.ac.at/Publications/Aussendung10.98/Personalisierung.htm, 1998.

[19] C. Miller, *Press Release: Over 2.5 billion Mobile Augmented Reality Apps to Be Installed Per Annum by 2017*. Retrieved September 14, 2013, from: http://www.juniperresearch.com/viewpressrelease.php?id=450&pr=334, August 2012

Elektronički račun u Republici Hrvatskoj i SEPA inicijativa

Danijel Matić*, Boris Vrdoljak**, Damir Kalpić**, Alan Čičak *
* InfoDom d.o.o., Zagreb, Hrvatska
** Sveučilište u Zagrebu Fakultet elektrotehnike i računarstva, Zagreb, Hrvatska
danijel.matic@infodom.hr, boris.vrdoljak@fer.hr, damir.kalpic@fer.hr, alan.cicak@infodom.hr

Sažetak - Elektroničko poslovanje bavi se poslovnim procesima unutar područja cjelovitog dobavnog lanca. Plaćanje je završni dio procesa nabave kada kupac transferira dobavljaču novčana sredstva naznačena na računu. Elektronički nalog za plaćanje je nalog koji korisnik daje elektroničkim putem za prebacivanje sredstava sa svog računa na drugi račun i mora sadržavati sve potrebne podatke za plaćanje. SEPA (engl. Single Euro Payments Area), jest inicijativa Europske Unije s ciljem uspostave jedinstvenog sustava platnog prometa do 2016. godine. SEPA definira pravila po kojima se odvijaju transakcije. Nacionalni odbor za platni promet (NOPP) je 2013. godine donio Odluku o osnivanju radnih tijela zaduženih za implementaciju SEPA projekta u RH, odnosno o osnivanju Hrvatskog odbora za SEPA koordinaciju (HOSK-a) i Hrvatskog SEPA foruma (HSF-a). U radu se analiziraju i opisuju procesi plaćanja elektroničkih računa te se predlaže način pozicioniranja informacijskog posrednika u procesima plaćanja prema inicijativi SEPA.

U ovom radu neće se zadirati u nadležnosti navedenih tijela već će koristeći javno dostupne informacije za analizu i pripremu preslikavanja e-računa po inicijativi SEPA.

I. UVOD

Elektroničko poslovanje je oblik poslovanja koji prilikom transakcija i razmjena informacija koristi informacijske i komunikacijske tehnologije između sudionika u poslovnom procesu. Cilj elektroničkog poslovanja jest automatizirati poslovne procese čime se smanjuju troškovi i povećava efikasnost. Elektroničko poslovanje pokriva poslovne procese cijelog dobavnog lanca, a plaćanje je završni dio procesa nabave kada kupac dobavljaču prenosi novčana sredstva naznačena na računu.

Europski gospodarski prostor nastanjuje više od 500 milijuna stanovnika i godišnje se provede više od 71,5 milijardi platnih transakcija. Dio tog prostora od 1. srpnja 2013. jest i Republika Hrvatska.

The Single Euro Payments Area (SEPA), jedinstveno područje plaćanja u eurima, je nastala kao inicijativa europske bankarske industrije s ciljem osiguravanja plaćanja u eurima u Europskom gospodarskom prostoru.

Cilj SEPA-e je omogućiti potrošačima, poslovnim subjektima i tijelima javne vlasti izvršavanje i primanje plaćanja u eurima pod jednakim uvjetima, pravima i obvezama, neovisno o njihovoj lokaciji.

Pristupanjem Europskoj uniji implementacija SEPA-e je obvezujuća i za Republiku Hrvatsku. U tu svrhu, na inicijativu Nacionalnog odbora za platni promet, pokrenut je nacionalni projekt SEPA. Glavno upravljačko tijelo projekta jest Nacionalni odbor za platni promet (NOPP) u čijem radu sudjeluje Hrvatska narodna banka, Ministarstvo financija, Hrvatska gospodarska komora (Sektor za bankarstvo i druge financijske institucije) i Hrvatska udruga banaka.

Nadležnosti nad provedbom SEPA projekta u Republici Hrvatskoj su kako slijedi:

- Nacionalni odbor za platni promet (NOPP) je na sjednici 21.5.2013. donio Odluku o osnivanju radnih tijela zaduženih za implementaciju SEPA projekta u RH, odnosno o osnivanju Hrvatskog odbora za SEPA koordinaciju (HOSK-a) i Hrvatskog SEPA foruma (HSF-a).
- Članovi HOSK-a su predstavnici Hrvatske udruge banaka (dalje HUB-a), HNB-a; Ministarstva financija, Fine i HGK-a; njime predsjedava HUB, a osnovna zadaća HOSK-a je: izrada nacionalnog plana migracije na SEPA-u, izrada i provedba nacionalnih SEPA platnih shema, tj. kreditnog transfera (SCT) i izravnog terećenja (SDD) u eurima i kunama, izrada i provedba rješenja potrebnih za izvršenje kreditnih transfera i izravnih terećenja u eurima.
- HOSK je osnovao radne skupine (RS) za kreditne transfere i izravna terećenja koje imaju zadatak izradu praktičnih rješenja radi ispunjenja gore navedenih zadataka nominiranih HOSK-u. Budući da radne skupine za SCT i SDD čine predstavnici svih banaka, HOSK je osnovao i operativne grupe za SCT i SDD koje imaju za zadatak razradu formata svih poruka klijent-banka i razmjene banka-banka kao i definiranja poslovnih pravila vezano za navedene razmjene.
- Članovi HSF-a su također predstavnici HUB-a, HNB-a; Ministarstva Financija, Fine i HGK-a; njime predsjedava HGK, a osnovna zadaća HSF-a je promoviranje SEPA platnih usluga krajnjim korisnicima s posebnim osvrtom na uključivanje IT kuća i velikih izdavatelja u implementaciju SEPA platnih usluga u RH odnosno informiranje javnosti o SEPA projektu, migraciji, formatima i promjenama koje će poslovni subjekti morati ugraditi u svoje poslovne sustave migracijom RH na SEPA platne instrumente odnosno do krajnjeg roka navedenog za 31.10.2016.

Ministarstvo gospodarstva, poduzetništva i rada RH osnovalo je 2008. godine Povjerenstvo za e-račun. Ovo povjerenstvo je 2009. godine predložilo informacijski model za e-račun [14] i korištenje norme Universal Business Language (UBL) 2.0 [2]. Krajem 2013 godine objavljena je nova verzija standarda UBL 2.1 [1] kao unaprjeđenje standarda UBL 2.0.

U sklopu projekta e-račun, Radna skupina HUB-a je prema HUB1 obrascu izradila elektronički obrazac e-HUB u XML formatu. Navedeni obrazac namijenjen je sudionicima platnog prometa koji su svoje poslovanje dogovorili i proveli kroz e-poslovanje.

U trenutku kada nadležne institucije pripreme i definiraju potrebne komponente za implementaciju SEPA inicijative u Republici Hrvatskoj bit će potrebno uspostaviti jedinstveni proces, tj. povezati proces izdavanja i obrade elektroničkog računa s novim procesom plaćanja. Trenutno dostupne norme i dokumenti vezani uz inicijativu SEPA omogućuju pripremu potrebnih elemenata s ciljem osiguranja spremnosti zainteresiranih strana u procesima vezanim za elektroničko plaćanje, tj. elektronički račun.

Zainteresirane stane, naročito informacijski posrednici (Informacijski posrednik je fizička ili pravna osoba koja u ime drugih obavlja otpremu, prijem i prijenos elektroničkih dokumenata.) morat će prilagoditi informacijske sustave i podržati nove načine provedbe i iniciranja elektroničkog plaćanja.

II. PREGLED NORMI, PRAVNOG OKVIRA I SMJERNICA

Cilj elektroničkog poslovanja je automatiziranje poslovnih procesa, a preduvjet za to je definiranje formata elektroničkih dokumenata i poruka. Format i dozvoljeni sadržaj elektroničkih dokumenata definiraju se normama. Za e-račun kao najrašireniji elektronički dokument koristi se paralelno veći broj normi.

U Republici Hrvatskoj preporučeno je korištenje norme Universal Business Language (UBL) 2.0 koja je implementirana u više europskih država i u mnogim projektima u području e-poslovanja, kao što je npr. sustav za javnu nabavu PEPPOL (Pan-European Public Procurement Online), projekt u koji je uključen velik broj država članica Europske Unije.

Osim razmjene e-računa, cjelokupni dobavni lanac uključuje razmjenu mnogih drugih vrsta poslovnih dokumenata, kao što su npr. narudžbenice. Ove vrste poslovnih dokumenata logički su povezane te se podrazumijeva da će račun automatski preuzimati velik dio podataka iz narudžbenice, ako ona postoji. Za e-narudžbenicu i ostale dokumente koji se razmjenjuju unutar dobavnog lanca (otpremnica, primka i slično) je stoga preporučljivo koristiti istu normu kao i za e-račun.

Norma Universal Business Language (UBL) se temelji na konceptualnom modelu informacijskih komponenti koje predstavljaju podatkovne elemente poslovnih dokumenata. Te podatkovne komponente spajaju se u dokumente kao što su narudžba ili račun koji se prema pravilima UBL Naming and Design Rules transformiraju u XML sheme.

Norma UBL 2.0, za koju se Povjerenstvo za e-Račun odlučilo kao podlogu za izgradnju dokumenata koji se razmjenjuju u postupku naručivanja i postupku fakturiranja, sadrži niz XML shema za 31 poslovnu poruku. Od studenoga 2013. godine za UBL je postala važeća verzija 2.1 koja ima 64 vrste poslovnih poruka. Budući da verzija UBL 2.1 predstavlja nadogradnju verzije 2.0, verzija 2.1 neformalno se smatra preporučenom normom za format e-računa u Republici Hrvatskoj.

E-plaćanje predstavlja posljednji korak u automatizaciji dobavnog lanca. Na temelju ispostavljenog računa, na strani kupca se generira nalog za plaćanje. Tim dokumentom inicira se obavljanje novčane transakcije gdje kupac dobavljaču transferira novčana sredstva naznačena na računu.

SEPA kao jedinstveni prostor plaćanja u eurima omogućuje da se plaćanja u eurima, bilo u zemlji ili inozemstvu, izvršavaju prema jednakim pravilima, neovisno o tome u kojoj se državi nalaze pojedini sudionici procesa. Pristupanjem Europskoj uniji, Hrvatska je dužna prilagoditi se SEPA platnoj shemi za euro platne transakcije.

Propis (EU) No 260/2012 Europskog parlamenta i Savjeta od 14. ožujka 2012. [15] [16] postavlja tehničke i poslovne zahtjeve za prijenose odobrenja i trajnih naloga u eurima

- identifikator računa uplatitelja mora biti IBAN
- norma za format poruka mora biti ISO 20022.

Financijske institucije definirale su normu ISO 20022 za razmjenu različitih vrsta poruka u formatu XML. Poruke ISO 20022 [9] [10] [11] [12] [13] pokrivaju kompletni lanac plaćanja: od kupaca prema bankama, od banaka prema bankama te za izvješćivanje; koriste se i za naloge investicijskih fondova, izvješća o cijenama i slično.

III. PROCESI PLAĆANJA I ELEKTRONIČKO POSLOVANJE

Elektroničko poslovanje pokriva poslovne procese cijelog dobavnog lanca, a plaćanje je završni dio procesa nabave kada kupac dobavljaču prenosi novčana sredstva naznačena na izdanom računu. Dobavljač ispostavom računa s kupcem stvara dužničko-vjerovnički odnos temeljem kojeg se provodi financijska transakcija.

Procesi plaćanja po SEPA inicijativi definiraju sljedeće načine plaćanja:

- Kreditni transfer (Credit transfer).
- Izravno terećenje (Direct Debit).
- Plaćanje karticama.

U radu se razmatra kreditni transfer i izravno terećenje, dok se plaćanje karticama neće razmatrati.

A. Pozicioniranje informacijskog posrednika u procesu plaćanja

Poslovni partneri mogu u međusobnom poslovanju ugovoriti elektroničku razmjenu poslovnih dokumenata. Bilateralno ugovaranje je relativno jednostavno kada je

poslovno okruženje jednostavno u smislu postojanja nekolicine partnera s kojim je uspostavljen neki oblik poslovnog odnosa. Međutim, u trenutku kada se broj poslovnih partnera značajno poveća, ugovaranje i razmjena elektroničkih dokumenata postaje vremenski intenzivno i kompleksno. Jedno od rješenja je poslove elektroničke razmjene dokumenata prepustiti vanjskom izvršitelju, tj. informacijskom posredniku.

Korištenje informacijskih posrednika po modelu 4-strane kao u [3] omogućuje ugovornim stranama prepuštanje ugovaranja i realizacije razmjene elektroničkih poslovnih dokumenata trećima, tj. informacijskim posrednicima. Svaka poslovna, ugovorna strana sklapa sporazum s informacijskim posrednikom po vlastitom izboru (ili s istim ili s različitim).

Informacijski posrednici osim razmjene elektroničkih dokumenta, obavljaju i funkcije elektroničke arhive, konverzije formata, digitalnog potpisivanja i označavanje dokumenta elektroničkim žigom. Nadalje, temeljem elektroničkih računa mogu i trebaju generirati instrukcije za provedbu plaćanja po različitim standardima.

Glavni sudionici procesa prema inicijativi SEPA za kreditni transfer [5] dijele se po nekoliko interesnih domena:

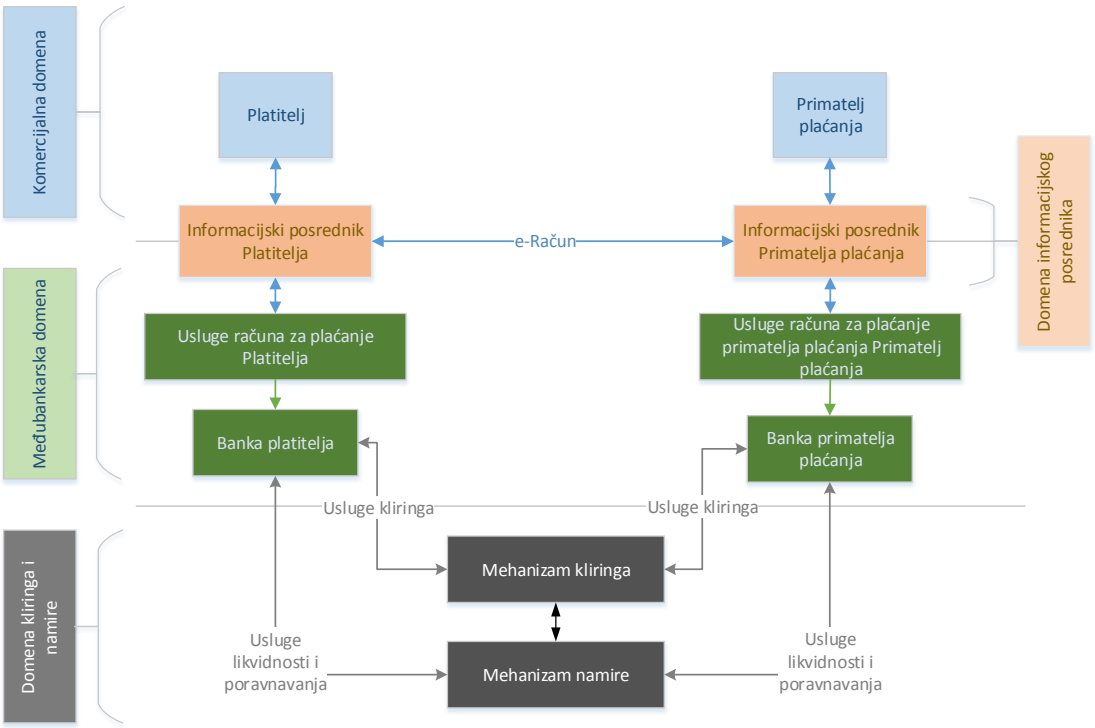
- Komercijalna domena (Commercial Space) – obuhvaća poslovne partnere generiraju financijske obaveze isporukom usluga i roba.
- Međubankarska domene (Inter-bank Space) – obuhvaća banke i potrebne servise za provedbu plaćanja i komunikaciju prema korisnicima.
- Domena kliringa i namire (Clearing & Settlement Space) –domena pokriva neophodne sudionike, procese i mehanizme za potrebe prebacivanja sredstava između banaka. Obuhvaća Clearing and Settlement Mechanism (nadalje CMS) čija je svrha pouzdana i sigurna razmjena informacija u procesima plaćanja.

Sukladno modelu 4-strane, informacijski posrednik preuzima dijelove poslovnih procesa u kojima se radi razmjena poslovnih dokumenata, u kojima je e-račun dokument kojim se stvara dužničko/vjerovnički odnos, a time i predstavlja podlogu za generiranje nalog za plaćanje. Informacijski posrednik temeljem zaprimljenog elektroničkog računa može generirati nalog za plaćanje i u ime primatelja računa pokrenuti plaćanje prema poslovnoj banci. Stoga pozicioniranje informacijskog posrednika je u procesima plaćanja između platitelja i njihovih poslovnih banaka. Na slici 2. prikazano je pozicioniranje sudionika u procesu plaćanja elektroničkog računa.

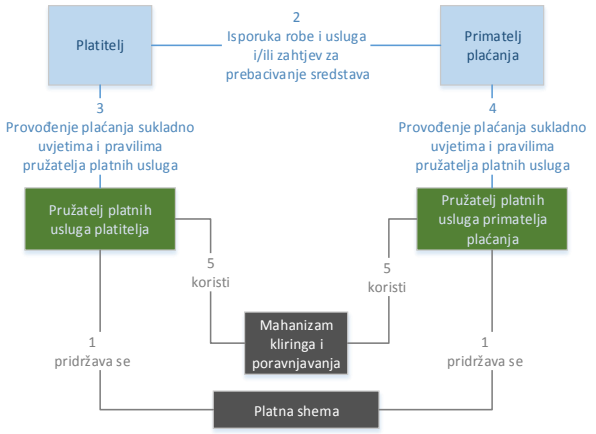
Pružatelj platnih usluga (PPU) platitelja, tj. banka u kojoj platitelj ima otvoren račun za plaćanje, mora imati sposobnost provođenja platnih transakcija prema pružatelju platnih primatelja plaćanja.

B. Proces plaćanja: Kreditni transfer

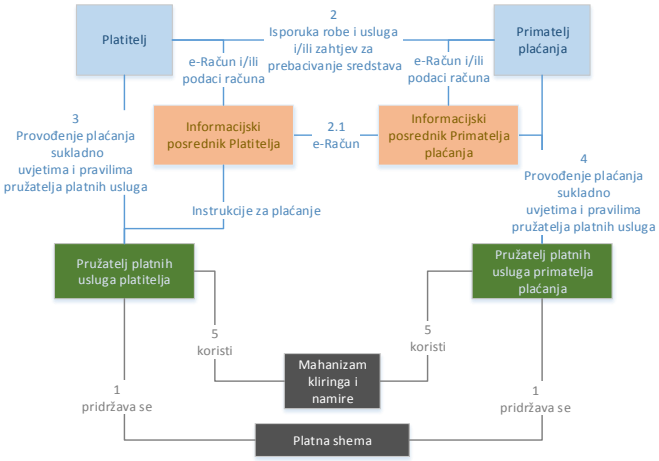
Kreditni transfer je nacionalna ili prekogranična platna usluga kojom se račun za plaćanje primatelja plaćanja odobrava za platnu transakciju ili niz platnih transakcija na teret platiteljevog računa za plaćanje od strane PPU-a kod kojeg se vodi platiteljev račun za plaćanje, na osnovi instrukcije koju daje platitelj [15] [16]. Poslovni odnos između platitelja i primatelja plaćanja je temelj kojim



Slika1. Pregled i pozicioniranje sudionika u procesu plaćanja elektroničkog poslovanja za kreditni transfer [5]



Slika 2. Kreditni transfer po modelu 4-strane [5]



Slika 3. Pozicioniranje informacijskog posrednika u virmanskom plaćanju po modelu 4-ugla

platitelj pokreće plaćanje putem izdavanja naloga za plaćanje svom pružatelju platnih usluga (engl. payment service provider, PSP). Platiteljev PPU prebacuje sredstva primateljevom PPU.

Uloge u procesu virmansko plaćanja [6]:

- Platitelj – fizička ili pravna osoba koja ima račun za plaćanje i koja da suglasnost za izvršenje nalog za plaćanje s tog računa.
- Pružatelj platnih usluga (PPU) platitelja– pružatelj platnih usluga platitelja, banka, koja zaprima naputke za plaćanje sukladno kojim u ime platitelja prebacuje sredstva u banku primatelja plaćanja, a u korist njegova bankovnog računa.
- Pružatelj platnih usluga (PPU) primatelja plaćanja – je sudionik procesa koji zaprima naputke za provođenje kreditnog transfera od PPU-a platitelja i prebacuje sredstva na bankovni račun primatelja plaćanja.
- Primatelj plaćanja - fizička ili pravna osoba koja ima račun za plaćanje i za koju su namijenjena novčana sredstva koja su predmet platne transakcije.

Isporukom roba ili usluga od strane dobavljača stvara se preduvjet za izdavanje i dostavu računa kupcu. Isporukom računa, stvara se dužničko-vjerovnički odnos između partnera sukladno [17].

U elektroničkom poslovanju dostavu računa, u ime dobavljača radi njegov informacijski posrednik. Informacijski posrednik od dobavljača može preuzeti ili već pripremljeni e-račun ili ga generirati iz dostavljenih podataka. Informacijski posrednik dobavljača dostavlja e-račun informacijskom posredniku kupca. Informacijski posrednik kupca, će zavisno od poslova koje za kupaca obavlja u kontekstu elektroničkog poslovanja, provjeriti ga (sintaksno i semantički), pohraniti u elektronsku arhivu, te dostaviti kupcu u njegov poslovni sustav.

Zaprimanjem e-Računa od strane kupca nastaje dužničko-vjerovnički odnos, kao i u klasičnom poslovanju, koji je temeljem za iniciranje plaćanja. Inicijator plaćanja je platitelj, tj. kupac. Informacijski posrednik, u ime platitelja, generira naloga za plaćanje i dostavlja ga PPU platitelja. PPU platitelja izvršava platnu transakciju prema PPU-u primatelja plaćanja. Nakon provedene platne transakcije PPU platitelja dostavlja informacijskom posredniku platitelja povratne informacije o statusu izvršenja platne transakcije. Informacijski posrednik platitelja temeljem zaprimanja statusa izvršenja transakcije informaciju dostavlja u poslovni sustav platitelja.

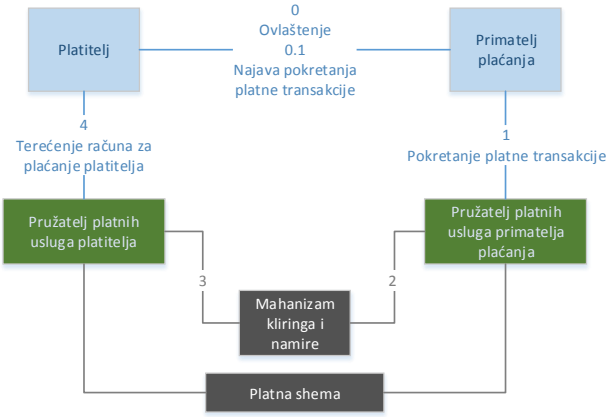
Informacijski posrednik primatelja plaćanja može zaprimiti informaciju o provedenoj platnoj transakciji. Poslovni procesi koji se pokreću zaprimanjem informacije o statusu provođenja platne transakcije ovise o ugovorenom odnosu između primatelja plaćanja i njegovog informacijskog posrednika i nisu predmet razmatranja u ovome radu.

C. Proces plaćanja: Izravno terećenje

Izravno terećenje je nacionalna ili prekogranična platna usluga za terećenje platiteljevog računa za plaćanje, kada platnu transakciju inicira primatelj plaćanja na osnovi suglasnosti platitelja [15] [16]. Platitelj daje ovlaštenje (engl. mandate) primatelju plaćanja za provedbu platne transakcije izravnim terećenjem. Primatelj plaćanja pokreće platnu transakciju izdavanjem naloga za plaćanje svom PPU. Izravno terećenje koristi se često za provođenje platnih transakcija za usluge koje se periodično isporučuju, npr. plaćanje komunalnih usluga. Također se koristiti i za jednokratne isplate. U prvom slučaju zahtijevaju ovlaštenje dok u drugom slučaju autorizaciju od strane platitelja.

Uloge u procesu plaćanje izravnim terećenjem su [8]:

- Primatelj plaćanja - prima i pohranjuje ovlaštenje za pokretanje izvršavanja izravnih terećenja. Temeljem pohranjenih ovaštenja vjerovnik može inicirati platne transakcije izravnim terećenjem.
- Pružatelj platnih usluga (PPU) primatelja plaćanja - je pružatelj platnih usluga u kojem primatelj plaćanja ima otvoren račun za plaćanje. Temeljem sporazuma s primateljem plaćanja (vjerovnikom)



Slika 4. Izravno terećenje model 4-strane [8]

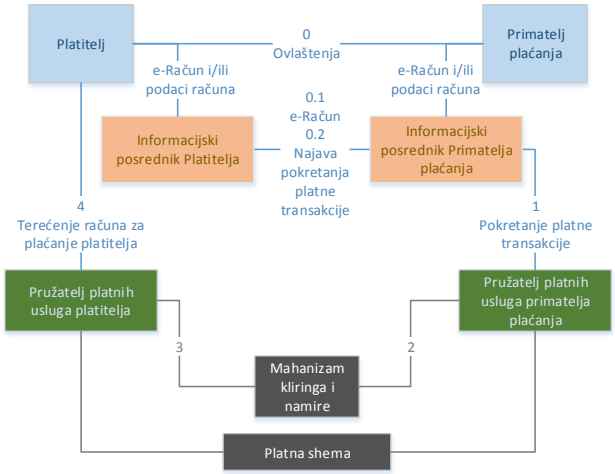
prima i izvršava nalog za plaćanje. Pokreće transakciju izravnog terećenja sukladno nalogu za plaćanje prema PPU platitelja.

- Pružatelj platnih usluga (PPU) platitelja – je pružatelj platnih usluga u kojem platitelj ima otvoren račun za plaćanje. Temeljem sporazuma s platiteljem pokreće terećenje računa platitelja, prebacivanje sredstava, koje je pokrenuto procesom izravnog terećenja od strane primatelja plaćanja.
- Platitelj – Njegov račun za plaćanje se tereti temeljem naloga za plaćanje poslanih od strane primatelja plaćanja. Izdaje suglasnost primatelju plaćanja za pokretanje transakcije izravnog terećenja.

Koraci koji prethode stvaranju dužničko vjerovničkog odnosa su isti kao kod kreditnog transfera. Zaprimanjem računa od strane kupca nastaje dužničko vjerovnički odnos temeljem kojeg se inicira platna transakcija. Inicijator platne transakcije je u ovome slučaju primatelj plaćanja, tj. dobavljač.

Prije pokretanja transakcije izravnog terećenja platitelj mora ovlastiti primatelja plaćanja za izvršavanje izravnih terećenja. Ovlaštenje sadržava sve potrebne informacije koje primatelju plaćanja daju ovlasti za izvršavanje izravnog terećenja. Informacijski posrednik primatelja plaćanja nakon slanja e-računa šalje najavu pokretanja transakcije izravnog terećenja platitelju, odnosno informacijskom posredniku platitelja. SEPA definira slanje najava za plaćanje barem 14 kalendarskih dana prije dana dospijeća [8]. Ovaj rok može biti i kraći ako je isti ugovoren od strane platitelja i primatelja plaćanja. Primatelj plaćanja može poslati informacije u PPU platitelja o iznosu sredstava koje je potrebno prebaciti transakcijom izravnog terećenja. Prijenos informacija u PPU platitelja se ne smije dogoditi prije gore definiranog roka za dostavu najave za plaćanje platitelju. Informacijski posrednik u ime primatelja plaćanja generira nalog za plaćanje, čime pokreće transakciju izravnog terećenja. Informacijski posrednik primatelja plaćanja šalje nalog za plaćanje u PPU primatelja plaćanja. PPU primatelja plaćanja pokreće platnu transakciju izravnog terećenja prema PPU platitelja. Nakon provođenja platne transkacije

Ovdje identificirajte konkretne sponzore. Ako nema sponzora, izbrišite ovaj tekstualni okvir. (sponzori)



Slika 5. Pozicioniranje informacijskog posrednika u modelu 4-strane izravnog terećenja

izravnog terećenja, PPU primatelja plaćanja dostavlja informacijskom posredniku primatelja plaćanja povratne informacije o statusu transakcije. Informacijski posrednik primatelja plaćanja dostavlja informaciju o statutu provedene transakcije u poslovni sustav primatelja plaćanja.

IV. PRESLIKAVANJE UBL 2.1. ELEMENATA NA ISO 20022 XML PAIN PORUKE

U prethodnom poglavlju opisani su procesi plaćanja koje inicira informacijski posrednik.

Da bi se uspješno proveli procesi plaćanja elektroničkim putem, potrebno je uspostaviti semantičku interoperabilnost između normi koje se koriste s jedne stanje u razmjeni elektroničkih računa, a s druge strane u sustavima koji provede prebacivanje financijskih sredstava. Sadržaj elektroničkog računa omogućuje generiranje instrukcija za plaćanje bez obzira da li se radi o kreditnom transfer ili izravnom terećenju.

U Republici Hrvatskoj norma UBL 2.1 [1] neformalno se smatra preporučenom normom za format elektroničkog računa, a SEPA preporučuje korištenje norme ISO 20022 za implementaciju procesa plaćanja.

U budućem radu potrebno je napraviti preslikavanje određenih elemenata između normi UBL 2.1 i ISO 20022 a sukladno s [4] [7] [8].

V. MOGUĆA POBOLJŠANJA I BUDUĆI RAD

Potrebno će biti definirati sheme temeljem koje će se informacije o modelu plaćanja i pozivu na broj prenositi kao jedinstveni podatak prilikom razmjene poruka u procesima ne samo plaćanja nego i razmjene e-računa.

Sa stanovišta primjene normi objavljene su nove verzije ISO 20022 pain poruka [9] [10] [11] [12] [13] pa je neophodno donošenje naputke od strane nadležnih radnih tijela za provedbu SEPA projekta u vezi korištenja istih, te primjena tih naputaka na iznesena preslikavanja.

Nastojati će se inicirati aktivnosti kojima je nositelj Ministarstvo gospodarstva Republike Hrvatske, a vezane za elektroničko poslovanje i elektronički račun radi objave

službenog dokumenta na temu preslikavanja elementa e-računa i elemenata poruke za inicijalizaciju plaćanja sukladno inicijativi SEPA.

VI. ZAKLJUČAK

Realizacijom inicijative SEPA u Republici Hrvatskoj elektroničko plaćanje u eurima će se izvoditi na isti način na koji se izvodi na području cijele Europske unije. Bankarski sektor će implementirati neophodne protokole i poruke za provođenje procesa plaćanja prema inicijativi SEPA.

Iz perspektive e-računa kupci i dobavljači će inicijalizaciju plaćanja prema bankama prepustiti informacijskim posrednicima kao i obradu povratih informacija od strane banaka.

Plaćanja u nacionalnim valutama će se vjerojatno, nakon realizacije SEPA-e, neko vrijeme provoditi postojećim nacionalnim protokolima. Za očekivati je da će bankarski sektor i za nacionalna plaćanja uvesti iste procese, poruke i sustave kao i za plaćanja u eurima zbog smanjenja troškova održavanja paralelnih sustava kao i zbog pojednostavnjenja sučelja prema entitetima iz komercijalne domene (kupci i dobavljači) ali i iz međubankarske domene (banke). Informacijski posrednici implementirat će potrebne protokole i poruke za komunikaciju prema bankama sukladno inicijativi SEPA i preporukama Nacionalnog odbora za platni promet, Hrvatskog odbora za SEPA koordinaciju i Hrvatskog SEPA foruma. Interoperabilnost na svim razinama (pravnoj, organizacijskoj, semantičkoj i tehničkoj) će se uspostaviti među svim sudionicima elektroničkog poslovanja cijelog dobavnog lanca.

Jasna definicija i usvajanje smjernica za postizanje organizacijske, a koja uključuje procesnu, i semantičke interoperabilnosti prikazanih u ovome radu kao de facto standard za inicijalizaciju plaćanja u procesima vezanim za e-račun u Republici Hrvatskoj će značajno ubrzati tranziciju od klasičnog prema elektroničkom poslovanju na svim razinama društva.

ZAHVALE

Izražavamo našu iskrenu zahvalnost direktoru Hrvatske udruge banaka dr. sc. Zoranu Bohačeku te gospođi Sandri Ruso, savjetnici u HUB-u. Svojim konstruktivnim savjetima pomogli su nam da ispravimo neke terminološke i faktografske pogreške te da oblikujemo članak za koji se nadamo da može barem malo pomoći u unaprjeđenju elektroničkog poslovanja u Hrvatskoj.

LITERATURA

[1] OASIS Universal Business Language Technical Committee, "Universal Business Language Version 2.1", OASIS (Organization for the Advancement of Structured Information Standards): OASIS Standard, http://docs.oasis-open.org/ubl/os-UBL-2.1/UBL-2.1.html, 04 January 2014

[2] OASIS Universal Business Language Technical Committee, "Universal Business Language Version 2.0", OASIS (Organization for the Advancement of Structured Information Standards): OASIS Standard, http://docs.oasis-open.org/ubl/cs-UBL-2.0/UBL-2.0.html, 12 January 2014

[3] CEN Workshops on eInvoicing, "CEN CWA 16050: Framework for the emerging network infrastructure of eInvoicing service providers throughout Europe", CEN (European committee for standardization): CEN CWA 16050, ftp://ftp.cen.eu/PUBLIC/CWAs/eInV2/CWA%2016050.pdf , 12 January 2014

[4] CEN Workshop on Business Interoperability Interfaces on public procurement in Europe Phase 3, "CWA 16073-2: Business Interoperability Interfaces for Public procurement in Europe - Part 2: Convergence and gap analyses", CEN (European committee for standardization): CEN CWA 16073-2, January 2010, ftp://ftp.cenorm.be/PUBLIC/CWAs/BII/Final/2010_16073_2.pdf , Annex 2.1 BII-ISO 2022 mapping", 12 January 2014

[5] European Payments Council Plenary, "SEPA Credit transfer scheme Customer-to-Bank Implementation guidelines version 7.0", European Payments Council, 20 November 2012 , http://www.europeanpaymentscouncil.eu/knowledge_bank_download.cfm?file=EPC132%2D08%20C2B%20CTIG%20V7%2E0%20Approved%2Epdf , 12 January 2014

[6] European Payments Council Plenary, " SEPA Credit Transfer Rulebook Version 7.1", European Payments Council, 27 January 2014, http://www.europeanpaymentscouncil.eu/knowledge_bank_download.cfm?file=EPC125%2D05%20SCT%20RB%20v7%2E1%20Approved%2Epdf , 12 January 2014

[7] European Payments Council Plenary, " SEPA Direct Debit Core Scheme Customer-to-Bank Implementation Guidelines Version 7.0", European Payments Council, 30 November 2012 , http://www.europeanpaymentscouncil.eu/knowledge_bank_download.cfm?file=EPC130%2D08%20SDD%20Core%20C2B%20IG%20V7%2E0%20Approved%2Epdf , 12 January 2014

[8] European Payments Council Plenary, "SEPA Direct Debit Core Rulebook Version 7.1", European Payments Council, 27 January 2014, http://www.europeanpaymentscouncil.eu/knowledge_bank_download.cfm?file=EPC016%2D06%20Core%20SDD%20RB%20v7%2E1%20Approved%2Epdf , 12 January 2014

[9] ISO (International Organization for Standardization), "Customer-to-Bank Message Usage Guide", ISO 20022, January 2009

[10] ISO (International Organization for Standardization), "Payments - Maintenance 2009", ISO 20022, March 2009, International Organization for Standardization, "Payments - Maintenance 2009", ISO 20022, March 2009

[11] ISO (International Organization for Standardization), "ISO 20022 Payments messages", ISO 20022 Standards. http://www.iso20022.org/payments_messages.page , 12 January 2014

[12] ISO (International Organization for Standardization), "Payments Initiation May 2013 Message Definition Report - Part 1", ISO 20022, January 2013, http://www.iso20022.org/documents/general/pain_PaymentsMaintenance2013.zip , 12 January 2014

[13] ISO (International Organization for Standardization), "Payments Initiation May 2013 Message Definition Report - Part 2", ISO 20022, January 2013, http://www.iso20022.org/documents/general/pain_PaymentsMaintenance2013.zip, 12 January 2014

[14] Povjerenstvo za e-Račun, "Specifikacija elektroničkog računa nacrt Verzija 0.1.",Ministarstvo gospodarstva, rada i poduzetništva Republike Hrvatske, Prosinac 2019, http://www.mingo.hr/userdocsimages/trgovina/Specifikacija%20e-Računa%202009-12-23.pdf, 12 January 2014

[15] EUROPSKI PARLAMENT I VIJEĆE EUROPSKE UNIJE, "Uredba (EU) br. 260/2012 Europskog parlamenta i Vijeća ", Hrvatska narodna banka: neslužbeni prijevod, 14. ožujka 2012, http://www.hnb.hr/platni-promet/propisi_eu_reference/h-uredba-260_2012.pdf , 12 January 2014

[16] Hrvatski sabor, "Zakon o obveznim odnosima", Narodne novine 035/2005, 4. ožujka 2005

Statistical Microdata – Production of Safe Datasets, Transparent Presentation of Contents and Advanced Services for Users through Metadata Authorization System

M. Poljičak*, H. Stančić**
* Central Bureau of Statistics/Information System Design and Development Department, Zagreb, Croatia
** Faculty of Humanities and Social Sciences/Department of Information and Communication Sciences, Zagreb, Croatia
*poljicakm@dzs.com
**hstancic@ffzg.hr

In the European Statistical System (ESS) there are more and more generic information systems (ISs) being developed and implemented for production and dissemination of official statistics (OS) macrodata datasets. These generic systems are based on metadata repositories providing, in return, a whole set of additional functionalities and services for users of the systems. Microdata dissemination, being the youngest and least developed discipline in official statistics dataset dissemination, currently still lacks appropriate organization of business procedures and compatible generic tools for the purpose of integration of all the ESS' subsystems into one central Data Service (DS) providing microdata datasets for international users and enabling preservation of microdata datasets. The authors analyze current situation regarding statistical microdata sets acquisition and preservation in digital archives (DAs), provide valuable considerations and propose recommendations to take into account when developing repositories and systems for presentation of available microdata datasets to users. Along with that, the authors introduce the way of usage of appropriate Statistical Disclosure Control (SDC) methods for production of safe microdata outputs. Finally, the authors recommend the development of a central administrative metadata repository that could be used for users' authorization, and that could have additional services for users based on available metadata about users in the system.

I. INTRODUCTION

Protecting the confidentiality of data in statistical surveys is one of the *Fundamental Principles of Official Statistics (OS).* OS should care about secure handling, preservation and dissemination of data. Setting authorized access levels as well as not publishing any data in any way that could reveal information, protected by legal or ethical restrictions, is necessary.

When it comes to data preservation in Official Data Archives (ODAs) different situations are present, at the moment, in different European countries. Some countries have a considerable level of cooperation between National Statistical Institutes (NSIs) and ODAs, and others still haven't developed satisfying and proper levels of cooperation and legal, administrative and technical solutions to tackle these needs.

It is of great importance to maintain the trust of the public in OS, and this is accomplished by setting extra efforts in regulating ways of accessing data, with proper security measures put in place when access is granted.

The problem of setting security measures and protocols, as well as secured ways of access to data, is complex especially if one takes into account a wide range of statistical data outputs available to users. These data outputs have different confidentiality levels and disclosure risks.

Statistical Disclosure Control (SDC) methods take into account legal and administrative boundaries to assure dissemination strategy aligned to disclosure risk and management of the same disclosure risk. Usage of appropriate IT tools and technical solutions, aligned with needs of legal compliance of procedures involved in dissemination of data, is still in the development phase in some areas. In these procedures a lot of manual work is done, involving time of skilled professionals in performing SDC methods, i.e. checking data output for statistical confidentiality. The area of microdata needs a better level of automatization of procedures in order to achieve a faster and simpler channel for production of safe data outputs for research community use.

Currently, new services, aligned to the new orientation of governments across the globe for opening access to different available government datasets, are calling for the new establishment of OSs' dissemination strategies. Organizations responsible for the development of OS in countries should invest efforts in establishing better ways of communication between NSIs, ODAs and research community (universities and research institutes in countries and abroad). These organizations should receive the support of political and other influential bodies in the country, as well as all necessary resources, to implement the procedures and prepare all protocols needed to raise the level of availability of microdata. Raised availability of microdata, could be useful and contribute to better de-

cision making and policies setting for the society. In effect they could bring new and advanced environment development and decision making concerning industry, economy and all other important sectors for empowering the society. Some of the examples of interested users of OS data are economists, demographers and social science researchers, as they could benefit from using vast resources created and preserved in conducting statistical surveys. All these users can find new and advanced secondary uses of extensive and expensive material already collected [12]. However, if data are released to these users, it must be under a proper disclosure control level in place for confidential statistical data.

The broad classification of statistical outputs usually distinguishes between two categories – *macrodata* and *microdata*.

Macrodata are aggregated statistical data (sums, averages, counting units), and these data are usually publicly available. However, some macrodata sets could be restrained from dissemination to public because the sample is too revealing and therefore public could infer conclusions about individual records in the macrodata dataset.

On the other hand, *microdata* are records collected as input data for the surveys. These records are confidential, subject to law restrictions and ethical standards set in place. However, under special conditions these data could be released, with more or less content (anonymized datasets), to special groups of users. Following current regulations, the users of microdata datasets could be any person belonging to the group of university researchers or members of other research institutions in the country, or even outside of the country, if they ask for accreditation. The procedure and the level of access rights are different in each European country, as are the ways of access. In the year 2013 a new regulation by the European Commission EC 557/2013 is set in place. By that regulation the process of application for statistical microdata access is made transparent for research community. "This Regulation establishes the conditions under which access to confidential data transmitted to the Commission (Eurostat) may be granted for enabling statistical analyses for scientific purposes, and the rules of cooperation between the Commission (Eurostat) and national statistical authorities in order to facilitate such access." [6]

A lot of work is still needed for the establishment of a successful Central Remote Access Data Center (RADC) for European microdata, providing the researchers with a simple and comfortable access to microdata and transparent presentation of data available. The service is currently being developed by the Eurostat and it should provide researchers with a central RA microdata access point in Europe.

In this paper the authors draw conclusions about several issues relating to establishing RADCs.

(1) How automated SDC procedures should be developed through RADC and aligned to the disclosure risk management for microdata datasets,

(2) How available resources should be transparently shown to the users, organized and made easily searchable,

(3) How researchers' authorization levels should be managed in a way to provide users with additional services aligned to their interests and needs concerning statistical microdata.

II. Related work

In providing public, and especially research community, with valuable statistical information NSIs need to set the balance between the level of the statistical outputs released and levels of protection of confidentiality of the individual data of respondents. In order to set an appropriate SDC method in place, i.e. to choose and implement appropriate software solutions, providers of official statistics microdata (as well as aggregated data outputs) should have in mind who their users are, as well as the possible uses of the released data. Also they need to consider possible further dissemination of the released data. For dissemination of the released data a proper ways of presenting available data should be defined. Also, users of the RADC, i.e. authorized researchers, should have an easy way of authentication and authorization procedures in order to manage data retrieval and to be able to get all other advanced services needed from the system.

A. Statistical Disclosure Control

In [1] "To protect the privacy of respondents a statistical office should prevent the disclosure of sensitive information on individual respondents by an intruder. As disclosure of sensitive information is possible only after an individual has been identified, a statistical office usually tries to prevent the identification of individual respondents."

As in [3] "Statistical Disclosure Control (SDC) or Statistical Disclosure Limitation (SDL) seeks to protect statistical data in such a way that they can be released without giving away confidential information that can be linked to specific individuals or entities." Also in [3] "a disclosure occurs when a person or organization recognizes or learns something that they did not know already about another person or organization via released data." And also in [3] "statistical disclosure control techniques can be defined as the set of methods to reduce the risk of disclosing information about individuals, businesses or other organizations." SDC methods minimize the risk of disclosure to an acceptable level while releasing as much information as possible. There are two types of disclosure risks as in [3], *identity disclosure* and *attribute disclosure*. Identity disclosure occurs with the association of a respondent's identity with the disseminated data record containing confidential information. *Attribute disclosure* occurs with the association of either attribute value in the disseminated data or an estimated attribute value based on the disseminated data with the respondent.

As mentioned in [3], microdata protection methods used for generating safe datasets can be divided into two categories: *masking methods* and *synthetic data generating methods*. *Masking methods* mask the original data to protect the dataset from the risk of disclosure. *Synthetic data methods* preserve some statistical properties of the original data, but on the other hand, are difficult to implement.

There are two types of SDC masking methods [3] – *perturbative and non-perturbative methods.* Perturbative methods falsify the data before publication by introducing an element of error purposefully for confidentiality reasons. In this way, unique combinations of scores in the original dataset may disappear and new unique combinations may appear in the perturbed dataset. Such confusion is beneficial for preserving statistical confidentiality. The used perturbation method should be such that statistics computed on the perturbed dataset do not differ significantly from the statistics that would be obtained on the original dataset. On the other hand, non-perturbative methods reduce the amount of information released by suppression or aggregation of data. Global recoding, local suppression and sampling are examples of non-perturbative methods." In that sense "*global recoding* method is a method where several categories of a variable are combined into single one, and *local suppression* method is a method where a value of a variable in a record is replaced by *missing*." [1]

In [3] it is showed that "a microdata set can be viewed as a file with *n* records, where each record contains *m* variables on an individual respondent. The variables can be classified in four categories which are not necessarily disjoint:

(1) *Identifiers.* These are variables that unambiguously identify the respondent (in Croatia, the example is OIB – Personal Identification Number, author's comment).

(2) *Quasi-identifiers or key variables.* These are variables which identify the respondent with some degree of ambiguity, but they can provide unambiguous identification of the individual record. Examples are name, address, gender, telephone number, etc.

(3) *Confidential outcome variables.* These are variables that contain sensitive information on the respondent. Examples are salary, religion, political affiliation, health condition, etc.

(4) *Non-confidential outcome variables.* These are variables which do not fall in any of the categories above."

Datasets anonymization could be done by avoiding dissemination of identifiers and quasi-identifiers since some of the general conclusions about the type of data are considered highly disclosing. Besides that, in anonymization procedures usually "categories of identifying variables with too significant identifying power are commonly aggregated into a single category. (...) Geographical information is a strongly identifying variable. (...) Release of date of birth is highly discouraged." [3] Also, when dealing with business data it is very easy to find the data belonging to a largest national company in some sectors, i.e. income or turnover of the company in telecommunications or national oil and gas company, so this data should be specially protected, especially in smaller countries where there is not a lot of suppliers of some services or goods in some sectors. "Examples of identifying variables are Age, Sex, Domicile and Occupation. Although each identifying variable is generally not sufficient to identify an individual when considered separately, a combination of values of identifying variables might be sufficient. When a combination of values of identifying variables is unique, i.e. occurs only once in population, then an intruder might identify the corresponding individual. (...) For preventing disclosure it is better to avoid the occurrence of combinations of values in the microdata set that are rare in the population, instead of trying to avoid only the population-uniques in the microdata set." [1]

The rules for handling disclosure risk take into account settings of the threshold, i.e. the minimum number of occurrence of the value of variable, considered to be the minimum for allowing the variable to be released in the output microdata dataset. "In case the frequency of a particular combination is at least the prescribed threshold value then this combination is considered safe. Otherwise the combination is considered unsafe and disclosure limitation measures should be applied." [1]

Access and overall dissemination of identifiers and quasi-identifiers should be avoided, especially taking into account possibility of merging these data with other available databases and greater risk levels concerned with possible disclosure of information in that scenario. Later in this research paper we will suggest providing free access to non-confidential outcome variables, with possibility of protecting confidential variables and/or confidential sets of variables. Restricted ways of access to confidential variables using SDC methods will be analyzed.

B. Presentation of data available in RADC

The second point of interest of this paper is the presentation of available content in an easy to search and explore ways, i.e. using navigation tools in RADC with possibility of learning about available datasets.

Developing a central RADC in Europe will not be an easy task. There are many countries with their NSIs involved in collecting, processing, preserving, and disseminating statistical datasets. Many of NSIs are either in the process of implementing standards for data collecting, documenting and preserving of the collected and processed data, or they have already done it. This should ease the situation in the future and enable a better connectivity between resources from different countries. Statistical surveys are also becoming standardized in the methodology for collecting, processing and disseminating data and in data structures.

C. Authorization system using metadata repository

The third point of interest of this paper is the ways of enabling advanced services for users of RA to microdata, as well as providing authorization mechanisms according to the authorization levels of users of the system and according to the authorization level needed to perform specialized operation on a dataset. It is obvious that in this DC, enabling access to confidential data, only authorized users should have access to data. The access should be granted according to all legal, technical and other measures taken into account and set in place for the proper level of data protection. This can be achieved by using metadata repository of users. Such repository should be placed in the core of the system responsible for documenting users, their rights and prescribed services available to a specific user.

III. MICRODATA SDC

A data dissemination strategy offers many different statistical outputs covering a range of different topics for many types of users. Different outputs require different approaches to SDC and different mixture of tools for handling it.

The microdata require a high level of disclosure control in order to protect confidentiality of available datasets. Usually anonymized datasets are available to users. Users can only perform statistical analysis on data – they cannot see the real data. The produced outputs are deposited in the system's repository. These outputs are first checked by NSI staff for statistical disclosure, and only if approved, the secondary data retrieved by using microdata can be released and used by the researcher.

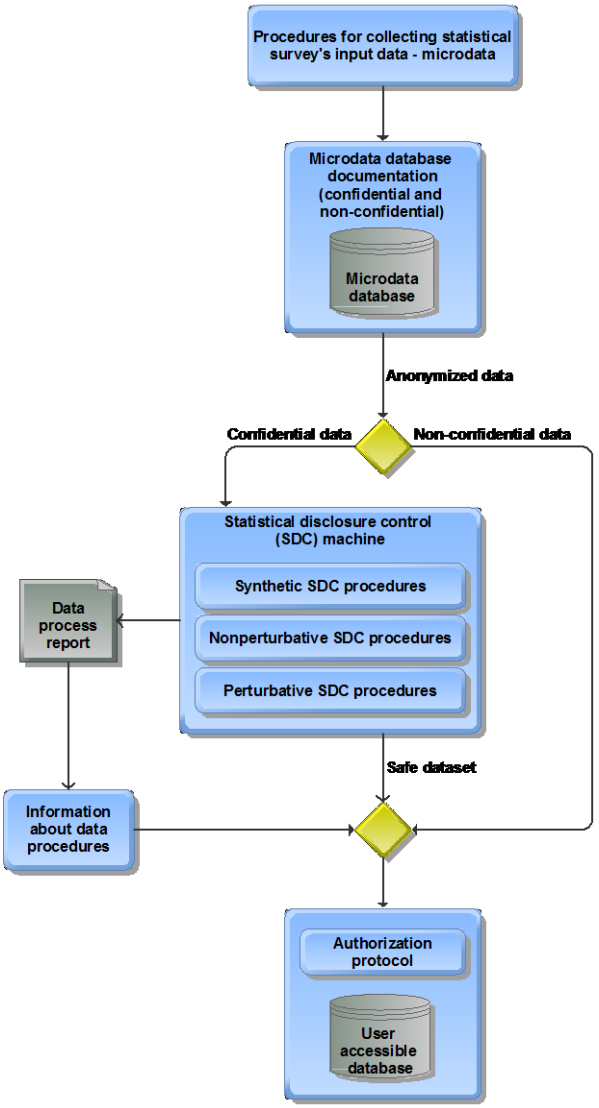
From the new Regulation introduced by the European Commission – EC 557/2013 – it can be concluded that the process of granting access to statistical microdata is becoming more transparent and procedures involved are more organized and centralized. In [6] "The Commission (Eurostat) shall publish on its website: (a) guidelines for the assessment of research entities, research proposals and access facilities; (b) the list of recognized research entities; (c) the list of accredited access facilities: (d) the list of datasets for research use with relevant documentation and the available modes of access."

However, for the NSI's staff, SDC procedures are still very time-consuming and burdening.

Therefore, we suggest building of a system for automatic creation of safe statistical outputs in the form of appropriate level of anonymization. Also, we suggest that variables and/or sets of variables considered to be revealing before being released through RADC, could be handled by development and implementation of software solution for enabling extraction of safe portions of datasets, as well as providing transparent documentation about applied SDC methods. Those datasets would then be made available to users in the form of a safe dataset with collection of non-confidential attributes and SDC protected confidential attributes which can then be used in a more user-friendly way, without creating barriers between users and the real datasets.

This system could be used for microdata dissemination strategy implementation for the research community. Accomplishing that, it is only a question of the extent of usability of the received datasets for research which are in accordance with the need for assuring the highest level of usability of data while at the same time protecting confidentiality of datasets.

However, there are, at the moment, software solutions for SDC. One example is µArgus, developed at Statistics Netherlands in 1995, using various SDC methods. The application is designed to work through several phases for ensuring production of safe data output. It works interactively with the user of the system. The metadata about data should be introduced to the system at the beginning of the process, and at the end the system produces a data process report for the data entering µArgus production of safe dataset.



Picture 1. Production of safe microdata outputs with documentation on applied SDC methods

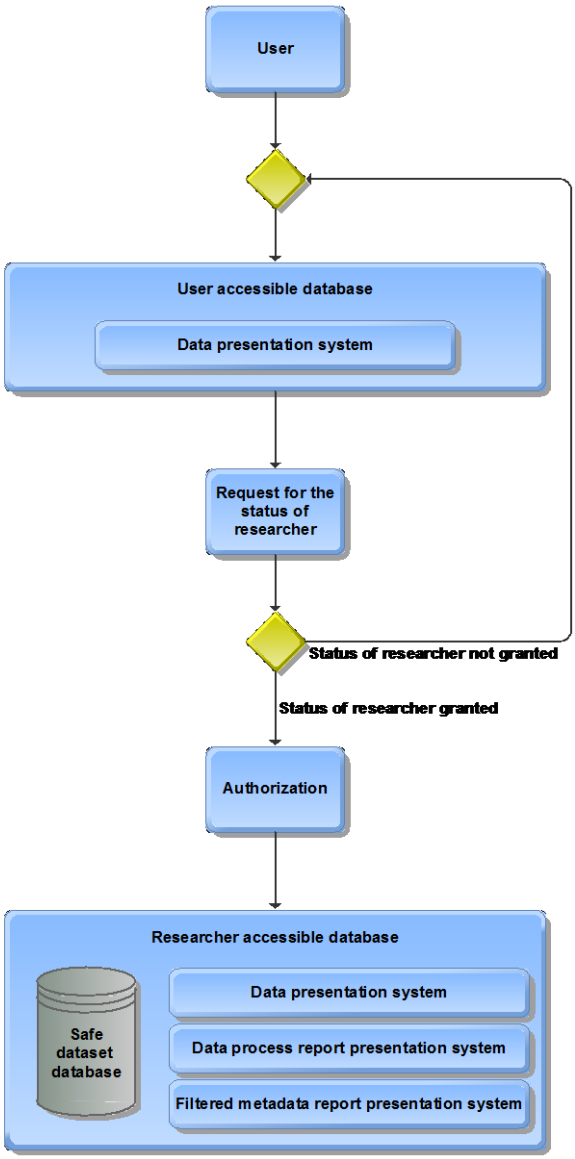
IV. TRANSPARENT PRESENTATION OF DATA IN THE USER ACCESS SYSTEM

When procedures for obtaining data collection and the documentation of the collected data in a Digital Archive (DA) become established for a statistical survey, then the same statistical survey data documentation should become available to the researchers through the DA interface that enables easy navigation and retrieval.

It is customary to use metadata repositories and semantics methodology for presentation of available resources in an interactive and easily retrievable way.

Metadata about surveys and datasets can help users learn about the available datasets and the survey methodology, and also help them to explore vast materials they are not familiar with.

The possible ways of showing the available contents are through interactive controls, like navigation trees, which should be prepared as a multilingual environment, thus enabling access and usage for possible users coming from different European countries. This goal is achievable if the system is generically built with resource-based approach to documenting all the data and metadata in a translatable repository of metadata. This repository can later be used to manage building multilingual controls for the user interface according to the available categories of data.



Picture 2. User access system

"Documentation is an essential part of any dissemination strategy both for auditing from external authorities and transparency towards users. The former may include descriptions of legal and administrative steps for a risk management policy together with the technical solutions applied. The latter is essential for a user to understand what has been changed or limited in the data because of confidentiality constraints. If a data perturbation method has been applied then, for transparency reasons, this should be clearly stated. (...) If a data reduction method has been applied with some local suppression (...) The released microdata should be obviously accompanied with all necessary metadata and information on methodologies used at various stages of the survey process (sampling, imputation, validation, etc.) together with information on magnitude of sampling errors, estimation domains, etc." [3]

V. USER AUTHORIZATION THROUGH METADATA REPOSITORY

By using a metadata repository for the purpose of building authorization system for diverse users, it is possible to develop all kinds of additional services for users, like subscribing to information about new datasets available, informing users about expiration of subscription to the data in advance, etc.

A metadata repository should hold the data about a user, surveys available to him/her to explore or use, and data released for the user exploration or additional analysis. This means that the repository will, in return, have a history of the user rights to datasets – preserved through time, and also it should be possible to record usage of the data to the user profile. This information could possibly be used in cases of confidentiality breech as well as for informing the user about some important information, or information concerning the deposit of new data and similar functions. In this way system's resources could be personalized in accordance with the users' interests and needs.

In spite of being generic and very versatile regarding available services for the user, this system should be easy to use. It should enable easy creation of new items in the database, as well as inheritance of all available modules. It should also be practical for the purposes of reuse and updating.

VI. CONCLUSION

The aim of this paper is to answer three questions important to have in mind when building either a central international RADC for statistical microdata access for researchers, or a national system for the same purpose in a national context.

Concerning the first question stated in the Introduction part of the paper, it can be concluded that SDC methods, aligned with the dissemination strategy for the data output, should be implemented through an automated system. Such a system should provide researchers with a satisfying level of data usability, and cover disclosure protection according to the level of risk of identity or attribute disclosure. This could be done by the proposed methods of anonymization and suppression of confidential values of variables where considered necessary, as well as masking of all confidential values of variables in the microdata output. Along with mentioned methods global recoding of confidential sets of variables in an automated system should be used in order to get the safe microdata dataset. Sophisticated methods, along with enabling a better level of data usability, should use perturbative methods for enabling non-disclosure scenario, because the data would be

"false", thus eliminating the disclosure risk, but at the same time "true" for obtaining exact aggregated statistical indicators.

The second question yielded a conclusion that the system providing transparent presentation of resources should be developed in a generic way, using metadata repository available in different languages, as well as using semantics methodology for building user interfaces that should enable easy exploration by users, as well as learning about the available datasets in overall transparent presentation of available resources in different languages. This system should certainly provide users with information about SDC methods used with the data before its dissemination.

Answers to the third question are in line with the usage of metadata repository for users' authorization levels implementation. They also enable the development of all kinds of advanced services for the users through personalized profile, developed on the grounds of users' metadata information. This system should be developed using a generic approach incorporating translatable metadata repository in the core of the system.

LIST OF ACRONYMS

DA	Digital Archive
DS	Data Service
ESS	European Statistical System
NSI	National Statistical Institute
ODA	Official Data Archives
OS	Official Statistics
RADC	Remote Access Data Centre
SDC	Statistical Disclosure Control
SDL	Statistical Disclosure Limitation

REFERENCES

[1] A.G. de Waal and L.C.R.J. Willenborg, "Optimal Local Suppression in Microdata", Journal of Official Statistics, 1998, http://www.jos.nu/Articles/article.asp

[2] A.Hunderpool and L.C.R.J. Willenborg, "m- and t- ARGUS: Software for Statistical Disclosure Control", Record linkage Techniques, Chapter 5, 1997, http://www.fcsm.gov/working-papers/hundepool.pdf

[3] A. Hunderpool et al., "Handbook on Statistical Disclosure Control, Version 1.2" ESSNet SDS, 2010, http://neon.vb.cbs.nl/casc/SDC_ Handbook.pdf

[4] B. Schouten and M. Cigrang, "Remote Access Systems for Statistical Analysis of Microdata," Voorburg/Heerlen July 2003, Discussion paper 03004, http://www.cbs.nl/NR/rdonlyres/C8B18053-7B3E-485F-851E-A0DE40350553/0/Discussionpaper03004.pdf

[5] Data without Boundaries, "Deliverable D7.1, Metadata Standards – usage and needs in NSIs and Data Archives," 2012, Work package 7, Standards Development http://www.dwbproject.org/export/sites/default/about/public_deliveraples/dwb_d7-1_metadata-standards-usage_report.pdf

[6] European Commission, "Commission Regulation (EU) NO 557/2013 on access to confidential data for scientific purposes," 2013, Official Journal of the European Union, L 164/16, http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2013:164:0016:0019:EN:PDF

[7] European Commission and Eurostat, "Guidelines for the assessment of research entities and research proposals," 2012, http://epp.eurostat.ec.europa.eu/portal/page/portal/microdata/documents/guidelines_assessment.pdf

[8] European Commission and Eurostat, "List of recognised research entities," 2012, http://epp.eurostat.ec.europa.eu/portal/page/portal/microdata/documents/Recognised_research_entities.pdf.

[9] F. Willis-Nunez, "Users of the statistical metadata system", Statistical metadata (METIS), 2012, http://www1.unece.org/stat/platform/display/metis/3.++Users+of+the+statistical+metadata+system

[10] G. Pongas G. and A. Wronski, "A distributed architecture for statistical data editing, processing and dissemination," Meeting on the Management of Statistical Information Systems (MSIS 2009), Oslo, Norway, 2009, Topic (iii): Architecture, http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2011/35_France-Germany-NL.pdf

[11] N. Ahmad, Koen De Bacher and Y. Yoon, "An OECD perspective on microdata access: Trends, opportunities and challenges," Statistical Journal of the IAOS 26(2009/2010) 57-63, http://donnageczi.com/pdfs/IOS-Press-Journal-Library/support-files/sji.pdf

[12] M. Poljičak and H. Stančić, "Proposing the Model for Croatian Remote Access Safe Centre for Statistical Microdata", in: A. Gilliland et al., Information Governance, Zagreb : Department of information and Communication Sciences, Faculty of Humanities and Social Sciences, 2013, pp. 137-146, http://infoz.ffzg.hr/INFuture/papers/4-02%20Poljicak,%20Stancic,%20Proposing%20the%20Model%20for%20Croatian%20RA%20Safe%20Centre.pdf

[13] R. Grim, P. Heus, T. Mulcahy and J. Ryssevik, "Secure remote access system for an upgraded CESSDA RI, metadata technology, Version Sep 2009" cessda ppp, http://www.cessda.org/project/doc/CESSDA_RI_SRA_FINAL.pdf

[14] R. Silberman, S. Bender and A. Hunderpool, "The need for networks on data access – Data Without Boundaries Project and the Workshop on data access," Joint UNECE/Eurostat work session on statistical data confidentiality, Tarragona-Spain, 26-28 October 2011, Topic (vii): Trans-border access to microdata, http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2011/35_France-Germany-NL.pdf

[15] S. Bender, J. Heining, L. Franconi and D. Ichim, "Microdata access: and international perspective," Tarragona-Spain, 26-28 March 2013, BLUE-Enterprise and Trade Statistics, SP1 – Cooperation-Collaborative Project, Small or medium-scale focused research project, http://www.unece.org/fileadmin/DAM/stats/documents/ecen/ces/ge.46/2011/35_France-Germany-NL.pdf

[16] U. Jensen, "Data and metadata extensions of the CESSDA RI (D8.3)," CESSDA PPP Project Deliverables 2010, http://cessda.org/project/doc/D8.3_Data_metdata_enhancement.pdf

[17] United Nations Economic Commission for Europe, "Managing statistical confidentiality & microdata access, Principles and guidelines of good practice," Conference of European Statisticians, New York and Geneva 2007, http://www.unece.org/fileadmin/DAM/stats/publications/Managing.statistical.confidentiality.and.microdata.access.pdf

[18] United Nations Economic Commission for Europe, "CMF Part A, Statistical metadata in a corporate context: A guide for managers," Conference of European Statisticians, Geneva 2009, http://www1.unece.org/stat/platform/display/metis/Part+A+-+Statistical+Metadata+in+a+Corporate+Context

Choosing broadband access solution for modern residential buildings in densely populated urban center

Jasenko Topić, Zlatan Begić, Izudinka Kapetanović Serdarević
BH Telecom d.d. Sarajevo Direkcija Tuzla, Tuzla, Bosna i Hercegovina
jasenko.topic@bhtelecom.ba, zlatan.begic@bhtelecom.ba, izudinka.kapetanovic@bhtelecom.ba

Abstract: Broadband access today is a result of many years of growing demand and represents the best choice for telecom operators to secure telecommunications market position. Data traffic is constantly increasing and brings higher profit to operators compared to voice traffic. Newest broadband services demand next generation broadband access infrastructure. Investment in better infrastructure is key for implementing new services. Telecom operator needs to decide how complex (and expensive) should its infrastructure be in order to serve as long – term broadband solution and at the same time make profit from providing broadband services, including more and more popular triple – play.

Choice rests on multitude of factors, which include existing infrastructure quality, population density, dominant building type and geographical household distribution. This article analyzed densely populated area with dominantly residential buildings, which is a feature of urban centers and residential – business zones. It is important to say that urban centers have increased demands for broadband services, a result of larger business subjects concentration, research and educational institutions from elementary to university level.

Article goal is to give answer to rising question: which technology is the best for telecom operators from aspect of implementation speed and profitability in given area, and that can rise up to challenge of providing ever demanding broadband services and provide a way for transition to FTTH infrastructure.

Keywords: fttx, telecom, triple – play, urban, xdsl

I. INTRODUCTION

Densely populated area is an area with average household density above 1000 per km^2. Areas with dominantly residental buildings are areas with more than 70% of households in residential buildings and the rest in residential houses. Modern infrastructure residential buildings have home wiring where measured parameters on more than 50% of twisted – pair copper wires support xDSL (*x Digital Subscriber Line)* standards [1] [2] and have adequate space for telecommunications equipment.

II. EXISTING STATE

Densely populated area with dominantly residential buildings is a feature of urban centers. Objects built in last 30 years usually have adequate space for active equipment and home wiring that supports xDSL standards for providing demanding broadband services.

As an example an urban center in Tuzla is chosen, with aerial photography given in fig. 1. Broadband service is being given through ADSL2+ (*Asymmetric Digital Subscriber Line 2+*) [1] DSLAM (*Digital Subscriber Line Access Multiplexer)* equipment over copper access network. Although analyzed area is located at just about 1000 m from active equipment, households have limited bandwidth, because of which there is a large number of pending requests for broadband services that telecom operator can not process. Exact reasons will be determined after terrain survey.

Analyzed urban center has surface area of 0,15 km^2 with about 1000 households. Household density is about 6666 per km^2, which means that this area is densely populated. With more than 70% of households in residential buildings, they present the dominant type.

Urban center has great potential because of population density. More detailed analysis of access network and residential buildings infrastructure is needed in order to choose the right broadband solution.

An exact copper cable position was determined after terrain survey. Primary and secondary access network consists of copper cable too dated to be efficiently used for providing broadband services, triple – play in particular. According to previous experience, a problem was found in a cable that leads to residential buildings, which reconstruction is time consuming and presents big financial investment compared to its positive effects.



Figure 1. City center example with modern residential buildings[1]

[1] Source: Google Earth satellite photography

Access network state will be analyzed for 6 residential buildings. Prior to the measurements an ADSL2+ modem is required to be placed in a household. That modem is connected via twisted – pair copper wire to main distribution frame in telecom company building, after which there is another 10 m of copper wire leading to modem in DSLAM port. After the modems are synchronized with ADSL2+ standard, xDSL parameters can be read from surveillance system shipped by manufacturer with DSLAM. Measured values are identical to those that would be measured on a household modem. Chosen xDSL parameters are measured according to standard [1] on twisted – pair copper wires that connect analyzed residential buildings to DSLAM equipment, results of which are shown in table 1.

Results are displayed graphically in fig. 2. According to obtained results it can be seen that measured values are relatively low for twisted – pair copper wires only about 1000 m long. Access network state can be improved with a new broadband solution.

III. RECOMMENDED BROADBAND SOLUTION

Urban center with residential buildings has large household concentration in small area. Telecom operators need to act quickly to eliminate all problems in providing service to households because of the market size and strong competition. Bigger changes to access technology when it comes to replacing telecommunications infrastructure are challenging in urban centers. It is imperative to choose and quickly deploy the right broadband solution.

In urban centers telecom operators usually have cable ducts in which optical fiber can be laid down up to residential building entrances or the distance from optical fiber highway is relatively short. Copper cable access network is of bad quality for broadband services. Residential buildings provide adequate space for active equipment. House wiring is relatively good, because we analyze modern residential buildings.

Because of this FTTB (*Fiber To The Building*) is the recommended broadband solution. DSLAM is the right choice for active equipment and other services can be deployed from existing equipment at telecom premises via old copper cables.

TABLE 1. AVERAGE XDSL PARAMETERS VALUES MEASURED ON COPPER WIRE TWISTED – PAIRS IN ANALYZED RESIDENTIAL BUILDINGS

Building ID	Measured xDSL parameters		
	Signal to noise ratio [dB]	*Line attenuation [dB]*	*Bandwidth [Mbit/s]*
1	12,18	12,73	15,14
2	11,18	17,37	9,99
3	9,39	18,56	11,90
4	12,94	16,73	13,06
5	14,41	19,73	11,61
6	16,93	22,00	12,70

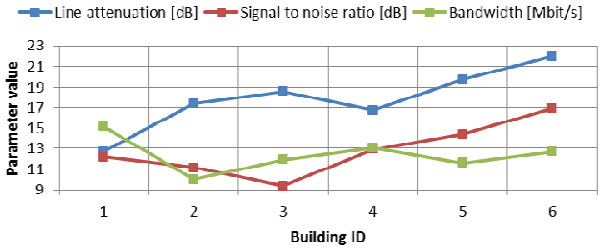


Figure 2. Graphical representation of measured xDSL parameters values

Significant improvement of xDSL line parameters is expected, because the length of twisted – pair copper wires is shortened from about 1000 m to less than 150 m of house wiring. Depending on number of households in residential buildings, telecom operators may choose FTTB DSLAM solutions for 8 – 96 households.

By using FTTB technology telecom operator eliminates copper access network entirely and leaves only copper house wiring. Telecom operator can not use existing copper access network because primary cable is dated and not adequate for xDSL broadband services. With relatively small and quick interventions telecom significantly improves broadband service quality and gains larger user base.

Compared to FTTC (*Fiber To The Cabinet*) in urban centers, FTTB has the following advantages:

- Predetermined active equipment location,
- Better elimination of copper in last mile,
- House wiring presents the only weak spot,
- Relatively short secondary copper access network length,
- Because only broadband access is changed, operator does not require MSAN (*Multi Service Access Node*) and keeps existing equipment,
- No issues with providing adequate room for active equipment and no building outdoor cabinets,
- Relatively quick optical fiber highway installation

FTTH (*Fiber To The Home*) solution compared to FTTB in urban centers is not acceptable because of:

- Complex and expensive home wiring replacement
- Telecom can always process eventual individual FTTH household requests
- optical fiber offers little advantage compared to vectorized VDSL2 (*Very-high-bit-rate Digital Subscriber Line 2*) on copper local loops under 150 m concerning bandwidth for future market demands
- FTTB is proven solution, telecom operator uses existing active equipment

House wiring in residential buildings should be adequate for deployment of bandwidth – hungry broadband services.

Telecom operator places FTTB DSLAM equipment in indoor cabinets in residential building entrances, which have enough space for active equipment, power source, optical fiber installation and copper main distribution frame. This is one of the reasons for choosing the FTTB solution. Figures 3., 4. and 5. graphically represent measured xDSL parameters values before and after deploying FTTB solution in given buildings. For comparision, before and after measurements were taken on identical ADSL2+ modems that limit bandwidth to 25 Mbit/s. It can be seen that access network is greatly improved given the previous state.

In fig. 6 are presented measured bandwidth values on VDSL2 modems [4] according to standard [2] for 9 households that use premium broadband services. This shows FTTB solution potential.

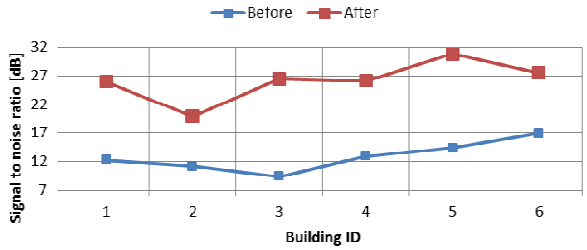


Figure 3. Before and after FTTB – signal to noise ratio

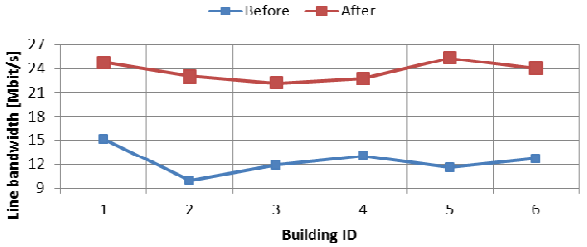


Figure 4. Before and after FTTB – line bandwidth

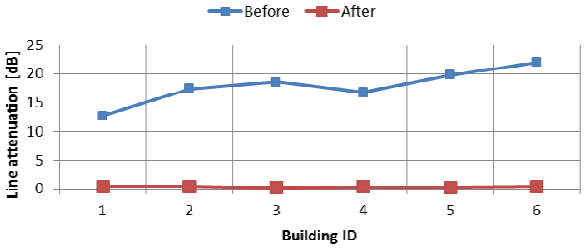


Figure 5. Before and after FTTB – line attenuation

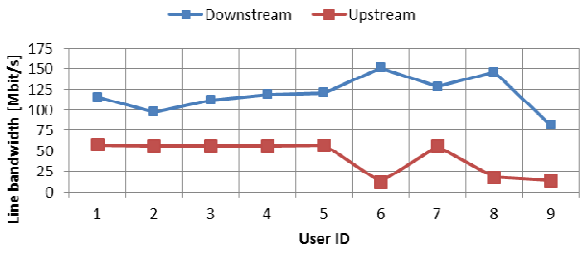


Figure 6. VDSL2 downstream and upstream line bandwidth

Copper wire downstream bandwidth is mostly above 100 Mbit/s and that is without vectoring. This bandwidth is enough for broadband services of today and in the future. For this reason operators still use copper access networks and do not go for optical fiber access networks.

Table 2. lists advantages and disadvantages of chosen FTTB solution.

Residential buildings are important to telecom operators for deploying FTTB solution because of small investment per user, high demand for broadband services and exclusivity in using home wiring thus eliminating competition.

In urban centers there is always strong competition and it is important to deploy broadband services quickly to households. Densely populated areas present large potential user base and by deploying FTTB solution, telecom operator offers better quality service compared to cable and satellite legacy services. Operator mostly uses existing infrastructure, offers technically superior solution compared to FTTC and investment – wise better than FTTH. Operator has the ability to offer 80 Mbit/s bandwidth on copper house wiring with lengths up to 150 m with VDSL2 technology [5], thus using full FTTB solution potential. According to analysis above, FTTB is recommended broadband solution.

IV. ALTERNATIVE BROADBAND SOLUTION

For FTTH solution deployment operator needs to change whole existing infrastructure. Compared to FTTB, operator needs to change house wiring from copper to optical fiber, which is more time consuming and financially challenging. FTTH solution is therefore best for greenfield locations, which have no previous copper wire broadband access network infrastructure. Cost of building optical fiber networks is currently relatively high and deployment time is also longer compared to FTTB/C solutions.

Todays household demands can be fulfilled with FTTB solution that uses line vectoring and bonding, thus achieving bandwidth up to 300 Mbit/s [5]. FTTH offers practically unlimited bandwidth, a potential unnecessary to households in the next decade.

TABLE 2. CHOSEN FTTB SOLUTION ADVANTAGES AND DISADVANTAGES

Advantages	Disadvantages
Eliminated primary copper access network	Deploying optical fiber to residential buildings
Bandwidth above 200 Mbit/s over house wiring with VDSL2 vectoring	Adaptation of building space for active equipment
Best local loop shortening	
Relatively simple deployment	
Unified technical solution	
No issues with equipment placement	

FTTC solution deployment requires good quality of existing copper access network or new copper infrastructure. Analyzed urban center has access network of poor quality, however, residential buildings in it have good quality house wiring to support demanding broadband services. FTTC solution would use primary copper network, which is of bad quality and for that reason FTTB is technically superior solution.

For an area with good quality access network located several km from active equipment FTTC can be used as recommended solution, given low household broadband demands. If longest twisted – pair copper wire would be less than 500 m, then 1 FTTC access node would be sufficient for broadband use. At that lengths operator can achieve minimum 80 Mbit/s per household [5].

Multiple FTTC nodes solution given the residential buildings have modern infrastructure is not recommended, because FTTB is technically superior and by deployment speed and cost is relatively the same. On one hand, FTTB node is more compact and can be placed in existing space indoors and on the other hand, FTTC node requires placement in outdoor cabinet and presents more expensive solution.

FTTC is recommended solution for dated residential buildings. Unpractical solution would be multiple FTTC nodes (fig. 7) and practical 1 FTTC node (fig. 8).

FTTC solution is not practical for urban centers with households in modern residential buildings. Entrances of these buildings present ideal location for active equipment (enough space for cabinet, power source and cooling). Operator shortens local loop ideally within these objects which makes VDSL2 the perfect broadband solution for deployment.

FTTC solution has longer local loop and therefore lower household bandwidth because of higher line attenuation. Multiple FTTC nodes is not recommended because of time and cost of building infrastructure for each FTTC node, which already exists in residential buildings.

V. CONCLUSION

Important factor in society informatization is telecommunications infrastructure of telecom operators that offers greater broadband connectivity with each year, improving services and building basis for development of society.

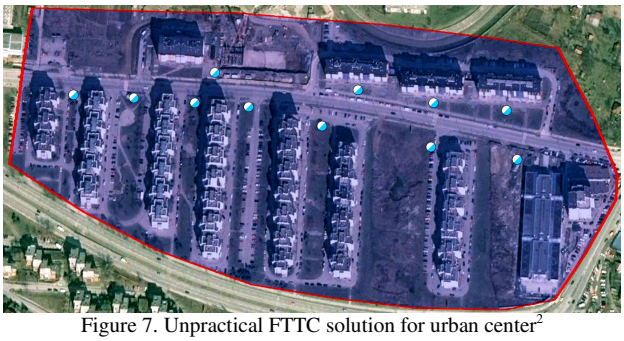


Figure 7. Unpractical FTTC solution for urban center[2]



Figure 8. Practical FTTC solution for urban center[3]

FTTB broadband solution is recommended in densely populated areas with dominantly modern residential buildings, a feature of urban centers with high informatization level. Technically and economically FTTB presents better solution compared to FTTC and in near future it is not expected for FTTH to be strong alternative. With FTTB it is possible to have *downstream* bandwidth in households up to 80 Mbit/s. Also, FTTB solution is designed for residential and other multiple – household buildings. Advantages lie in relatively quick deployment and in creating basis for later FTTH deployment when households demand more than 100 Mbit/s *downstream*.

REFERENCES

[1] ITU-T, "G.992.5 Asymmetric digital subscriber line transceivers 2 (ADSL2) – Extended bandwidth (ADSL2plus)", ITU-T Recommendation, April 2009.

[2] ITU-T, "G.993.2 Very high speed digital subscriber line transceivers 2 (VDSL2)", ITU-T Recommendation, December 2011.

[3] Broadband Forum, "TR-100: ADSL2/ADSL2plus Performance Test Plan", 2. edition, August 2012.

[4] Broadband Forum, "TR-114: VDSL2 Performance Test Plan", 2. edition, November 2012.

[5] J. Williams, "How Enhanced DSL Technologies Optimize the Last Copper Mile", JDSU Whitepaper, May 2013.

[2] Source: Google Earth satellite photography

[3] Source: Google Earth satellite photography

Odabir rješenja širokopojasne pristupne tehnologije u gusto naseljenom području sa dominantno individualnim stambenim objektima

Jasenko Topić, Zlatan Begić, Izudinka Kapetanović Serdarević
BH Telecom d.d. Sarajevo Direkcija Tuzla, Tuzla, Bosna i Hercegovina
jasenko.topic@bhtelecom.ba, zlatan.begic@bhtelecom.ba, izudinka.kapetanovic@bhtelecom.ba

Sažetak – Kako je širokopojasni pristup značajan faktor društvenog razvoja, telekom operater treba temeljito razmotriti odluku koju tehnologiju implementirati uzimajući u obzir brzinu realizacije i isplativost, a koja može kvalitetno odgovoriti na rastuće zahtjeve korisnika. Za najnovije širokopojasne usluge potrebne su pristupne infrastrukture nove generacije. Ulaganje u kvalitetniju infrastrukturu otvara put razvoju novih usluga.

Operater treba donijeti odluku koliko složenu, a time i skupu tehnologiju odabrati, kako bi dugoročno riješio širokopojasni pristup, a ujedno i ostvario profit od korištenja sve popularnijih *triple – play* usluga.

Izbor ovisi o nizu faktora, koji uključuju kvalitet postojeće infrastrukture, gustinu naseljenosti, dominantni tip stambenih objekata i prostorni raspored domaćinstava. U ovom radu analizirano je gusto naseljeno područje sa dominantno individualnim stambenim objektima karakteristično za ruralne sredine, odnosno prigradska naselja.

Cilj rada je odgovoriti na pitanje koja tehnologija je perspektivna operateru sa gledišta brzine realizacije i isplativosti na predmetnom području, a koja može kvalitetno odgovoriti na rastuće zahtjeve korisnika i omogućiti postepeni prelazak prema konačnom FTTH rješenju.

Ključne riječi: fttx, ruralno, telekom, *triple – play*, xdsl

I. UVODNA RAZMATRANJA

Gusto naseljeno područje je takvo područje, koje na svojoj površini ima srednju vrijednost broja domaćinstava iznad 250 po km^2. Područja sa dominantno individualnim stambenim objektima su područja sa više od 70% domaćinstava u individualnim, a ostatkom domaćinstava u kolektivnim stambenim objektima.

II. POSTOJEĆE STANJE

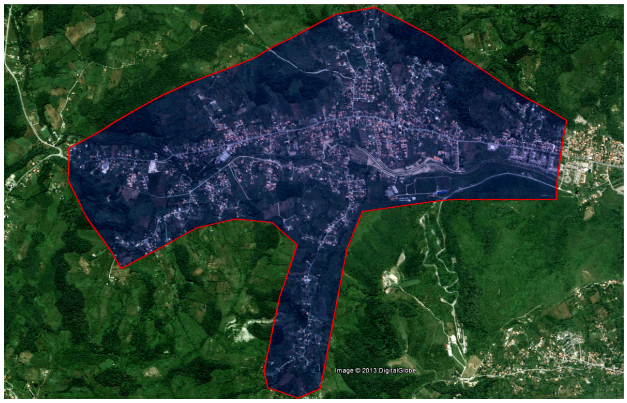
Na području sa relativno velikom dužinom parice za ADSL2+ (*Asymmetric Digital Subscriber Line 2+*) (iznad 1200 m) [1] telekom operater ima problem realizacije kvalitetne širokopojasne usluge, posebno *triple – play*. Iz tog razloga postoji velik broj smetnji kod postojećih i odbijenih zahtjeva potencijalnih korisnika zbog nepostojanja tehničke mogućnosti, problem koji operater treba riješiti.

Kao primjer ovakvog područja usvaja se naselje u Tuzli sa obuhvatom datim na sl. 1. Širokopojasna usluga realizuje se ADSL2+ tehnologijom [2] DSLAM (*Digital Subscriber Line Access Multiplexer*) uređajem po bakarnoj pristupnoj mreži.

Dati obuhvat ima površinu 3,3 km^2, a ima oko 950 domaćinstava. Gustina naseljenosti je oko 280 domaćinstava po km^2 i zaključuje se da je područje gusto naseljeno. Sa više od 70% domaćinstava u individualnim stambenim objektima, zaključuje se da je dominantni tip stambenih objekata individualni. Kablovski pravac je velike dužine. Za odabir tehničkog rješenja potrebno je izvršiti obilazak terena i ispitivanje pristupne mreže.

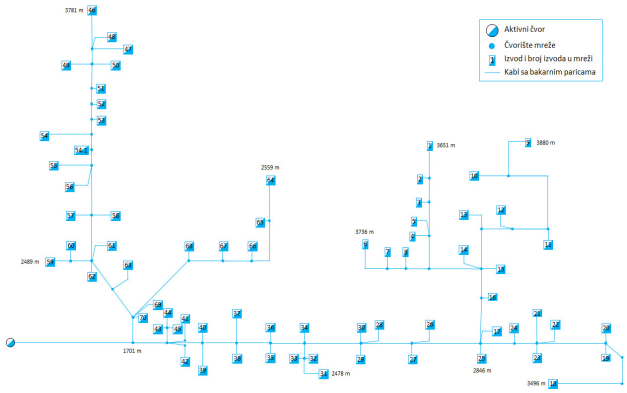
Obilaskom terena utvrđen je položaj bakarnog kablovskog pravca uz koji je postavljen i optički spojni put za spajanje pristupne aktivne opreme na distribuiranu *backbone* mrežu.

Primarni i sekundarni dio pristupne mreže sastoji se od bakarnog kabla za zahtjevne terenske uvjete. Često se za xDSL (*x Digital Subscriber Line*) usluge koristi kabl tipa TK59 GM [3]. Primjer topologije pristupne mreže dat je na sl. 2.



Slika 1. Primjer obuhvata područja sa dominantno individualnim stambenim objektima[1]

[1] Izvor: Google Earth satelitska fotografija



Slika 2. Primjer topologije pristupne mreže

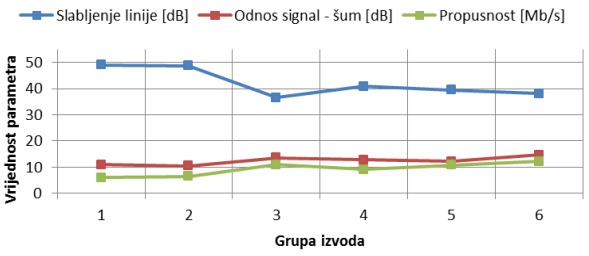
Primjer topologije pristupne mreže sadrži bakarne segmente, aktivni čvor i korisničke izvode. Mreža je podijeljena u dva segmenta približne dužine. Aktivni čvor se nalazi na početku kablovskog pravca sa dužinama parica iznad preporučenih za ADSL2+ tehnologiju [1].

Putem nadzornog sistema za DSLAM izmjereni su odabrani xDSL parametri na korisničkim portovima i izračunata je njihova prosječna vrijednost po izvodu [1]. Rezultati mjerenja prikazani su u tabeli 1. Grafički prikaz rezultata mjerenja dat je na sl. 3.

Kao što je i očekivano, najudaljeniji izvodi imaju najveće slabljenje, najmanji odnos signal – šum za datu uslugu, kao i najmanju propusnost na liniji [4]. Vrijednosti mjerenih parametara relativno su loše zbog velike dužine parica. Najveća propusnost je 12,3 Mbit/s i nije dovoljna za buduće širokopojasne usluge.[2]

TABELA 1. PROSJEK MJERENJA XDSL PARAMETARA NA BAKARNIM PARICAMA NA GRUPAMA IZVODA SMJEŠTENIM U DATOM OBUHVATU

Grupa izvoda	**Mjereni xDSL parametri**		
	Odnos signal – šum [dB]	***Slabljenje linije [dB]***	***Bitska propusnost linije [Mbit/s]***
1	11,05	49,07	6,02
2	10,56	48,78	6,53
3	13,56	36,49	11,03
4	12,80	40,81	9,15
5	12,21	39,52	10,78
6	14,75	38,06	12,29



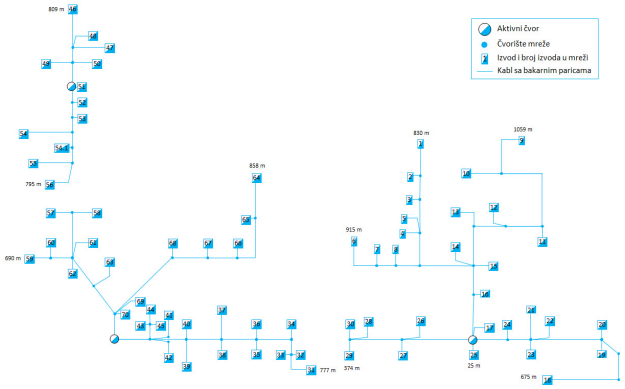
Slika 3. Grafički prikaz rezultata mjerenja xDSL parametara

[2] HD IPTV (*High Definition Internet Protocol TeleVision*) 8 Mbit/s, brzi Internet pristup 10 Mbit/s

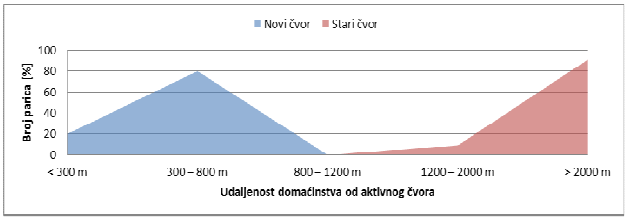
III. TEHNIČKO RJEŠENJE

Područje je gusto naseljeno domaćinstvima koja širokopojasnu uslugu ostvaruju preko parice relativno velike dužine. Ovakvo područje sa otprilike 1000 domaćinstava telekom operateru predstavlja značajnu bazu korisnika. Preporučeno je FTTC (*Fiber To The Cabinet*) rješenje sa 3 DSLAM aktivna čvora.

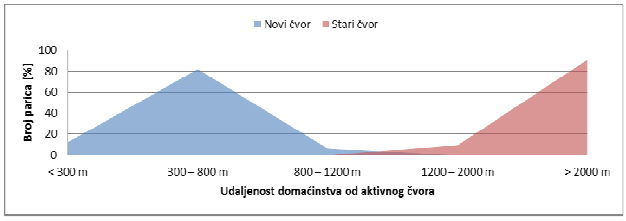
Nova topologija pristupne mreže prikazana je na sl. 4. Većina domaćinstava ima paricu kraću od 800 m. Najveća dužina parice iznosi 1059 m, pa su ostvarive brzine najmanje 30 Mbit/s za VDSL2 (*Very-high-bit-rate Digital Subscriber Line 2*) [4]. Na sl. 5, sl. 6. i sl. 7. grafički su prikazani efekti skraćenja parice izraženi u procentima broja parica u datom opsegu.



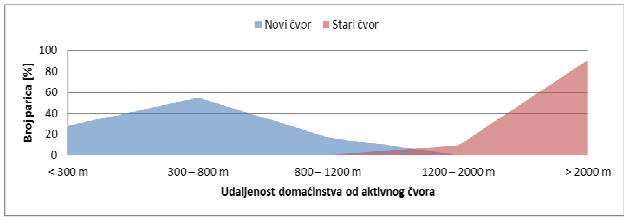
Slika 4. Nova topologija pristupne mreže



Slika 5. Efekti skraćenja parice za 1. aktivni čvor



Slika 6. Efekti skraćenja parice za 2. aktivni čvor



Slika 7. Efekti skraćenja parice za 3. aktivni čvor

Tabela 2 prikazuje prednosti i mane odabranog tehničkog rješenja.

TABELA 2. PREDNOSTI I MANE ODABRANOG FTTC TEHNIČKOG RJEŠENJA

Prednosti	Mane
Omogućena kvalitetna širokopojasna usluga cijeloj pristupnoj mreži	Udaljenija domaćinstva imaju relativno male pristupne brzine za VDSL2 tehnologiju [4]
Sa vektoriziranim VDSL2 moguće 100 Mbit/s do 400 m [4]	
Iskorišten postojeći bakarni kabl	
Uklonjene stalne smetnje na kablu	
Minimalna ulaganja	
Dokazano rješenje u praksi	

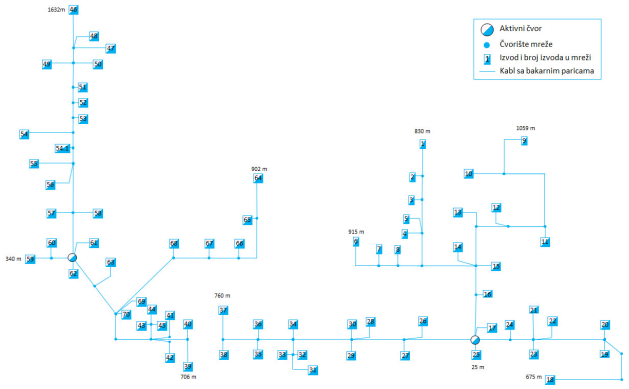
Na analiziranoj mreži do sada nije bilo tehničke mogućnosti za kvalitetnu širokopojasnu uslugu uglavnom zbog velike dužine parica. Postojeći korisnici osjetit će poboljšanje kvalitete usluge, imati veće pristupne brzine na parici, a osjetit će se i znatan porast broja korisnika širokopojasne usluge. Također će se datim tehničkim rješenjem trajno ukloniti smetnje uzrokovane velikim slabljenjem i malim odnosom signal – šum na paricama.

Implementacijom preporučenog rješenja očekuje se povećanje odnosa signal – šum iznad 20 dB za postojeće usluge, čime je moguće korištenje i zahtjevnijih širokopojasnih usluga. Očekivana slabljenja na liniji su ispod 20 dB. Najveće ostvarive pristupne brzine na parici ovisno od udaljenosti domaćinstva približit će se praktičnoj granici za ADSL2+ od 25 Mbit/s [5], a domaćinstva udaljena do 300 m od aktivnog čvora imat će pristupne brzine do 150 Mbit/s sa vektoriziranim VDSL2 [4]. Zbog gustine naseljenosti preporučuje se postavljanje više aktivnih čvorova, jer velik broj domaćinstava dobija kvalitetniju uslugu. Ne očekuje se dalja pojava smetnji uzrokovanih dužinom parice, a povećanjem zadovoljstva korisnika raste i prihod telekom operateru.

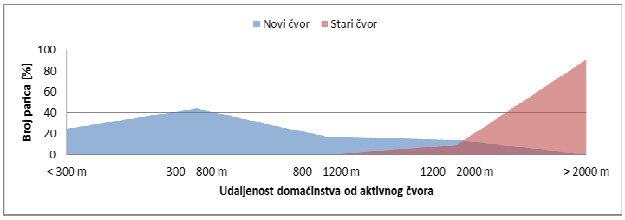
IV. ALTERNATIVNO RJEŠENJE

Za razliku od preporučenog, alternativno rješenje je FTTC sa 2 DSLAM aktivna čvora. Operater bira nove 2 lokacije za aktivne čvorove kako bi prosječne dužine parica bile što kraće. Na sl. 8 prikazana je nova topologija pristupne mreže, a na sl. 9 i sl. 10 prikazani su efekti skraćenja parice.

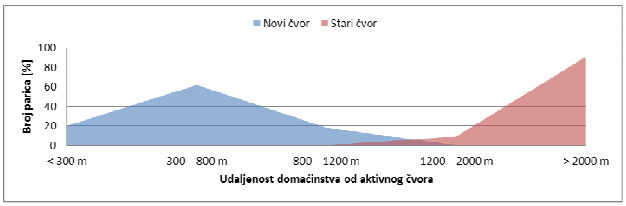
Ako se posmatraju obuhvati aktivnih čvorova, najveća dužina parice prije skraćenja iznosi oko 3800 i 3700 m, a nakon skraćenja iznosi oko 1600 i 1000 m, respektivno.



Slika 8. Nova topologija pristupne mreže – alternativno rješenje



Slika 9. Efekti skraćenja parice za 1. aktivni čvor – alternativno rješenje



Slika 10. Efekti skraćenja parice za 2. aktivni čvor – alternativno rješenje

Sa 2 aktivna čvora 75% parica u pristupnoj mreži kraće je od 800 m na kojim operater osigurava kvalitetnu širokopojasnu uslugu sa VDSL2 brzinama (najmanje 40 Mbit/s) [4]. Ostatak parica, njih 25%, duže je od 800 m, a pripada relativno malom broju domaćinstava koja će imati pristupne brzine ispod 40 Mbit/s.

Područje predmetne pristupne mreže gusto je naseljeno, gdje su bolji tehnički i ekonomski efekti postavljanja dodatnih aktivnih čvorova. Riječ je o individualnim stambenim objektima i pristupnoj mreži koju operater iskorištava FTTC rješenjem. Smanjenjem broja aktivnih čvorova u odnosu na preporučeno rješenje povećava se prosječna dužina parice, čime domaćinstva imaju lošiju tehničku mogućnost za širokopojasne usluge, a operater manja početna ulaganja. Domaćinstva imaju manje pristupne brzine u odnosu na preporučeno rješenje. Potrebno je iskoristiti potencijal ovog područja, a alternativnim rješenjem operater ne može osigurati kvalitetnu širokopojasnu uslugu, čime se ono odbacuje. Time je FTTC sa 3 DSLAM – a preporučeno rješenje.

V. ZAKLJUČAK

U gusto naseljenom području sa dominantno individualnim stambenim objektima, karakterističnim za ruralne sredine i prigradska naselja, preporučuje se FTTC rješenje za realizaciju širokopojasnog pristupa. Zbog gustine naseljenosti postavljanjem više FTTC aktivnih čvorova ostvaruje se veći potencijal korištenja zahtjevnih širokopojasnih usluga. Uglavnom se koristi postojeća bakarna pristupna mreža i za operatera sa ekonomskog gledišta predstavlja bolje rješenje u odnosu na FTTB (*Fiber To The Building*) rješenje. Ovakav tip naselja ne iskorištava puni potencijal FTTB rješenja, a FTTC rješenjem moguće je realizirati gotovo sve zahtjeve domaćinstava, jer date dužine bakarne parice omogućuju najmanje 30 Mbit/s sa VDSL2 tehnologijom [4]. U određenim slučajevima operater treba naći novo tehničko rješenje za postojeće usluge koje koriste analiziranu infrastrukturu. Iz ovog razloga bitno je vršiti najmanje moguće izmjene postojećeg stanja, a ostvariti što veće ekonomske efekte i postaviti temelje prelaska domaćinstava na FTTH rješenje u budućnosti.

LITERATURA

[1] Broadband Forum, "TR-100: ADSL2/ADSL2plus Performance Test Plan", 2. izdanje, august 2012.

[2] ITU-T, "G.992.5 Asymmetric digital subscriber line transceivers 2 (ADSL2) – Extended bandwidth (ADSL2plus)", ITU-T Recommendation, april 2009.

[3] www.novkabel.com/katalog/kablovi_i_provodnici_novkabel.pdf, Katalog kablova firme Novkabel, Novi Sad, 2014.

[4] J. Williams, "How Enhanced DSL Technologies Optimize the Last Copper Mile", JDSU Whitepaper, maj 2013.

[5] D. Kagklis, "A comparative Performance Evaluation of the ADSL2+ and ADSL Technologies", GESTS Int'l Trans. Computer Science and Engr., Vol.19, No.1, oktobar 2005.

E-poslovanje zasnovano na internetu u funkciji nabave

Zaim Buljubašić, Nermin Buljubašić, Izudinka Kapetanović Serdarević
BH Telecom.d.d.Sarajevo,RD Tuzla

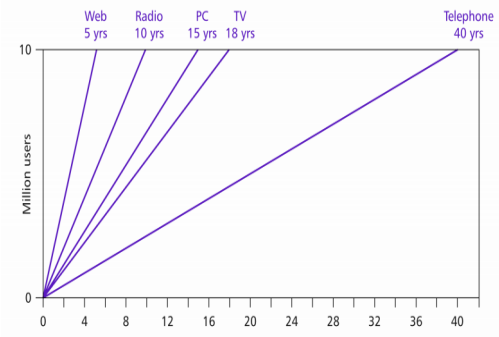
zaim.buljubasic@bhtelecom.ba
nermin.buljubasic@bhtelecom.ba
izudinka.kapetanovic@bhtelecom.ba

Sažetak - Danas tehnologija upravlja cjelokupnim ljudskim životom i radom. Klijenti žele pristup proizvodima i uslugama u bilo koje doba dana. Ulaganje u tehnologiju je neminovno kako bi se našao način za stvaranje novih poslovnih mogućnosti, bržem osvajanju novog tržišta. Internet danas određuje poslovnu strategiju u najvećim svjetskim kompanijama, ali i u onim malim koje su svjesne njegove prednosti, a sve to ima za cilj održivost u biznisu.
Nepostojanje razvijenog sustava za elektroničku razmjenu podataka, poslovnih informacija i transakcija predstavlja glavnu prepreku za poboljšanje poslovanja. Nužno je ulaganje u IT infrastrukturu za E-nabavu koja će pokriti sve faze procesa nabave. Uvođenjem sistema E-nabave znatno bi se uštedjelo vrijeme i novac kroz smanjenje troškova unutrašnje administracije upravljanjem postupkom kupovine putem Interneta.
Prednosti ovog pristupa su ušteda financijskih sredstava zbog pristupa velikom broju dobavljača koji se natječu za posao, te ušteda vremena u procesu nabave. Nedostatak ovog pristupa je rizik koji je sastavni dio poslovanja s nepoznatim dobavljačima, možda čak i inozemnim, pokretanje pitanja o pravnom statusu, troškovi dostave, itd.

Cilj rada je dokazati da E-poslovanje može poduzeću pružiti osjetnu prednost među konkurencijom. Pristupom E-poslovanju učvrstio bi se položaj poduzeća na tržištu, ostvarile bi se nove poslovne mogućnosti i poboljšala stopa dobiti.

I. UVOD

Elektronsko poslovanje se mijenjalo kao što se mijenjao i Internet. Kompanije su na početku Internet koristile isključivo za svoje predstavljanje, ali to danas više nije dovoljno. Nakon faze predstavljanja, počelo se razmišljati o ponudi roba i usluga, te su počele i prve narudžbe putem e-maila ili elektronskih formulara, a na kraju su se razvili B2C (*business to consumer*) i B2B (*business to business*).



Slika 1. Vrijeme potrebno različitim tehnologijama da dostignu milijun korisnika [1]

Posljednjih godina u svijetu se sve više razvijaju elektronske tržnice (*e-marketplace*) koje postavljaju nove smjernice u elektronskom poslovanju, slika 1 [1]. Elektronske tržnice omogućavaju poslovanje između kompanija virtualnom tržnicom gdje kupci, dobavljači, distributeri i prodavači pronalaze i izmjenjuju informacije o dobrima i uslugama u realnom vremenu. Prednosti E-tržnice su u pružanju bolje transparentnosti tržištu pri čemu daju uvid u dostupnost, cijenu, dostavu i uvjete plaćanja proizvoda, te pojednostavljuju način poslovanja na lokalnom, regionalnom, nacionalnom i globalnom tržištu. Kompanije sve više svoje poslovanja sele u sferu virtualnog, a tu uveliko pomažu B2B centri koji danas u svijetu okupljaju hiljade tvrtki i nude kompleksan skup usluga - posredovanje prilikom kupovine i prodaje, iznajmljivanje aplikacija, održavanje aukcija, planiranje i izvršavanje marketinških akcija, sve do prodaje informacija o konkretnoj privrednoj grani. Upravo zbog relativno velike vrijednosti pojedine narudžbine, kao i zbog mogućnosti jednostavnog spajanja kupaca i prodavca (stvaranje tržišta), B2B centri su nosioci rasta cjelokupne B2B ekonomije. Prvi oblici elektronskog poslovanja datiraju od početka sedamdesetih, kada su banke uspostavile sustave elektronskog prijenosa gotovine, dok je primjena u poslovanju i državnoj upravi počela nekoliko godina

kasnije uvođenjem sustava elektronske razmjene podataka (*Electronic Data Interchange* – EDI) bazirane na EDI normama.

II. ELEKTRONSKA DISTRIBUCIJA PODATAKA - EDI

EDI je razmjena poslovnih ili administrativnih poruka (narudžbina, faktura, obavještenja o plaćanju, stanju zalihe), između računala dva ili više poslovnih partnera uz primjenu dogovorenih normi za strukturiranje transakcija ili podataka. Analitičari ocjenjuju da trenutna vrijednost poslovanja preko mreže uporabom EDI iznosi preko 150 milijardi USD [2]. Također, očekuje se da će se ta brojka u sljedeće tri godine udvostručiti. Razlozi primjene EDI-ja su višestruki: direktna razmjena informacija, brža komunikacija, nema potrebe ponovljenog unosa podataka na strani primaoca, manje grešaka, niži troškovi. Korisnici EDI-ja su izračunali da se troškovi pripreme i izrade poslovne dokumentacije u odnosu na klasičan način smanjuju do 3% vrijednost robe u domaćem, a do čak 10% u međunarodnim okvirima. EDI tehnologiju koristi mali broj uglavnom velikih poduzeća i institucija zbog visokih troškova uvođenja (nabava podatkovno – telekomunikacijske opreme, prihvaćanje tehničkih normi, povezivanje računala između poslovnih partnera, obrazovanje korisnika i drugo). Budući da pojedine zemlje, ali ni međunarodno pravo, nisu regulirali takav način poslovanja, poslovni partneri, da bi osigurali obavezujuću snagu poslova zaključenih putem računala, sklapaju interne sporazume o upotrebi EDI-ja u međusobnom poslovanju (*Electronic Data Interchange Agreement*). EDI je zatvorena mreža, komunikacija se odvija između poznatih poslovnih partnera, pa su problemi sigurnosti manje izraženi. Masovnom primjenom Interneta, omogućena je jednostavnija i jeftinija komunikacija između poslovnih partnera, pa tako i male i srednje firme mogu koristiti prednosti elektronskog poslovanja.
Razlozi za primjenu elektronskog poslovanja su:

- niži troškovi poslovanja u području,
- nabavka roba i usluga,
- skladištenje robe,
- marketing,
- bolje usluživanje kupaca,
- prilagođavanje zemlje promjenama u svijetu što je uvjet poslovanja s inostranstvom.

III. POČETAK PRIMJENE INTERNETA U POSLOVANJU

Dolaskom Interneta elektronsko poslovanje je postalo dostupno i najmanjim domaćim firmama, što je dovelo do pravog procvata i eksponencijalnog rasta elektronskog poslovanja. Firme svih veličina sada mogu elektronskim putem komunicirati, i to preko javne mreže (Internet), preko mreža namijenjenih samo kompanijama (Intranet), tj. mreža namijenjenih kompanijama i njihovim poslovnim partnerima (Ekstranet) i preko privatnih mreža. Najvažnija područja elektronskog poslovanja su: poslovanje između kompanija (*business to business*) i firme i krajnjih kupaca (*business to consumer*). Najveći udio poslovanja između kompanija i kupaca zauzima *on-line* kupovina potaknuta stalnim rastom broja korisnika.
Elektronsko poslovanje je na Internetu do sada prošlo nekoliko faza:

- predstavljanje poduzeća na mreži,
- prikaz kataloga robe i usluga,
- prikupljanje zahtjeva kupaca,
- preuzimanje porudžbina,
- prodaja robe stanovništvu, pojedincima (B2C),
- poslovanje između firmi (B2B).

Elektronska trgovina omogućava ljudima *on-line* nabavku proizvoda i razmjenu informacija o poslovnim transakcijama. Iako je primjena Interneta kao platforme za poslovne transakcije relativno nova, Internet je svakako ostvario najzapaženiji uspjeh u odnosu na druge platforme za poslovanje prisutne u posljednjih nekoliko decenija. Elektronska trgovina omogućava kompanijama povećanu efikasnost i pouzdanost poslovnih procesa kroz automatizirane transakcije.
Jedna vrsta gotove poslovne aplikativne programske podrške jesu ERP rješenja. Proizvođači modernih ERP rješenja su uočili postojanje poslovnih funkcijskih područja koja su zajednička nekoj skupini poslovnih djelatnosti (npr. proizvodnja, uslužne djelatnosti, financijski sektor, javni sektor itd.), te su tako dobivena opća poslovna aplikativna programska rješenja primjenjiva u različitim poslovnim djelatnostima (opće ERP rješenje). No, poslovne djelatnosti, bez obzira da li pripadaju istoj ili različitim skupinama, imaju i neka specifična područja koja u općem ERP rješenju nisu podržana (tada se govori o vertikalizaciji ERP rješenja). Vertikalizacija ERP rješenja se može definirati kao postupak proširenja, dorade ili prilagodbe općeg ERP rješenja specifičnostima neke poslovne djelatnosti. Rezultat tog postupka je specijalizirano ERP rješenje koje podržava poslovne procese konkretne poslovne djelatnosti.

IV. STRATEGIJA E-POSLOVANJA

Strategija predstavlja skup aktivnosti koje upućuju kako ostvariti najočitije ciljeve poduzeća. U globalnim relacijama figuriraju se neki pojmovi koje je potrebno opisati i definirati:

A. Misija poduzeće

Predstavlja poseban, autonomni segment u okviru sustava strategijskih ciljeva poduzeća. Misija je najočitiji strategijski cilj koji prožima sve aktivnosti poduzeća na svim nivoima poslovanja. Misija nema naglašenu operativnu funkciju već određuje globalno usmjerenje i

daje odgovor o svrsi postojanja ili se ponekad naziva i vrhovnom strategijom .

B. Kredo (slogan)

Misija može biti eksplicitno izražena pomoću slogana koji osigurava prepoznatljivost poduzeća i njegovih proizvoda u okruženju. Na osnovu ispitivanja 269 japanskih poduzeća slogane su najčešće činili sljedeći pojmovi:

- dobar proizvod, razumna cijena,
- prosperitet, stalni napredak,
- razvijanje povjerenja s*takeholdera*,
- blagostanje zaposlenih.

C. Korporativna strategija

Odnosi se na cjelokupnu svrhu i obim poslovanja korporacije.

D. Strategija poslovne jedinica

Poslovna jedinica kao dio većeg sustava mora imati svoju strategiju koja daje odgovore kako se uspješno takmičiti na određenom tržištu.
Ostaje pitanje gdje je mjesto i koliki je značaj strategije elektronskog poslovanja. Postojala su gledišta da je strategija elektronskog poslovanja jedna vrsta funkcionalne strategije ili da je ona dio strategije podatkovnih sustava. Uspjeh Internet kompanija doprinijelo je da se strategija elektronskog poslovanja sve više može smatrati dijelom korporativne strategije.

V. E-POSLOVANJE U FUNKCIJI E-NABAVE

Uspješnost poslovanja i sposobnost stvaranja vrijednost u tvrtkama usko su povezani s razvojem naprednih informacijskih sustava. Osobna računala, Internet i mobilna telefonija iz temelja su promijenili životna kretanja, uključujući i način rada, obrazovanje i sve oblike interakcije. Temeljna strategija suvremenog uspješnog poslovanja je E-poslovanje (*E-business*), a unutar E-poslovanja tehnologije se nalazi E-nabava. Razmjena ideja, vizija, znanja i iskustava nikada nije bila jednostavnija, brža i jeftinija. Prednosti koje svakoj organizaciji donosi međusobno povezivanje računalne opreme čini informatičke prometnice nezamjenjivim dijelom poslovne infrastrukture. Približavanje biznisa i tehnologije uspješno je u uvjetima kada je tehnologija shvaćena kao pogonsko gorivo, a tržište kao vodilja biznisa. Tehnologija pomaže u promjeni društva. Pametna informatička mreža je funkcionalna tehnologija u službi unapređenja poslovanja, drugim riječima, inteligentna mreža je sjedište sustava koji unapređuje poslovanje. Internetska tehnologija današnjice jedno je od najmoćnijih čovjekovih dostignuća. Primjena Interneta kao poslovnog rješenja u kompanijama dovodi do toga da se one odvajaju od konkurencije i postaju vodeće na tržištu. Izazovi za implementaciju E-nabava poslovnih rješenja su:

- osigurati održivi dramatični rast poslovanja,
- osigurati visokokvalitetnu podršku korisniku,
- stvoriti i zadržati vodeću poziciju u industriji i na tržištu.

E-nabava je B2B (*business to business*) sustav izvršavanja svih elemenata poslovnih aktivnosti ekonomskog procesa elektroničkim putem. U E-nabavi dvije ili više registriranih kompanija komuniciraju web servisima kako bi se naručio proizvod bez nepotrebnih odlazaka u zastupništva kompanija, već naručivanjem preko weba. E-nabava je poslovna transformacija temeljena na:

- Udruživanju poduzeća (integracija),
- Procesu suradnje (kolaboracija),
- Globalnom mrežnom povezivanju,
- Upotrebi Interneta kao medija.

Internetsko i elektroničko poslovanje preko E-nabava sustava omogućuje:

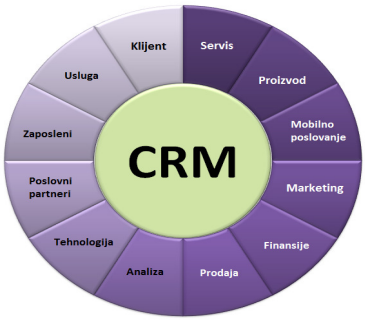
- Povezivanje dobavljača,
- Brzu, trenutnu razmjenu informacija,
- Značajno snižavanje troškova, a time i cijena.

VI. PRENOSTI I NEDOSTACI E-NABAVE

Prednost E-nabave je u tome što ono može poboljšanim marketingom povećati prodaju, kao i veličinu prodajnog programa, te smanjiti troškove vezane za upite kupca. E-poslovanje proširuje tržište poslovanja i time povećava različite mogućnosti kupovine. Kupcima su informacije o proizvodima i uslugama dostupnije, imaju veći izbor proizvoda po nižim cijenama, a kupovina im postaje praktičnija i prilagodljivija zbog dužeg radnog vremena, tj. dostupnosti proizvoda 24 sata dnevno, sedam dana u tjednu. E-nabavom smanjuju se transakcijski troškovi kojima su izloženi i kupci i proizvođači. Ono smanjuje transakcijske troškove tako da poboljšava tok informacija i poboljšava koordinaciju aktivnosti. Jednom kad je prednost E-nabave prepoznata, unutar pojedine kompanije, najvažnije što menadžeri i tehničko osoblje trebaju razmotriti jest infrastruktura internetski zasnovanog poslovanja, od pojedinačnog internetskog poslužitelja što je povoljno za manju kompaniju, pa sve do sustava informacijski intenzivne *on-line* obrade transakcija kakve grade velike zrakoplovne kompanije i turističke agencije. Nedostaci E–poslovanja i E-nabave se kriju u prebrzom mijenjanju tehnologije i međunarodnim kulturološkim i zakonskim poteškoćama. Također, neki proizvodi nisu prikladni za ovaj vid prodaje (lako kvarljivi proizvodi). E-nabava je vizija uspješnog poslovanja budućnosti. Poslovanja u kojem dominira suvremena tehnologija povezana s poslovnom strategijom kompanije koja ju implementira. Smatram da je E-nabava esencijalno za kompaniju koja želi svoje proizvode prodavati diljem svijeta i tako proširiti granice svog poslovanja bez nepotrebnih troškova.

VII. CRM

Razvoj informatičkog društva, koje se bazira na povećanju pristupačnosti znanju, informacijama i servisima putem različitih mreža i konvergenciji različitih medija, uzrokuje značajne promjene u poslovanju. Informatička i komunikativna infrastruktura ima poseban značaj za unapređenje poslovanja, jer omogućava pristup udaljenim tržištima, veliko povećanje brzine poslovnih transakcija, značajno smanjenje troškova poslovanja, kontakt s korisnicima i personalizaciju usluga. Tehnologija, informacije i znanje postali su odlučujući faktori za konkurentnost kompanija u svim sektorima poslovanja.



Slika 2. Model CRM [6]

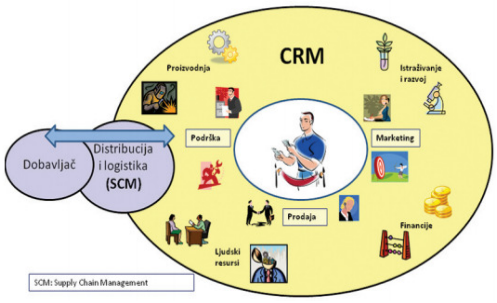
Elektronsko poslovanje omogućava povećanje efikasnosti rada kompanija, a samim tim povećava njihovu konkurentnost na tržištu. Elektronsko poslovanje također omogućava i stvaranje novih kompanija utemeljenih na suvremenoj podatkovnoj i komunikativnoj tehnologiji. Za razvoj elektronskog poslovanja potrebno je stvoriti podatkovnu i komunikativnu infrastrukturu kao i zakonske pretpostavke za elektronsko poslovanje.

Tijekom proteklog perioda, upravljanje odnosima s korisnicima (*Customer Relationship Management*) je postao jedina popularna inicijativa i visoko rangiran u procesu konzultiranja kompanija. Utvrđeno je da tri četvrtine kompanija koriste usluge CRM, odnosno uvode ili planiraju CRM rješenja. Usprkos svojoj popularnosti CRM i dalje ostaje enigma za mnoge kompanije. To je uglavnom zato što mnogi direktori kompanija, menadžeri, dobavljači software-a, industrijski analitičari i konzultanti za menadžment imaju mnoštvo pogleda na CRM, svaki oblikovan prema njihovim specifičnim stanovištima, situacijama, iskustvima i naravno, motivima. U skladu s tim, o CRM-u se može razmišljati, diskutirati, i prihvatiti ga, na tri različita nivoa: tehničkom, funkcionalnom i filozofskom.

CRM sve više stavlja na prvo mjesto analizu korisnika, a onda razrađuje neophodne funkcije u smislu podrške, bez obzira o kojem se poslovnom nivou radi. Iako je ovaj pristup bolji od onog tehnološki usmjerenog, ovaj funkcionalni pogled ne može da obuhvati suštinu CRM- a u potpunosti.

Danas je CRM uglavnom prihvaćen kao poslovna filozofija koja promatra osnovnu suštinu korisnika, ili preciznije podatke o njemu. Ona predstavlja fundamentalni preokret u načinu na koji kompanije obavljaju posao. Preokret u perspektivi je zabilježen i na globalnom planu poslovanja - prelazak s usmjerenja ka proizvodu na usmjerenje ka korisniku.

Mnoge konzultantske kompanije i industrijski analitičari zalažu se za ovu filozofsku perspektivu CRM-a. Oni reorganiziraju organizacije kako bi se ostvarila dugoročna vrijednost CRM-a da se prihvati strategija koja se odnosi na poslovanje u cjelini, a teže da se organizacija okrene pristupu na nivou kompanije. U stvari, Gartner je definirao pravi CRM kao: „ *Poslovnu strategiju koja maksimalizuje profitabilnost, zaradu i zadovoljstvo korisnika tako što se organizira oko korisničkih segmenata, njeguje ponašanje koje zadovoljava korisnika i sprovodi procese usmjerene ka korisniku.*“



Slika 3. Kupac u ishodištu svih poslovnih procesa [5]

CRM i njegova primjena, ako se gleda sa filozofskog stajališta u biti je primjenjiv i u sferi nabave, nekako bi se moglo reći po principu povratne sprege. Ovaj put kupac primjenjuje skoro isti algoritam na kategoriju dobavljača, stvarajući sebi povoljnije inpute i uvećavajući svoju konkurentnost.

VIII. CRM PROCESI U POSLOVNOM OKVIRU IT KOMPANIJE

Važan aspekt CRM-a je njegova orijentacija na korisničke procese (ili životne cikluse korisnika). Korisnički proces je niz aktivnosti koje korisnik mora da obavi da bi zadovoljio potrebu ili riješio specifičan problem. Zato, korisnički proces odlučuje koji su to potrebni proizvodi i usluge koje kompanija mora da ponudi da bi potpuno pokrila korisnički proces. Zato što najveći broj kompanija ne može potpuno da pokrije korisnički proces, ili ne želi da ga pokrije zbog fokusa na neke suštinske kompetencije, neophodna je suradnja s ostalim kompanijama koje nude komplementarne proizvode i usluge. Među njima razlikuju se tri kategorije CRM procesa:

- ***CRM procesi isporuke***: Procesi direktnog kontakta s korisnicima koji su osmišljeni tako da pokrivaju dio korisničkog procesa (upravljanje kampanjom, upravljanje prodajom, upravljanje uslugom, kontrola žalbe),

- ***CRM procesi podrške:*** Procesi direktnog kontakta s korisnicima koji nisu osmišljeni da pokriju dio korisničkog procesa, već da ispune funkcije podrške u okviru CRM konteksta. (istraživanje tržišta, kontrola lojalnosti),

- ***CRM procesi analize*****:** Procesi koji konsoliduju i analiziraju saznanja o korisnicima koja su sakupljena u okviru drugih CRM procesa.

Rezultati analize se prebacuju na CRM procese isporuke i podrške, ali se koriste i za inovaciju usluge i procese unapređenja usluge zbog efektivnosti (uvećanje broja korisnika, smjernice, profiliranje i segmentacija korisnika, povratna sprega (*feedback*) i upravljanje znanjem).

IX. ODLIKE CRM-A U IT SEKTORU

CRM predstavlja složeni skup poslovnih procesa i tehnologija za upravljanje relacijama s postojećim i potencijalnim korisnicima i poslovnim partnerima, u marketingu, prodaji i podršci, preko svih raspoloživih kanala komunikacije. Osnovni cilj CRM-a je da optimizira zahtjeve korisnika i partnera, uveća učešće na tržištu, profit i poslovnu efikasnost, gradeći što je moguće snažnije relacije na nivou organizacije.
Informatički sektor je primjenom suvremenih tehnologija stvorio novi vid poslovanja. Kompanije su stvorile poslovni model koji se sastoji od:

- načina segmentiranja i izbora korisnika,
- diferenciranja svojih ponuda,
- oblikovanja svojih resursa,
- nastupanja na tržištu i
- stvaranja vrijednosti baziranih na strateškom razumijevanju prioriteta i sklonosti korisnika.

Stvoren je strateški koncept koji omogućava top menadžmentu razumijevanje i predviđanje vrijednosti poslovanja. Prednost je razumijevanje potreba korisnika i stvaranje inovativnog poslovnog modela koji je u stanju da zadovolji te potrebe - CRM.

Korištenje prednosti CRM sustava uvjetovano je:

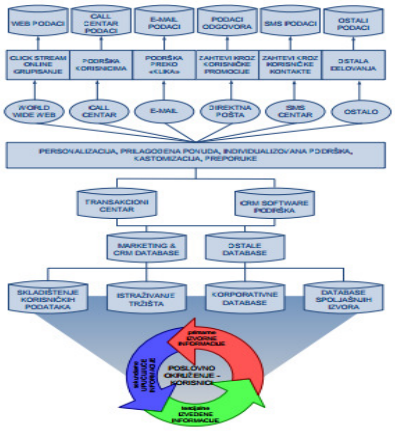
- prikupljanjem i skladištenjem podataka iz kompanije,
- analizom prikupljenih podataka i pronalaženjem pravila ponašanja korisnika,
- uspostavljanjem kontakt centara (telefon, fax, E-mail, SMS, MMS, chat, itd.) između kompanije i korisnika.

Upravljanje odnosa s korisnicima baziran na principima primjene CRM-a nalazi primjenu u mnogim kompanijama. Kompanije koje generiraju ili primaju mnoštvo podataka sigurno su one u kojima se CRM može iskazati na pravi način. Potrebno je prikupiti, skladištiti, analizirati i generirati informacije potrebne za usavršavanje proizvoda/usluga koje se pružaju korisnicima. Principi CRM-a u oblasti elektronskog poslovanja daju veliki efekt jer se svaka konekcija/transakcija odvija preko Interneta i ostavlja određene podatke koji imaju šansu da se pretvore u informaciju. Podatak postaje informacija samo ako u sebi sadrži upotrebljivost.

X. OPERATIVNI MODEL CRM STRATEGIJE

CRM sustavi se mogu klasifikovati u sljedeće tri kategorije:

- Operativni CRM sustavi unapređuju efikasnost CRM procesa isporuke i podrške. Oni obuhvaćaju solucije za marketing, prodaju i automatizaciju usluge, na primjer: sustavi i za upravljanje kampanjom i sustavi za interaktivne centre za korisnike.
- Analitički CRM sustavi skladište i vrednuju znanje o korisnicima zbog boljeg razumijevanja svakog korisnika i njegovog ponašanja. Stoga, oni pružaju podršku CRM analizi. Primjeri su skladištenje podataka, analitičko *on-line* procesuiranje (*OLAP*) i sustavi za „ rudarenje“ podataka.
- Kolaborativni (*collaborative*) CRM sustavi kontroliraju i sinhronizuju točke korisničke interakcije i kanala komunikacije. CRM sustavi suštinski procesuiraju dobro strukturirane informacije, na primjer podatke o kontaktima s korisnicima i master podatke. Suprotno tome, sustavi za upravljanje znanjem pružaju podršku zbirci manje strukturiranih informacija, razmjeni i primjeni kao što su dokumenti (eksplicitno znanje) i implicitno znanje zaposlenih. Ovo znanje je posebno neophodno u okviru CRM procesa isporuke radi zadovoljenja potreba korisnika. Na primjer, upravljanje prodajom zahtjeva detaljno znanje o dostupnim proizvodima da bi konsultovanje o ponudi prema korisnicima bila efektnija u odnosu na njihove potrebe.



Slika 4. Operativni model CRM strategije [10]

XI. ZAKLJUČAK

Zahvaljujući E-poslovanju mijenjaju se poslovni modeli, tradicionalni odnos kupac-prodavač se polako napušta, te kupci polako počinju sudjelovati u kreiranju novih roba i usluga, odnosno, počinje se koristiti interaktivnost u poslovanju. Zahvaljujući Internetu, danas se sve više i više koristi *on-line* kupovina, tako tvrtke moraju biti u korak s vremenom i pokušati naći rješenja kako zadržati kupca i kako mu omogućiti poseban osjećaj kupovine.

Povećana je brzina i reakcija tvrtki na globalne promjene zahvaljujući E-poslovanju. Globalno tržište je postalo mjesto za tržišno natjecanje, a udaljenost je prestala biti presudan faktor u poslovanju. Glavne prednosti E-poslovanja su mogućnost prodaje roba i usluga, suradnja s partnerima 24 sata na dan, te se tako omogućava veća konkurentnost na tržištu.

U Bosni i Hercegovini postoje sustavi elektronskog plaćanja koji se obavljaju preko banaka. Nedovoljna informiranost u elektronske sustave plaćanja kod građana izazivaju određeno nepovjerenje i dozu straha od prevara na globalnoj mreži.

Kada je u pitanju efikasno obavljanje većine poslovnih procesa u poduzećima potrebno je da se izvrši i njihovo planiranje. Danas u svijetu velike kompanije imaju ugrađene sustave koji im omogućavaju da planiraju i organiziraju svoje poslovne procese i da na osnovu dobre baze podataka mogu donositi i dobre poslovne odluke. Osnovna karika koja osigurava ICT podršku poslovnih procesa je upravo poslovni informacioni sustav kompanije. Primjenom E-trgovine, elektronskog sustava plaćanja i planiranja poslovnih procesa može unaprijediti uvjete poslovanja kompanija u Bosni i Hercegovini.

CRM sustav pojednostavljuje zatvorenu petlju marketinga u kojoj se podaci o korisniku transformiraju u informacije korisne za sustav. Ove informacije transformiraju se u marketing strategiju i strategiju upravljanja korisnicima. CRM strategija u informatičkom sektoru je niz aktivnosti koje se obavljaju da bi kompanija potvrdila i sprovela vlastitu korisničku strategiju, fokusiranje na identifikaciju, prioritetnost i implementaciju novih biznis mogućnosti okrenutih ka korisniku i praksi koja će poboljšati operativnost i segmente profitabilnosti. Informatički sektor uz primjenu operativnog modela CRM strategije sprovodi ciljane akcije usmjerene na zadržavanje postojećih i privlačenje novih korisnika korištenjem informacija u smjeru uobličavanja proizvoda i usluga prema potrebama raznih profila korisnika s različitim ustaljenim navikama.

Uvođenjem CRM-a u IT sektor kompanije dobivaju korporativnu prednost u odnosu na svoju konkurenciju u pogledu efikasnosti kompanije s korisnicima.

LITERATURA

[1] www.emarketer.com

[2] B. Latinović, "Elektronsko poslovanje", Banja Luka, 2007.

[3] M.Milosavljević,G.Grubar,M.Veinović, "Poslovna informatika", Univerzitet Singidunum, Beograd, 2009.

[4] "Elektronsko poslovanje", Banja Luka, Septembar 2010.

[5] M. Perčić, „Poslovna funkcija nabave", Veleučilište u Varaždinu, Varaždin 2011.

[6] B. Žilbert, „ Strateška nabava" , Mate d.o.o, Zagreb 2007.

[7] J. Dyche, "The CRM Handbook: A Business Guide to Customer Relationship Management", Pearson Education Corporate Sales Division, USA, 2001.

[8] D.K. Rigby, F.F. Reichheld, F.F., P. Shefter, " Avoid the Four Perils of CRM", Harvard Bussiness Review, 2002.

[9] D. Peppers, M. Rogers, " Managing Customer Relationship, John Wiley and sons, Inc, Hoboken, New Jersey, 2004.

[10] J. Reynolds, " A Plactical Guide to CRM, CMP Books", New York,2002.

e-plaćanje u uslovima ne postojanja potpunog sistemskog rješenja e-poslovanja

Zaim Buljubašić, Nermin Buljubašić, Izudinka Kapetanović-Serdarević
BH Telecom.d.d.Sarajevo,RD Tuzla

zaim.buljubasic@bhtelecom.ba
nermin.buljubasic@bhtelecom.ba
izudinka.kapetanovic@bhtelecom.ba

Sažetak: U Bosni i Hercegovini kao i većini drugih zemalja Jugoistočne Evrope ne postoji sistemski, niti pravno oubličen sistem funkcionisanja e-poslovanja, a dinamičnost savremenih tržišta i sistema poslovanja primjenu takvog koncepta apsolutno zahtijeva. Tako dio poslovnih subjekata pogotovu onih koji su izloženi jačoj konkurenciji kako na domaćem, tako i na globalnom međunarodnom tržištu su primorani tražiti neka hibridna rješenja, koja u temelju se baziraju na uvažavanju međunarodnih standarda i preporuka usvojenim od strane kompetentnih institucija iz ove oblasti.

Elektronsko plaćanje je uvedeno u više banaka u Bosni i Hercegovini, ali ne u svim i koriste ga samo pojedini poslovni subjekti. Rješenja koja se najčešće nude su vezana za elektronsko (bankarstvo) plaćanje u 10 zemalja, uštedu vremena, veću kontrolu novčanih sredstava, jer odgovorno osoblje ima pristup računu 24 sata na dan i na taj način može kontinuirano i odmah pratiti sve prilive i odlive sa računa, smanjenje troškova (papira, tonera, goriva), pregled izvoda. Benefiti koje donosi primjena ovakvog sistema plaćanja također se odnose na: veći nivo finansijske discipline, bolju komunikaciju sa poslovnim partnerima i sl.

Ključne riječi: e-poslovanje, e-plaćanje, hibridni sistemi, standardi, ICT

Uvod:

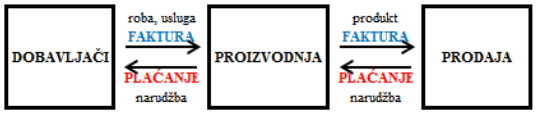
Ekonomska globalizacija sa svim svojim posljedicama, kako u organizaciji nacionalnih i internacionalnih regionalnih, svjetskih tržišta, tako i u uspostavljanju jačih, širih, i u većem broju kooperativnih proizvodnih procesa stvorila je pretpostavku primjene visokih tehnologija, kompjuterizovanih sistema proizvodnje, skraćenje životnih ciklusa proizvoda i značajnog rasta pomoćnih i uslužnih aktivnosti. Sve te promjene u ambijentu i novom načinu poslovanja zahtijevale su promjene u organizacijskoj strukturi koja je do tih vremena bila glomazno postavljena sa krutom vertikalnom hijerarhijom u menadžerskom odlučivanju.

Novi razvoj je zahtijevao fleksibilnu organizaciju sa horizontalnom strukturom odlučivanja, jer u uslovima globalne konkurencije pravovremeno se može reagirati na tržištu samo pod tim uslovima. Proizvodni sistemi koji su do tada bili pravi divovi, a nisu blagovremeno izvršili restruktuiranje svojih poslovnih procesa, vrlo brzo su počeli da na tržištu gube svoju poziciju. Restruktuiranje privrede u svim djelatnostima i na svim nivoima je nužnost, kako bi se blagovremeno moglo odgovoriti na sve prefinjenije zahtjeve potrošača. Nove informaciono komunikacijske tehnologije su sastavni dio poslovnog instrumentarija i važna osnova za kvalitetno odvijanje ovih procesa. Sa tako razvijenom organizacijom postavljenoj na znatno fleksibilnijim odnosima, uspostavljaju se odnosi kooperacije i partnerstva. Kooperacija i partnerstvo se traže i kod drugih sistema, najčešće na relaciji dobavljač-proizvođač-kupac, kako bi se na tržištu ojačale pozicije i ostvario maksimalan profit. Bez IK tehnologija to bi bilo izuzetno teško, moglo bi se reći nemoguće. Kompjuterski potpomognuta proizvodnja rezultira proizvodima različitog asortimana uz uvažavanje srodnih dizajnerskih rješenja. Vrijeme prelaska sa jednog dizajna na drugi, odnosno vrijeme podešavanja mašina je kratko i sastoji se samo od identifikacije programa.

«Svako u organizacionoj strukturi mora komunicirati i sarađivati sa svima na nivou jedne organizacione ravni, kao i između organizacionih ravni»[1]. Kooperativni odnosi se mogu razvijati između firmi iste djelatnosti ili različitih djelatnosti. Ako se govori o odnosima na relaciji proizvođač-dobavljač, onda

[1] Novičević, B.: Globalizovani poslovni ambijent i izazovi upravljačkog računovodstva, Zbornik radova, Revicon, 2003 godina, strana 29.

proizvođač zajedno sa dobavljačem želi da kreira svoje inpute u pogledu kvaliteta, mjesta i uslova nabave, cijene, vremena isporuke i sve druge značajne elemente. Takođe prodavci na tržištu najbolje znaju šta žele potrošači, pa zajedno sa proizvođačima kreiraju proizvod koji tržište prihvata.[2]



I u jednoj i u drugoj interakciji, između proizvodnje i dobavljača i proizvodnje i prodaje, javljaju se vrlo kompleksni odnosi koji svakim danom zahtijevaju saradnju i razmjenu najnovijih znanja, iskustava i informacija na velikom broju frontova. Ono što je zajedničko za obje interakcije je ispostavljanje računa (faktura) i izvršenje potrebnih plaćanja. Kako ove vrlo značajne funkcije u ukupnom sistemu poslovanja digitalizirati i po tom osnovu ostvariti konkurentske prednosti, u ovom radu će biti izvršen pokušaj davanja određenih prijedloga, pogotovu u uslovima kad koncept e-poslovanja nije sistemski zaživio. Koje su prednosti uvođenja e-računa i e-plaćanja biće prezentirano na istraživanjima u nekim kompanijama u BiH.

1.Pravni okvir za e-poslovanje, odnosno e-plaćanje

Informacione tehnologije su svojim inovacijama izvršile redefinisanje poslovnih procesa na taj način što su smanjile neefikasnost u radu unutar samih privrednih subjekata, kao i neefiksasnost u odnosima sa drugim poslovnim subjektima. U današnje vrijeme firme se suočavaju sa stalnim pritiskom da poslove obavljaju bolje, brže, jeftinije, dok održavanje visokog nivoa rezultata i dalje ostaje prioritet. U aktuelnim odnosima trgovinsko poslovnih saradnji, standardni poslovni proces je jednostavno ne adekvatan, pa je zbog toga sklapanje ugovora preko Interneta ne zaobilazan činilac za izgradnju novih i boljih poslovnih odnosa.
Međutim, ubrzan razvoj informacionih tehnologija i razvoj Interneta kao medija između subjekata u poslovnim procesima otvorio je novo pitanje kada je ugovorno poslovno pravo u pitanju. To je uslovilo donošenje Direktive od strane Evropske unije o elektronskim potpisima. Pomenuta Direktiva predstavlja pravno utemeljenje elektronskog potpisa. Na osnovu nje su donijeti zakoni o elektronskom potpisu u svim zemljama Evropske unije, kao i u većini ostalih zemalja Evrope.
Dvije najvažnije stvari koje trebaju biti istaknute u Zakonu o elektronskom potpisu su propisivanje uslova pod kojima je elektronski potpis pravno ekvivalentan svojeručnom potpisu i uslova koji moraju da ispune certifikacijska tijela koja izdaju kvalificirane certifikate za verifikaciju kvalificiranih elektronskih potpisa. Zakonom o elektronskom potpisu jasno treba biti napravljena razlika između elektronskog potpisa i kvalificiranog elektronskog potpisa što predstavlja bitno razgraničenje odlučujućih elemenata pri korišćenju ovog načina poslovne komunikacije. Ako želimo da objasnimo šta je elektronski potpis onda dolazimo do zaključka da je to skup podataka u elektronskom obliku, koji su pridruženi ili su logički povezani sa elektronskim dokumentom i oni služe za identifikaciju potpisnika, a dok je kvalificirani elektronski potpis onaj elektronski potpis kojim se pouzdano garantuje identitet potpisnika**,** integritet elektronskih dokumenata i onemogućava naknadno poricanje odgovornosti za njihov sadržaj i koji ispunjava uslove utvrđene Zakonom o elektronskom potpisu.
Tekst Zakona o elektronskom potpisu tačno navodi sve uslove koje potpis u elektronskom obliku treba da zadovolji da bi se smatrao kvalificiranim elektronskim potpisom i da bi imao isto pravno dejstvo i dokaznu snagu kao i svojeručni potpis, odnosno svojeručni potpis i pečat, u odnosu na podatke u papirnom obliku. On upućuje da je potpis povezan sa potpisnikom i da ga sa sigurnošću identifikuje.
Digitalni potpis (digital signature) je elektronska verzija potpisa, na osnovu kog se **i**dentifikuje pošiljalac i dokazuje vjerodostojnost poruke. Može se shvatiti kao način obezbjeđivanja servisa integriteta u smislu potvrde ispravnosti ili tačnosti. Baziran je na asimetričnom kriptografskom modelu ili šifrovanju sa javnim ključem**.** Osnovni dijelovi tog kriptografskog modela su šifrovanje, dešifrovanje i ključ.
Šifrovanje se sastoji od transformacije čitljivog teksta podataka u oblik nečitljiv za onoga kome taj tekst nije namijenjen, a dešifrovanje je postupak vraćanja šifrovanog teksta u čitljiv oblik. Javni ključ bi predstavljao početnu vrednost algoritma kojim se vrši šifrovanje. U ovakvom modelu šifrovanja poruke, sam postupak šifrovanja teksta može načiniti bilo ko, bitno je da tekst može pročitati ili dešifrovati jedino osoba koja ima odgovarajući privatni ključ. Jedino osoba sa privatnim ključem može digitalno potpisati neki tekst, dokument, a svako ko ima javni ključ može proveriti ispravnost potpisa. Postoji set osnovnih normativnih akata, prije svega zakona koji uređuju ovu oblast. Tako npr. R Hrvatska ima solidna

[2] Novičević, B.: Globalizovani poslovni ambijent i izazovi upravljačkog računovodstva, Zbornik radova, Revicon, 2003 godina, strana 29.

rješenja koja se očituju kroz slijedeće zakone: Zakon o elektroničkoj trgovini, Zakon o elektroničkom potpisu, Zakon o elektroničkoj ispravi, Zakon o tajnosti podataka, Zakon o informacijskoj sigurnosti, Zakon o zaštiti osobnih podataka, itd. Treba također istaći da su svi ovi zakoni u znatnoj mjeri usaglašeni sa rješenjima u EU. To je logično jer EU želi jedinstven informacijski prostor, pa se s razlogom može preporučiti i drugim zamljama koje pretenduju i streme ka članstvu u ovoj asosijaciji da blagovremeno vrše ta usaglašavanja. U čemu je onda problem? U ne dovoljnoj afirmaciji i primjeni koncepta e-poslovanja, a posebno e-plaćanja koje nudi prostor ogromnih ušteda u obavljanju ove vrlo značajne funkcije u ukupnim poslovnim procesima i propuštanju šanse ubrzavanja ovih procesa i njihove kvalitetne integracije na aplikativnom nivou sa ostalim procesima.

2. e-račun i e-nalog za plaćanje

E-fakture se iz informacionog sistema izdavaoca (ERP), putem e-banking kanala, prenose u e-banking rješenje primaoca (kupca). Primalac, bilo da je u pitanju pravno ili fizičko lice, dobija e-fakturu u svoje e-banking rješenje, što mu omogućava da samo jednim klikom prenese podatke iz e-fakture u platni nalog kojim vrši plaćanje.

Izdavalac u svom informacionom sistemu (ERP) priprema fakturu u elektronskom obliku i uvozi je u svoje e-banking rješenje. Iz e-banking programa, baš kao i platne naloge, šalje fakturu svojoj banci putem e-banking kanala.

Banka izdavaoca prima e-fakturu. Na osnovu tekućeg računa primaoca i BIC koda, koji služe kao njegova adresa, banka izdavaoca, preko e-banking kanala, šalje e-fakturu odgovarajućoj banci korisnika. Za tu svrhu nije potrebno da izdavalac i primalac imaju otvorene račune u istoj banci. Banka može, na osnovu finansijskog ugovora sa izdavaocem, zajedno sa arhivistom da proslijeđuje izdate e-fakutre u arhiv '1-2-3 U ARHIV!' .

Banka primaoca prima e-fakturu. Na osnovu tekućeg računa primaoca i BIC koda banke, banka primaoca šalje dalje fakturu u e-banking rješenje primaoca. Primalac može da sklopi sa arhivistom poslovni ugovor o elektronskom arhiviranju prenijetih e-faktura. Na taj način, slanjem e-faktura primaocu, banka istovremeno prosljeđuje fakturu i u arhivu.

Primalac prima fakture u svoje e-banking rješenje. U slučaju da su primaoci pravna lica, oni imaju mogućnost uvoza e-faktura u svoje informacione sisteme (ERP) za potrebe računovodstvenih evidencija. Osim toga, na osnovu e-fakture primaoci mogu u e-banking rješenju samo jednim klikom kreirati platni nalog sa podacima iz fakture. Platni nalog se vraća istim putem izdavaocu e-fakture, koji zajedno sa prilivom dobija ispravne i odgovarajuće podatke o uplati (referenca).

Primjer iz prakse zemlje članice EU: Jedna informatička kuća je, u saradnji sa naručiocem, razvila cjelovitu informatičku podršku za potrebe pet slovenačkih javnih preduzeća za distribuciju električne energije. Značajan dio cjelokupnog sistema je podsistem za obračun isporučene električne energije korisnicima, među kojima je 100.000 poslovnih korisnika i više od 800.000 domaćinstava. Kod tako velike količine faktura koje mjesečno šalju korisnicima veoma je značajna optimizacija postupka slanja faktura. Vlastiti sistem za pripremu, izdavanje i slanje e-faktura je veoma zahtjevna investicija. Njihov izazov bio je pronalaženje načina uvođenja što efikasnijeg sistema izdavanja faktura koji bi pojednostavio i pojeftinio proces distribucije.

Poslije razmatranja različitih mogućnosti i opcija, Informatika se odlučila za infrastrukturu e-banke sa kojom je već upoznata većina slovenačkih preduzeća. Elektronsku banku već upotrebljava oko 80% slovenačkih preduzeća, koja su ujedno i korisnici električne energije. Pored pravnih lica korisnika, elektronsku banku – u znatno manjem obimu, ali u većem broju – upotrebljavaju i domaćinstva. Postojeća sigurna i stabilna infrastruktura e-banke može da se koristi i za slanje elektronskih faktura. »Infrastruktura je već na raspolaganju, korisnici je upotrebljavaju i veoma dobro poznaju. To svakako skraćuje vrijeme i olakšava implementaciju izdavanja e-faktura. Pored toga infrastrukturu za e-bankarstvo održavaju banke i za to mi ne moramo brinuti« rekao je Franci Klauzner, stručni savjetnik direktora sektora za razvoj. Oni korisnici koji nemaju e-banku, a željeli bi uprkos tome da primaju račune za utrošenu električnu energiju u elektronskom obliku, Informatika šalje e-fakture na svoj portal e-Usluge (e-Storitve), preko koga korisnici mogu da pregledaju i mijenjaju određena mjesta. U Informatici su se za slanje e-faktura putem kanala e-banke odlučili iz više razloga. Prije nego što pošalju fakturu u elektronskom obliku, oni moraju od primaoca da dobiju saglasnost za to. Primaoci tu saglasnost mogu da daju putem portala Informatike ili svoje e-banke. Sistemom e-banke se može obezbijediti da elektronske fakture izdavača, nakon slanja primaocu, budu usmjerene u

elektronski arhiv. E-fakture, koje izdaje Informatika putem sistema e-bankarstva, su digitalno potpisane i imaju digitalni vremenski pečat pa su, time, ispunjeni svi zakonski uslovi i kao takve su spremne za arhiviranje. I kakav je postupak slanja računa? Informatika u svom sistemu priprema digitalno potpisanu e-fakturu u određenom obliku (odlučili su se za e-SLOG) i zajedno sa dokumentom u PDF formatu uvozi u sistem e-banke, koji sigurnosnu komunikaciju bazira na digitalnom potpisu. Po Zakonu o elektronskom poslovanju i elektronskom potpisu (ZEPEP) e-faktura je jednaka papirnoj ako je digitalno potpisana. Primalac prima e-fakturu u svoju e-banku i, ako je pravno lice, uvozi je direktno u svoj informacioni sistem. Svi primaoci, kako pravna lica tako i domaćinstva, nalog za plaćanje mogu pripremiti u e-banci ili automatski iz e-fakture. Primalac e-fakture je oslobođen naknadnog prekucavanja podataka, a izdavač fakture uplate prima sa pravilnim povratnim informacijama o izvršenom plaćanju. Pravilni podaci, kao što su 13-cifreni poziv na broj i iznos plaćanja su ključni pri usklađivanju primljene uplate i izdate fakture.[3]

3. Mogući efekti primjene e-plaćanja i e-računa u BiH

Analizirajući ova i druga slična rješenja u praksi, proučavajući mogućnost njihove implementacije makar kao hibridnih rješenja u realnu praksu u BiH, intervjuišući veći broj informatičkih kuća i sučeljavanju njihovih stavova, te uvažavjući principe naučne metoda analogija, došlo se do spoznaje da postoji realna mogućnost primjene ovakvih rješenja u praksi BiH, pa i drugih zemalja Jugoistočne Evrope koja nisu u cjelosti primijenila ovaj ili neki drugi koncept u svojoj praksi. Da bi se stvorila slika o mogućim efektima primjene ovih rješenja, u nastavku će biti prkazane procjene na samo jednom dijelu kompanija koje se bave uslužnim djelatnostima i čiji je država u većini slučajeva većinski vlasnik. Radi se o kompanijama (nisu sve obuhvaćene, ali značajan dio jeste) koje se bave distribucijom električne energije, telekomunikacijama, uslugama centralnog grijanja, vodosnabdijevanja i drugim komunalnim uslugama.

Godišnje ove kompanije kumulativno obrade i distribuiraju više od sto milona računa (faktura) uz prosječnu cijenu distribucije „klasičnim“ putem od ≈1€, što znači da ti troškovi iznose oko sto miliona € uz skoro iste troškove koje izaziva funkcija obrade i realizacije platnih naloga, odnosno za samo manji dio ove funkcije obzirom da su analizirani slučajevi u relativno malom broju preduzeća, koštaju preko 200 miliona €. Analizirajući oficijelne cjenovnike iz ponuda e-plaćanja od strane nekoliko banaka može se zaključiti slijedeće: Cijena distribucije e-računa i obrade platnih e-naloga u zbirnom iznosu ne prelazi cifru od 0,25-0,30 €, što pored toga što ove transakcije čine bržim, sa manje grešaka, sa 24/7/365 radnim vremenom, eliminirajući sve vremenske i prostorne razlike, čine ih i jeftinijim za 6-7 puta. Ako bi se na ovom planu napravio iskorak za barem 30-ak%, ta bi ušteda imala veći efekat nego što iznosi proračun osnovnog, srednjeg i visokog obrazovanja najmnogoljudnije županije u BiH – Tuzlanske županije.

Uzme li se u obzir da ovo čini veoma mali broj od ukupnog broja transakcija svih subjekata, a producira procijenjene ovakve efekte, s pravom se može postaviti pitanje granica efekata primjene integralnog koncepta ovih funkcija, pogotovu u zemljama sa ograničenim investicijskim resursima i uvjetima niske uposlenosti. Prema nekim istraživanja svaki subjekt koji u toku jedne godine producira barem 60.000 ovakvih transakcija u biti ima značajne efekte u njegovoj implementaciji. Treba također reći da prema istražvanjima EC pozitivni efekti od uvođena ovakvih sitema poslovanja samo u ovim funkcijama produciraju pozitivne efekte preko 300 mld. €.(Strategija razvoja e-poslovanja R Hrvatske)

Zaključak:

Ne postavlja se dilema opravdanosti implementacije ovakvih rješenja u poslovnu praksu, nego se prije nameće pitanje zašto to već odavno nije zaživjelo, pogotovu u zemljama sa oskudnim investicijskim i drugim razvojnim resursima? Mnoštvo je različitih odgovora među kojima dominiraju politički motivi svih oblika, ali se oni nisu analizirali. Realna potreba kompanija u uvjetima ne zaokruženog sistema poslovanja na principima e-poslovanja, nameće nužnost primjene hibridnih (mada to grubo zvuči) rješenja, jer unutrašnje i međunarodno tržište nije statično i ono nameće nužnost primjene ovih rješenja. Kao i svaki sistem tako i ovaj sistem će doživjeti svoju evoluciju i razvoj u svim zemljama i gospodaskim sustavima, pa tako i u analiziranim područjima Jugoistočne Europe. Što se prije sistemski i organizirano pristupi ovom poslu, to će i efekti biti brži i sa multipritirajućem dejstvu. Ono što svakako treba istaći je da infrastruktura ili postoji ili je dostupna na zavidnom nivou, pa se ne vidi niti jedan racionalan razlog za intenzivnijim započinjanjem ovog procesa.

[3] preuzeto sa www.halcom.rs/index

LITERATURA

[1] B. Latinović, "Elektronsko poslovanje", Banja Luka, 2007.

[2] M.Milosavljević,G.Grubar,M.Veinović, "Poslovna informatika", Univerzitet Singidunum, Beograd, 2009.

[3] M. Perčić, „Poslovna funkcija nabave", Veleučilište u Varaždinu, Varaždin 2011.

[4] B. Žilbert, „ Strateška nabava" , Mate d.o.o, Zagreb 2007.

[5] J. Dyche, "The CRM Handbook: A Business Guide to Customer Relationship Management", Pearson Education Corporate Sales Division, USA, 2001.

[6] D.K. Rigby, F.F. Reichheld, F.F., P. Shefter, " Avoid the Four Perils of CRM", Harvard Bussiness Review, 2002.

[7] D. Peppers, M. Rogers, " Managing Customer Relationship, John Wiley and sons, Inc, Hoboken, New Jersey, 2004.

[8] J. Reynolds, " A Plactical Guide to CRM, CMP Books", New York, 2002.

[9] Turban et al. Electronic Commerce – Managerial Perspective, 2010.

[10] Gričar, Jože, Innovative Cross-border eRegion Development: Possible Directions and Impact. eRegion Development, Thematic Issue of the Journal 40(2007)V2.

[11] Kalakota, Robinson – e-Poslovanje, Mate Zagreb, 2002

[12] Novičević, B.: Globalizovani poslovni ambijent i izazovi upravljačkog računovodstva, Zbornik radova, Revicon, 2003

[13] www.halcom.rs/index

ERP Systems in Public Sector

R. Kelemen *
* Varazdin County/County Governors Office, Varazdin, Croatia
robert.kelemen@vz.t-com.hr

Abstract - ERP systems are tended to be implemented in the public sector. In this paper the "industries" of public sector are identified which are suitable for ERP implementation. The systematic review method is used to give an overview of state of the art ERP system implementation in the public sector worldwide in the last five years as well as possible directions of future research. Finally, the critical review is given regarding potential of ERP system implementation in the public sector.

I. INTRODUCTION

The vast majority of ERP system implementation in production systems has been completed and consequently the needs of that market have been satisfied. Some other potential areas of ERP systems implementation were in focus of ERP systems manufacturers. One of the emerging markets is the public sector. Implementation of ERP systems in the public sector has already began regardless of some well-known ERP implementation problems such as enormous investment and risk of failure in implementation itself. Numerous published scientific papers deal with various aspects of ERP systems, but the number of papers regarding ERP systems implementation in the public sector is relatively small. That particular area of the public sector is becoming increasingly interesting for ERP systems manufacturers, and researchers are interested in the areas of public services where the ERP systems have already been implemented.

Vrček and Brumec [6] define ERP system as "set of logically connected programs which support organization's business processes and use central and unique database, … it is not developed for any particular organization, regarding its functions it is general but simultaneously adoptable for specific needs of majority of organizations dealing with similar activities".

International Data Corporation (IDC) [7] defines that Enterprise Application Suite must fulfill the following criteria:

1. Minimal Functionality – not less than four basic functionalities:
   a. financial accounting modules,
   b. procurement modules,
   c. order management modules,
   d. inventory management modules.
2. Integrity – design based on simple rules and use of metadata, provides users access to all functional modules using one interface, it can store and retrieve data in one common database.
3. Relatively high price per user which differentiate s it from low-end book accounting software.
4. Multiuser platform – application (program) is designed for server/host and client/server environment.

According to Tomičić-Pupek [8] the general ERP system classification has not yet been accepted, but some criteria for ERP system classification are defined:

1. According to organization's line of business (a so called industry),
2. According to the set of business activities in one industrial area (vertical industry),
3. According to the business area (e.g. Accounting, Finances, Human Resources, Sales, Procurement, etc.),
4. According to the method of implementation,
5. Accordingto the ERP system adjustment approach to the organizational needs.

Jakupović, Pavlić and Fertalj [9] recommend ERP systems classification according to 70 business operations sorted into the three basic categories (Finances and public services, Production and Services). That approach is based on analyses of commercial ERP system manufacturers' classifications compliant with Global Industry Classification Standard (GICS) and EUROSTAT classification. The authors use the IDC definition of ERP systems, explained earlier.

Classification according the business functions - "industries" is commonly used in ERP system manufacturers' offers and it is concurred with noneconomic service of public interest definition.

The second area of interest in this paper is the public sector. It is important to define which services are offered by public sector and what in fact is the public sector.

Đulabić defines public sector services [1] as activities used to satisfy public needs, whereby the fundamental values of modern democratic societies are achieved. The concepts of "Public Service" and "services of public interest" describes the principle of same activity [1]. The concept of "Public Service" is more often used in continental European countries and the same concept but under the different name as "municipal services" or "public interest" can be found elsewhere [2]. The European Parliament defines public service [2] as permanent

economic activity which is in the public interest, and which is created and controlled by some public institution and at the same time it is under certain degree of legal protection beyond the regular legislative regardless the private or public authority which provides that service.

Public services are based on following principles [2]: equality, continuity, adaptability. Privilege of founding some public service agency (founding, setup and control) are given to [2]: central government (national public services), federal states and local authorities (regions, province, municipalities).

The new standards for public services general interest are established by the new European legislative regulation. The European Union differs economic public services (e.g. network industries) and non-economic public services like social security and healthcare [3]. The economic public services are big industries and network industries like telecoms, post offices, gas distribution, electrical energy distribution, water distribution, etc [4]. The non-economic public services are: education, culture, healthcare, social services, etc, everything which is not commercially based [3].

Koprić [4] indicates that term "services of general interest" isn't used in EC Agreements, and the use of that term spread in the late 1990ties. After intensive and long public discussion, the wider and narrower terms have been defined in a series of public documents. The wider term is defined as "public services of general interest" and the narrower term is defined as "economic public services of general interest". In addition, the specificities of certain categories of public services are recognized, so the terms like "social services of general interest" or "healthcare services of general interest" are also used [4]. According to Koprić [4] the intention of European Union is to harmonize the public services using policies and legislative (White Book, Bolkestein Directive).

Based on the above facts, this research aims at non-economic public services focused on the areas shown in Table I.

In this research it is not important at which level of the government these services are provided, since the research takes into consideration all levels of government.

The research question is defined as follows:

In which direction the ERP systems in non-economic public services have been developing over the last five years.

Research method: A method of systematic literature review as described by B. Kitchenham [5].

TABLE I. NON-ECONOMIC PUBLIC SERVICES

Code	Non-economic public service
PS1	Education
PS2	Culture
PS3	Health Care
PS4	Social Services
PS5	Law Enforcement
PS6	Military – defense
PS7	Public library
PS8	Public Security
PS9	Physical Planning / Town Planning
PS10	Public sector

II. SYSTEMATIC LITERATURE REVIEW

A systematic literature review is a means of evaluating and interpreting all available research relevant to a particular research question, topic or phenomenon of interest [5].

A. Planning

The scientific databases with full text paper, and the other available scientific articles in the field of social sciences were used in the research: Academic Search Complete, Business Source Complete, Directory of Open Access Journals – DOAJ, Regional Business News, Science Direct. Springer Link, IEEE Xplore, Google schoolar, HRČAK – portal znanstvenih časopisa, http://elf.foi.hr/.

All scientific and other papers and works written in the time span from 2008 to March 2013 were taken into account in the results selection.

Based on the analysis of services of general public interest which is described in the introduction to this paper, following keywords will be used to query the database: ERP system, government, public sector, public services, public utilities, public interest, service of general economic interest, education, culture, healthcare, social services, law enforcement, military, defense, public security, physical planning and public administration.

Using the above keywords the queries in Table II. are defined.

The article selection was performed in three stages (according to Jurišić [21]):

1. selection based on the analysis of titles and keywords,
2. selection based on the analysis of the abstracts,
3. selection based on the analysis of the full text.

At all stages of the selection articles related to the areas shown in Table I were considered.

TABLE II. RESEARCH QUERIES

Code	Query
Q1	'ERP system' AND 'government'
Q2	'ERP system' AND 'public sector'
Q3	'ERP system' AND 'public services'
Q4	'ERP system' AND 'public utilities'
Q5	'ERP system' AND 'public interest'
Q6	'ERP system' AND 'service of general economic interest'
Q7	'ERP system' AND 'education'
Q8	'ERP system' AND 'culture'
Q9	'ERP system' AND 'healthcare'
Q10	'ERP system' AND 'social services'
Q11	'ERP system' AND 'law enforcement'
Q12	'ERP system' AND 'military'
Q13	'ERP system' AND 'defense'
Q14	'ERP system' AND 'public libraries'
Q15	'ERP system' AND 'public security'
Q16	'ERP system' AND 'physical planning'
Q17	'ERP system' AND 'public administration'

At each stage of selection articles that are too general or those who are not focused on the research areas listed in Table 1 were rejected.

Identical articles were identified and duplicates were discarded at each stage of selection.

B. Conduction

After rising queries with keywords in databases and other available sources, according to the strategy specified in the research plan, 1046 articles have been discovered.

Having conducted the first phase - selection based on analysis of titles and keywords from the results of the previous stage - 170 articles have been extracted.

In Phase 2, 170 abstracts of articles resulting from phase 1were analyzed and 67 articles was allocated.

In the third stage of selection, after detailed reading, 67 articles were analyzed. 36 articles of these were eliminated after the content analysis, because they were not related to the topic of research. Some of the reasons were:

- Articles describing the upgrading of the existing ERP system with some new module or a separate decision,
- Articles describing systems that have already been analyzed,
- Articles not dealing exactly with the description of ERP systems, they were related to management, organizational culture, the reasons of successful or unsuccessful implementation, evaluation of investment ...

Full text analysis resulted with a set of 31 articles, 4 of them describing research conducted over more respondents from the public sector [30], [33], [8] [42], while the other 27 refer to one specific ERP system which has been applied to an institution of the public sector. 7 articles do not relate directly to the purpose of the research but generally describe some aspects of the ERP system in the non-economic public services.

C. Results and analysis

Analysis of 31 article shows that the largest number of articles related to the introduction of ERP systems in the field of education (11), then the area of functioning of the public sector (cities, states, regions, agencies) (10), the area of defense (5) area of healthcare (3), and (2) articles describe the introduction of ERP systems in several institutions.

Analyzing separate articles regarding the country of implementation, it was found that the largest number of articles describing the implementation of ERP systems refers the U.S. - 17, an unexpectedly large number of articles describing the implementation of ERP in Croatia - four, while the ERP implementation in Italy, Egypt, Hungary, Portugal, Saudi Arabia, the Netherlands, Denmark and Finland are described with one article each. Two articles describe ERP implementation in several countries.

Regarding the ERP system manufacturer, most articles - 11 describe SAP, 6 articles describes multiple ERP systems at once, 2 articles describe Oracle, 2 articles describe Microsoft Dynamics AX, 1 article describes each of the following manufacturers: Infor, BAAN, CGI AMS, Datatel Colleague, People Soft, MADAR, MUNIS, "Ericsson Nikola Tesla alliance", Custom solution, while in one article the manufacturer is not even listed.

1) Education

According to the classification of business operations created by Jakupović Pavlic and Fertalj [9], education as part of the public services is divided into three activities: schools, higher education and research. As a result of this research related to education, there were 11 articles discovered describing the ERP modules implemented in educational organizations and they are presented in Table III. Tomicic-Pupek [8] in doctoral dissertation refers to the research of module types, manufacturers and recent clients of the ERP system implementation in higher education institutions in the U.S. where 23 of 66 higher education institutions have introduced full ERP. In the European higher education institutions such register was never found.

2) Public Sector

In the field of public sector, research has identified 10 articles from the level of the city administration up to regional, state and federal government with ERP modules implemented as shown in Table IV.

3) Defence

German army uses the ERP system called "SAP DFPS" (SAP for Defense Forces and Public Security) [16] with ERP modules stated in Table V. In the same article the Logistic information system - ISL (Information System for Logistics) of the Czech Republic Department of Defense and the Czech Republic Army is described [16] developed by Czech producers AURA with the modules presented in Table V.

U.S. Department of Defense 's (DoD - Department of Defense) uses four primary ERP systems [46] and several smaller specialized ERP systems. The largest four are: Integrated Personnel and Pay System (IPPS-A) it is a system of human resource management, the Global Combat Support System for tactical logistics (GCSS-Army), Logistics Modernization Program (LMP) of the National Logistics, General Funds Enterprise Business System (GFEBS) is a global financial system. Manufacturer of the IPPS-A is a PeopleSoft Company, SAP is a manufacturer of the other three.

In addition to the U.S. military, the ERP system was introduced in Denmark army. Christiansen et al. describes the results of the introduction of the SAP ERP system in a special unit of Danish Navy [19]. The main purpose of the system implemented in Denmark army was to meet accounting standards set by the Danish National Audit Office, and modules are presented in Table V.

Hyovnen et al. [20] describe the case of the Finnish Defense Forces. By the end of 2005, the new ERP system replaced 19 different administrative information systems.

TABLE III. PS1 - EDUCATION

Ref.	ERP Modules Implemented	ERP Manufact.
[35]	Finacial accounting, HT mgt, Material Mgt, Project Mgt, Record Mgt, Faze II: Payroll, Asset Mgt, Procurement Mgt	SAP
[36]	Financial, Human Capital, Pension, Logistics and Technical	SAP
[37]	full ERP	Datatel® Colleague®
[38]	finance, procurement, human resources, campus solution - student modules	PeopleSoft
[39]	full ERP	MADAR
[40]	human capital management, financial management and payroll processing	Munis
[41]	full ERP	Oracle
[8]	different ERP componenets or full ERP implemented in different institutions	multiple manufacturers
[12]	full ERP	proprietary developed ERP "Is_UVN"
[42]	financial system, business processes of budget execution, financial accounting	SAP
[43]	finance, human resources (HR) and student information system software	not specified

SAP R/3 system was introduced with modules specified in Table V, and it was the first major implementation of SAP R/3 system in the Finland public sector [20]. Modules project management, maintenance and human resources are planned for later introduction, the time of implementation has not been determined.

4) Healthcare

Boonstra and Govers [15] in their paper have described the implementation of SAP ERP system in the medium - sized general hospital in the western part of the Netherlands. The hospital consists of six main units: 1 clinical part, 2. outpatient, 3. medical part, 4. facilities, 5. staff, 6. IT and finance. ERP system has been implemented over 5 years and the following modules were introduced: patient management, finance, material management, human resources, laboratory system, management support system, presented in Table VI.

The management has concluded that after three years of implementation, the expected improvements in the business haven't been achieved and that the ERP system is poorly implemented. Implementers wanted, due to fixed implementation costs, to implement the system quickly and therefore they had underestimated the role and the importance of medical practice that is required to get the system set up in accordance with the existing processes of medical practice. Implementer has tried to adapt processes found in hospital system to the ERP system, instead of adapting the ERP system processes to the hospital. Due to poor implementation, the hospital has lost at least 10% of annual revenue and was very close to bankruptcy [15].

The literature review has discovered a practical example of the introduction of an integrated hospital information system (IBIS) at University Hospital "Dubrava" in Croatia.

TABLE IV. PS10 – PUBLIC SECTOR

Ref.	ERP Modules Implemented	ERP Manufact.
[22]	financial management, procurement, supplier order management, contract management and requsitions, human resources, payroll and employee/manager self service	New York-based Infor
[23]	SB (Basic System): Document and work flow management; SCI (Integrated Accounting System): Accounting and controlling ; HR (Human Resources)	SAP
[24]	HR	SAP
[25]	Accounting, Budgeting, Payroll, Human Resources, and Procurement	SAP R/3
[26]	Full ERP, The ERP system integrates all of the city's business functions, including procurement, payroll, finance, inventory and human resources	Oracle
[27]	financial, procurement, personnel, budget, inventory control, grants and project management wifh fully integrated document management and workflow	SAP
[29]	finance, sales	BAAN
[31]	financials, controlling, human resources, logistics. The software is to assist companies by providing support for sales, customer relationships, inventory, operations, financials and human resources.	Microsoft Dynamics Ax
[32]	budgeting solution	CGI - AMS Advantage ERP Suite
[34]	General Ledger, Purchasing and Workflow. These were complemented with Tyler's Payroll and Cashiering applications	Microsoft Dynamics AX 2012 core ERP

This example is described in two articles [44] and [14] stating that all modules are planned in an integrated hospital information system "Ericsson Nikola Tesla alliance" (company "Ericsson Nikola Tesla" from Zagreb with partners: "Grad" from Pula and "Adacta", "Samson", "HrPro" and "Vams" from Zagreb). It is expected to improve the overall system in the KBD:

- transparent monitoring of hospital operations,
- Fast and efficient invoicing costs
- Maintaining comprehensive costs under control.

The introduction of the ERP system was started in 2003 and in 2005 all modules were not introduced. Only complete system for administrative monitoring of patients has been implemented [44] presented in Table VI.. The project is currently in the phase of standstill.

Authors of the article Motivations Underlying the Adoption of ERP Systems in Healthcare Organizations: An Analysis from the "Success Stories" in their study identified 180 successfully introduced ERP systems in the field of health care in the world. From the perspective of the software manufacturer in this research, the largest market share has Oracle with 100 references , followed by SAP with 38, 37 references has Cerner, 5 references have all other manufacturers [30].

TABLE V. PS6 – MILITARY – DEFENSE

Ref.	ERP Modules Implemented	ERP Manufact.
[18]	Integrated Personnel and Pay System (IPPS-A),Global Combat Support System for tactical logistics (GCSS-Army), the Logistics Modernization Program (LMP), the General Funds Enterprise Business System (GFEBS).	GFEBS - SAP, GCSS-Army - SAP, LMP - SAP R/3, IPSS-A - PeopleSoft, Navy ERP - SAP R/3, GCSS-MC - Oracle, IPPS-Navy - PeopleSoft, DEAMS - Oracle, ECSS - Oracle, AF-IPPS - PeopleSoft
[17]	GFEBS, GCSS-Army, DEAMS, ECSS	GFEBS - SAP, GCSS-Army - SAP, DEAMS - Oracle, and ECSS - Oracle
[19]	human resources, controlling, logistic	SAP R/3
[20]	material management, procurement, asset management, budget planning, payment, and financial management	SAP R/3
[16]	SAP DFPS human resources, finance, operations and business processes, module for modeling units ISL - procurement, supply, logistics, equipment maintenance and ammunition.	SAP DFPS, AURA -ISL

Successful implementation (180) relates to 31 different countries, 92 were in North America (U.S. 89), 68 in Europe (UK 17, Germany 15, Netherlands 8 - all other 28), 12 implementations in Asia, 5 in Latin America and 3 in Oceania. The study noted that the following functionalities (modules) were implemented: 145 administrative systems, clinical information systems 38, 36 clinical support systems and 22 systems for patient management. Administrative systems have modules: accounting, human resources, payroll costs to diseases, material management.

TABLE VI. PS3 – HEALTH CARE

Ref.	ERP Modules Implemented	ERP Manufact.
[30]	multiple ERP systems	multiple manufact
[15]	patient management (appointments, registration, admission, discharge andtransfer);financial management (accounts payable, accounts receivable, general ledger andfixed assets); materials management (inventory and purchase); human resource (scheduling, training and payroll);•ancillary services (dietary, laboratory and pharmacy); and management information (reporting on and providing statistics on various issues).	SAP
[14]	planned: patient management (appointments, registration, admission, discharge and transfer); ancillary services (dietary, laboratory and pharmacy); Radiolgy IS, operations, intensive care; materials management (inventory and purchase); financial management (accounts payable, accounts receivable, general ledger and fixed assets); management information (reporting on and providing) implemented: patient management	Hospital informatio n system "Ericsson Nikola Tesla alliance"

Clinical information systems have modules like emergency services, surgery, nursing services, etc. The clinical support system has modules like pharmacy, etc. The system for managing patients: patient admission, transfer and discharge of patients [30]. The research described earlier is focused on the motivation for the introduction of ERP systems in health care facilities, so that it was not possible to determine which modules were combined together from that article.

III. CONCLUSION

Based on the analysis of the articles presented, regarding the area of non-economic public services shown in Table I, it can be concluded that ERP systems have been introduced in the areas of education, health care, defense, and municipalities, regions and states to support the public administration activities. There were no articles related to non-economic public services in the field of culture, social services, justice, libraries, public safety and urban planning. With regard to the ERP systems classification prepared by the authors Jakupović Pavlic and Fertalj [9], it is evident that there is a full agreement between classifications and systems implemented in practice, and ERP systems are often introduced in the areas of education, public administration, defense and health. As seen from the analyzed articles, despite the fact that European legal regulations recognize some other non-economic public services [3], the implementation of ERP systems in those hasn't been accomplished yet. In addition, even ERP software manufacturers haven't classified such systems.

Some of the general definitions of ERP systems were described in articles [6], [7] and [8]. The articles [8], [10], [11] and [12] give a set of modules which should include ERP systems in education. Complete ERP systems in the field of health should include modules according to the authors [13], [14], [44] and [15]. ERP systems in the field of defense are described in the articles [16], [17] and [18].

Vrček and Brumec [6] in their paper display modules (functions) of ERP system and group them as business support, networking and value chain.

According to Spathis and Constantinides [51] and Spathis [52] research in organizations that have implemented ERP systems, prevalent modules were: finance and accounting, and management support systems for management and control. The authors conclude that organizations introduce ERP system beginning the modules of finance and accounting, with a desire to integrate the accounting processes.

In the field of public administration, according to the data from the articles [22], [23], [24], [25], [26], [27], [29], [31], [32] and [34] and the results obtained by the Microsoft and GFOA [28] and [33], it can be concluded that ERP systems have been introduced solely to support the business .

In the field of education, according to Tomičić-Pupek [8] 23 of the 66 higher education institutions in the United States introduced a full ERP. As seen from the articles of this research, a full ERP in the field of education is described in the articles [37], [41], [12] and [39]. ERP systems to support business operations are covered in the

articles [35], [36], [40] and [42] and it is shown that the institutions have introduced ERP to support the business. ERP systems with different modules that are in the value chain have established institutions that are described in [38], [43].

In the area of defense a full logistics and financial ERP systems were introduced, as described in the articles [17], [18], [16], [19] and [20].

In the field of healthcare according to a study by Poba-Nzaou et al., Described in the article [30] it is noted that in 180 implementations of ERP systems in 31 different country the following functionalities (modules) have been introduced: 145 administrative systems, 38 clinical information systems, 36 clinical support systems and 22 systems for patient management. The full ERP systems implementations introduced or planned in the area of healthcare were described in the articles [15] and [14].

It can be concluded that existing research shows that ERP systems in non-economic public services are generally introduced in the areas of health, education, defense and public administration.

In the area of Healthcare, the most common ERP modules were introduced to support the business processes, while a lesser extent represents modules of the value chain.

In the field of defense, exclusively full ERP systems have been introduced that cover the entire target area: logistics, human resources and finance.

Vrček and Brumec [6] define ERP system and conclude that ERP system is not intended for the implementation in one business function ("island"), but on the contrary, ERP system connects all business functions involved in performing the key processes. Justification for the introduction of the ERP system in the organization is questionable if this system does not include all the necessary modules for specific area from which the organization derives.

Based on the above, it can be concluded that in the area of non-economic public services like public sector, education, healthcare and defense, there is still some space for exploration, especially with respect to the implementation of ERP modules in the value chain of each area, taking into account the specifics with regard to the objectives of the organization and the level of development of the state.

REFERENCES

[1] V. Đulabić, "Socijalne službe u konceptu službi od općeg interesa," Revija za socijalnu politiku, vol. 14, no. 2, pp. 137–162, May 2007.

[2] F.-X. Camenen, Public Undertakings and Public Service Activities in the European Union. European Parliament, 1996.

[3] I. Koprić, "Od javnih službi do službi od općeg interesa: nova europska regulacija i njezin odraz u modernim upravnim sustavima," Zbornik radova Drugog skopsko-zagrebačkog pravnog kolokvijaSkopje: Pravni fakultet Justinijan Prvi Sveučilišta Sv. Kiril i Metodij, Skopje, Makedonija, pp. 27–49, 20.11 2008.

[4] I. Koprić, A. Musa, and V. Đulabić, "Europski standardi regulacije službi od općeg interesa:(kvazi) nezavisna regulacijska tijela u izgradnji modernog kapitalizma," Hrvatska javna uprava, vol. 8, no. 3, pp. 647–688, 2008.

[5] B. Kitchenham, "Procedures for performing systematic reviews," Keele, UK, Keele University, vol. 33, p. 2004, 2004.

[6] N. Vrček and J. Brumec, "Nastavni materijali za predmet Stanje i razvoj ERP sustava," Sustav za učenje FOI, 2013.

[7] "IDC ERP Forum Adriatics 2010." [Online]. Available: http://www.idc-adriatics.com/?showproduct=38243&content_lang=HR. [Accessed: 05-Mar-2013].

[8] K. Tomičić-Pupek, "Sustavi planiranja i upravljanja resursima visokih učilišta," Doktorska disertacija, Sveučičište u Zagrebu, Fakultet organizacije i informatike, Varaždin, 2011.

[9] A. Jakupović, M. Pavlić, and K. Fertalj, "Analysis and Classification of ERP Producers by Business Operations," CIT. Journal of computing and information technology, vol. 17, no. 3, pp. 239–258, 2009.

[10] G. Sabau, M. Muntean, and A. R. Bologa, "Analysis of Integrated Software Solutions Market for Romanian Higher Education," Economic Computation & Economic Cybernetics Studies & Research, vol. 43, no. 1, p. 197, 2009.

[11] R. Alt and G. Auth, "Campus Management System - Springer," Business & Information Systems Engineering, vol. 1/2010, no. 2, pp. 187–190, Apr. 2010.

[12] N. Novak and M. Balen, "Primjena informacijsko komunikacijske tehnologije u preobrazbi sustava upravljanja visokim učilištem," presented at the 12. Međunarodni znanstveni skup "Društvo i tehnologija 2005.", Zadar, Hrvatska, 2005, pp. 16–17.

[13] G. Galijašević, "Koncept integriranog bolničkog informacijskog sustava," MEDIX, vol. 10, no. 54/55, pp. 96–101, Oct. 2004.

[14] D. Vuković, "Uvođenje integriranog bolničkog informacijskog sustava," MEDIX, vol. 10, no. 54/55, pp. 104–106, Oct. 2004.

[15] A. Boonstra and M. Govers, "Understanding ERP System Implementation in a Hospital by Analysing Stakeholders," Social Science Research Network, Rochester, NY, SSRN Scholarly Paper ID 1487800, Oct. 2009.

[16] M. Dronjak and M. Milenkov, "Development of logistics information systems," Vojnotehnički glasnik, vol. 60, no. 4, pp. 147–164, 2012.

[17] ..., "DOD Financial Management, Implementation Weaknesses in Army and Air Force Business Systems Could Jeopardize DOD's Auditability Goals." GAO Reports;3/29/2012, preceding p1, 29-Mar-2012.

[18] ..., "DOD Financial Management: Reported Status of Department of Defense's Enterprise Resource Planning Systems." GAO Reports;3/30/2012, following p1, 30-Mar-2012.

[19] U. Christiansen, A. Kjærgaard, and R. K. Hartmann, "Working in the shadows: Understanding ERP usage as complex responsive processes of conversations in the daily practices of a Special Operations Force," Scandinavian Journal of Management, vol. 28, no. 2, pp. 173–184, Jun. 2012.

[20] T. Hyvönen, J. Järvinen, J. Pellinen, and T. Rahko, "Institutional Logics, ICT and Stability of Management Accounting," European Accounting Review, vol. 18, no. 2, pp. 241–275, 2009.

[21] M. Jurišić, "Transition between process models (BPMN) and service models (WS-BPEL and other standards): A systematic review," Journal of Information and Organizational Sciences, vol. 35, no. 2, pp. 163–171, Dec. 2011.

[22] ..., "Savannah Automates Finance Activities," American City & County, vol. 127, no. 6, 2012.

[23] A. Spano, D. Carta, and P. Mascia, "The Impact of Introducing an ERP System on Organizational Processes and Individual Employees of an Italian Regional Government Organization," Public Management Review, vol. 11, no. 6, pp. 791–809, 2009.

[24] D. J. McCubbrey and C. Fukami, "ERP at the Colorado Department of Transportation: The Whistle Blower's Dilemma," Communications of the Association for Information Systems, vol. 24, no. 1, Jan. 2009.

[25] W. Wagner and Y. L. Antonucci, "The ImaginePA Project: The First Large-Scale, Public Sector ERP Implementation," Inf. Sys. Manag., vol. 26, no. 3, pp. 275–284, Jun. 2009.

[26] J. Cable, "St. Petersburg purchasing department doesn't sweat the small stuff," St. Petersburg purchasing department doesn't sweat the

small stuff, 01-Dec-2008. [Online]. Available: http://govpro.com/technology/content/st_petersburg_purchasing/. [Accessed: 04-Mar-2013].

[27] N. Hall, "It's all in the numbers in South Carolina," Govpro, 01-Aug-2009. [Online]. Available: http://govpro.com/mag/commodity-coding-benefits-200908-09/. [Accessed: 04-Mar-2013].

[28] ..., "The real Impact of ERP Systems in the Public Sector." Microsoft and GFOA Consulting Research Paper, 11-Jun-2012.

[29] L. Jack and A. Kholeif, "Enterprise Resource Planning and a contest to limit the role of management accountants: A strong structuration perspective," Accounting Forum, vol. 32, no. 1, pp. 30–45, Mar. 2008.

[30] P. Poba-Nzaou, S. Uwizeyemungu, L. Raymond, and G. Pare, "Motivations Underlying the Adoption of ERP Systems in Healthcare Organizations: An Analysis from 'Success Stories'," in Proceedings of the 2012 45th Hawaii International Conference on System Sciences, Washington, DC, USA, 2012, pp. 2927–2936.

[31] I. Orosz and T. Orosz, "Business Process Reengineering project in Local Governments with ERP," in 2012 7th IEEE International Symposium on Applied Computational Intelligence and Informatics (SACI), May, pp. 371–376.

[32] "State of Vermont selects CGI's public sector budgeting solution to modernize statewide budget operations | CGI.com," 21-Dec-2011. [Online]. Available: http://www.cgi.com/en/cgi-ams-advantage-state-of-vermont. [Accessed: 19-Mar-2013].

[33] V. Botta-Genoulaz and P.-A. Millet, "An investigation into the use of ERP systems in the service sector," International Journal of Production Economics, vol. 99, no. 1–2, pp. 202–221, Jan. 2006.

[34] "Tyler Technologies Completes First-ever Deployment of Microsoft Dynamics® AX 2012 With New Public Sector Functionality At the City of Redmond, Wash. | Business Wire," 20-Sep-2011. [Online]. Available: http://www.businesswire.com/news/home/20110920005393/en/Tyler-Technologies-Completes-First-ever-Deployment-Microsoft-Dynamics%C2%AE. [Accessed: 19-Mar-2013].

[35] V. Mornar, K. Fertalj, and D. Kalpić, "Introduction of SAP ERP system into a heterogeneous academic community," AIP Conference Proceedings, vol. 1239, no. 1, pp. 388–395, Jun. 2010.

[36] "Outsourcing an ERP System." [Online]. Available: http://www.districtadministration.com/article/outsourcing-erp-system. [Accessed: 19-Mar-2013].

[37] M. Day, "N.C. Financial Aid Departments Boost Productivity, Accuracy With Systemwide Technology," Community College Week, 2008.

[38] M. Bradford, "North Carolina State University: Implementing ERP Student Modules," Issues in Accounting Education, vol. 26, no. 3, pp. 507–520, Aug. 2011.

[39] S. Dezdar and S. Ainin, "The influence of organizational factors on successful ERP implementation," Management Decision, vol. 49, no. 6, pp. 911–926, Jun. 2011.

[40] "Burlington School District of Vermont Reports Immediate Benefits From Tyler Technologies' ERP System.," Mena Report, Sep. 2012.

[41] "Florida Keys Community College Gains Full ROI with Toad for Oracle Within One Week of Deployment.," Education Letter, Nov. 2011.

[42] D. Kalpić, L. Luić, and Z. Stanić, "Development of an Integral Business Information System of Higher Education," in EUNIS 2008, 14th European University Information Systems Conference "VISION IT - Vision for IT in Higher Education" / Sidelmann Jorgensen, Dorte ; Hvid Stenalt, Maria (ur.). - Arhus : NAT Reprocenter, University of Arhus , 2008. 120-128 (ISBN: 978-87-91234-56-9)., Arhus, Danska, 2008, pp. 120–128.

[43] N. Kulick, M. Jeffery, T. Riitters, S. Abbott, D. Papp, T. Schad, J. Wallace, and J. Wiemann, "The San Diego City Schools: Enterprise Resource Planning Return on Investment." Kellogg School of Management, 01-Jan-2006.

[44] M. Pretovečki, D. Raunić, D. Cej, and D. Vuković, "Problematika uvođenja integriranog bolničkog informacijskog sustava u postojeću bolničku organizaciju -iskustvo Kliničke bolnice 'Dubrava' u Zagrebu," Acta Med Croatica, vol. 59, pp. 219–224, 2005.

[45] "Higher Education Software Vendors." [Online]. Available: http://www.edustrat.com/vendors.html. [Accessed: 20-Mar-2013].

[46] "How the Army Normalizes Disparate ERP systems – and Gets Better Data | ERP Executive." [Online]. Available: http://www.erpexecutive.com/2011/04/how-the-army-normalizes-disparate-erp-systems-%E2%80%93-and-gets-better-data/. [Accessed: 21-Mar-2013].

[47] "GFEBS." [Online]. Available: http://www.gfebs.army.mil/. [Accessed: 06-Mar-2013].

[48] "A User's Perspective of GCSS-Army," Army Sustainment, Apr-2012. [Online]. Available: http://www.almc.army.mil/alog/issues/MarApril12/A_Users_Perspective_GCSS.html. [Accessed: 06-Mar-2013].

[49] "LMP," SAP. [Online]. Available: http://www.sap.com/community/webcast/2008_05_SAPPHIRE_US/2008_05_SAPPHIRE_US_OR3824.pdf. [Accessed: 06-Mar-2013].

[50] M. W. Jones, "Implementation Challenges for DoD Logistics Enterprise Resource Planning IT Systems," Sep. 2010.

[51] Spathis C. and Constantinides S., "Enterprise resource planning systems impact on accounting processes," Business Process Management Journal, vol. 10, no. 2, pp. 234–247, 2004.

[52] [52] C. Spathis, "Enterprise systems implementation and accounting benefits," Journal of Enterprise Information Management, vol. 19, no. 1, pp. 67–82, Jan. 2006.

[53] A. J. A. Abugabah and L. Sanzogni, "Enterprise Resource Planning (ERP) System in Higher Education: A literature Review and Implications," presented at the International Conference on Information Systems (ICIS'10), 2010.

Oslonjivost sustava e-poslovanja u oblaku

Mihael Buković, Milan Davidović, Pavle Špoljarić
VSITE – Visoka škola za informacijske tehnologije, Zagreb, Croatia
mihael.bukovic@vsite.hr, milan.davidovic@vsite.hr, pavle.spoljaric@vsite.hr

Sažetak – Utvrđivanje oslonjivosti sustava e-poslovanja (eng. *dependability*), kojom se dokazuje u kojoj se mjeri mogu dionici osloniti na njegove usluge, obuhvaća praćenje karakteristika i analizu prijetnji, te određivanje sredstava za postizanje i održavanje potrebne oslonjivosti. Oslonjivost sustava e-poslovanja u oblaku analizira se sa stanovišta davatelja i korisnika usluga, s tim da je naglasak na utvrđivanju mogućnosti koje stoje na raspolaganju korisniku glede pouzdanosti, raspoloživosti, sigurnosti, povjerljivosti i integriteta. S obzirom da je pouzdanost ključni faktor oslonjivosti sustava poslovanja predložen je pristup analizi pouzdanosti sustava e-poslovanja u oblaku.

Ključne riječi: sustav e-poslovanja, oslonjivost, pouzdanost, računalni oblak, sustav tolerantan na neispravnosti

I. UVOD

Računarstvo u oblaku (eng. *Cloud computing*) omogućit će poduzećima, posebno malim i srednjim, implementirati sustav e-poslovanja brže i uz manje troškove implementacije, pogona i održavanja u odnosu na vlastitu informacijsko-komunikacijsku infrastrukturu. Cijena programske podrške za e-poslovanje u oblaku može biti niža, jer se plaća usluga i to prema vremenu i broju korisnika koji ju koriste. Pri tom je dostupna zadnja verzija programske podrške, tako da korisnik usluge nema troškova nadogradnje i održavanja softvera (jasno, pri tom treba adoptirati dizajn procesa e-poslovanja koje podržava softver, a adaptacije softvera su nemoguće ili preskupe). Također nema direktnih troškova nabave i održavanja potrebnog hardvera i licenci, niti troškova instalacije, konfiguracije i održavanja hardvera, sistemskog softvera, softvera za baze podataka i servere elektroničke pošte i dr. U cijenu usluge je uključena i antivirusna zaštita, a kao posebna usluga može se ugovoriti pohranjivanje, te izrada i čuvanje sigurnosnih kopija određenih, za poslovanje poduzeća ključnih podataka. Pristup do resursa u računalnom oblaku je moguć samo ovlaštenom korisniku iz bilo koje lokacije koja ima pristup na Internet [1][2].

I pored svih navedenih dobrobiti nameće se pitanje u kojoj mjeri se korisnik može osloniti na sustav e-poslovanja u računalnom oblaku i koja mu sredstva stoje na raspolaganju da bi dobio ne samo jeftiniji, već i sustav e-poslovanja u koji se može pouzdati. Analizu oslonjivosti tog sustava trebalo bi provesti prije ugovaranja usluga i periodički ju provoditi nakon ugovaranja u sklopu ocjenjivanja izvršavanja ugovora. Pristup analizi oslonjivosti sustava e-poslovanja u računalnom oblaku dat je u t. II-IV.

II. KARAKTERISTIKE OSLONJIVOSTI E-POSLOVANJA U OBLAKU

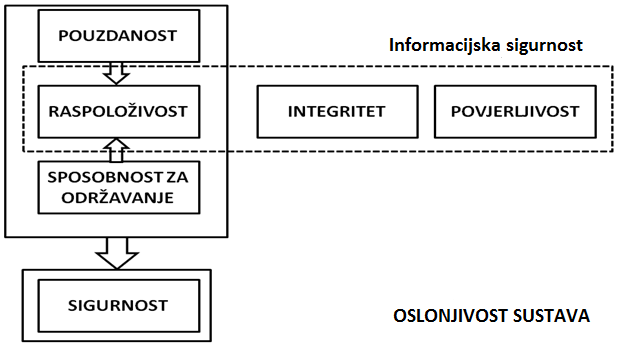
S obzirom da korisnik ovisi o uslugama informacijskog sustava, on želi znati u kojoj mjeri se može osloniti na informacijski sustav. **Oslonjivost informacijskog sustava** (eng *information system dependability*) je mjera opravdanosti korištenja informacijskih usluga. Što se aktivnosti korisnika više oslanjaju na informacijsko-komunikacijsku podršku to će korisnik biti ovisniji o informacijskom sustavu. Što je veća ovisnost (eng *dependency*) korisnika o informacijskim uslugama, to veća bi trebala biti oslonjivost informacijskog sustava [3]. Isti informacijski sustav mogu različiti korisnici ocjenjivati kao potpuno oslonjiv, manje oslonjiv ili potpuno neoslonjiv, ovisno o tome da li se u realizaciji svojih aktivnosti mogu u potpunosti ili djelomično osloniti na usluge informacijskog sustava ili se uopće ne mogu osloniti na njegove usluge [4].

Polazeći od ove definicije oslonjivosti proizlazi da će **oslonjivost sustava e-poslovanja** ovisiti o tome

- U kojoj se mjeri mogu menadžment i ostali dionici (kupci, dobavljači, zaposlenici, dioničari) u provedbi svojih aktivnosti, koje su usmjerene na povećanje sigurnosti razmjene dobara, usluga i informacija, osloniti na informacijsko-komunikacijsku potporu;
- U kojoj mjeri se menadžment može osloniti na informacijsko-komunikacijsku potporu u izgradnji novih poslovnih odnosa unutar organizacije i s poslovnim partnerima ili
- Koliki utjecaj na poslovni uspjeh imaju preustroj organizacije i redizajn poslovnih procesa koje je potrebno provesti radi djelotvornog korištenja informacijsko-komunikacijske potpore, te
- Kolika je sigurnost elektroničkih transakcija koje prate razmjenu dobara i usluga.

Kvantitativno možemo oslonjivost sustava e-poslovanja u oblaku definirati kao vjerojatnost da će sustav biti **raspoloživ** da zaprimi zahtjeve korisnika za uslugom i biti **operativan** (pouzdan) dok je usluga potrebna korisniku, što uključuje i zahtjeve kao što su **povjerljivost** ili **tajnost** (eng *confidentiality*), **cjelovitost** ili **integritet** (eng. *integrity*), **zaštita okoline** (eng. *environmental safety*) i **sposobnost održavanja** ili **popravljivost** (eng. *maintainability*). Što je veća vjerojatnost da će se sustav e-poslovanja u oblaku ponašati u skladu sa zadanim specifikacijama u zadanim uvjetima u okruženju to se korisnik može jače osloniti na njegove usluge.

Glavna obilježja oslonjivosti sustava e-poslovanja u oblaku su **pouzdanost** (eng *reliability*) i **raspoloživost** (eng *availability*). Raspoloživost određuju pouzdanost i sposobnost održavanja sustava (v. sliku 1). Računalni sustavi koji se koriste za pohranjivanje i obradu podataka, sigurnosne kontrole i komunikacijski kanali za pristup podacima u oblaku moraju biti raspoloživi da bi bili raspoloživi podaci potrebni za e-poslovanje. **Raspoloživost**, **integritet i povjerljivost podataka** tvore trijadu CIA [1] osiguravanja informacija (eng *information assurance*) koju određuje **model sigurnosti** (eng *security model*) koji primjenjuje sustav e-poslovanja u oblaku čije se usluge koriste.



Slika 1 Odnos glavnih obilježja oslonjivosti sustava
Izvor: Davidović [3]

Pouzdanost R(t) (eng. *Reliability*) je uvjetna vjerojatnost ispravnog rada sustava u zadanom vremenskom intervalu [0,t). Sustav ispravno radi ako obavlja sve zadaće u skladu sa zadanim specifikacijama u zadanim radnim uvjetima. Kao pokazatelj pouzdanosti koristi se i srednje vrijeme između kvarova (eng. *MTBF – Mean Time Between Failures*) [3][5]. Pouzdaniji sustav e-poslovanja ima veću raspoloživost i sigurnost, a time i bolju zaštitu od vanjskih napada i ostalih prijetnji (v. t. III). Da bi se postigla potpuna oslonjivost sustav e-poslovanja u oblaku trebao bi biti tolerantan na neispravnosti (v. t. IV).

Usluga sustava e-poslovanja u oblaku je **korektna** ako sustav ispravno obavlja sve zadaće potrebne za njenu realizaciju u skladu sa zahtjevima korisnika. Kod **zatajenja sustava e-poslovanja** (eng *e-business system failure*) isporučena usluga razlikuje se od ugovorom definirane korektne usluge, jer se zadaće ne obavljaju u skladu s ugovorenim specifikacijama zbog neke trajne, povremene ili prolazne **neispravnosti ili kvara** (*eng permanent, temporary, transient fault)* ili ako specifikacije sustava ne opisuju adekvatno uslugu (na primjer, neispravan dizajn) [4][5]. Sustav je **siguran** (eng *safety*) ako nema katastrofalnih posljedica zatajenja za korisnika i njegovu okolinu [4].

[1] CIA – akronim od *Confidentiality, Integrity, Availability*

Usluge neće biti korektne kad sustav nije raspoloživ zbog aktivnosti preventivnog održavanja, nadogradnje, vanjskih utjecaja i drugih razloga zbog kojih se ne može realizirati zahtjev korisnika za obavljanjem usluge iako je sustav ispravan. Usluge su nekorektne i kad sustav nema potrebne (ugovorene) performanse u trenutku zahtjeva korisnika za uslugom t. **Izvršljivost** (eng *performability*) P(I,t) je vjerojatnost da će performanse sustava u trenutku t biti iznad razine I [5]. Degradacijom performansi ispod razine I usluge postaju nekorektne, a sustav neoslonjiv.

Obnavljanje sustava (engl. *system restore*) obuhvaća otklanjanje neispravnosti i pogrešaka koje su uzrokovale zatajenje sustava ili promjenu specifikacija, redizajn funkcija i njihove tehničke izvedbe da bi sustav pružao korektnu uslugu. [3] Sustav e-poslovanja u oblaku mora posjedovati sposobnost otklanjanja skrivenih softverskih defekata koji mogu biti meta vanjskih napada ili mogu pod određenim okolnostima izazvati zatajenje sustava ili ugroziti njegovu sigurnost (v. t. III). Glavne karakteristike **sposobnosti održavanja ili popravljivosti** sustava e-poslovanja su vjerojatnost M(t) da će sustav biti obnovljen u vremenskom intervalu [0,t), intenzitet obnavljanja sustava μ [obnavljanja/h] i srednje vrijeme obnavljanja sustava MTTR (eng. *Mean Time To Repair*), MTTR=1/ μ ako je μ =konst. [3][5]

Raspoloživost sustava A(t) je vjerojatnost da će sustav biti ispravan u trenutku t ili ako je bio u intervalu [0,t) bio neispravan da će do trenutka t biti obnovljen. Kao pokazatelj oslonjivosti koristi se granična vrijednost funkcije raspoloživosti K_A ili koeficijent raspoloživosti

$$K_A = \frac{MTTF}{MTTF + MTTR} \qquad (1)$$

MTTF je srednje vrijeme do kvara (eng *Mean Time To Failure*). MTTR je srednje vrijeme obnavljanja. [3][5]

U stanju neraspoloživosti sustava e-poslovanja ne mogu se na siguran način razmjenjivati dobra, usluge i informacije, jer se ne mogu realizirati elektroničke transakcije koje ih prate, tako da su mogući vrlo visoki gubici u poslovanju zbog zatajenja i slabe sposobnosti održavanja sustava e-poslovanja. Iz (1) proizlazi da bi bili oslonjivi, sustavi e-poslovanja u oblaku moraju imati visoku pouzdanost i sposobnost održavanja.

Sigurnost oblaka utječe na percepciju oslonjivosti sustava e-poslovanja u oblaku. Problem jest što još ne postoje sigurnosni standardi za računalne oblake. Sustav e-poslovanja može se implementirati u računalnom koji ima mehanizme koji omogućavaju korisnicima da se mogu pouzdati da se u oblaku s njihovim podacima upravlja na jednako siguran način kao u standardnom IT okruženju [1]. Računalni oblak treba pružati usluge sigurnosnih kontrola prijenosa podataka i skladištenja

podataka u oblaku, zaštite od neovlaštenog pristupa podacima korisnika i druga sredstva za otklanjanje potencijalnih prijetnji.

Integritet ili cjelovitost je sposobnost sustava da spriječi neregularno ažuriranje podataka/informacija, što uključuju njihovo dodavanje, brisanje i izmjene. Pristup podacima uskladištenim u računalnom oblaku, koji se ostvaruje preko računalne mreže, podložan je vanjskim napadima. **Povjerljivost** je sposobnost sustava da podatak/informaciju otkriva samo onim korisnicima (ljudima ili drugim informacijskim sustavima i uređajima) koji imaju pravo pristupa i potrebu za informacijama/podacima. Do donošenja standarda za infrastrukturu i sigurnosnu arhitekturu računalnih oblaka potrebnih za postizanje potpune oslonjivosti sustava e-poslovanja u oblaku, povjerenje između davatelja i korisnika usluga može se uspostaviti i održavati primjenom AAA kontrola[2] i upravljačkih mehanizama za kripto ključeve kojima će se osigurati da podacima u oblaku mogu pristupiti samo autorizirani korisnici [1].

III. PRIJETNJE OSLONJIVOSTI E-POSLOVANJA U OBLAKU

Prijetnje oslonjivosti sustava e-poslovanja (eng *threat*) su neispravnosti, pogreške i zatajenja zbog kojih njegove usluge nisu korektne i korisnik se na njih više ne može osloniti. Otklanjanjem prijetnji održava se potrebna oslonjivost sustava. Utjecaj korisnika na otklanjanje prijetnji njegovom e-poslovanju u oblaku je minimalan, jer korisnik nema na raspolaganju sredstva za održavanje oslonjivosti. Samo davatelj usluga raspolaže sa sredstvima za prevenciju, predviđanje, toleriranje i otklanjanje neispravnosti aplikacija, platformi za izvođenje aplikacija i informacijsko-komunikacijske infrastrukture. Da bi se spriječilo stvaranje lanca prijetnji u oblaku između podsustava e-poslovanja, koje mogu u krajnjoj konsekvenci dovesti do prekida e-poslovanja, u računalnom oblaku treba uspostaviti sustav e-poslovanja koji je tolerantan na neispravnosti.

Kod odlučivanja o izvedbi sustava e-poslovanja u računalnom oblaku, osim dobrobiti koje donosi oblak, poduzeće treba ocijeniti i sve rizike zbog nedostupnosti podataka ili servisa, odnosno resursa računalnog oblaka. Osim poznatih prijetnji (kvarovi hardvera, pogreške i neispravnosti softvera, vanjski napadi na oblak i sl) nedostupnost podataka može biti uzrokovana i legislativom zemlje u kojima se nalaze resursi oblaka. Sve to treba regulirati u ugovoru o kvaliteti usluge (SLA – *Service Level Agreement*) između davatelja i korisnika usluga (dostupnost sigurnosnih kopija u slučaju ispada sustava, visina odštete za gubitke zbog prekida rada sustava i dr) [1]. Jasno je, da će cijena usluga biti to veća što se traži veća pouzdanost i raspoloživost sustava.

[2] eng. AAA – Authentication, authorization, accounting

IV. POUZDANOST SUSTAVA E-POSLOVANJA U OBLAKU

Sustav e-poslovanja jednog poduzeća uspostavlja se u računalnom oblaku, tako da se za odabrani aplikacijski okvir e-poslovanja osiguraju svi potrebni resursi, što smanjuje ukupne transakcijske troškove poduzeća. U slučaju stvaranja zajedničkog lanca vrijednosti dio funkcija e-poslovanja može se izdvojiti u posebne oblake (na primjer, u oblak nabave ulaze svi stalni ili povremeni dobavljači roba i usluga, a u oblak prodaje svi stalni ili povremeni kupci). Velike tvrtke će većinu svojih resursa držati u privatnom oblaku, a dio poslovnih funkcija mogu preseliti u javni oblak (e-pošta, Web maloprodaja, marketing i sl). Privatni oblak može korisniku osigurati veći stupanj kontrole performansi, prijetnji i sigurnosti sustava e-poslovanja, ali je takvo rješenje skuplje od korištenja usluga u hibridnom ili javnom oblaku [1][2].

Pouzdanost sustava e-poslovanja u oblaku je vjerojatnost

$$R(t,\tau) = R_{in}(t,\tau)\cdot R_{pl}(t,\tau)\cdot R_{sw}(t,\tau)\cdot R_{cn}(t,\tau) \qquad (2)$$

gdje su

- $R_{in}(t,\tau)$ – pouzdanost infrastrukture (hardver),
- $R_{pl}(t,\tau)$ – pouzdanost platformi aplikacija e-poslovanja,
- $R_{sw}(t,\tau)$ – pouzdanost aplikacija e-poslovanja,
- $R_{cn}(t,\tau)$ – pouzdanost mrežnih servisa,
- τ - vrijeme potrebno za elektroničku transakciju.

Relacija (2) vrijedi ako su vjerojatnosti ispravnog rada hardvera, softvera i mrežnih servisa međusobno neovisne.

Pouzdanost hardverske komponente računalnog oblaka $R_{in}(t,\tau)$ je sa stanovišta korisnika pouzdanost usluge korištenja hardvera i drugih dodatnih resursa potrebnih za realizaciju sustava e-poslovanja korisnika (virtualni strojevi i diskovi, vatrozid, virtualne lokalne mreže i dr). Korisnici u pravilu unajmljuju samo aplikacije, tako da ne vide infrastrukturu na kojoj se aplikacije izvode, tako da se $R_{in}(t,\tau)$ može uzeti kao mjera utjecaja infrastrukture na pouzdanost sustava e-poslovanja korisnika. Na isti način se tretira **pouzdanost mreže i mrežnih servisa** $R_{cn}(t,\tau)$, pa se mreža i hardver sa stanovišta korisnika mogu promatrati kao jedinstvena cjelina čiji je ukupan utjecaj na pouzdanost e-poslovanja $R_{in}(t,\tau)\cdot R_{cn}(t,\tau)$.

Pouzdanost platformi aplikacija e-poslovanja $R_{pl}(t,\tau)$ uključuje pouzdanost operacijskih sustava, baza podataka i Web poslužitelja. Pouzdanost tih usluga, napose pouzdanost baza podataka i Web poslužitelja, valja promatrati zasebno, odvojeno od aplikacija e-poslovanja. Pouzdanost softvera e-poslovanja $R_{sw}(t,\tau)$ je vjerojatnost obavljanja korektne usluge pomoću unajmljenih aplikacija e-poslovanja. Korektnost te usluge je određena s pouzdanošću softvera koja je jednaka vjerojatnosti da će softver u vremenskom periodu $(t,t+\tau)$ u zadanim radnim

uvjetima izvesti svoje zadaće u skladu sa specifikacijama bez pogreške i u skladu s ugovorenim performansama. Aplikacije e-poslovanja mogu biti sustav planiranja resursima poduzeća (ERP – *Enterprise Resource Planning System*), sustav upravljanja odnosima s kupcima (CRM – *Customer Relationship Management System*), sustav upravljanja dobavnim lancem (SCM – *Supply Chain Management System*) i dr.

Kritičan faktor oslonjivosti sustava e-poslovanja je pouzdanost softvera. To znači da u slučajevima kad se traži visoka pouzdanost i raspoloživost pojedinih aplikacija e-poslovanja nije dovoljno u oblaku osigurati rezervnu infrastrukturu i platforme na kojima se izvode te aplikacije, već treba osigurati više verzija programa. Nakon zatajenja aplikacije korisnik može koristiti druge verzije aplikacije u istom računalnom oblaku ili u drugim računalnim oblacima istog ili drugih davatelja usluga (eng *Cloud Computing Service Provider*, *CCSP*) koji se mogu nalaziti na različitim (geografskim) lokacijama.

Da bi se korisnik zaštitio od prijetnji kao što su velike prirodne katastrofe na određenom geografskom području, potrebno je osigurati i neovisan pristup do rezervnog podatkovnog centra u slučaju ispada ili zatvaranja primarnog podatkovnog centra.

Kad je sustav e-poslovanja korisnika tri-modularni redunantni TMR 3-2-1 sustav (eng *Triple Modular Redundant System*), u kojem je bar jedan sustav ispravan, njegova pouzdanost je

$$R = \binom{3}{3}p^3 + \binom{3}{2}p^2(1-p) + \binom{3}{1}p(1-p)^2$$

odnosno

$$R = p^3 - 3p^2 + 3p \qquad (3)$$

gdje je p pouzdanost sustava e-poslovanja određena prema izrazu (2) [3].

Realno je za pretpostaviti da se ovi sustavi vladaju po eksponencijalnom zakonu pouzdanosti, pa je srednje vrijeme do kvara redundantnog sustava

$$MTTF_{TMR} = \frac{1}{3\lambda} - \frac{3}{2\lambda} + \frac{3}{\lambda} = \frac{11}{6\lambda} = 1{,}83\,MTTF\,(h) \qquad (4)$$

gdje je MTTF srednje vrijeme do kvara osnovnog sustava, MTTF=1/λ [h], a λ [h^{-1}] je ukupni intenzitet kvarova osnovnog sustava [3].

Sustav e-poslovanja u računalnom oblaku s tolerancijom neispravnosti treba imati više pristupa preko različitih prijenosnih medija (eng *Transmission Media*) do davatelja internetskih usluga (eng *Internet Service Provider*) preko kojih se pristupa do računalnog oblaka davatelja usluga s kojim korisnik ima sklopljene ugovore (SLA) [1].

Predmet ugovora s davateljima usluga može biti i izvještavanje o raspoloživosti i pouzdanosti sustava u oblaku koje obuhvaća sve podatkovne centre na svim geografskim lokacijama na kojima davatelj usluga čuva podatke i izvodi aplikacije za e-poslovanje korisnika.

Pouzdanost virtualnih strojeva određuje se na temelju mjerenja vremena između uključivanja i zatajenja virtualnog stroja. Interval od ispada virtualnog stroja do podizanja novog virtualnog stroja je vrijeme obnavljanja tog čvora [6].

Pouzdanost prijenosnih sustava i sustava za pohranjivanje je razmjerno visoka zahvaljujući njihovoj izvedbi koja je tolerantna na neispravnosti [7]. No, ako se ne eliminiraju zavisni kvarovi, čiji uzrok su prirodne katastrofe i drugi vanjski utjecaji primjenom više transmisijskih medija, može doći do zatajenja sustava e-poslovanja s katastrofalnim posljedicama za poslovanje korisnika.

V. DALJNJA ISTRAŽIVANJA

Provedena analiza oslonjivosti sustava e-poslovanja u oblaku ukazala je na potrebu uspostavljanja mjerenja i praćenja oslonjivosti sustava e-poslovanja kao nove usluge koju pruža davatelj usluga pored izvođenja aplikacija e-poslovanja u oblaku tolerantnom na kvarove.

U daljnjim istraživanjima bi trebalo, s obzirom na mrežnu strukturu sustava e-poslovanja u oblaku, proširiti model pouzdanosti za primjenu u detaljnim aposteriornim analizama korektnosti usluga i pouzdanosti sustava e-poslovanja u računalnom oblaku koji će koristiti korisnici. Na temelju tog modela moglo bi se pristupiti i analiziranju odnosa performansi, troškova i oslonjivosti sustava e-poslovanja u računalnom oblaku sa stanovišta korisnika i davatelja usluga.

VI. ZAKLJUČAK

Razvoj tehnologije računalnih oblaka omogućava razvoj sustava e-poslovanja u računalnom oblaku, što je posebno zanimljivo za male i srednje tvrtke. S obzirom da korisnik, nakon uvođenja e-poslovanja, u potpunosti ovisi o informacijskoj i komunikacijskoj tehnologiji, potrebna su sredstva za praćenje oslonjivosti sustava e-poslovanja koja će obuhvatiti informacije o prijetnjama oslonjivosti sustava, pokazatelje pouzdanosti, raspoloživosti, popravljivosti i izvršljivosti, te raspoloživosti, integritetu i povjerljivosti informacija u računalnom oblaku.

Za potpunu oslonjivost sustava e-poslovanja potrebno je primijeniti metode povećanja pouzdanosti, uključivši i N modularnu redundanciju softvera, zbog čega poskupljuje

korištenje usluga sustava e-poslovanja u računalnom oblaku, ali raste i povjerenje korisnika u kvalitetu usluga. Sustavi e-poslovanja u računalnom oblaku tolerantni na neispravnosti (eng *Fault Tolerant e-Business System*), koji imaju visoku izvršljivost i osiguravaju potrebnu raspoloživost, integritet i povjerljivost podataka u oblaku, mogu osiguravati veliku vjerojatnost da će usluge obaviti korektno i tako ispuniti očekivanja svojih korisnika.

Nema jedinstvenog rješenja za oslonjivost informacijskih sustava, pa kroz razvoj oslonjivih sustava e-poslovanja u računalnom oblaku treba neprekidno prepoznavati (nove) prijetnje oslonjivosti, iznalaziti učinkovita sredstva za povećanje oslonjivosti i metode za specifikacije potrebne oslonjivosti za različite grupe korisnika, te neprekidno unapređivati metode mjerenja oslonjivosti sustava kako sa stanovišta korisnika, tako i sa stanovišta davatelja usluga.

REFERENCES

[1] K. Zuva, T. Zuva, K.O.M. Mapoka: "Open Cloud Computing with Fault Tolerance", World Academy of Science, Engineering and Technology, 78, 2013.

[2] Q. Zhang, Lu Cheng, R. Boutaba: "Cloud computing: state-of-the-art and research challenges", J. Internet Serv. Appl. (2010) 1, p. 7-18, 2010

[3] M. Davidović: "Pouzdanost informacijskih sustava", skripta, VSITE Zagreb, 2013

[4] A. Jakupović:"Utjecaj oslonjivosti informacijskog sustava na poslovne organizacije", Zbornik Veleučilišta u Rijeci, Vol 1(2013), No. 1, pp 165-178

[5] V. Sruk: "Pouzdanost računalnih sustava", predavanja (prezentacija), FER, Zavod za elektroniku, mikroelektroniku, računalne i intelegentne sustave, Zagreb, 2011.

[6] S. Malik, F. Huet: "Adaptive Fault Tolerance in Real Time Cloud Computing", 2011 IEEE World Congress on Services, 2011.

[7] Y.-M. Hsieh, Y.-M. Chen:"Reliability Assesments on a Distributed-Computing Based Monitoring Information System", 14th International Conference on Computing in Civil and Building Engineering", Moscow, 2012.

Comparative analysis of tourist satisfaction and online booking services usage for incoming tourists in Zadar County

Božena Krce Miočić, Gabrijela Vidić, Tomislav Klarin
Department of Tourism and Communication Studies
University of Zadar; Franje Tuđmana 24i, HR-23000 Zadar
krceb@unizd.hr; gvidic@unizd.hr; tklarin@unizd.hr

Summary – In a digital environment that is now present in tourism and tourist transport, information and reservation systems are the basis for the realization of businesses. Besides reducing operating costs and 24-hour service availability, an additional benefit can be seen in dynamic pricing of services. Airlines first noticed the advantages of information-communication technologies (ICTs) and developed a whole new concept of business in the form of low-cost airlines (LCAs).

LCA tourists have to use information technologies in order to use LCA services, therefore it is expected that the Internet usage of low-cost airline tourists significantly differs from the Internet usage of average tourists.

Over the last several years Zadar County has strongly encouraged low-cost airlines to link its airport with as many European destinations as possible, which enabled the connection with remote outbound markets and the increase of city breaks tourism.

This paper will show the results of a research which was conducted on a sample of 900 low-cost airline (LCA) tourists and 1697 average tourists who stayed in Zadar County. Different methods of descriptive statistics were used in order to determine and compare the level of usage of information technologies of these two types of tourists.

I. INTRODUCTION

Carriers and tourism economy make up an inseparable whole. Development of means of transport resulted in the development of tourism as we know it, while the development of airlines played a key role in a creation of a global tourism market as it made any destination reachable in a short period of time. Air transport cost component had made this transport exclusive until low-cost airlines (LCAs) appeared on the market [1]. Researches show a significant positive influence LCAs have on tourism in the destination [2]. At the same time, airlines and workers in the hotel industry are among first users to recognize the advantages of the use of information-communication technologies in their business (ICTs). They see the implementation of ICTs in their business both as a new sales channel and a powerful marketing tool realizing its advantages at large scopes of target population through promotional actions of high level response, thus optimizing the use of tourism capacities. Moreover, auction system of price determination allows LCA users to see LCA advantages [3] while LCA can follow tourists' post-purchase actions.

LCA tourists are Internet usage-oriented in the purchasing process, therefore it is expected that − in gathering information and accommodation reservation in the destination – Internet usage of LCA tourists is more significant than the Internet usage of average tourists.

II. THE USE OF INFORMATION-COMMUNICATION TECHNOLOGIES IN TOURISM

Contemporary trends in tourism show a complete affirmation of ICTs in tourism business, offering different possibilities for changes and innovations. With the development of the Internet, ICTs have significantly influenced tourist offer meeting market demands as well as they enabled easier accessibility and tangibility of tourist products and services. Along with this, ICTs have changed information channels. Therefore, physical media such as brochures, tourism magazines and catalogues have partially been replaced with electronic information characterized by its easy accessibility and quality timeliness. Tourist service information has become easily verifiable with unlimited possibilities of distribution. As a result of this, today tourists are more sophisticated and more informed, with more and more complex demands and common requests for individual tourist journeys.

Gathering information is a starting point when making decision about a tourist journey. Within this process, information about tourist destination, means of transport, accommodation, cultural activities, entertainment, and other, is found and selected. Therefore, being informed presents a complex multi-stage process including one or more persons getting relevant information from different sources – gathering information depends on the composition of the group going on a journey, on presence of friends and relatives in the tourist destination to be visited, earlier staying at the destination as well as on the destination's popularity [4]. Fodness and Murray [5] point out three main factors influencing tourists' behaviour and their decision on going on a tourist journey: motivation, determinants and sources of information. Gursoy and McCleary [6] have developed a theoretical model explaining tourists' behaviour while gathering information, and certain studies have been conducted, related to tourists' sources of information, pointing out some of the categories in gathering information: informal sources, such as friends, relatives, colleagues or other tourists; earlier experience; tourist agencies or tour operators; tourist offices at the level of states, regions, or towns; various brochures, catalogues and tourism magazines; airline web sites; other web sites [7].

The Internet has changed tourism behaviour significantly [8], also stated by Buhalis and Law [9] in their retrospective review of ICTs progress and use in

tourism in their paper which stresses the development and implications of the Internet usage. Various tools, such as online reservation and online purchase, have replaced mediary services of tourist agencies and tour operators [10], and the Internet has become the basic marketing tool [9]. Certain researches show that tourists who gather information via the Internet spend more money in the destination than the tourists who gather information from other sources [11]. Global online agents, such as Lastminute.com, Expedia or Orbitz, are of great importance here. With their simple and clear web sites, they enable tourists getting all information needed, as well as making online reservation and purchase, all in one place. Low-cost airlines become more and more popular, as they provide, along with online ticket purchase, certain information about tourist destinations marked in their airlines [12]. The Internet has put a tourist into the center of planning and organization of his/her journey. However, certain researches show 55% of guests, before making a reservation of accommodation, read one to five online reviews [13]. The Internet is not an only source of information – information can be found in other electronic forms or in word-of-mouth referrals [14], and researches show tourist combine the Internet and other sources, such as brochures and catalogues [15], that is, getting information depends on online as much as on offline sources of information [16].

III. LOW-COST AIRLINES AND THEIR EFFECT ON TOURISM

During the past thirty years, there has been a significant increase and changes in the world airlines. An increase of 5.2% was reported in 2013 compared to 2012, despite high fuel prices and relatively slow economic growth. Air transport expansion is linked to economic expansion of emerging markets [17].

Although it is generally thought that low-cost airlines enabled more people to travel, the researches of Civil Aviation Authority (CAA) have shown there were no significant changes in passenger structure compared to 1990s. There was an increase in the number of passengers due to low-cost airlines, but LCA users are still mostly the ones with medium to high purchasing power, who can now travel more often and go on shorter trips [18]. The first so-called low-cost airlines occurred in the USA in 1978, after adoption of US Deregulation Act. After that, they occurred in Australia in 1990, Great Britain and Ireland in 1995, Canada and New Zealand in 1996, Japan in 1998, the EU in 1999, while in the rest of the world they occurred after 2001 [19].

Common features of the majority of low-cost airlines are: high aircraft utilization, internet booking, use of secondary airports, minimum cabin crew, lower wage scales, lower rates of unionization among employees, one class of seating thus allowing more seats per aircraft than traditional airlines, short ‘on the ground’ turn around times, no cargo carried to slow down turn around times, a simple fare structure and pricing strategy, e-ticketing, no seat allocation, passengers having to pay for food and drink, flexible working terms and conditions for employees relative to traditional airlines, point to point services and no connections [19].

An additional factor enabling cost reduction for airline companies, and thus lower ticket prices for passengers, is the fact that there is only one class of seating and therefore only one type of service; also, airline companies do not depend on travel agencies and do not pay commitment fees, telephone or online reservations can be made, point-to-point tickets are sold, no connections are possible (simplifying luggage handling and planes turnaround), no tickets are printed; there are no frequent flyer programs, and non-integration into alliances is the rule. Moreover, when a flight is cancelled or delayed, travellers usually do not get compensation. Since many low-cost airlines are not members of IATA (International Air Transport Association), they are not obliged to meet the quality standards imposed upon other airline companies [20].

Liberalization of European air transport resulted in a fast increase in the number of low-cost airlines because lower costs of air transport enabled a development of new tourist markets [21]. According to UNWTO data, more than a half of tourists (52%) in 2012 (the total number of tourists being 1.035 million) arrived in the destination by means of air transport. It is estimated that air transport will still increase, with a faster rate than land and water transport [22]. At the same time, there has been an increase in the number of young people and students who travel by air, which is an occurring trend among the young. In 2007, 16% of them travelled by air, while there was a significant increase in 2012/2013, with 26% of the young travelling by air, which was also a result of the growth in the LCA offers [23].

According to the data provided by European Low Fares Airline Association, from June 2012 to July 2013 Ryanair carried 79.9 million passengers (39%), the total number of passengers being 206.8 million carried by 10 LCAs. Second best air carrier according to the number of passengers was easyJet with 60.2 million of passengers. In June 2013 the average number of flights for Ryanair was 1600, to 180 destinations and 26 countries [24].

LCAs stimulate the development of city tourism, weekend tourism and short-break tourism as well as the development of new potential destinations [25].

A research by O'Conell and Willams has shown that LCAs attract a large number of young travellers (unlike other companies), their clients are more often groups than individuals, the self-employed, and main reasons for their travelling are visiting friends and relatives (VFR), vacation and short trips. Also, the research has shown that 31% of Ryanair's passengers stayed in hotels. The European Travel Trade explains this by stating that passengers with lower-priced tickets spend their savings on a more luxurious accommodation, such as hotel [26]. This results in an increase of various motivational factors for travelling and more common journeys for private reasons (vacation), but also in acquisition of new market niches, such as the market for cultural tourism, second home market, pensioner journeys, etc.

Competitive advantage of LCAs is seen also through the use of direct sales channels, such as web sites. Researchers have shown that passengers search available reservation channels for purchase of lower-priced tickets

even in countries with no high usage of Internet [26]. A research on the use of IT technologies in air transport from 2013 has shown that more than 90% of passengers think new technologies help them with their journeys. In 2013, 69% of passengers made an online ticket reservation, and 20% used a kiosk check-in on the day of the journey [27].

Passengers want to use IT during their journey, especially if this will make their journey easier or their costs reduced. Therefore, LCAs will definitely expand their market gradually. Along with this, it is obvious how LCAs' competition is unable to implement new technologies and IT innovations.

IV. RESEARCH METHODOLOGY

The aim of the research was to define the profile of the tourists in Zadar County and the profile of tourists using the LCA services, to compare their features as well as ways of their Internet usage for gathering information about destination and for organization of journeys. We have chosen Zadar Airport for an analysis of LCA users.

Zadar Airport and Zadar County have been chosen due to their LCA-user-oriented tourism. Zadar Airport is one of Ryanair's bases, which enabled business development of this airport. LCA started their business in Zadar Airport in 2002. Statistical data of Zadar Airport show a more significant increase in air transport only after 2007, with the occurrence of Germanwings and Ryanair.

The data available show an increase in the number of Zadar Airport passengers from 39.244 for 2001 to 472572 for 2013 (12 times larger number). There was also an increase in the number of airplanes, from 1 226 in 2001 to 4 050 in 2013 [28].

For the purpose of this research, a structural questionnaire was given to tourists (no matter which transport they used to reach to the destination) in Zadar County (including most cities and villages of the county). The research was conducted from August to September 2012, and then from May to August 2013, using random sampling including 1.697 tourists. The questionnaire was offered to tourists in eight languages. All questionnaires were filled out in the presence of a researcher.

An additional research using a structured questionnaire in English, German and Croatian was conducted on a sample of 900 LCA tourists from May to October 2013 on outbound flights at Zadar Airport. All questionnaires were filled out in the presence of a researcher.

The results obtained in the primary research were analyzed at two levels: in terms of the total number of tourists a segment of LCA tourists was also observed, and for the complete information on LCA tourists the data were obtained through an additional examination of only this segment of tourists.

The data obtained were processed in an SPSS program using methods of descriptive statistics.

V. RESEARCH RESULTS

In its strategic determinants, Zadar County has recognized the importance of LCA for its airport. Therefore, according to internal data of Zadar Airport, depending on the planned number of passengers, Zadar County invests additional financial resources of the overall tourist offer of the County in promotion of the destination on Ryanair web sites, in the form of joint marketing.

The research has shown 11.3% of total number of tourists in Zadar County chose LCAs for their air carriers on the way to their destination, while most of the tourists came to the County by car (53.4%). Average tourists (29.4%) as well as LCA users (28.1%), who chose Zadar County for their vacation, point out the main motive for doing this was enjoying the sun and the sea: with regard to the primary motive, the basic difference between these two categories is the motive of nautics. Nautics is the primary motive for travelling for only 0.7% of average tourist, while even 7.7% of LCA users point out nautics as their primary motive for travelling. Many researches [29] find that nauticians are guests with a higher purchasing power, thus, we can conclude that LCA use is not conditioned by reduced financial resources for vacation. Also, we should not neglect the fact that 3.3% of LCA tourists more than average tourists decide to go to this destination because of the favorable offer. In the framework of the research of secondary motives for the choice of destination, it is shown how even 13.3% of LCA tourists consider the destination of Zadar to be near, although the most of the respondents were from Great Britain (41.1%).

When analyzing both categories, word-of-mouth referrals present the most important channel in getting first information. Although previous researches showed LCA tourists obtained first information about the destination of Zadar mostly via Internet (50.78%) [30], today the primary source is the information obtained through relatives and friends (57.1%), while 22.6% of respondents point out the Internet as their primary source of information.

Both categories analyzed mostly obtain information about the destination via Internet. An average tourist obtained the information about Zadar County in 56.9% cases, while 76.6% of tourists using LCA services obtained the information via Internet. LCA tourists only in 4.2% cases say they needed no information about the destination, thus presenting the most informed group of tourists. Compared to this, 11.7% of tourists using land public transport believed they did not need information about the destination upon coming to the destination. Both groups mostly visited web sites of Croatian National Tourism Board at all levels (from state level to local level) (Figure 1).

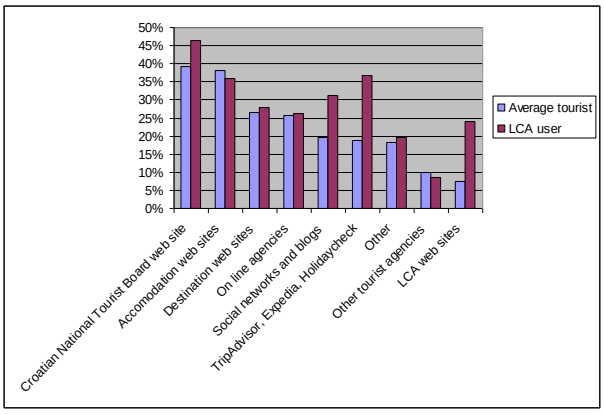


Figure 1. Web sites used for gathering information about the destination

Mostly, in the title of web sites of tourist offices there is the name of the destination, which makes tourists visit the web sites more often as these destinations are the first to appear when entering the name of the destination in Google browser, which shows other web sites visited. The data obtained are different from those from Miller's research, according to which 61% of tourists use the Internet often or always in planning their journeys, a relatively small number of tourists visit the destination's web sites, and the most visited web sites are the ones for searching accommodation in the destination [31]. Graham and Shaw [18] point out that some of the fundamental principles in LCA business are sophisticated web sites with detailed information about destination. Among all respondents, LCA tourists more often visit LCA web sites (21.5%) in order to gather information about destination compared to all other modes of transport and compared to average tourists (7.5%). Most visited LCA web sites are Ryanair's sites (77.0%), as expected because of the flight structure in Zadar Airport.

A great increase in the category of social networks and blogs during past five years has opened up new possibilities of the destination in its communication with tourists.

Among all respondents, there can be seen a significant difference in the usage of social networks and blogs in terms of modes of transport. 50.1% of LCA tourists gather information about the destination via social networks, while 20% of all users of other modes of transport (camp tourists - 8%) gather the information via social networks. An additional research among LCA tourists has shown that relatives and friends are the most important source of information about accommodation, too (53.6%), while the Internet is their second source (including digital media - 39.5%). Web sites of tourist offices are the most common source of information about accommodation here too (31.7%), followed by web sites of the airport (8.7%). The percentage of other answers (59.7%) is divided among various other web sites: booking.com, Ryanair, hostel world, hostel booking, tripadvisor, etc.

Gathering information about the destination opens up possibilities for attracting tourists or for usage of the overall tourist offer of a destination. LCA tourists have significantly higher net monthly incomes than average tourists. While monthly incomes of an average tourist visiting Zadar County are mostly (43.8%) below 1 000.00 EUR, monthly incomes of LCA tourists are mostly (27.7%) above 3 000.00 EUR.

Vacation organization has changed completely during past 30 years. While organized journeys were typical of 1980s, today the most common type of a journey is an individual journey. One of the preconditions for such a trend is definitely the development of the Internet.

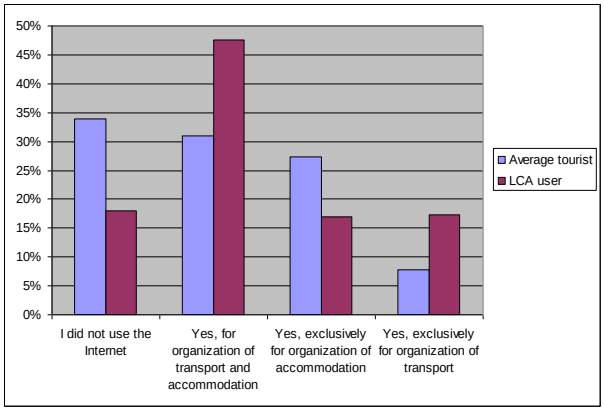


Figure 2. The usage of the internet in the organization of a vacation

As Figure 2 shows, LCA tourists use the Internet for the purpose of their vacation organization more often than average tourists. Only 18.1% of LCA tourists did not use the Internet in the organization, compared to 33.9% of average tourists not using the Internet. LCA tourists not using the Internet are not decision-makers as they are a journey accompaniment. An additional research among LCA tourists has shown that 20% of the users did not make accommodation reservation before their arrival. The largest number of the tourists (36.6%) made an online reservation via web site of accommodation objects. 16.2% of the users made an online reservation via some other web site, and 13.2% via e-mail. These data show a strong prevalence of the Internet as a distribution channel in accommodation tourism objects.

The Internet as a medium has enabled a simpler journey organization, therefore it is expected for the tourists with an Internet access to feel more satisfied with the destination, to recommend the destination and to stay loyal to it.

Table 1 shows indexes of satisfaction and desire to return to the destination according to the level of Internet usage in organizing the journeys. The significance of the differences among groups with various levels of the Internet usage in organizing accommodation and the overall satisfaction with the destination, destination recommendation and return to the destination has been verified by a variance analysis. F-ratios are significant in terms of the overall satisfaction with the destination ($F=11.178$, $p=0.000$), return to the destination ($F=5.385$; $p=0.001$), and destination recommendation ($F=38.889$; $p=0.000$). Therefore, it is concluded that there are

significant differences in terms of the desire for return to the destination and destination recommendation according to the level of Internet usage in organizing the journeys.

TABLE 1. RATIO OF SATISFACTION AND LOYALTY IN TERMS OF THE INTERNET USAGE IN ORGANIZING THE JOURNEYS

	Usage of the Internet for the organizaation	*Mean*	*Std. Deviation*	*Std. Error*
Overall satisfaction with the destination	Accommodation and transport	83.00	14.84	0.65
	Accommodation	84.37	15.16	0.71
	Transport	84.11	15.16	1.34
	I did not use it	85.15	15.81	0.66
	Total	**85.21**	**15.43**	**0.38**
Destination recommen-dation	Accommodation and transport	85.88	15.61	0.68
	Accommodation	85.06	15.53	0.73
	Transport	84.72	16.34	1.45
	I did not use it	88.66	16.73	0.70
	Total	**86.51**	**16.10**	**0.39**
Return to the destination	Accommodation and transport	42.75	29.27	1.29
	Accommodation	51.50	30.69	1.44
	Transport	44.79	33.81	2.99
	I did not use it	63.18	35.12	1.48
	Total	**52.22**	**33.18**	**0.81**

T-test analysis has shown significant differences between Internet users and non-users in organizing the journeys as well as in the categories of the overall satisfaction with the destination, desire to recommend the destination and loyalty to the destination. An analysis of the overall data shows Internet non-users are more prone to return to the destination (63.18 index points) than Internet users (46.6 index points).

Furthermore, it is expected that using a cheaper and faster air carrier makes journey easier and increases the satisfaction with staying in the destination, desire to recommend the destination and to return to the destination.

TABLE 2. AN ANALYSIS OF THE USAGE OF THE INTERNET IN ORGANIZING THE JOURNEYS AND SATISFACTION WITH AND LOYALTY TO THE DESTINATION, T-TEST

Internet usage		**Variances**		**t-test for Equality of Means**				
		F	***Sig.***	***t***	***df***	***Sig. (2-tailed)***	***Mean Differ.***	***Std. Error Differ.***
Overall satisfaction with the destination	Yes	16.18	0.00	-5.61	1645	0.00	-4.46	0.79
	No			-5.52	1082	0.00	-4.46	0.80
Destination reccomme-ndation	Yes	5.71	0.01	-3.90	1648	0.00	-3.26	0.83
	No			-3.82	1061	0.00	-.326	0.85
Return to the destination	Yes	37.09	0.00	-9.87	1645	0.00	-16.57	1.67
	No			-9.45	1000	0.00	-16.57	1.75

Table 3 shows indexes of satisfaction and desire to return to the destination in terms of modes of transport. Moreover, the significance of the differences between the groups in terms of modes of transport used to get to the destination and the overall satisfaction with the destination, destination recommendations and return to the destination has been verified by variance analysis. F-ratios are significant in terms of return to the destination (F=6.896; p=0.000) and destination recommendations (F=4.342; p=0.005), while there is no significance in terms of the overall satisfaction with the destination (F=2.199; p=0.096). Therefore, it is concluded there are significant differences in the desire for return to the destination and destination recommendation in terms of the modes of transport used to get to the destination.

TABLE 3. RATIO OF SATISFACTION AND LOYALTY IN TERMS OF MODES OF TRANSPORT TO THE DESTINATION

		Mean	*Std. Deviation*	*Std. Error*
Overall satisfaction with the destination	Own transportation	85.46	15.80	0.47
	Airplane - LCAs	86.13	11.78	0.85
	Land public transportation	82.90	16.86	1.09
	Other public transportion	85.62	14.52	1.23
	Total	**85.19**	**15.46**	**0.37**
Destination recommen-dation	Own transportation	86.35	16.30	0.49
	Airplane - LCAs	89.34	12.04	0.86
	Land public transportation	84.08	18.47	1.18
	Other public transportion	88.07	13.68	1.16
	Total	**86.51**	**16.06**	**0.39**
Return to the destination	Own transportation	55.04	33.24	0.99
	Airplane - LCAs	48.35	31.98	2.29
	Land public transportation	47.90	32.62	2.11
	Other public transportion	45.22	35.44	3.02
	Total	**52.44**	**33.37**	**0.81**

VI. CONCLUSION

LCAs have introduced significant changes in the overall market of transportation services. This has especially reflected on tourist journeys now characterized by acceptable prices of air transportation and easy accessibility of any destination. LCA business system is oriented towards Internet as a distribution channel. The research has shown that LCA tourists use the Internet more often than average tourists when it comes to gathering information about the destination. Mostly visited web sites are those of tourist offices, and LCA tourists gather information via social networks more than average tourists. LCA tourists are guests with above-average monthly incomes, and their motives for the journey are the same as the motives of average tourists, except in the segment of nautics, considered to be the main motive for LCA tourists' visit to the destination. LCA tourists make online accommodation reservations more often than average tourists. Mostly, they visit the web sites of the accommodation tourism object. Both categories analyzed point out a high level of the overall satisfaction with the destination, while in terms of destination accommodation and return there is a lower

level of agreement but more significant differences among categories of tourist coming to the destination by different modes of transport. An analysis of the level of satisfaction and loyalty of the tourists who use the Internet for the purpose of journey organization shows a significantly higher Internet non-users' level of satisfaction.

REFERENCES

[1] T. Bieger, T. and A. Wittmer, "Air transport and tourism - perspectives and challenges for destinations, airlines and governments," *Journal of Air Transport Management*, vol. 12 (1), 2006, pp. 40-46.

[2] B. Rey, R. L. Myro and A. Galera, "Effect of low-cost airlines on tourism in Spain. A dynamic panel data model," *Journal of Air Transport Management*, vol. 17 (3), 2011, pp. 163-167.

[3] K. Button, Transport Economics, Cheltenham: Edward Elgar Publishing, 2010.

[4] D. Snepenger, K. Meged, M. Snelling and K. Worrall, "Information search strategies by destination-naive tourists," *Journal of Travel Research*, vol. 29 (1), 1990, pp. 13-16.

[5] D. Fodness and B. Murray, "Tourist information search," *Annals of Tourism Research*, vol. 24 (3), 1997, pp. 503-523.

[6] D. Gursoy and K. McCleary, "An integrative model of tourist's information search behaviour," *Annals of Tourism Research*, vol. 31 (2), 2004, pp. 353-373.

[7] D. Snepenger and M. Snepenger, "Information search by pleasure travelers," 1993, in: *Encyclopedia of hospitality and tourism*, M. A. Kahn, M. D. Olsen and T. Var, New York: Van Nostrand Reinhold, pp. 830-835; B. Dey and M. K. Sarma, "Information source usage among motive-based segments of travellers to newly emerging tourist destinations," *Tourism management*, vol. 31, 2010, pp. 341-344; D. Gursoy and W. T. Umbreit, "Tourist information search behaviour: cross-cultural comparison of European union member states," *Hospitality Management*, vol. 23, 2004, pp. 55-70; B. Sparks and G. W. Pen, "Chinese Outbound tourists: Understanding their attitudes, constraints and use of information sources," *Tourism management*, vol. 30, 2009, pp. 483-494; I. Bizirgianni and P. Dionysopoulou, "The influence of tourists trends of Youth Tourism through Social media (SM) & Information and Communication Technologies (ICTs)," Procedia – Social and Behavioral Sciences, vol. 73, 2013, pp. 652-660.

[8] J. Mills and R. Law, "Handbook of consumer behaviour, tourism and Internet," New York: Harworth Hospitality Press, 2004, in: D. Buhalis and R. Law, "Progress in information technology and tourism management: 20 years on and 10 years after the Internet – The state of eTourism research," *Tourism Management*, vol. 29, 2008, pp. 609-623.

[9] D. Buhalis and R. Law, "Progress in information technology and tourism management: 20 years on and 10 years after the Internet – The state of eTourism research", *Tourism Management*, vol. 29, 2008, pp. 609-623.

[10] A. M. Morrison, S. Jing, J. T. O`Leary and A. C. Lipping, "Predicting usage of the Internet for travel bookings: An exploratory study," *Information Technology & Tourism*, vol. 41 (1), 2001, pp. 15-30.

[11] M. A. Bonn, H. L. Furr and A. M. Susskind, "Using the Internet as a pleasure travel planning tool: An examination of the sociodemographic and behavioural characteristics among Internet users and non-users," Journal of Hospitality & Tourism Research, vol. 22 (3), 1998, pp. 303-317; M. Luo, R. Feng and L. A. Cai, "Information search behaviour and tourist characteristics: The Internet vis-à-vis other information sources," *Journal of Travel & Tourism Marketing*, vol. 17 (2/3), 2004, pp. 15-25.

[12] S. Klein, "Web impact on the distribution structure for flight tickets," in: *Information and communication technologies in tourism*, K. W. Wöber, A. J. Frew and M. Hitz, Eds. New York: Springer Wien, 2002, pp. 219-228.

[13] Trip Advisor, http://www.tripadvisor.com/PressCenter-c7-Survey_Insights.html, accessed February 2nd 2014.

[14] D. G. Pearce and C. Schott, "Tourism distribution channels: the visitors' perspective," *Journal of Travel Research*, vol. 44, 2005, pp. 50-63.

[15] C. Seabra, J. L. Abrantes and L. F. Lages, "The impact of using non-media information sources on future use of mass media information sources: the mediating role of expectations fulfilments," *Tourism management*, vol. 28, 2007, pp. 1541-1554.

[16] O. Gronflaten, "Predicting travellers` choice of information sources and information channels," *Journal of Travel Research*, vol. 48 (2), 2009, pp. 230-244.

[17] Air Passenger Market Analysis, IATA Economics, http://www.iata.org/publications/economics/Documents/passenger-analysis-dec2013.pdf, accessed February 6th 2014.

[18] B. Graham and J. Shaw, "Low-cost airlines in Europe: Reconciling liberalization and sustainability," Geoforum, vol. 39, pp. 1439-1451, 2008, pp. 1442.

[19] G. Francis, I. Humphreys, S. Ison and M. Aicken, "Where next for low cost airlines? A spatial and temporal comparative study," *Journal of Transport Geography*, vol. 14, 2006. pp. 83-94.

[20] F. Dobruszkes, "An analysis of European low-cost airlines and their networks," *Journal of Transport Geography*, vol. 14, 2006, pp. 249-264.

[21] P. Forsyth, "Martin Kunz Memorial Lecture. Tourism benefits and aviation policy," *Journal of Air Transport Management*, vol. 12, 2006, pp. 3-13.

[22] UNWTO Tourism Highlights, Edition 2013, http://dtxtq4w60xqpw.cloudfront.net/sites/all/files/pdf/unwto_highlights13_en_lr_0.pdf, accessed February 14th 2014.

[23] Tourism Research and Marketing and Laura Daly, "New Horizons III – A global study of the youth and student traveller," WYSE Travel Confederation, 2013, http://wysetc.files.wordpress.com/2013/12/newhorizonsiii-v7-execsummary-v4s.pdf, accessed February 14th 2014.

[24] European Low Fares Airline Association (ELFAA), *ELFAA members' statistics are updated every six months, in June and December of each year, http://www.elfaa.com/statistics.htm, accessed February 14th 2014.

[25] B. Graham and J. Shaw, "Low-cost airlines in Europe: Reconciling liberalization and sustainability," *Geoforum*, vol. 39, 2008, pp. 1439-1451.

[26] J. F. O'Conell and G. Willams, "Passengers' perceptions of low cost airlines and full service carriers: A case study involving Ryanair, Aer Lingus, Air Asia and Malaysia Airlines," *Journal of Air Transport Management*, vol. 11, 2005, pp. 259-272.

[27] Air Transport Industry Travel Insights, The Passenger IT Trend Survey, SITA, Air Transport World, 2013, http://www.sita.aero/surveys-reports/industry-surveys-reports/passenger-it-trends-survey-2013, accessed February 14th 2014.

[28] Internal data of Airport Zadar, http://www.zadar-airport.hr/media/pdf/02-Passanger-HR.pdf, accessed February 14th 2014.

[29] Stavovi i potrošnja nautičara u Hrvatskoj - *TOMAS NAUTIKA Jahting 2012*, http://www.iztzg.hr/UserFiles/Pdf/novosti/2012-Tomas-nautika JAHTING.pdf, Institut za turizam, Zagreb, 2012.

[30] B. Krce Miočić, V. Sušac and T. Milković, "Interaction between Low-Cost Airlines and the Development of Tourist Destination," 2nd Advances in Hospitality and Tourism Marketing & Management Conference. Corfu, 2012.

[31] D. Miller, "Destination websites – Learning what consumer want," Presentation at the Travel & Tourism Research Association Conference, New Orleans. 2005, in: S. Pike, *Marketing turističkog odredišta*. Zagreb: M Plus. 2010, pp. 342.

Economic and financial analysis of investments in information security

Zvonko Čapko[1], Saša Aksentijević[2], Edvard Tijan[3]

[1] University of Rijeka, Faculty of Economics
Ivana Filipovića 4, Rijeka, Croatia
Tel: +385 51 35 51 52 Fax: +385 212 268 E-mail: zvonko.capko@efri.hr
[2] Aksentijević Forensics and Consulting, Ltd.
Gornji Sroki 125a, Viškovo, Croatia
Tel: +385 51 65 17 00 Fax: +385 51 65 17 81 E-mail: axy@vip.hr
[3] University of Rijeka, Faculty of Maritime Studies
Studentska 2, 51000 Rijeka, Croatia
Tel: +385 51 33 84 11 Fax: +385 51 33 67 55 E-mail: etijan@pfri.hr

Abstract - The paper analyzes information security investments from both economic and financial points of view. Authors research special characteristics of information security investments and the difficulties which appear during the determination of input parameters of the economic analysis. Challenges in application of internal rate of return method are also analyzed. The constructed model of cash flow analysis of information security solution investments shows the feasibility of calculating the net present value, internal rate of return and the return on capital when using such an approach.

I. INTRODUCTION

Within the usual paradigm of business financing, the investment value of assets (which are owned or controlled by an enterprise and which produce certain revenue) is evaluated against the related costs (for example, maintenance, or procurement of raw materials). Assets are usually divided into material assets (machinery, buildings), non-material assets (patents, software, goodwill) and a special form of assets capable of intrinsic reproduction, called financial assets.

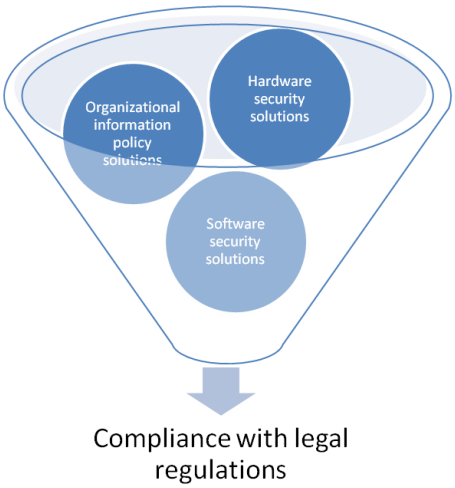
One of the tools most often used by financial managers is determining whether the Net Present Value of assets is positive by using the marginal capital cost, relevant for a certain type of business or certain type of entrepreneurial undertaking.

Information Security Management Systems, or ISMS, (technically speaking) consist of the following four components, organized in a hierarchical manner:

1. Organizational forms ensuring the compliance with legal regulations
2. Organizational information policy, or the knowledge of users and management regarding the functioning and managing of ISMS's, resulting in adequate application of risk removal techniques by using hardware, software and orgware, often formalized by security certification (e.g. ISO 27001:2005),
3. Computer hardware (servers, switches, computers, network devices, routers),
4. Computer software and applications.

The relation between the above mentioned components is shown in Figure 1.

Figure 1. Interaction between different solutions used to achieve the goals of information security



source: authors

All of the above mentioned components are related to a certain level of cost, where there is a high level of solution substitutiveness, meaning the treatment of capital investment or operative cost. Almost all modern ICT solutions, especially in complex corporate environments, can be obtained in a form which can be considered either as an investment or as operative cost [1]. For example, instead of purchasing computer applications, they could be rented or leased as a SaaS (Software as a Service) model. Instead of purchasing computer hardware, it could be rented or leased as a Cloud. Several issues can be identified while tackling difficult choice of investing in own information security assets in comparison to leasing the solution:

1. On a technical level, such divergence of possible solutions could create problems in terms of constant demands for additional education, arising from quickly changing technologies.

2. On an economic level, the total investment cost can further be obscured.

3. On an operative level, a whole set of additional costs exist, related to the setup and functioning of ISMS (application costs, ICT infrastructure etc.).

Furthermore, other costs related to information security can occur in enterprises, addressing the technical disciplines, such as the cost of training the personnel which implement information security technologies, the cost of training the employees for using the new technologies, etc. In terms of capital investments, the basic investment forms are related to security hardware, software, security telecommunication solutions, investments in ICT infrastructure, security studies and organizational blueprint implementations (among them also the applications of best practice models of ICT security).

The scope of this paper is only evaluation of the specific characteristics of the investment in ISMS, when such an investment is a part of enterprise's capital budgeting activity.

II. DIFFICULTIES IN DETERMINING THE INPUT PARAMETERS OF ECONOMIC ANALYSIS

When moving away from technical, technological and organizational aspects of ISMS and when approaching the investment aspect of such systems, a number of difficulties can arise when trying to determine the input parameters of financial analysis [2]. Some of the difficulties can be identified as follows:

1. The decisions about ISMS investments depend upon the risk assessment as a professional/specialist activity [3]. Practical risk assessment of information security is generally void of quantitative financial indicators, and is usually based on experiential assessment of the probability of unwanted occurrences and their impact on business [4], without quantifying possible damages. Risk assessment is an abstract technical assessment, providing the indicators which guide information security experts towards the fields which should be taken into special consideration in order to achieve the set business goals of information security [5]. Paradoxically, because of the lack of the detailed financial analysis (apart from the basic cost analysis), the management that deals with investments (without possessing the specific technical knowledge) often shows aversion towards the ISMS investments and implicitly takes unreasonably high levels of risks. [6] The reverse scenario is also possible, overinvesting in ISMS. This specific situation results from the traditional separation of information security, control and risk management processes from the corporate ICT management, which is generally more oriented towards the system efficiency assessment and towards ensuring the adequate return on investments [7].

2. The high level of substitutability of ISMS investments with the costs that can be considered as operative costs often complicates investment decisions. [8] Generally, such "alternative" solutions may seem acceptable, but superficial analysis often do not include numerous hidden costs, such as customer education costs, technical personnel training costs (the people which will work on solution implementation and maintenance [9]), etc.

3. Software, hardware and telecommunication solutions obtained as long-term investments by the enterprise usually imply the necessity of maintenance contracting [10]. Such contracts offer customer support through the lifetime of the purchased information security solution, but even more often it is impossible to use the solution without signing such contracts. In case that such contract has not been signed, certain hardware or software product may not be able to function properly (for example lack of security patches or similar problems which may occur during work). Regarding antivirus/antispam solutions or network devices with security characteristics, it would be impossible to receive instant definitions of certain attacks or related security profiles and definitions. Furthermore, it would be difficult to integrate a certain solution into ISMS, because its settings might not allow the integration or application to increased number of users or workplaces [11]. For that reason, when performing the initial input parameter analysis it is necessary to discern which part of the initial costs represents investments, and which part represents yearly maintenance. This should be done in order to correctly determine the total cost of implementation and exploitation of a certain ISMS.

4. It is difficult to correctly predict the real residual value of certain information security investments [12]. Some such solutions have a long lifespan (for example telecommunications and network solutions) and could be used even after being fully depreciated. The functionality of some systems might depend on the regular payments to the supplier, who is able to continuously upgrade and modernize a certain solution (which is especially applicable to software security solutions).

5. Small and medium enterprises often lack the specific knowledge necessary to adequately assess the influence of information security investments on enterprise performance [13].

III. CHARACTERISTICS OF INFORMATION SECURITY INVESTMENTS' ECONOMIC ANALYSIS

When analyzing the economic flow of ISMS solution, the structure of the source of financing is ignored (own assets or loan/financial lever) because the loan interest does not reduce the solution's economic potential, but enterprise's economic potential as a whole [14].

The classic investment economic flow analysis (or in this case the investment into information security infrastructure) calls for the initial assessment of effects reduced to market prices from the first year of the project duration [15]. Those items which reduce the economic potential of the project or solution are considered expenses. In this context, the expenses may be:

- Initial investments in information security solution or project
- Cost of project or solution maintenance
- Material expenses for using the solution (electricity, utilities)
- External services related to the solution (consulting)
- Training costs for solution implementation (permanent employees)
- Training costs for solution usage (permanent employees)
- Gross salary for employees in charge of solution implementation (reduced to full time equivalents)

When using the economic cash flow analysis of the information security solution, the problem of determining the income against determining the expenses can arise. Positive cash flow of such investments usually consists of the overall income during the solution lifetime and the estimated residual value at the end of the lifetime. [16] Such incomes can be properly defined only in enterprises who mainly provide security solutions to others, or who provide services to others and undertake the investment activity themselves.

All other enterprises do not derive income from information security solutions, but negative cash flow instances in cash flow analysis could be directly compared with the potential damage caused to the enterprise in case of occurrence of some security lapse during the usage of the security solution.

From the static point of view of such a model of modified economic flow, initially justified security investment would be the one in which the total, cumulative benefits (or the avoided cost of security problems increased by the residual value) is larger than the cost of security solution implementation [17]. Such situation is shown in Table 1.

Table 1: The analysis of modified economic flow of information security investments

	structure/period	1	2	3	...
1.	AVOIDED EXPENSES	...	...	...	...
2.	RESIDUAL VALUE	-	-	-	...
3.	EXPENSES	2.1+2.2+2.3+2.4+2.5+2.6+2.7	...	...	...
3.1	Security solution investment	...	...	...	...
3.2	Maintenance costs	...	...	...	...
3.3	Material costs	...	...	...	...
3.4	External costs	...	...	...	...
3.5	Training costs for solution implementation	...	...	...	...
3.6	Training costs for solution usage	...			...
3.7	Gross salary	...	...	...	...
4.	NET EFFECT	1.+2.-3.			

Source: authors

IV. CHARACTERISTICS OF INFORMATION SECURITY INVESTMENTS' FINANCIAL ANALYSIS

Similar to the economic flow method, unmodified financial flow method [18] could also be applied to information security solution investments only if the enterprise holds such solutions as assets and leases them to other enterprises. Unlike the economic flow method, financial flow method also takes into account sources of financing [19] represented by cash inflow and obligations towards sources of financing (outgoing interest), shown in table 2.

Table 2: Modified financial flow method when investing in information security

	structure/period	1	2	3	...
1.	AVOIDED EXPENSES	...	...	...	...
2.	RESIDUAL VALUE	-	-	-	...
3.	FINANCING SOURCES	3.1+3.2	...	...	...
3.1	Own sources	...	...	...	...
3.2	Loans	...	...	...	...
4.	Expenses	2.1+2.2+2.3+2.4+2.5+2.6+2.7	...	...	...
4.1	Security solution investment	...	...	...	...
4.2	Maintenance costs	...	...	...	...
4.3	Material costs	...	...	...	...
4.4	External costs	...	...	...	...
4.5	Training costs for solution implementation	...	...	...	...
4.6	Training costs for solution usage	...			...
4.7	Gross salary	...	...	...	...
4.8	Installment (annuity)	...	...	...	...
5.	NET EFFECT	1.+2.+3.-4.			

Source: authors

The problem which appears when using the financial flow method is similar to the economic flow method – determining avoided costs. In the dynamic analysis, classical dynamic methods [20] could be used, as follows:

1. **Return on investment rate** – representing the simplest criterion of financial decisions regarding real investments. The method calculates the number of years necessary for the project (in this context, information security asset or system) to return the invested funds. In case that methodology calls for elimination of lack of incorporation of time value of money, discounted cash flow method will be used [21].
2. **Discounted period of return method** – the variant of the above method in which time value of money is incorporated in calculation.
3. **Net present value method** of investment in information security, represents the sum of present values of resultant cash flows related to the same information security solution.
4. **Internal rate of information security investment return** – representing the internal rate of return (RoR) reducing the pure cash flows of using the information security solution to the value of their investment flows [22].
5. **Profitability index** could be used as an added criterion of investment decisions. It represents the ratio of discounted pure cash flows of information security solution during its lifetime and the investment costs.

V. CHALLENGES IN APPLICATION OF INTERNAL RATE OF RETURN (RoR) METHOD

The classic analysis of the internal rate of return demands that the rate of return (leveling the investment with the pure cash flows) should be larger than the reference discount rate which is dependent on the risk level and the capital cost. [23]. Common problem with the multiple internal rates of return is usually not present in information security investments. According to the premise that the information security risks are present throughout the period of usage of the information security solution, and that the distribution of cash flows is identical (there is no change in cash flows, and last cash flow is not negative), the classic method of internal rate of return could be used.

However, when using the internal rate of return method in the analysis of information security investments, several facts should be taken into consideration [24]:

- This method could not be used when analyzing or comparing investments into multiple information security solutions, only when analyzing single investments due to the fact that the obtained results are not comparable
- The internal rate of return implies reinvesting the positive cash flow in projects or solutions that have the similar rate of return, whether it is the case of reinvesting in similar solutions or other comparable solutions. For that reason, the internal rate of return method will be used in evaluating those projects in which the reinvested cash flow is directed into projects with lesser rate of return [25]. This is especially true for security solutions or projects with high rates of return, because enterprises have difficulties in finding comparable reinvestment projects with equally attractive rates of return.

- As a rule, cash flows do not change from positive to negative and vice versa, and the last cash flow is never negative. Therefore, the problem of multiple internal rates of return should not exist.
- The internal rate of return method will only provide a relative calculation of return for a given security project or solution, not absolute.

The starting assumptions for the set model are as follows:

1. The initial investment in information security solution has been executed,
2. No opportunity costs exist,
3. Minimal residual value of the information security solution exists at the end of the solution lifetime,
4. The proportional depreciation method is being used,
5. Generated cash flow is estimated as the consequence of avoiding the occurrence of security incidents and connected direct and costs,
6. There is no tax credit,
7. Other investments related to introduction, implementation and usage of information security solution are treated as having fixed value and they are not depreciated during life time of exploitation,
8. It is estimated that variable cost exists during the lifetime of the security solution, expressed in relative terms compared to income,
9. The income tax is set as fixed during the observed period,
10. It is assumed that the working capital during the lifetime of the solution is annulled or fully depreciated (amounts to zero),
11. The discount rate is used as a direct discount rate. It is possible to calculate the discount rate within the model as it is calculated in the Capital Asset Pricing Model. [26] When doing so, risk free return rate, premium based on assumed market risk, cost of borrowing the capital and debt leverage have to be taken into account,
12. It is assumed within the model that the benefits from implementing the information security solution do not change through the years, in other words the benefits are proportionally distributed,
13. Fixed costs of solution maintenance during its lifetime do not increase; they are contained in the initial investment.

In this paper we will outline a possible practical application of the cash flow analysis in information security solution.

Input parameters of the proposed model consist of initial investment, represented by the value of investment in ICT security solution, opportunity cost of the capital, life span of the solution, scrap value, depreciation method, tax credit and non depreciated capital investments. Furthermore, required working capital is taken into consideration, so are generated cash flow, selection of the discount rate and alternative risk-free market going rate. Also, a possibility of loan and market capital borrowing rate is also anticipated - for simplicity reasons in the outlined model, investment is 100 % owned and external leverage is not used.

Finally, growth rates for implicit revenues derived from information security solution and fixed cost are set to zero, meaning they are omitted from the model. These parameters present input values of the model of cash flow analysis of information security solution investment project. The constructed model shows that by using the initial assumptions of the model it is possible to calculate the net present value, internal rate of return and the return on capital when using such a solution. Further simulations show that the internal rate of return is particularly sensible to the following input parameters [27]:

1. Time (duration) of using the information security solution,
2. Perceived (set in advance) ability of the information security solution to generate the positive cash flow during its lifetime,
3. Discount rate used.

Such model could be used in practice if the input parameters are correctly estimated. Judging from experience, the largest problem in forecasting is connected to the ability to assess the generation of the positive cash flow by avoiding costs related to the occurrence of security incidents.

VI. CONCLUSION

Enterprises and organizations which operate in the modern environment (especially in the service sector) are faced with the challenges of the modern dematerialized economy. Because of that, one of the new business functions arose in the last few decades – the business function of information security, augmented by the increased usage of networking and communication technologies. The roots of information security are in the defense sector and high technology sector, and it was firstly implemented by those enterprises which possess the advantage over their competitors because of new ideas, inventions and patents. Due to the increased number of threats, and the existence of information assets which could be compromised by those threats, even the smallest companies nowadays have to pay due attention to information security to avoid the occurrence of security incidents which could compromise the competitiveness and even the existence of such companies. Therefore, securing of the normal functioning of the enterprise by means of information security is no longer reserved for the corporate sector, it is equally important for small companies, all the way down to the level of individual entrepreneurs.

The demands for ensuring the satisfactory level of information security are also set by the legal system, trying to regulate and to positively influence the companies which have a strategic interest to the state, primarily the financial sector. Such demands are the legal requirements which those companies have to meet. As a rule, such

demands are not imposed on other companies, they have to assess themselves (according to their business plans, financial and human resources and risk assessment) the profitability of information security investments. When doing so, the benefits of information security system implementation are primarily compared to implementation and maintenance costs. According to this evaluation, investment decisions should be made. The estimation of total cost of implementation and maintenance of the information security solution is further hampered because of the different ways of achieving such goals: it can be achieved both by capital investments (real investments) in information security solutions and by presenting information security solutions as "running costs" or maintenance costs. Suppliers of information security solutions are trying to adapt the solution functionalities to fit various customer needs and their financial abilities at the given moment. In any case, information security solutions carry with them substantial maintenance costs.

The main problem when deciding about investing in information security solutions is how to decide between the two opposite investment viewpoints. The first, techno centric viewpoint, is often advocated by information security experts employed in an enterprise. Their primary focus is on risk assessment, so when trying to minimize security risks they sometimes suggest uneconomic information security investments. On the other hand, enterprise management as a rule lacks understanding of the technical aspects of information security and would prefer that such investments are avoided, because the necessary input for argumented decision making is often not presented to them, while they are concerned primarily about benefits of information security solutions for the enterprise as a whole. A key factor when investing in information security solutions is to find the balance between the techno centric and financial perspectives. In reality, the scales are usually tipped to one side, and the misbalance will either cause reduced business performance of an enterprise or implicit exposing to high levels of risk related to information security.
as shown in this paper, when deciding about information security investments, classic methods of economic and financial (cash) flow can be used. A problem of determining input parameters arises, and similar to usual models (investing into manufacturing equipment, project investments), the largest problem is determining the "income part", which can be identified as "avoided damages" in the case of security incidents which could happen if the information security solution is not implemented. Such a model has been proposed and empirically documented using proprietary data. Related data is often not publicly available, and even the specialized consulting firms do not have them, because such data is extremely dependant on the type of business and type of enterprise. It might be only partially available for financial institutions, because of legal regulations regarding financial information security. The method which could be used is a modified Internal Rate of Return, adapted to suit the information security paradigm, with the input parameters modified to specific demands of investing in information security.

REFERENCES

[1] Hirshleifer, J., Riley, John G., "The Analytics of Uncertainty and Information", Cambridge Surveys of Economic Literature, Cambridge, UK, 1992.

[2] Shim, K. Jae; Siegel, Joel G.: "Handbook of Financial Analysis, Forecasting and Modeling", CCH, 2007.

[3] Ostrom, T. Lee; Wilhelmsen C.: "Risk Assessment: Tools, Techniques, and Their Applications", Wiley, 2012.

[4] Hiles, A: "Enterprise Risk Assessment and Business Impact Analysis: Best Practices", Rothstein Associates Inc., 2002.

[5] Aksentijević, S.: "Integral protection and information security management system – Saipem Mediteran Usluge d.o.o, Rijeka", unpublished, 2008.

[6] Maddock, V: "IT Induction and Information Security Awareness: A Pocket Guide",It Governance Ltd., 2010.

[7] Phillips, Pulliam P., Phillips, J. J.: "Return on Investment (ROI) Basics (ASTD Training Basics)", ASTD Press, 2006.

[8] Brcar, F., Bukovec, B., "Analysis of Increased Information Technology Outsourcing Factors", Organizacija, Volume 46, Number 1, January-February 2013.

[9] Dempsey, P.: "Hidden Costs: The iceberg that could sink your company: Kaizen and Copq, Relentless Opportunity", Xlibris, Corp., 2012.

[10] Palmer, R.: "Maintenance Planning and Scheduling Handbook", Xlibris, Corp., 2012.

[11] Larman, C., Vodde, B.: "Scaling Lean & Agile Development: Thinking and Organizational Tools for Large-Scale Scrum", Addison-Wesley Professional, 2008.

[12] Pinto, J.E. et. al.: "Equity Asset Valuation", Wiley, 2010.

[13] George, J.A., Rodger, J.A.: "Smart Data: Enterprise Performance Optimization Strategy", Wiley, 2010.

[14]Conyngton, H.R.: "Financing an enterprise", Ulan Press, 2012.

[15] Mian, M.A.: "Project Economics and Decision Analysis, Volume 1: Deterministic models", PennWell Corp., 2011.

[16] Yescombe, E.R.: "Principles of Project Finance", Academic Press, 2002.

[17] Malik, Krishan A., „Petroleum Project Evaluation & Investment Decision Making", Institute for Petroleum Development, Austin, Texas, 2011.

[18] Mulford, C.W.: "Creative Cash Flow Reporting: Uncovering Sustainable Financial Performance", Wiley, 2005.

[19] Marks, K.H.: "The Handbook of Financing Growth: Strategies, Capital Structure, and M&A Transactions", Wiley, 2009.

[20] Zečević, Z.: "Methods for financial effectiveness evaluation of investment projects", http://www.scribd.com/doc/75998229/Metode-Za-Ocjenu-Financijske-Efikasnosti (25.11.2012.)

[21] Voss, J.A.: "Valuation Techniques: Discounted Cash Flow, Earnings Quality, Measures of Value Added, and Real Options", Wiley, 2012.

[22] Bojanc, R., Jerman-Blazic B.,"Quantitative Model for Economic Analyses of Information Secutiy Investment in an Enterprise Information System", Organizacija, Volume 45, Number 6, November-December 2012.

[23] Pratt, S.P., Grabowski, R.J.: "Cost of Capital: Applications and Examples", Wiley, 2010.

[24] Tipton, H.F., Krause, M.: "Information Security Management Handbook", 6th Edition, CRC Press, 2007.

[25] Elton, Edwin J.; Gruber, Martin J.,, Brown, Stephen J.,. Goetzmann, William N, „Modern Portfolio Theory and Investment Analysis“, VIII ed., Wiley, Bognor Regis, West Sussex, 2009.

[26] Tapiero, C.S.: "Risk Finance and Asset Pricing: Value, Measurements, and Markets", Wiley, 2010.

[27] Phillips, Pulliam P., Phillips, J. J.: "Return on Investment (ROI) Basics (ASTD Training Basics)", ASTD Press, 2006.

Seaport cluster information systems - a foundation for Port Community Systems' architecture

Edvard Tijan[1], Saša Aksentijević [2], Dragan Čišić[3]
[1,3] University of Rijeka, Faculty of Maritime Studies
Studentska 2, 51000 Rijeka, Croatia
Tel: +385 51 33 84 11 Fax: +385 51 33 67 55 E-mail: etijan@pfri.hr, dragan@pfri.hr
[2] Aksentijević Forensics and Consulting, Ltd.
Gornji Sroki 125a, Viškovo, Croatia
Tel: +385 51 65 17 00 Fax: +385 51 65 17 81 E-mail: axy@vip.hr

Abstract - Numerous stakeholders are involved in operations of modern Port Community Systems: container terminals, customs, border police, coast guard, government ministry in charge of sea traffic and infrastructure, seaport agents, port authorities etc. Traditionally, all of these entities use separate and heterogeneous business information systems. Integration of these complex systems presents a challenge: the common goal should be the creation of one interconnected system, able to provide all vital information to various stakeholders while being efficient, financially viable and avoiding data and process duplication. In this paper, the authors will present main characteristics of existing seaport cluster information systems, their possible overlapping dependencies and initial component analysis, as a basis for decision-making related to architecture of future Port Community Systems.

I. INTRODUCTION

This paper is an introductory research into characteristics of seaport clusters, based on technologies and organization used in Republic of Croatia at the time of research (2013.). Various stakeholders (container terminals, customs control, police, port authorities, port agents etc.) utilize different ICT systems that should be thoroughly analyzed before integration into integral business information systems of the port clusters (PCS – Port Community Systems). In order to fully understand the complexity of the different systems, it is important to outline and study them [1]:

- Concession owners use heterogeneous ICT systems and their inclusion in one single system can incur elevated levels of risk imposed on information and integral security.
- Port authority and concession owners have to be organized in a way to be aligned with ISPS code (International Ship and Port Facility Security Code). ISPS code development is caused by increased number of hijackings and attacks on vessels and ports.
- Handling of dangerous cargo, its packaging, labeling, storage, separation and emergency procedures have to be in charge with IMDG code (International Maritime Dangerous Goods Code).
- Port authority is performing tasks related to safety and oversight in those segments that are related to uninterrupted traffic flow in areas of their responsibility and it is involved in search and rescue operations and inspections.
- Traffic tracking in port area is accomplished by using modern VTMS system that utilizes GPS equipment.
- Service user oversight is performed by usage of ID cards and integrated business information systems.

Integration of ICT systems in complex environments like seaports is a challenge for all subjects involved in port operations. It involves not only the creation of physical and logical networks and connections, lifeware and orgware, but also control from a single point over all delivered services, while respecting all basic principles of disaster recovery and business continuity.

II. THE PURPOSE OF SEAPORT INFORMATION SYSTEMS

Port information systems serve the following purposes [2]:

- To increase the speed of cargo loading and unloading operations.
- To adjust the duration of physical and administrative operations in order to provide services to users and to speed up the traffic and enable optimal usage of the port's infrastructure.
- To provide information systems support to stakeholders, equally for basic applications and extranet portal used by external stakeholders.
- To create relations toward external databases.
- To manage the information flow by enabling connections between stakeholders and using different information sources while ensuring information security and confidentiality.
- To manage the infrastructure by optimizing critical resources.
- To optimize cash flow by simplifying the accompanying documents.

Preliminary research has shown that the achieved level of development of port community systems in Croatia is modest in terms of their use in management of port facilities. Identified reasons for this condition are:

- The general issue of inadequate rate of ICT technology implementation in comparison with the speed of its overall development.
- Inadequate level of knowledge and adjustment ability of the entrepreneurs and the management caused by rapid scientific and technological development.

- Lack of knowledge and understanding of ICT technologies as major economic resources.
- Inadequate level of employee education and low levels of ICT investments.

However, most stakeholders inside the seaport cluster own and use information systems as part of their daily and routinely executed operations. These systems are diverse, mainly because of size of the company, operations and characteristics of the branch the company is operating in.

III. CONTAINER TERMINAL ICT SYSTEMS

Information flow and exchange in modern container terminals is complex. Cargo agents are exchanging all necessary messages with the container terminal operators, who are in turn exchanging messages with the authorities (customs, port authority). Information is accessible also to cargo haulers, trucking and railway cargo companies and other terminals the cargo will reach in the future. Information systems of container terminals are in fact systems used to operate the transport of goods stored in containers that involves computers, clients, servers and networks, with intention to collect the data about transport process, information processing, decision making, creation of messages and documents, order and document transmission and actions. In order to facilitate flawless transport process, it is necessary to understand crane and container timing and bottleneck analysis. This role is nowadays entrusted to management information systems that enhance automatization of the container terminals. Enhancement of the cargo transport inside terminals can be achieved in the following ways [3]:

- Introduction of the communication network inside terminals with a complete centralized database.
- Correct evaluation of duration of cargo transport and loading-unloading operations.
- Development of the automatic operations management (driverless vehicles, cranes without operators).
- Inclusion of the system buffers.

The most important function of ICT in the container terminal operations is creating the possibility to connect container terminals with other port community stakeholders, therefore creating an electronic community of the whole seaport system (Port Community System). Container management system that is well envisaged and possesses the proper architecture can increase terminal efficiency, with the main benefits being [4]:

- Faster container unloading and loading with increase of productivity,
- Better container storage oversight,
- Higher percentage of correct data and valid information.

Container terminals typically use TOS (Terminal Operating Systems), computerized systems used to enhance the process of cargo management. TOS basis is several decades old, but it is still in use today, even though heavily modified and modernized. Usage of the TOS system enables the terminal management to follow up individual container status (size, weight, type, content, special instructions), resources (free operative area, equipment location), limits (characteristics of the operative area and required equipment) and processes (optimal container placement and loading priorities).

ICT systems typically used in container terminals are the following [5]:

- Gate System, used to control and identify containers, perform X-ray checks and detect trucks,
- Corporate System, used for business functions, to analyze human resource deployment, create accounting and financial reports for the management,
- Engineering System, used for implementation of technological innovations on transport utilities and malfunction diagnostics,
- Anciliary System, used for management of empty areas and container repair,
- OCR Handling System, the system used for container tracking based on optical recognition and tag reding,
- Equipment Control System, follows up the equipment, current position of the cranes, controls RFID readers,
- Equipment PLC/SCADA System, used to track and control cranes and AGV (automatic guided vehicles) [6],
- Information Technologies Analysis and Design System, top-level system that manages all other elements of the hardware and software, and
- Community System that connects all stakeholders in the seaport cluster by exchanging information and electronic messages.

As a conclusion, advantages of introduction of ICT systems and electronic data exchange systems in container terminals are timely information collecting, data organization, ability to accept large amount of information, possibility of implementation even in smaller terminals, simple data disclosure, less document errors and decrease of volume of paper documents and required staff support.

IV. CUSTOMS CONTROL ICT SYSTEMS

Customs control documents were traditionally filled manually and stamped by persons in charge, resulting in slow data processing and occurrence of errors. In 1995. the customs has started importing data using magnet media (floppy disks). Since year 2000, the goods that are a part of customs procedure can be processed using electronic data exchange. The first step to implement this type of data and form management was creation of the Web application for customs declarations, that has enabled participation of importers, exporters and haulers. Every such declaration can be submitted even outside of customs working hours, and is formally and logically controlled before submission.

Further step in development of the customs informations system was completed as a part of PHARE 2005. project that has strengthened abilities of Croatian customs to achieve alignment with EU standards. Three such systems were subjects of development that ensured full compatibility and interactivity with EU information systems [7]:

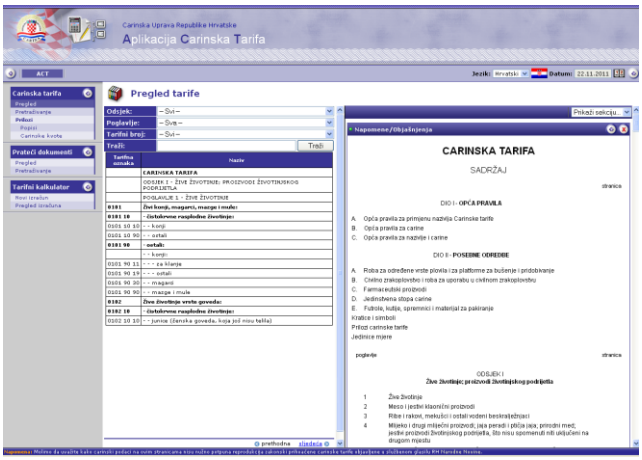
1. TARIC (acronym for French „TARif Intégré de la Communauté“) is a database of all customs fees related to imports, exports and applicable tariffs dependent on the tariff code of the goods
2. NCTS (eng. New Computerized Transit System), necessary for insurance of transit goods, and
3. EMCS (eng. Excise Movement Control System), required for efficient control of goods subjected to excises (alcohol, cigarettes, fuels and electric energy).

Development of these systems was done in six stages and three twinning projects. Execution of Phare 2005. project started in September 2007. and was completed by July 2009.

Phare 2006. project had a goal to ensure full compatibility and functioning of the information systems with customs systems of the EU in the areas of customs tariffs, customs transit control and export of goods subjected to excises.

One of the main results of the above mentioned projects was the creation of “Customs tariff” application, which is shown in Fig 1.

Fig. 1: Customs tariff application in Republic of Croatia



The customs has also created a procedure named Rules of usage for electronic data exchange system and service delivery of the e-business of Ministry of finance, customs (e-Customs), that governs the basic principles for data exchange and delivery of services of Ministry of finance, customs and all users of the electronic date exchange system, based on laws and regulations related to customs, excise, taxation and official statistics. Since 1st November 2011., NCTS, the first electronic system of business operations on the national level has started operating [8]. It enables uninterrupted transit of goods under customs supervision, and it is iat the same time the first interoperable trans-European customs system. Development and implementation of a fully functional NCTS system was one of the main conditions EU Commission has assigned to all EU member states, so Republic of Croatia has opted for its own NTA (National Transit Application) that supports all NCTS functionalities. With implementation of the NCTS system, entered data is the only legally accepted evidence, in comparison to old system where paper declaration was treated as a legal warrant.

V. BORDER (MARINE) POLICE ICT SYSTEMS

ICT systems used by the border (marine) police are specific by nature. They have to adhere to the highest information security standards, and use a number of legal, organizational, logical and technical measures and standards in order to protect information confidentiality, integrity and availability. They are executed through five recognized areas of information security: background checks, physical security, data security, information system security and business cooperation security). Implementation of these measures ensures alignment with NATO and EU security politics [9].

NBMIS (eng. National Border Management Information System) is a new tool for efficient border control that at the same time ensures flawless traffic of people and goods. It was built by using EU CARDS program (project 2002. NBMIS). The project that is enhancing this system has started in 2010 and was completed at the beginning of 2012. The goal of the project was to upgrade and expand functionalities of the NBMIS application and develop a mobile platform. New equipment was installed in 25 international border crossings (11 road crossings, 5 sea crossings, 4 airport crossings, 4 railway crossings and 1 river crossing). The system operates by using the principle of reading bar code on travel documents and enables police officers to check the passengers in a quick and efficient manner.

ANPR (eng. Automatic Number Plate Recognition) is an additional subsystem that during the border crossing checks whether the vehicle was stolen and whether it is properly registered. The system takes a photo of the vehicle and checks all data related to the vehicle. If some irregularities are identified, the system will alert the police officer.

Another important system that will have to be implemented is SIS (eng. Schengen Information System). SIS is a database of European countries that stores information about people and goods. The purpose of this system is to safeguard national security, control borders and support the rule of law. Second generation of the SIS (SIS II) will contain warnings about people and objects and it will be used by the border police, customs control and state bodies issuing visas [10]. SIS II consists of the central system (Central SIS II), national systems (National SIS II) in all member states that will exchange data with Central SIS II and communication infrastructure between central and national system that will ensure encrypted data exchange. Every member state has to provide a national S.I.Re.N.E. office (Supplementary Information Requested at the National Entries), according to preset conditions. The role of these offices is consulting, notification exchange and implementation of checks and safeguards in order to avoid multiplication of warrants and priority setting.

Also, a manual was issued in form of set of guidelines, general and special procedures that have to be followed by authorized instances in order to exchange information about entry or residence denials, arrests and extraditions, missing persons, legal procedures, discrete and specific checks and objects that have to be temporarily confiscated because they might be used as evidence [11].

VI. MINISTRY OF MARITIME AFFAIRS, TRANSPORT AND INFRASTRUCTURE ICT SYSTEMS

Ministry of maritime affairs, transport and infrastructure is managing several information systems, whose purpose if the oversight and management of the maritime traffic, its safety, collection and distribution of the meteorological information. One such system is the VTS (Vessel Traffic Service), established in order to enhance traffic safety and efficiency and protect sea environment. This system has to provide information to all who participate in sea traffic in the area of VTS surveillance. This role can be divided in three different categories:

- provisioning of the information to ships in the area of VTS (information services)
- provisioning of the assistance in navigation through VTS area (navigational assistance services), and
- organization of the sea traffic in the area of VTS supervision (traffic organization services)

VTS consists of a set of Croatian and international regulations that ensure safe traffic on rivers, lakes and seas, practically executed by VTMIS (Vessel Traffic Management and Information Service). This service has been put in the spotlight of the system by a law passed in 2008. Also, establishment of the VTMIS service was the obligation of the Republic of Croatia, as a part of accession process to the EU.

VTMIS is a complex technical and information system whose purpose is to track, manage and organize overall sea traffic. It consists of Coastal Automated Identification of Ships System, VTS Radar System, maritime radio-communication system and other systems that enable insight in sea conditions and interaction with all involved stakeholders.

Project documentation shows that VTMIS has to contain one or both of the following elements:

- electronic data exchange and service provisioning towards the immediate environment, region or remote subjects (horizontal data exchange),
- electronic data exchange with other sea traffic subjects – officials (vertical data exchange).

VTMIS system is financed within PHARE 2006 program of the EU in amount of 75 % of the total project cost. Republic of Croatia has started VTMIS activities since 2004. In 2005. and 2006. the first phase was successfully completed when AIS system was put in service. This system is used to automatically identify ships. Full coverage was achieved in 2006. by creation of five AIS base stations. In second phase, additional 13 base stations were made functional and the system has achieved total area coverage. Data collected by VTS operators is forwarded to the police and customs, but only the Ministry is operating the system.

AIS (Automatic Identification System) is based on SOLAS convention [12]. It operates on maritime VHF band and transmits static and dynamic route data. Static data is entered manually, and it includes MMSI ship identification number, IMO number, name and call sign of the ship, its length and width, type, and antenna location. Dynamic data is entered automatically, and it requires integration with ship's navigation equipment installed on the bridge. This data includes ship's position, course, speed, several sets of technical data and navigation status. The purpose of AIS system is ship identification, tracking assistance, simplified data exchange and collision avoidance. AIS system is primarily aimed towards enhancement of the VTS service.

VII. PORT AGENT ICT SYSTEMS

Port agents are shipping company's representatives and they perform all formalities related to ship being in the port. Port (shipping) agent is in fact a shipping company's representative in seaports where the company does not have its branch or office. Agent conducts all business in the port on behalf of the shipping company. Port agents use information systems which are usually based on server/client technology. ICT equipment of the system can be integrated in the port information system, so data and information are distributed using standardized documents.

Agent information system consists of several modules or databases:

- Docking database, contains data about ship's docking, from its announcement, to accounting and archiving,
- Cargo database, includes separate tracking of announced and really performed offloading cargo operations, including related data: operations, cargo type, units, quantity, code and description
- Database of advances, contains data about requested and paid advances with related data: date, obligated party, requested amount, paid amount, type of payment, bank, document number, date of payment, notification
- Database of ships, contains data related to the ship, its code, flag, owner, port of origin and the following data: Call Sign, Gross Tonage, Net Tonage, Deadweight, Length Over All, Summer Draft, Max. Draft, Air Draft, Vessel Type, Built, Capacity: Grain, Bale, TEU, IMO No., notification.
- Code database, contains data about the flag according to ISO standard, port according to UN's port code list, information about cargo, currency, type of docking, business unit and packing type
- Event database, contains all data related to important events on the ship and in its surroundings that are neccessary to create document called Statement of Facts: date and hour of the beginning of the event, code and description, responsibles, warehouses and notifications
- Database of costs, follows up all costs and accounting of the ship and contains also agent costs that will be transferred to the Disbursemenet Account.
- Business partner database, a database of all business partners that work together with the agent.

Agent business information system usually has a graphical user interface (GUI) for data entry and applications used to print various documents that agent process produces, like cargo and freight manifest, invoicing, credit notes, statistic reports and other.

VIII. PORT AUTHORITY ICT SYSTEMS

Port authorities in Republic of Croatia use several ICT systems, where some of them (like VTS and VTMIS) are very similar to those used by the relevant Ministry. These systems are usually controlled from Port control centers, that become centers in charge of traffic oversight improvement and preventive environment protection. Except VTS and VTMIS, systems used to control sea traffic and ship communication can contain electronic charts (ECDIS) over which data from radar systems and automatic ship identification systems is integrated [13]. These systems can be further upgraded by using communication radio VHF system, meteorological data acquired from stationary sensors and meteorological buoys, video surveillance of the sea and port area using CCTV. All these systems are based on TCP/IP protocol using different communication links (telephone lines, ISDN, optical cables, broadband, microwave and satellite links).

The main difference between port VTS and VTMIS systems and those managed by the state systems is that they oversee exclusively the seaport area, while state managed systems are concentrated on sea ways and state see territory.

Efficient functioning of the seaport system is achieved by development of information management tools that allow seaport system management adequate support when executing management functions. These systems are called PMIS – Port Management Information Systems. Such systems are in fact composed of several information systems: data processing and office applications, information systems of external business units and information systems related to the environment of the ports. Management information systems enable communication between different functions within port clusters, by using the common database. Management reports is a separate functionality inside PMIS used by the management to extract information about business events and transactions and to enable control of the past transactions and plan the future ones. Using different data sources, it is possible to analyze the balance sheets, cash flow, income and costs, key performance indicators, trends, analysis of the reports through several time periods, structure of the long term and short term assets and debts etc.

IX. INTEGRAL BUSINESS INFORMATION SYSTEMS IN SEAPORT CLUSTERS

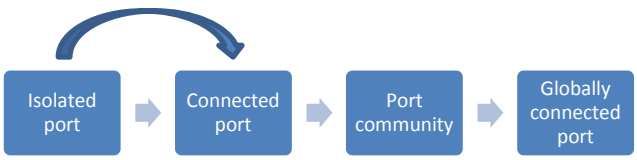
Integral business information systems in seaport clusters (Port Community Systems – PCS) are focusing all information, data and messages that are part of business information systems of various stakeholders in the port cluster. While individual business information systems that relate to individual stakeholders process and store only data and messages that are relevant for them, PCS can exchange data that is useful for a wider number of users. PCS can be defined as a system delivering information to supply chain members operating in the port. PCS are used to deliver, control, distribute and safeguard data. Its electronic platform connects various information systems used by different organizations that together make port community. PCS are holistic, geographically bound information centers in global supply hubs that serve interests of the heterogeneous collective that consists of various business entities. Main goal of PCS is to gather all stakeholders at a virtual table and establish virtual coordination by providing exact information about arrival and departure operations and cargo loading/unloading.

It is possible to identify four different phases of introduction of ICT and business information systems in seaport communities:

1. Isolated port. Stakeholders operating within seaport clusters are isolated, ICT systems are used only for internal processes of individual stakeholders.
2. Connected port. Introduction of ICT and business information systems spreads outside boundaries of internal cohesion and opens up to wider community and users of port services. This is achieved by usage of common Electronic Data Interexchange (EDI) system used to exchange documents and information.
3. Port community. ICT and business information systems are connected with land transport and shipping companies. New technologies like RFID (Radio Frequency Identification) are being used. Subjects of the seaport community are connected by Web applications.
4. Globally connected port. This is the highest level of integration and implementation of ICT technologies in the seaport community. On this level, all paper documents are replaced by electronic documents, logistic management is modernized, full integration with external subjects is achieved and all cooperation between ports is accomplished by using ICT technologies. This phase of implementation usually uses RFID and optical recognition (OCR) technologies, service oriented architecture and solutions based on XML (Extensible Markup Language).

Four different phases or development and integration of ICT technologies in seaport communities and achieved level in Croatian ports are shown in Fig. 2. [14].

Fig. 2: Phases of port community integration



Croatian seaports are currently between first and second phases of development and integration of ICT technologies in seaport community. There are two main reasons for this state of affairs: partially, the situation is caused by lowered investments in sea ports during 1980s, when ports in neighboring countries were heavily investing in port infrastructure and technology development. In 1990s, because of war events, large portion of cargo was lost to

other ports. After the war, ports have started investing in renewal of their infrastructure by using World Bank loans.

Implementation of PCS in Croatian ports is a complex and demanding task that requires active cooperation of all stakeholders in seaport clusters. The process of PCS implementation has to be based on adequate standards that will ensure efficiency, simple data and document exchange and information security. This process also includes implementation of relevant Data Classification and Information Lifecycle Management processes and measures for Disaster Recovery (DR) and Business Continuity Management (BCM).

X. CONCLUSION

Seaport clusters bring together various stakeholders - container terminals, customs, police, port authorities, shipping companies and agents - each of them having a complex set of different information systems serving their own purposes. All this makes the interaction between those systems very complex and elevates overall levels of the integral risk. Different stakeholders in Croatian seaport clusters had to implement various business information systems in order to be compliant to requirements imposed by the European laws and legislations. Majority of those were funded by applicable EU programs and World Bank loans. Due to lowered investments and recovery of traffic that was lost in the past two decades, Croatian ports are now in transition between isolated and connected ports, as the first two stages in the creation of integrated port community and eventually, a globally connected seaport system. Further efforts should be undertaken to align Croatian seaport business information systems with modern requirements and therefore increase efficiency, productivity, decrease costs and facilitate the underlying logistics chain.

REFERENCES

[1] Hlača, B., Aksentijević, S., Tijan, E.: Influence of ISO 27001:2005 on the Port of Rijeka security, Journal of Maritime Studies 2 (2008), p. 248.

[2] Dvorski S., Novak S.: Uloga informacijsko–komunikacijskih tehnologija u špediterskoj logistici, Suvremeni promet, Zagreb, 2010. br. 1-2

[3] Dundović, Č.: Proces usavršavanja komunikacijskih sustava kontejnerskih terminala, Suvremeni promet, 16 (1996), 5, p. 377.

[4] Jolić, N: Luke i ITS, Fakultet prometnih znanosti, Zagreb, 2008, p. 104.

[5] Martin, J.: International port training conference: developing skills for terminal automation, www.iptconline.net/scroll/documents/JeffMartin.pdf (25.12.2013.)

[6] Port Technology International: Modern Interface between TOS and Crane Control System, Edition 41, 2009, p. 54.

[7] http://www.carina.hr/EU/OProjekti.aspx (25.12.2013.)

[8] NCTS 4.0 upute za gospodarstvenike, v2.3, p.5. http://www.carina.hr/Dokumenti/Download.aspx?args=4EjqCjp+5QXseGAwl6Z8lPdHzWr816cDlxmjyrKg4Rs= (25.12.2013.)

[9] http://www.mup.hr/main.aspx?id=108775 (25.12.2013.)

[10] Regulation (EC) No 1987/2006 of the European Parliament and of the Council of 20 December 2006 on the establishment, operation and use of the second-generation Schengen Information System (SIS II) (25.12.2013.)

[11] Perić Hadžić, A.: Elektroničko poslovanje u funkciji racionalizacije upravljanja sustavima morskih luka, Master's final thesis, Ekonomski fakultet u Rijeci, Rijeka, 2008., p. 68.

[12] Safety Of Lives At Sea convention, Chapter V/19-1, Rule 19

[13] Kasum, J., Bićanić, Z., Karamarko, A.: Predvidivi razvoj tehnologije izrade pomorskih karata i publikacija, Naše more 52(1-2)/2005, p.55.

[14] De la Guia and Llop, Valencia PCS Development Strategy & Practice, Autoridad Portuaria de Valencia, Gijon, p.6.

Development of e-clusters

Milan Davidović
VSITE – College for information technologies, Zagreb, Croatia
milan.davidovic@vsite.hr

Abstract – In this paper key presumptions and prerequisites of e-cluster organization and sustain development in TIMES technology sectors (TIMES - Telecommunication, Information technology, Multimedia, Entertainment, Security) are analyzed and how can integrate small and medium enterprises (SMEs) in the e-cluster. E-clustering is macroeconomic policy and strategy with crucial impact on horizontal and vertical integration of SMEs in TIMES sectors and regional development. Possibilities of vertical e-clusters and building new value added e-chains as opposed horizontal e-clusters and their entering in large supply chain are compared. Innovation and knowledge are presented as drivers of horizontal and vertical e-cluster development.

Key words: e-cluster, e-clustering, knowledge management, innovation management, intellectual capital

I. INTRODUCTION

Organizational and technological grow are key of business success and sustainability of clusters, especially in TIMES sectors. Building and further development of e-cluster is a project of continuous organizational and technological growth of TIMES clusters.

E-clusters are information and communication technology (ICT) based organization structures derived from TIMES clusters (see Chapter II) with respect to macroeconomic policy and university-cluster alignment (see Chapter III). Organizational development of e-cluster is cyclical process that consisting four stages (foundation, growing, stagnation and declining) and growth, recovering and crisis cycles. Innovations and knowledge are drivers sustain growth of e-clusters and their members (see Chapter IV). Cluster member orientation and relationships in cluster determine horizontal and vertical integration TIMES companies and their partners, and organization of cooperation process and e-supply chains (see Chapter V).

ICT support fast and effective linking TIMES companies and their partners in e-clusters and e-supply chains, but only cluster member's positive expectations and absolute trust make interrelationships in e-cluster and e-supply chains sustain and synergetic. E-cluster is more than e-logistics of TIMES cluster; this is an example how create and work in world of digital economy.

II. FROM TIMES CLUSTER TO E-CLUSTER

A. *TIMES cluster*

TIMES clusters integrate independent organizations that strive for economic sustainability, so initiate and push the networking of TIMES companies, universities, research institutions, customers, and other stakeholders in a region [1][2][9]. Typical representatives of TIMES clusters are IT and communication clusters, film, TV production and other clusters in digital content industry. Their main togetherness is

- composed digital content and products,
- digitalized processes of creating (production), storage, transmission and using of digital content and products,
- knowledge, creativeness, innovativeness and
- Intellectual property (IP) rights.

Process of cooperation in TIMES cluster is a value added chain which involves creative activities, protecting of intellectual property rights (IPR) activities, digitalized production and services, and e-distribution (see Fig. 1).

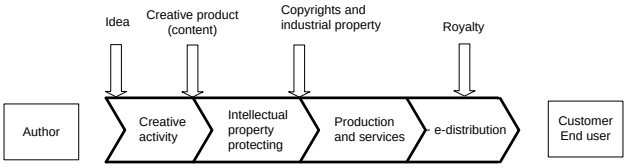


Fig. 1 Organization of creative activities and intellectual property protecting in cluster

Source: Adapted by [8]

Creative activity is transformation a new idea in digital content of creative products what intellectual property is (copyright work or industrial property). These properties need protect before their distribution (alone or as content of digital product). Created intellectual properties and IP rights are new economic values and inputs of digitalized production and services.

TIMES clusters organize value-added chains with different topologies (see Chapter V) that continuous improve and innovate it, so create add value for cluster members and customers at the same time (see Fig. 2). Innovations are new, improved or upgraded products, services and processes that have new commercial or/and social value. It is transformation knowledge and ideas in

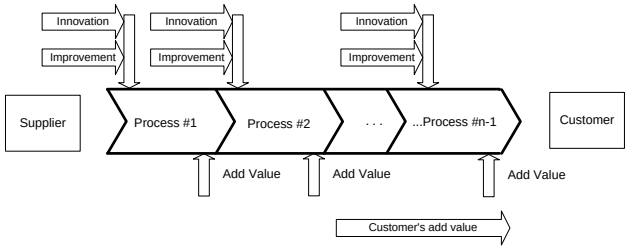


Fig. 2 Cluster's value-added chain

Source: Adapted by [8]

new added value. E-clusters have innovation management system which supports of innovation process from the idea fixing up, research and development, prototyping, pilot construction to implementation of innovation.

Presumptions and expectations cluster members and their benefits and trust, human resource abilities and other cluster potentials are base of cooperation (see Fig. 3).

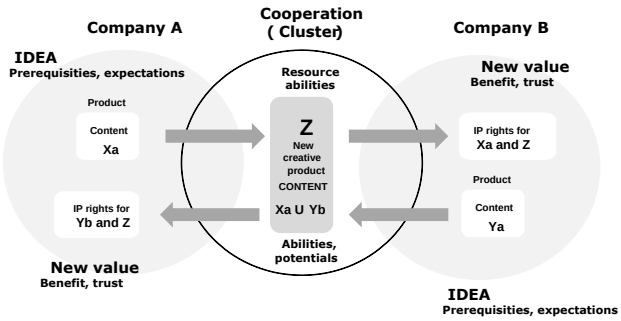


Fig. 3 Cooperation in the TIMES cluster

Source: Adapted by [8]

Digital content is intellectual property. TIMES cluster agreement and contracts between members, owners and users of their intellectual properties, should regulate acquiring and using IP rights and IP management in the cluster. IPM need incorporate in cooperation, innovation, and knowledge management processes (see Fig. 4).

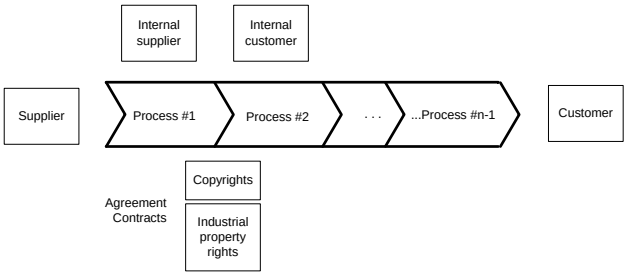


Fig. 4 Usage of intellectual property rights in cluster

Source: Adapted by [8]

Intellectual property management extends the process of cooperation in the TIMES cluster. This process begins before the creative activity (IP rights contracting) and lives long time after (e.g. licensing, royalty accounting). The intellectual property value chain exists in TIMES clusters and amplifies benefits of cooperation (see Fig. 5).

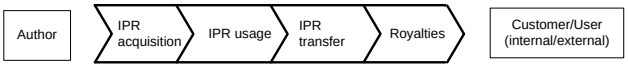


Fig. 5 Intellectual property value chain

Source: Adapted by [8]

Intellectual property rights are intangible property which stay in the TIMES cluster when authors gone or goes with authors if didn't contract elsewhere.

Intellectual capital is hide property which create added value. Hide property of the cluster are also:

- Cluster organizational knowledge,
- Relationship with suppliers and customers,
- Brand loyalty, and
- Market position.

B. Key presumptions and prerequisites of development TIMES cluster and e-clusters

TIMES clusters are human resource and innovation based clusters. Only creative individuals can produce digital contents and products, so human capital is key presumption for organizing TIMES clusters. **Human capital** is creative people with knowledge and abilities required for digital products and content creation, and sustainable growth of TIMES cluster based on creativity, productivity, profitability and competitiveness.

Creative people need organized infrastructure and processes for new values creating, as well as communication with other creators of digital products and contents. **Infrastructure or structure capital** is key prerequisite for organizing of human capital in TIMES companies and clusters, because it includes:

- Organizational culture and others declarative and procedural cluster organizational knowledge,
- Knowledge basis and knowledge management system,
- Intellectual property (e.g. trademarks, domains, patents, licenses and other intellectual property rights),
- Intellectual property management,
- Information basis and basis of documents, content management system, document management system and e-business systems.

Structure capital is everything what remain in TIMES cluster and cluster members) when people gone.

Human and structure capital are intellectual capital. TIMES cluster grows if his intellectual and creative capitals increase. Intellectual capital is knowledge and abilities that should convert in new economic value.

The cluster has **creative capital** if creativeness treats as capital property. Creativeness is realized when intellectual capital (own and/or strange) is used, new (individual) value added and new original product, process or service created.

Cluster organizational capital is ability of cluster that intellectual capital transforms into process of creating the new economic or/and social value.

Declarative knowledge are cluster politics and strategies, product specifications, organization models, document and knowledge basis. Procedural knowledge constitutes cluster organizational culture, business, production and logistics processes, procedures, applications, e-business cases and guidelines [3].

C. E-clusters

E-cluster is TIMES cluster with integrated processes of cooperation, innovation and knowledge management. It is a e-business network that horizontally or vertically integrates companies in TIMES sector and builds procurement and distribution e-channels and e-supply chains, provides product and processes innovation and applies knowledge, intellectual properties and other local strengths for growth of productivity and attainment global competitive advantage [2][3].

The geographic area covered by e-clusters can vary. They are usually regional clusters, but also may be national and even international (industrial clusters, clusters of excellence, union of clusters e.g.).

III. E-CLUSTERING

A. Main cluster development factors

Development of clusters determines several internal and external factors. Severe impact on founding and development of clusters has political, economic, social, technological, environmental and legislative factors. Political, economic, social, technologies, legislative and environment (PESTLE) analysis can indicate trends, possible long-term orientation and sustainability of TIMES clusters in the region. Key external factors are

- Macroeconomic policy (e-clustering) and
- Growth of TIMES technologies.

Key internal factors are

- Concept and e-cluster strategy,
- Connections (trust-based relationships between cluster members and their partners) and
- Cluster's intellectual capital.

B. Impact of macroeconomic policy on e-clusters

Macroeconomic policy has crucial impact on development and growth of clusters, because determines their main resources and competitiveness (availability of skilled human resources, labor and transaction costs, local market protection, investment in education, science and research). **E-clustering** is long-term macroeconomic policy that initiates and pushes founding and development of e-clusters in regions.

E-cluster composition is based on costs, resources or innovations, and their combination [4]. **Cost-based e-clusters** reduce transaction costs in TIMES sectors, **resource-based e-clusters** integrate human resources and services, and **innovation-based e-clusters** share know-how and other knowledge created in research and development of digital products and contents.

C. E-clustering as new strategy of TIMES cluster

E-clusters accelerate the distribution of knowledge, reduce transaction costs, produce increasing economies scale, and stimulate competition and cooperation. [2][9] So **e-clustering** is new TIMES cluster management strategy oriented to innovation and intellectual capital trough cooperation, growth and employment. E-cluster is a vision and strategic aim of established TIMES cluster.

D. University and cluster alignment

Growth of the cluster intellectual capital is key prerequisite of successful development of e-cluster. Universities are source of research and technology, knowledge, and creative workforce. Their impact on TIMES cluster growth is dominant, especially on cluster focused on innovative products and complex projects.

Foundation of innovation-based clusters is limited in region with low research intensity. When cluster concentration is high in the region, clusters are focused on building research and development units.

Establishing and development of innovation-based and human resource-based e-clusters is e-clustering strategy in regions with high research intensity and dominant universities. When clusters and universities are aligned, the focus is on efficiency and transfer of TIMES technologies and knowledge from university to e-clusters. Universities participate in innovation process in e-clusters from idea, research and development up to pilot construction and implementing the new digital products, digitalized processes or TIMES technologies.

Universities is industry-oriented and active regional engaged if new TIMES technologies and knowledge transfer to TIMES or e-clusters and other regional economy. Measures of success are licensing income, percentage of research and development funded by industry, start-ups per million EUR or USD, and full-time employees in the technology transfer. [11]

IV. ORGANIZATIONAL DEVELOPMENT OF E-CLUSTER

A. Organizational growth of cluster

Key cluster performance and organization growth indicators are employees, size (number of members and partners), added value, revenue and profit, average wage and intellectual capital or number of customers. Lifecycle of cluster's organization has four stages:

- Foundation,
- Growing,
- Maturity (stagnation), and
- Declining or crisis (see Fig. 6).

Foundation is the first stage of growth. At this embryonic stage, relationships and cooperation process need establish (see Pert B). Embryonic or emerging e-cluster is a group of TIMES SMEs and other actors in a region that want perform common in market. They start to cooperate and realize common opportunities through their linkages.

The cluster is established if have a room for further growth. Established or developing cluster continuously grows up to point of culmination when begin stagnate.

Cluster at stage of stagnation is maturity if its growth stable. It is at declining stage when growth declines.

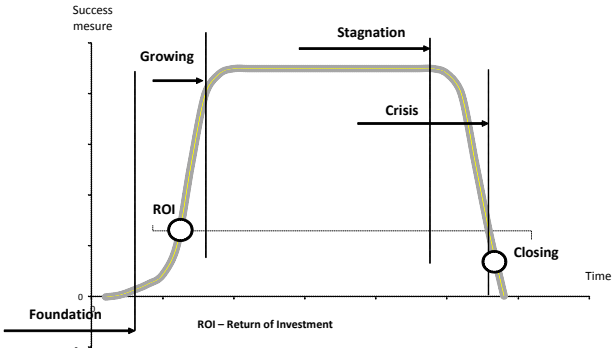


Fig. 6 The growth of e-cluster organization
Source: Author

Organizational development stages are detail described in Part C.

B. Foundation of e- cluster

In region with agglomeration of TIMES SMEs and other actors may be organize TIMES clusters with long-term perspective. Foundation of cluster is the end of complex process which involves

- Trust of potential founders creating,
- Linkages between actors building,
- Vision and strategic directions defining and then
- Undertake action with stakeholders.

Group of TIMES companies and other actors in one region can establish Marshallian, hub-and-spoke, satellite platform or state anchored TIMES cluster according to relationships to each other. Reasons of TIMES company for entering a TIMES cluster are costs or resource and innovation sharing. From this perspective TIMES cluster is cost, resource or innovation based.

The TIMES cluster organizational development planning should base on business environment indicators e.g. students in pre-vocational and vocational programs, lifelong learning, research and development personnel, household broadband access, part-time employment, human resources in science and technology, vacancy, skilled emigrants, and population aged from 15-34. [6] Building of e-cluster encloses next projects:

- Digitalization and integration process of cooperation,
- Innovation management system integration,
- Implementation of knowledge management system,
- Implementation of IP management system.

C. Organizational development stages and cycles

Unsuccessful embryonic e-clusters need close (see Fig. 7). Established e-clusters maintain growth and so stay in the growth cycle until the point of culmination. The culmination may be discarding with:

- Projects of new cluster's growth,
- The cluster development projects (organization, information systems, competitiveness, human resources, cluster management) and
- Projects for elimination of culmination.

Projects of new cluster's growth concerns new digital product and e-service implementation, innovations, new market accessing, investments, technology, new distribution e-channel and new e-supply chain.

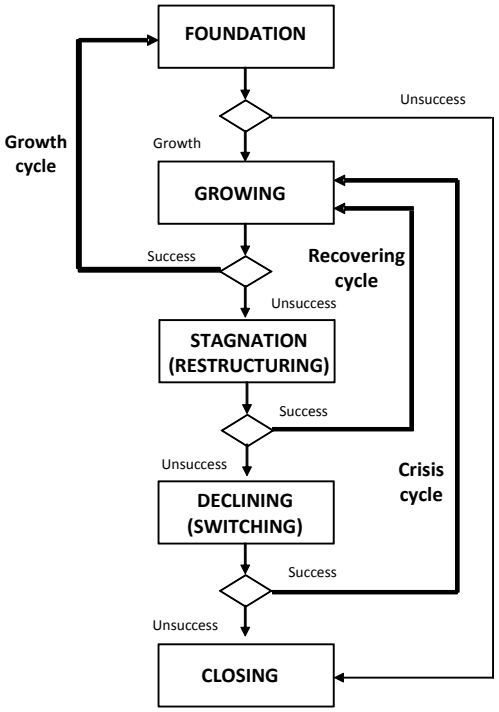


Fig. 7 Lifecycle of cluster's organization
Source: Author

Successful restructured maturity e-clusters exit out of the stagnation phase and return at the growing stage again. The e-cluster remains in the recovering cycle while growth stagnates. Restructuring can enclose products or projects portfolios restructuring, cooperation process innovation and redesign, e-channels and e-supply chain restructuring or redesign, and reorganization.

The e-cluster is at declining stage after unsuccessful restructuring and entering in the crisis cycle. The cluster is going out of the crisis with successful switching strategy that radically reducing costs of cooperation and only profitable products and projects leave. If switching strategy is unsuccessful then e-cluster need close.

D. Innovation and knowledge as development drivers

Innovation and knowledge are development drivers of e-cluster (see Fig. 8). Knowledge management supports acquisition, diffusion and applies cluster organizational knowledge (know-what, know-why, know-how, know-who, know-where and know-when) and so improves the e-cluster performance in e-market, and induce innovations of processes, digital products and e-services. [2]

Innovations increase cluster productivity and profitability, and improve products, services, cooperation processes and e-supply chains. They originate spiral growth of cluster knowledge in research and development phase. Innovation strategy activates creative capital for

the purpose of the cluster's strategic position improvement. It enables design and redesign products and e-services, and new TIMES technology and e-business systems implementation.

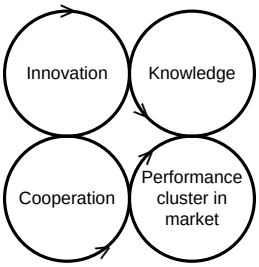


Fig. 8 E-cluster organizational development drivers
Source: Adapted by [1]

Creativeness and organizational capital are key e-cluster's development drivers in digital content industry.

V. ORGANIZATION OF COOPERATION IN E-CLUSTERS

A. Marshallian horizontal e-clusters

In Marshallian clusters there are no dominant members (leaders). Group of TIMES SMEs can establish successful Marshallian cluster if workforce in region is stable, focused on this industry and members want finance development of the cluster. This bottom-up development of TIMES cluster is sustain if cluster has:

- Low costs,
- Attractive resources (human resources, market, intellectual property, IP rights) and/or
- Competitive innovation and creative potential.

Governments initiate the founding of Marshallian TIMES clusters if TIMES sectors are target of macroeconomic policy and regional development.

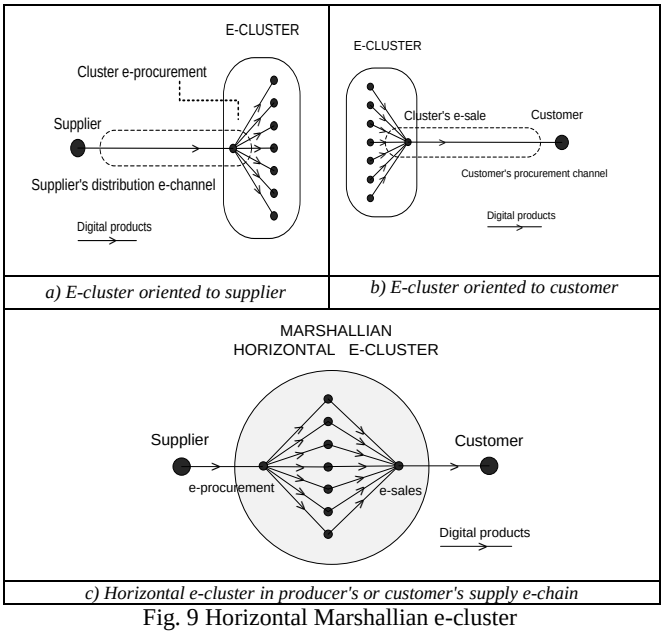


Fig. 9 Horizontal Marshallian e-cluster

Source: Adapted by [1] and [2]

Development of Marshallian e-cluster depends on the cluster members' orientation. The change of member's orientation is opening a new cluster's development cycle.

Members of horizontal Marshallian cluster perform common in the market, have similar characteristics, their products and services are complementary, or use same inputs. They are oriented to own business and few suppliers or the known customer.

Horizontal e-clusters oriented to supplier build e-procurement e-channels or enter in supplier's distribution e-channels (see Fig. 9a). E-clusters oriented to known customer build e-sales channels or enter in customer's supply e-chain (see Fig. 9b). Horizontal e-clusters develop procurement and sales e-channels and enter in suppliers' or/and customers' supply e-chains (see Fig. 9c). Transaction costs of cluster members are lower [5].

Among horizontal e-cluster members there is no digital product flows.

B. Cooperation in Marshallian vertical e-clusters

Vertical Marshallian e-clusters share innovation, knowledge and intellectual properties among members. Human and structure capital stays at cluster members. In e-cluster is only common structure capital.

Members of embryonic e-cluster build e-chains and integrate them in embedded e-clusters (see Fig. 11a, b). Transformation of horizontal in the vertical e-cluster starts when the first e-chain built. Topology of cooperation process is tree in e-clusters oriented to products, and direct graph (network schedule) in e-clusters oriented to projects (see Fig. 11c, d).

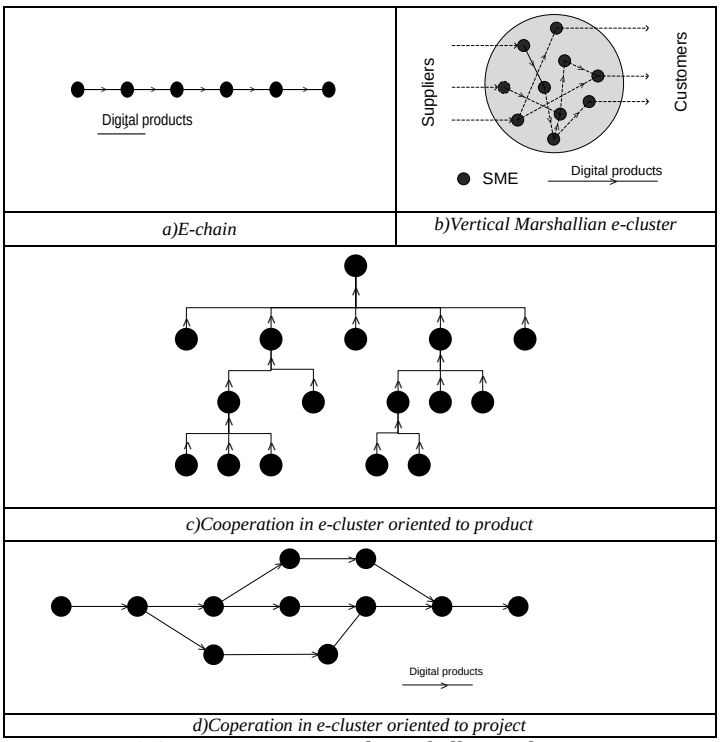


Fig. 10 Cooperation in vertical Marshallian e-cluster

Source: Adapted by [1][2]

Marshallian e-clusters may be support by e-logistic locate in public or private cloud. This solution is cost-effective, because cluster pay only used e-services.

C. Cooperation in other e-clusters

The hub-and-spoke TIMES clusters are organized by one or few large companies. Hubs are leaders and dominant member of cluster. They establish cluster in regions with low transaction cost, attractive resources and/or higher innovation and creative potentials. Cooperation between hubs and spokes are very intensive, but among spokes are little or no exist (see Fig. 11a). Finance capital of hubs will go elsewhere for better investment [7]. When hubs leave region, in region remain only human capital. Structure capital departs with hubs.

Cooperation between hub and their cluster members is supported by hub's e-business system. If cluster member participates in digital product or service development then enters in hub's knowledge management system. Hub can retain all copyrights and industrial property rights.

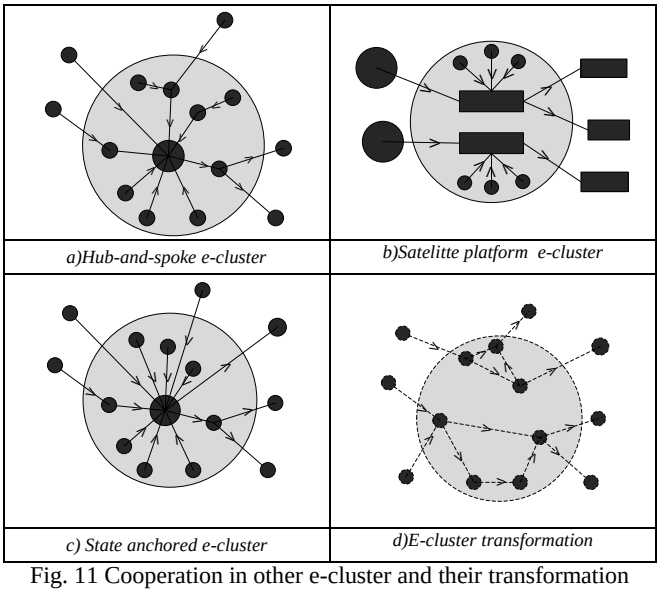


Fig. 11 Cooperation in other e-cluster and their transformation

Source: Author based on Markusen [2][10][12]

Large multinational companies establish satellite platform TIMES clusters in regions (see Fig. 11b) through government incentives such as free or inexpensive land, direct subsidies, or tax holidays. The government develops infrastructure and stimulates the large companies' coming and staying in the region [7]. These clusters will dissipate when government incentives dry up or favors development in other regions and industrial sectors. Infrastructure, workforce and small companies without own structure capital stay in the region.

Government institutions (university, research institute, military base, or capital) organize state-anchored clusters. The anchor institution cooperates with supplying companies. Cooperation among supplying firms is very little. Finance capital gone when better opportunity finds, but human capital remains (see Fig. 11c).

Basic topology of cooperation process in hub-and-spoke, satellite platform and state anchored e-clusters are star (see Fig. 11a-c). When hubs, founders of satellite platform and state anchor gone then cluster can transform in one or more Marshallian clusters (see Fig. 11d).

VI. CONCLUSION

E-clusters are cost, resource and innovation based TIMES cluster with integrated processes of digitalized cooperation, and innovation, knowledge and intellectual property management. Key factors of foundation and development of e-cluster are macroeconomic policy, development of TIMES technologies and cluster's intellectual capital (human and structural).

Knowledge and innovations are drivers of e-cluster growing, so e-clusters should cooperate with universities and research institutes and build research and development units. E-clusters can be supported by e-business and other management system located in private or public cloud.

REFERENCES

[1] Davidović, M.: "Building e-clusters", XIII Business Logistic in Modern Management, Osijek,October, 2013.

[2] Davidović, M.:"Od sustava e-poslovanja klastera do e-klastera" ("From cluster e-business system to e-cluster"),36. MIPRO, 2013.

[3] M. Davidović: "Uloga upravljanja znanjem u nastupu klastera na domaćem i stranom tržištu" ("Role of the knowledge management in the performance of business cluster performance in domestic and international market"), XII Business Logistic in Modern Management, Osijek, October, 2012.

[4] E.L.Seeley:"A New View on Management Decision that Lead to Locating Facilities in Industrial Clusters", The Journal of Business Inquiry, ISSN 2155-4056 (print)/ISSN 2155-4072 (online), pp 81-94, October, 2011

[5] M. Davidović: "Distribucijski e-kanali klastera","The Business Cluster's Distribution e-channels"), XI Business Logistic in Modern Management,Osijek, October, 2011.

[6] J. Stejskal:"Analysis of applicability of selected methods for industrial clusters identifying", International Journal of Systems Applications, Engineering & Development, Issue 3, Vol. 5, 2011

[7] J. He, M.H. Fallah:"The typology of technology clusters and its evolution – Evidence from the hi-tech industry", Technological Forecasting & Social Change 78 (2011) p. 945-952, 2011

[8] M. Davidović:"Innovation process in cluster and intellectual property protecting", (croat. orig. "Proces inovacija u klasteru I zaštita intelektualnog vlasništva"), workshop for cluster managers, GTZ, Osijek, 2010.

[9] U. Hansen, "Innovation through process oriented knowledge management", 45th Congress of the Regional Science Assotiation, Amsterdam, August, 2005

[10] T. Anderson et all:"The Cluster Policies Whitebook", IKED, ISBN 91-85281-03-4, August, 2004

[11] J. Paytas et all:"Universities and the development of industrial clusters", Carnegie Mellon Center for Economic Development, No 99-07-13822, 2004.

[12] A. Markussen:"Sticky Places in Sleepery Space: A Typology of Industrial Districts", Economic Geography, Vol. 72, p. 293-313, July 1996.

E-consultations in Croatian central government

Klasinc, Anton-Jan*
* Vertex projekt d.o.o.
ajklasinc@gmail.com jan.klasinc@mail.inet.hr vertexprojekt@inet.hr

Abstract - So far citizen consultations and participation in policy-making has been mostly the feature of local governments and local administrations. This paper aims to explore the usage of public participation in policy making via e-consultations with interested public in central government, namely in two Croatian ministries - Ministry of Science, Education and Sport (MoSES) and MInistry of Culture. This topic is mainly related to the problems of accountability and transparency, but also citizen cynicism and digital divide. Moving beyond this general approach, there is an underlying issue of defining and measuring accountability and establishing the role of e-consultations in the strengthening of accountability principle and framework within governmental organisations. The main problems analysed are the creation of curricula in elementary and secondary school in MoSES and policy-making concerning mainly the distribution of subsidies (establishment of public cultural needs), theaters, and electronic media in the Ministry of Culture. The paper explores the type of comment left on the web sites of those ministries, their quality and general tone, by the means of content analysis, but also the type of organisation and/or inndividual leaving the comment in order to assess the motivation and impact of the coment. On the other hand, it explores and analyses the structures enabling accountability and transparency in those ministries in order to establish the role and potential of e-consultations in strengthening them and possible improvements of e-consultation methodology and practice.

I. INTRODUCTION

This paper attempts to analyse e-consultations in Croatian central government on the basis of conducted e-consultations in two ministries: Ministry of Culture and Ministry of Science, Education and Sports. Although the analysis of e-consultations has been conducted on the local level (e.g. The paper on e-consultations in the city of Pula), such analysis has never been conducted on central government level, although the framework for "consultations with the interested public" (savjetovanje sa zainteresiranom javnošću) has been defined by a Code of Regulations – a legal document – that has prescribed the necessary and minimum parameters for such consultations, and e-consultations as well. The reasons may be multiple, first, it is unlikely that Croatian central government operates on the same level in all ministries, and therefore we have chosen the aforementioned ministries as examples of novel practices, since the ministry involved with science and education is the most likely to promote the new technologies, and one involved with culture is most likely to be sensitive to citizen engagement issues. Therefore, those two ministries could be percieved as current examples of best practice in Croatia. However, this is not necessarily the case and more studies are needed to establish this fact. Also, the volume of conducted e-consultations by the Ministry of Science has forced us to narrow our research to just two of them, when it comes to analyisis of the actual comments left by citizens – the consultations on Health Curriculum in schools and the Act on Higher Education in the MoSES and the consultations on Theater Act and the Electronic Media Act in the Ministry of Culture.

The control variables were the number of consultations conducted, the number of consultations that have resulted in policy change, the number of comments, the quality of interested public (physical person, legal person, institution or whether they are directly involved in the proposed regulation by their activity or not). The comments were analysed with respect to their constructiveness towards the proposed topic, and were therefore divided into two groups – constructive and destructive/cynical commentaries, and each of those groups has been subdivided into further two categories.

On the other hand, we have analysed the involvement of broader public and media and resulting policy changes, as well as structures enabling accountability in those ministries and their relation to policy changes and policy making.

The total number of e-consultations carried out in both ministries is rather different. In MoSES, there was a total of 43 e-consultations of which 2 are currently open, and in the Ministry of Culture there was a total of 22 e-consultations, most of which were in 2013 (13).

Table 1 Nunber of consultations in MoC and MoSES from 2008 to 2014

Year	Number of consultations - MoC	Number of consultations - MoSES	Total
2008	1	-	1
2009	2	-	2
2010	1	2	3
2011	2	2	4
2012	2	9	11
2013	13	28	41
2014	1	2	3
Total	22	43	65

The e-consultation in MoC that attracted the largest interest of the public concerned the Draft of the ammendments of the Act on Culture Councils (OJ 48/04 and 44/09)[1] and in the case of MoSES it was the Consultation on Health Education Curriculum [2]. The growth in the number of e-consultations can be seen even in 2012, but only in 2012, whereas it has reached a dramatic peak in 2013 and is continuing on this level in 2014. The reason behind this was probably not the Code (Kodeks) for the consultations with the interested public, which the government [3] (of the former prime minister Jadranka Kosor) passed on November 21. 2009, but the resolution of the new government elected in 2011 to respect the Code and put it into practice. This has seen numerous obstacles. One of them is the deeply rooted party mentality in Croatia that is not really open to citizens and their suggestions, and the other is probably ideological, which means that we are dealing with a "war of culture" or opposing mentalities rather than dialogue of equals, albeit with different ideas and opinions. As the third problem, there appear to be numeorus vested interests of different groups, (e.g. NGOs, Catholic Church and organisations connected to it, etc.) that are constantly trying to have upper hand in the society no matter what the results of elections are or regardless of any government policies.

This situation makes the problem of e-consultations with respect to the culture of transparency and accountability of government officials rather complex. The normal policy cycle of the cycle of public policy making [4], consisting of 1) agenda making, 2) analysis 3) policy making, 4) implementation, 5) monitoring is undermined by different social groups fighting for different agendas that are really just the camouflage for the power play and spoils system of dividing the proceeds of government "businesses" among themselves. Therefore it appears that some of the main goal of e-democracy or electronic consultation and public participation in policy making is not met, namely the increasing of e-engagement of citizens from all walks of life. In theory, the reasons for better engagement of citizens in the process of policy making are:

- better policies
- building of trust and acceptance of policies
- sharing of responsibilites for policy making [5]

Within this context, the ICTs should enable better disemination of technology enabled information, consultation and participation in order to improve the process of policy making by the means enabling:

- reaching and engagement of wider public
- provision of relevant information in an aceesible and user friendly format
- enabling thorough consultations and support to deliberation online
- facilitating the analysis of inputs in order to help policy makers and improve policies.[5]

II. METHODOLOGY

The qualitative analysis is based on the comments left on the web pages in consultations conducted in Croatia in the Ministry of Culture (MC) and Ministry of Science, Education and Sport (MoSES) which have resulted in policy change and/or significant media coverage. The control quantitative variables are the total number of consultations conducted in those ministries in 2013. and the total number of consultations from the beginning of the introduction of this practice (2008), the total number of consultations that have resulted in policy change, the number of comments left in each consultation that resulted in policy change. Also we have analysed the quality control variables: the type of interested public (private citizen or physical person, legal person, institution or whether they are directly involved in the proposed regulation by their activity or not). On the basis of this analysis it is possible to determine which variables crucially affect policy change due to the consultations with interested public. The main criticism of the central government public consultations is that they are conducted just formally, in order to fulfil a legal obligation, and not in order to really consult the public in order to change the policies and strengthe accountability. It is sometimes possible to find arguments for this conclusion, such as too short terms for consultations and poor informing of the public about the consultations, which goes against the better practice and still fulfils the legal obligations to provide consultations.

In the quality of the comments we will try to find proof whether it is the comments themselves that resulted in the policy change or some other factor and to establish whether it is due to faulty accountability structures in the govenrment. A large number of cynical and unconstructive comments would indicate that the comments were not crucial for the change, and that instead some other factor contributed to it, such as media coverage or public outcry, or some other form of political action, such as demonstrations on the streets, involvement of the opposition political parties, etc., which would then imply that the accountability structures of the government are not working. Also, the accountability structures themselves will be analysed according to an already established methodology by the Galway Accountability project (2009). In the quantitative control variables we shall seek to establish whether the number of consultations that resulted in some degree of policy change is statistically significant enough so that the e-consultations can be justly referred to as a means to policy change and not just an empty formality. For this we shall analyse also the total number of comments left, the number of comments taken into further consideration and the number of comments that were partially accepted. The number of consultations in time series will help us to setablish whether the situation is progressing or deteriorrating, and in which way it is progressing or deterriorating – in the sense of enabling more or less constructive and more or less significant impact of the public in policy making or just increasing or decreasing the quantity of consultations that do not result in any

significant impact or policy change. Finally, we will conduct the qualitative analysis of the type of interested public in order to determine whether it was the vested interests that motivate the public to participate actively in the consultations, or the genuine desire or the interested public to improve the quality of legislation and the accountability of central government structures.

III. E-CONSULTATIONS IN THE MOSES

Two main e-consultations in MoSES that are analysed are the Consultation on Health Education Curriculum, which was provoked by the large interest of the public and was a rerun of the already conducted consultations, where the whole issue was provoked by the organised citizen group "In the name of the family" connected to the Catholic church and which has resulted in the first citizen initiated referendum; and the other is the consultation on javna rasprava o the Draft of the Code of Regulations on appointment and work methodology of expert councils for the assessment of textbooks and additional educational tools for elementary and secondary school (Nacrt prijedloga Pravilnika o sastavu, imenovanju i načinu rada stručnih povjerenstava za prosudbu udžbenika i pripadajućih dopunskih nastavnih sredstava za osnovnu i srednju školu). The first one has attracted large interest of the public and even resulted in a citizen referendum, for which it was necessary to collect 700000 signatures of citizens all over Croatia, and the other one concerned also the whole public (or at least those with children going to elementary and secondary school) and small vested interest groups of publishers of elementary and secondary school textbooks and educational tools.

In the analysis, the comments were divided into two groups: negative/cynical (without reference to the issue addressed, without supporting arguments) and constructive/positive (directly referred to the document, with supporting arguments and informed opinion, applicable solutions and offering alternative solutions and approaches).

Table2. Comments in e-consultations in MoSES, 2013

Title	Constructive	Cynical	Total comments
Health curric.	cca 10%	cca 90%	1962
Expert Councils	-	-	0
Stud. food cost	7	1	8

In the first case, a large percentage of comments were destructive and cynical, although some pointed out the need to have public consultations that are better suited for this issue, and that the consultations were already carried out on this topic, which made the consultations illegal. The number of comments was rather high and difficult to analyse, and only the abbreviations with the essence of the meaning were conveyed in the final analysis document. In the second case, the number of comments was exactly zero (0) reflecting probably the vested interests of publishers. The consultation did not reach wider audience nor did it affect considerably the policy outcome, although it is actually important which textbooks will be used in schools, but the topic is not politicised as such. The period for consultations was only 2 weeks, which is considered a minimum period in the Code. However, some issues regarding publishing of textbooks were widely debated in the newspapers, implying that the issue was not really resolved by deliberation or e-consultations. The third is a typical consultation with 7 comments. The total number of providers was high, probably due to high interest attracted by the issue of Health curriculum – 3034, and the number of comments taken into consideration was equally high – 3035. However, the number of accepted comments was only 287 with partially accepted 108 comments, which means that only about 10% of comments created policy changes. This confirms that the increase of public interest and comments not made purely out of vested interests do not yield more constructive comments, nor does it provide for proportionally more impact of the public in policy making. However, in absolute sense, it may provide for more impact, but also due to other factors, such as media campaign and political action (political parties, demonstrations, protests, referenda, etc.) The ratio of physical (125) to legal (35) persons in the "Textbook standard" consultation shows that the public has genuine interest in the issues important for accountability.

IV. E-CONSULTATIONS IN THE MINISTRY OF CULTURE

The Ministry of culture has attracted wide attention of the public with two of its e-consultations with the interested public. The first one was the consultation about the ammendments of the Act on Culture Councils, which has attracted a lot of attention from groups with vested interests, as the Councils are the bodies within the MoC that distribute subsidies. The Act that was to be ammended was not very transparent, and it appeared that the main goal was to enable the minister to have complete control over the decisions of the Councils. Therefore, a lot of constructive, but also destructive comments appeared and the image of the minister of culture in public media, and especially newspapers was considerably eroded. However, that did not really matter much to the minister, that continued its practice as usual.

The other consultation concerned the ammendments to the Theater Act that was in fact designed to enable the minister to have better control over the appointment of the Principal of Croatian National Theater and other theaters that are also partly or totally funded by local governments. This, however, was not immediately attacked by groups with vested interests, probably due to the fact that it was a "clash of titans" between the local government mayor of Croatian capital, Zagreb, and the minister of culture, in which the majority of public interested in culture and its groups with vested interests supported either one or the other, but not so much on principle, but more due to vested interests and different points of view (traditional/conservative or liberal/left-wing). Therefore, the only comment received in this particular case was the comment of the principal and owner of a small childrens' theater "Mala scena", despite the fact that the conflict between the minister and the mayor was the topic of daily newspaper headlines for months leading up to the consultations. In addition to this, the comment was mostly constructive, although it did not reach wider audience nor did it affect policy making in this particular case. The

comment pointed out the lack of overall strategy and previous cases of ad hoc decision making that led to disastrous results and mistakes that had to be corrected due to the fact that legal acts were changed only to accomodate ephemere whims of political figures and enable power play.

Table3. Comments in e-consultations in MoC 2013/14

Title of the consultation	Constructive comments	Cynical comments (or useless)
Culture Councils	11	2
Theater Act (2014)	1	-
Pula film festival	1	-
Pr. status of librarians	16	2
Non-profit media grant	4	6
Electronic media	17	3
Theater Act I (2013)	31	6
Croatia House	3	1

In an overall statistics for 2013, it is stated that there was a total of 108 providers of comments, and that 172 were taken in consideration (the total number of comments is unknown), and that 34 of these were accepted with 1 partially accepted which is about 20%. This statistics confirms that in the Ministry of Culture there is generally more openness to suggestions of the interested public, which also goes in line with the fact that the comments are more constructive than in the MoSES and reflect vested interests of the interest groups (such as publishers, theater owners, actors, librarians, etc.); however a large portion of consultations received less then 2 comments (0 or 1). In the case of more comments, ratio of physical to legal persons was sometimes about 3:1 or even more, except for media and el. media (almost all legal persons). Another problem that was pointed out by many comments is the lack of thorough preparation and information phase in such consultation and the short period for consultation, which was even shorter than the minimum of 2 weeks that was prescribed by the Code for consultations with the interested public (2009). Such practice is not unusual in Croatia, nor is it a surprise, but it is unusual that it is condoned by a declaratively liberal minister and the fact that the Code is not obeyed is justified that Codes are not binding legal acts, so they don't have to be respected by the public administration, although they serve "as a benchmark".

V. STATISTICAL DATA ON E-CONSULTATIONS

The Government of the Republic of Croatia adopted the Report on the implementation of the consultations with the interested public in 2013 in the procedures of adopting acts, regulations and other legal norms, on its session No 143, held on 27/02/2013. In 2013 there was a total of 374 e-consultations. This is an increase of almost 160% more consultation than in 2012, when there were only 144 and 675% more than in 2011, when there were only 48 consultations. In the consultation procedures there was a total of 8.299 legal and physical persons that provided a comment to the draft of an act, or another regulation or legal act to the responsible government body. The number of the providers of comments increased by 73% with respect to 2012, and 4697% with respect to 2011. The individual number of comments was 9270, of which 2391 were taken into further consideration, 579 were partially accepted, and the reasons for not accepting the comments were explained in 149 reports about the results of consultations that are prepared pursuant to the obligations of the Government Code of regulations. Of 348 consultations, 272 lasted less than 30 days, which is less than is provided by the Act on Right to Access to Information, due to emergency procedures of legal act adoption.

It can be seen that on the basis of those numbers, the 2013 was the first year that significantly implemented the code for the consultations with the interested public, although it had existed 4 years before. This increase means that the impact of the public has increased, but not necessarily when weighted against the increase of the quantity of consultations, meaning that there were possibly more consultations, but fewer comments leading to some policy change. In fact, if the increase in consultations was almost 160% more than in 2012 and the number of providers of comments increased only by 73%, it would mean that the number of providers per consultations has decreased by more than 50%. However, for the previous year, also the number of providers per consultations has increased, as there was an increase of 4697% in providers and increase of 675% in consultations (when there were only 48), so that the number of providers increased 7 times, with respect to 2011. Also, on the qualitative side, it could be seen that in consultations, especially those leading to some policy change, there was an increase in the number of public that did not have vested interest in the consultations, but some kind of genuine interest to increase the quality of legal acts, whereas before 2012 almost exclusively persons with vested interest provided comments. On the other hand, it means that the time series for such analysis are too short to yield statistically significant results, and could perhaps be used only for the ministries with some tradition in consultations (MoSES and MC). [10]

Table 4 – Central government ministries and other organizations with more than 10 consultations in 2013

No	Organization	No of e-cons.
1	Ministry of Environment Prot.	52
2	Ministry of Finance (MoF)	50
3	Croatian Agency for Post and El. Comm.	39
4	Ministry of Health	36
5	Ministry of Maritime Affairs	29
6	Ministry of Science (MoSES)	28
7	Ministry of Labour and Pension System	16
8	Cr. Agency for Supervision of Fin. Serv.	16
9	Ministry of Culture	13
10	Ministry of Economy	13
TOTAL		292

VI. E-CONSULTATIONS AND ACCOUNTABILITY

When discussing accountability there is a problem of definition of accountability and methods of achieving accountability in organisations. One such method is given in the 2009 Accountability project (Galway) [6]:

In 2009, Phase I of the Accountability Project (Galway) articulated a set of essential elements of accountability. It is against these elements that an organisation's accountability would be established. They are as follows: (1) Organisation commitment to accountability and adoption of internal policies consistent with external criteria. (2) Mechanisms to put (privacy) policies into effect, including tools, training and education. (3) Systems for internal, ongoing oversight and assurance reviews and external verification. (4) Transparency and mechanisms for individual participation. (5) Means for remediation and external enforcement.

As for these elements it is difficult to determine whetеher such elements exist or are in place in the MoC or MoSES. However, it may be established that the commitment to the Code for consultations is an element of the commitment to accountability, as it enables two way communication with citizens and that it may also affect actual policies and policy making. However, the systems for internal ongoing oversight and assurance reviews are certainly lacking, and the policy makers are certainly not evaluated on the basis of comments collected by e-consultations. There are no systems for external oversight and verification, nor for remediation or external enforcement. Therefore, the public is forced to protest by referenda or public outcries in media or other methods that may make government officials and policy makers more accountable. In order to upgrade e-consultations as a tool for enforcing accountability, their visibility and flexibility should be greatly improved, as well as its reach toward general public. Also the creation of policy institutes of some scope, authority and influence may serve as an external evaluation and oversight structure for government policy makers.

VII. CONCLUSION

In conclusion, it may be said that the policy making by e-consultations in Croatia suffers from the lack of broader perspective and strategy and is mostly reduced to power play and clash of different mentalities and points of view, without much referrence to concrete arguments and issues that are debated and deliberated on. The number of destructive comments in the issues that reach the attention of wider public shows that the culture of deliberation is missing and that the main goal of referenda and public engagement and involvement in policy making is not just to change or improve policies, but to attack the government as a whole or to impose a different general point of view (e.g. conservative or liberal), and to try to stop policy changes that are perceived as threatening to that mentality, general point of view or vested interests of certain social and professional groups.On the other hand, the number of constructive comments on issues that are relevant for policy making or interesting to certain smaller interest groups or even general population (such as school curricula and textbooks) don't attract wider attention, and are mostly ignored by policy makers. The only way that such policies can attract wider attention is when they contradict general points of view and mentality of larger portion of people or powerful institutions, such as Catholic church, in which case the consultations and other instruments for citizen engagement are used for waging "culture wars" or trying to gain political upper hand or score some points in the political arena, without paying much attention to the policies that are actually involved or deliberated on by such consultations.

Therefore, we must conclude that the culture of deliberation and dialogue is defective in Croatia, and that it also affects the accountability structures in the government, as they don't feel responsible for concrete policies and their outcomes and benefits for general population, but mostly for ideological issues and points of view that are important for scoring political points with the voters. The accountability structures in ministries also follow those lines, and the need to maintain e-consultations is also mostly dictated by ideological reasons and the stimulus coming from the EU and its culture of dialogue, accountability and transparency, rather then from some inherent desire to create better and more citizen oriented policies.

Another problem with e-consultations in Croatia may be that the policies should be better formulated and conveyed to the citizens in a more appropriate form, reaching wider population, more inclusively, so that citizens are informed about the issues that affect their lives even when they don't affect their religious beliefs and general outlook or world view. In addition to this the tools for public participation and consultation should offer more possibility for dialogue and two way communication with government in real time, such as electronic fora, and alike, and the results of such consultations should be widely disseminated and made public.

LITERATURE

[1] Savjetovanje sa zainteresiranom javnošću: Nacrt prijedloga Zakona o izmjenama i dopunama Zakona o kulturnim vijećima (NN 48/04 i 44/09)

[2] http://public.mzos.hr/Default.aspx?art=12808&sec=3493

[3] http://narodne-novine.nn.hr/clanci/sluzbeni/2009_11_140_3402.html

[4] Caddy, Joanne et al: "Citizens as Partners", OECD, Paris, 2001, p. 41

[5] McIntosh Anne, at al.: "Promisses and Problems of e-Democracy", OECD, Paris, 2004

[6] The Centre for Information Policy Leadership LLP., Demonstrating and Measuring Accountability, a discussion document", Paris, 2010

[7] Bebić, Domagoj, Zenzerović, Bernard, Vučković, Milica: "Online Participation in Croatia: Movingbeyond citizens' cynicism – the case of city of Pula", European Journal of ePractice · www.epracticejournal.eu Nº 16 · June/July 2012

[8] Marc Gramberger: "Citizens as Partners – OECD handbook on information, consultation and public participation in polcy making", OECD, Paris, 2001

[9] http://www.min-kulture.hr/default.aspx?id=8407

[10] Vlada RH: Izvješće o provedbi savjetovanja u 2013 godini http://www.uzuvrh.hr/userfiles/file/Izvje%C5%A1%C4%87e%20o%20provedbi%20savjetovanja%20u%202013-nacrt-final.pdf

EVALUATION OF SHARED DIGITAL ECONOMY ADOPTION: CASE OF AIRBNB

Ljiljana Zekanović-Korona
Department of Tourism and Communication Sciences
University of Zadar
Franje Tuđmana 24i
ljkorona@unizd.hr
Jurica Grzunov
Department of Tourism and Communication Sciences
University of Zadar
Franje Tuđmana 24i
jgrzunov@unizd.hr

Abstract - Electronic market is a new economic space created by the development of the Internet and it has more than a billion of potential customers. In addition to all of the elements that define a classic market, e-market generates entirely new ways of doing business, such as the shared economy.
Shared economy is a sustainable economic system built around the sharing of private assets. This rather new system mostly relies on information technology (P2P) to empower individuals and other profit and non-profit organizations with ways of sharing excess capacity in goods, knowledge and services. Considering the fact that reliance and safety are crucial factors when it comes to sharing private assets, the value of those assets is increased just by sharing information about them. Airbnb.com, as an outstanding example of the new e-business model, enables individuals to rent their unoccupied living space and other short-term accommodation to guests. Besides being a rental mediator, Airbnb is also a social network used by its users to exchange experiences and interests related to traveling.
The research was conducted by an online survey that was posted on the Airbnb Facebook wall. The objective of this research is to determine the structure of Airbnb users and certain advantages and disadvantages of this type of e-business. The data acquired by this research was processed by the methods of descriptive statistics.

Keywords: digital economy, Internet, information technology, e-business, accommodation.

I. INTRODUCTION

Socio-economic phenomenon of tourism is based on the needs of people for a temporary change of residence. Today's modern concept of tourism causes significant changes in the tourism business and tourism development. Modern and sophisticated technology has resulted in improvement of communication, and the crucial role is played by the Internet. By abandoning the outdated forms of communication, the Internet has become an inevitable medium and communication channel and thus has brought distant places closer, facilitated travel organization, and improved the creation and marketing of tourist attractions. The development of the World Wide Web has affirmed the notion that the Internet can be used as a medium in performing various tasks and as an active and attractive business environment. Information and communication technology - ICT is developing rapidly and its change in use in everyday affairs, private, public and other market activities is expected. It eventually became the largest possible market of everything - goods, capital, labor and information content. More precisely, it is a new global economic space and virtual reflection of economic globalization. By the beginning of the 21st century the Internet was profiled as the segmented economic space within which four main economic sectors are now clearly distinguishable [1]:

- sector of producers of Internet infrastructure elements
- sector of providers of Internet services and applications
- sector of Internet entrepreneurs
- sector of intermediaries in the Internet business.

A plenty of people and organisations with innovative ideas of better or even completely new ways of doing business, ofering services and sales of tangible and intangible assets find their place in the sector of Internet entrepreneurs. There are such businesses which are not feasible in a traditional tangible business premises, but in the virtual space of the Internet we can notice a trend of specialization to conduct them. In this sector the most visible and the most interesting are the business benefits of e-business. The youngest sector is the one of intermediaries in electronic business, which brings together companies that organize and implement the provision of new types of intermediary services aimed at the better linking of supply and demand in the Internet market. These companies add value to the basic offer and therefore are called *value-added services*. The users of such companies do not perceive them as a "necessary evil" but as an added value that facilitates the achievement of business goals.

A direct consequence of the development of modern technology and ICT is the emergence of entirely new products and services on the Internet. It really is a global computer and communication "network of networks", which with its capabilities, features and amenities have enabled truly great acceleration of

virtualizing the market of private, public and other business services and activities, both in the social domain and the domain of economic relations [2]. Nowadays, in order to respond to the movement of modern world trends many enterprises transfer their operations from the traditional to the virtual space. This undoubtedly requires a modernization of business processes, so business is becoming faster, more efficient and profitable. The modernization of business processes makes the company extremely competitive in the global market exchange and in everyday competition.

The rapid development of information technology over the past fifty years has determined the information revolution. Today it is at the stage of maturity and has become the focus of business activities. The focus is increasingly shifting from technology to information and to the contents created by information technology. The growth of content and services related to information technology is inevitable as it opens new opportunities and encourages the creation of entirely new contents and services.

Digital Economy - DE is the modern way of doing business with the help of ICT, especially the Internet. Digital networking and communications infrastructure provide a global platform where people and organizations can cooperate, communicate, search for and obtain information as much as exchange digital products and services. Digital economy has led to the transition from the industrial economy to the economy characterized by information, intangible assets and invisible values and a new form of organization. The digital economy is characterized by the globalization, networking, mobility, integration, digital products and services and e-business. Every company faces with three kinds of pressure: the pressure of the market, technology and the pressure of society. The pressure of the market includes intense competition, global economy, changing workforce and powerful clients. The pressure of technology is characterized by technological innovation, accumulation and obsolescence of information technology. The pressure of society includes legislation, i.e. regulations and laws. Application of ICT enables companies to cope with everyday pressures. However, implementation of ICT does not ensure success, but it is necessary for the application of ICT to pursue the objectives, strategy and business processes of the company.

Today tourism is largely reliant on technology. ICT has become the support to many functions of tourism and has opened the door to innovation. Many tasks in the tourism sector, such as direct contact with clients, entertaining and services for clients, communication with customers and partners is being done through ICT. Since the early sixties, information technology and tourism have significantly influenced the mutual development. The first significant application of information systems, databases and computer networks occurred by introducing tourism booking systems, today indispensable in the tourism business. Tourism represents a significant application in the field of electronic business based on the B2C relations (business-to-consumer).

In the process of creating tourist attractions and its adjusting to the demand, various tourism products and services have been developed for the purpose of segmented business and creating specific forms of tourism. Most of the products and services result from an act of free innovation, while others appear as a by-product or a surrogate of indirect intentions.

The impact of the Internet on the tourism sector is reflected in the following. Suppliers, intermediaries and consumers (tourists) are the participants in the domain of the tourism. Suppliers (hotels, airlines, cars and vessels rental companies) provide the tourism and travel products. Tour operators and travel agencies, as intermediaries, create travel packages, and provide services of administration and sales to customers. Tourists as customers are playing role of the end users. The emergence of the Internet has led to major changes in tourism:

- prompted disappearance of intermediaries
- emergence of new mediators on the Internet with changing roles of traditional ones
- enabling direct access to end-users or consumers
- providing currently transmitting reliable information
- enabling display of financial position in real time
- decrease in organizational and distribution costs
- increase in convenience and flexibility.

The Internet has increased the possibility of suppliers' and customers' influence on prices, as well as on increase of the rivalry among existing competitors. It has also offered new direct sales channels in the travel industry. In this environment, the new company must be innovative to survive. At the same time, the tourism sector is forced to adapt to new technology.

Tourism today is a global activity which encompasses many areas of human activity including information technology, in addition to economics, ecology, sports and culture. Unlike other areas, tourism has always been open to new technologies. It wanted to meet the demanding needs of individual end users who, with the help of information and communication technologies, are not looking for the raw information only. Individual users want to become travel agents of their own in order to plan and organize their trips independently by using various forms of travel services. At the same time, the Internet has become an important medium of communication with the growing number of users in the world, and certainly the means of achieving the goals of end users who use the Internet not only for booking or buying plane tickets, but also for collecting relevant information for their possible travel arrangements. With the exception of the tourism offer that attracts tourists with its information, there are other types of information which, besides informing, encourage and guide the thinking of tourists, thus directly influencing their decision on taking trips. Useful information is often found on various social portals, blogs and networks, where individuals share their experiences and interests. Such forms of communication and information are increasingly serving the tourist industry, which places them on their websites and uses them as an additional way of attracting tourists.

The Internet is rapidly taking the characteristics of social media, which are:

- openness of communication

• reliance on community
• ease of connectivity
• active participation
• development of the spirit and atmosphere of cooperation.

In the first decade of the 21st century the Internet was transformed into a first-class social media, which in no way diminishes its business value and role. From a business standpoint online virtual space becomes even more interesting with the growth of its variety and diversity, and the interest of its users in what is happening in this space is increasing.

II. AIRBNB ONLINE SERVICE

Going on a trip that does not involve costly agency arrangements and hotel accommodation has become a frequent mission of tourists. The number of online services that enables it is increasing. Through them, one can book a desired type of accommodation in the desired destination and with the price range that is most acceptable to the traveller, in either modest or luxurious accommodation.

For this reason over the last few years, a multitude of Internet services have emerged based on the principle of shared economy or collaborative consumption or peer-to-peer (P2P) marketplaces, which are engaged in renting and sharing accommodation and various facilities and activities. An excellent example of such a service is Airbnb.com, a website that offers exclusive accommodation in a house, apartment, boat, room or something less conventional like a tree house or an igloo. Through Airbnb, accommodation can be found in almost every part of the world, and its advantage is that it is fast, clear, simple and secure. It is this latter reason why it is used by more and more tourists, but also by the renters.

The biggest problem of all these services, including Airbnb, is to establish trust with owners who open the doors of their homes. Airbnb was established in 2007 by the designers Brian Chesky, Joe Gebbia and Nate Blecharczyk. At that time, roommates Chesky and Gebbia could not afford the rent for their apartment in San Francisco, so they made their living room into a bed and breakfast, accommodating three guests on air mattresses and providing homemade breakfast. After renting their living room they earned enough to pay their own rent. They soon realized the potential in growing this market and joined with the programmer Nathan Blecharczyk to start Airbnb and changed the way people think about travel [4]. When they found the site, their goal was to create a base of private accommodation where guests can feel at home, but also to allow them to discover the local aspect of the destination, to connect with those who have already been there and to know what is not mentioned in guidebooks. Airbnb functions as a service used to host people who choose to travel cheaper around the world in an informal and friendly way. Airbnb offers comments and advices of tourists who have already stayed in a particular town, city or village and the users can find out if the accommodation is really as presented. Given that it works as a kind of social network, in a short time and firsthand one can find out where the nearest café or a shop is, as well as all the other practical tips. In order to protect their customers from unpleasant situations while still being able to plan additional facilities and activities, the approximate location of accommodation is marked on the map and the system reveals the exact street and house number only after the reservation is confirmed.

To avoid manipulation in the accommodation rating system, the reviews on Airbnb are purely qualitative. The unwritten rule is that after every staying in accommodation, travellers leave a comment, good or bad, so that future users of accommodation have better insight into the accommodation. But the site does not only assess the hosts and their apartments. As all users must create their own profile, the hosts can also assess their guests, which provide security to future hosts.

The Airbnb guests pay the full price increased for 6-12% which is the Airbnb guest fee, and Airbnb holds the money on its account until the customer appears. In this way, the host is insured in a case of cancellation by the guest, and the guest is insured in a case the accommodation is not as described on the website. In a case the guest does not show up his account will be charged and he is entitled to a refund only if the cancellation is made on time, but with a fee. Hosts are additionally insured by the deposit that Airbnb retains 48 hours after the guest's departure in a case something is broken. The day after the guest arrives and confirms that the accommodation is as described on the Airbnb site, the owner shall be paid the total amount reduced by 3% of the Airbnb charge. If guests are not satisfied with the accommodation and confirm it through the system, on the same day they must vacate the property and then they can get their money back or ask Airbnb to divert money into an alternative accommodation [6]. The advantage of Airbnb is extremely simple communication with the owners. Once the guest has selected some of the accommodation offered, via e-mail or social networks that are integrated into Airbnb site, he contacts the host and sends him his request. Since it is not obligatory to accept every guest, the host must first accept the request, and only then follow the booking and payment. Thus, the system of Airbnb's business is to a greater extent adjusted to hosts than to guests. Guests do not have much control in the process because the host can cancel their reservation at any time. On the other hand, Airbnb is investing a lot of effort in preventing the misuse of their services in terms of direct negotiation between guests and renters, because in that way the landlord could lower the price for the amount otherwise taken by Airbnb, and thus convince the customer to stay with him. Because of these actions, the system automatically disables the visibility of contact numbers or e-mail, until the booking has been confirmed.

In order to protect their users, Airbnb also started the offline verification during which the guest and host are identified by using identification documents such as passport, driver's license or ID card, and in adition to that a voice and video verification systems. Verification serves as a guarantee that their data on the profile are correct and accurate. Renting accommodation is possible only to those owners who have online verification of their profile. Airbnb also introduced a 50,000 dollars physical-

property guarantee for hosts to feel more secure about renting their accomodation [7].

In addition to online users, Airbnb pays a great deal of attention to the users of mobile phones. Over a quarter of its users visit Airbnb's site via their mobile apps. Today's 'mobile' generation of users finds the experience of knowing the environment more important compared to renting rooms. For example, instead of looking for a shop on the Internet, they will rather ask someone who lives there through the application and get an answer in a few seconds. Besides the world leading destinations, using Airbnb one can easily find accommodation in Croatia. By typing in Croatia in the desired destination, thousands of proposals are offered.

Until a few years ago this American startup was known only to the immediate community of Silicon Valley. After a year of business, its founders have almost given up on it. However, by 2014 Airbnb has reached a figure of over 11 million bookings. The site now offers over 600,000 units in 34,000 cities in 192 countries [3]. Besides being useful for tourists, Airbnb is one of the hottest technology startups in the U.S., and is valued at about 10 billion dollars [5]. Online site meant for finding accommodation around the world is employing over 600 people in 12 offices in the Americas, Europe, Asia and Australia. Expansion beyond the borders of the United States began in May 2011 when Airbnb bought German rival Acceleo. This lead to the opening of the first international Airbnb's office in Hamburg, and later of another one in London. Given the increasing number of its service users, in 2012 Airbnb has opened six more offices including Paris, Milan, Barcelona and Moscow. At the end of the 2012 the company announced that it would focus on Australia, the second largest market after the U.S., as well as Thailand and Indonesia. Therefore, Airbnb opened an office in Sydney, and the Australians now constitue one tenth of the Airbnb base. A few weeks after the announcement of aggressive expansion in Australia, they also announced expansion into the Asian market, and their goal is to reach 2 million accommodation units.

III. The questioning method

The aim of the research is to show the possible criteria for evaluating the structure of the content, manner of presentation of information and communication, internal and external motivation, and the simplicity and accessibility of the Airbnb database.

Testing was conducted through an online questionnaire which was originaly placed on the official Airbnb Facebook wall so that it could be accessible to the entire population of Airbnb users. Unfortunately, due to strict moderation policy the link was removed from their official Facebook wall very soon, which reflected negatively on the amount of collected data, ie. on the number and structure of respondents who filled out the survey. The online survey was opened during February 2014. and during that period a random sample of 118 respondents filled it out (77 females, i.e. 65.25% of the sample and 41 male respondents, i.e. 34.75% of the sample with the average age of 25 years). The online questionnaire consisted of 13 questions and all the questions in the survey were closed type with multiple choice answers.

IV. Results and discussion

Analyzing the survey results, we conclude that most of the respondents were from Croatia (54.23%) and Italy (18.21%), followed by a less significant number of those from France, Germany, Great Britain and other countries. Most respondents (38.1%) have a monthly income of less than 1,000 Euros, which was expected considering the age structure of respondents and the standard of living in the countries where most respondents come from. The average monthly income of the respondents amounted to 1529.66 Euros (SD. 1284.34). With the further analysis of the results, we see that 41.53% of the respondents travel 2-3 times a year, 21.12% of the respondents once a year. Ultimately, the respondents travel on average 2-3 times a year.

In the questionnaire there was a set of questions from which the TRI-index (*Technology Readiness Index*) was calculated for every respondent. Technology Readiness Index (TRI) is used to calculate the willingnes of respondents to adopt and use new technologies. This index is based on the examination of feelings that the respondent expresses for new technologies. Components, i.e. feelings covered by the TRI-index are: optimism, innovativeness, discomfort and insecurity. The above feelings are measured by Likert scale as a series of statements in which the respondents can assess how much they agree or disagree with a given statement on a scale of 1 to 5. The obtained result is the assessment of how much a particular individual or respondent is willing to accept new technologies. Readiness assessment includes the evaluation range of -20 to +20. If the above assessment scale is divided into five equal parts, it is possible to distinguish between the following categories of respondents:

1) people reluctant to use new technologies (from -20 to -13)
2) paranoids (from -12 to -5)
3) skeptics (from -4 to +4)
4) pioneers (from +5 to +12)
5) researchers (from +13 to +20).

Therefore, if the level of TRI-index is higher, the tendency of respondents to the acceptance and use of new technologies is greater.

Given the age of the respondents we get the expected structure of the respondents with respect to the TRI-index, i.e., we have 36.44% of sceptics, 61.86% pioneers, and 1.69% of researchers. 12.71% of the respondents have found out about the Airbnb service through newspapers and television, 36.44% via the Internet, 38.14% from friends or relatives, and 12.71% from other sources. 44% of respondents use this service to rent their own accommodation, 35.59% are planning to do so, while 20.33% of them are not attracted to the option of renting their own accommodation through the Airbnb service. Our respondents spend 3-5 days (51.7% of respondents) or one week (46.6%) on a trip in average.

Table 1. presents the average and the most common assessments of the Airbnb service elements arranged in groups according to the TRI-index of respondents.

TABLE 1. Advantages of Airbnb

Advantages of Airbnb	All		Skeptics		Pioneers	
	Average	Mod	Average	Mod	Average	Mod
Ease of use	**4,364**	**5**	**4,139**	**5**	**4,493**	**4**
Graphical design	**3,949**	**4**	**3,558**	**4**	**4,173**	**5**
Search options	**4,322**	**5**	**4,162**	**5**	**4,413**	**4**
Payment methods	**3,898**	**4**	**4,023**	**4**	**3,826**	**3**
Payment security	**3,720**	**4**	**4,000**	**4**	**3,560**	**4**
Information about the host	**4,296**	**5**	**3,883**	**3**	**4,533**	**5**
Guest and host service fee amounts	**3,712**	**4**	**4,023**	**4**	**3,533**	**4**
Communication with the hosts	**3,923**	**3**	**3,046**	**3**	**4,426**	**4**
Accommodation availability check	**3,347**	**3**	**3,604**	**4**	**3,200**	**3**
Accommodation price transparency	**3,992**	**4**	**4,441**	**4**	**3,733**	**5**

Table 2. presents the average and the most common assessment of the Airbnb service components by groups with respect to the TRI-index of respondents.

TABLE 2. Components of Airbnb

Components of Airbnb	All		Skeptics		Pioneers	
	Average	Mod	Average	Mod	Average	Mod
The amount of attention that is paid to the needs of users	**3,347**	**3**	**3,000**	**3**	**3,547**	**3**
The amount of data on the host and accommodation	**3,169**	**3**	**2,442**	**2**	**3,587**	**3**
The possibility of changing the booked date	**3,707**	**4**	**3,581**	**4**	**3,786**	**4**
The functionality of the service (search speed)	**3,881**	**4**	**4,140**	**5**	**3,733**	**4**
The possibility of changing accommodation on arrival due to dissatisfaction with accommodation	**3,338**	**3**	**4,116**	**5**	**2,893**	**3**
The utilization of information and communication technologies for booking	**4,067**	**4**	**3,605**	**4**	**4,333**	**4**

Table 3. presents the average and the most common assessments of the Airbnb service shortcomings divided in groups according to the TRI-index of respondents.

TABLE 3. Drawbacks of Airbnb

Drawbacks of Airbnb	All		Skeptics		Pioneers	
	Average	Mod	Average	Mod	Average	Mod
The necessity of checking the availability of accommodation with the host	**3,822**	**2**	**3,674**	**1**	**3,907**	**2**
The inability of paying in installments or in cash	**3,602**	**3**	**4,698**	**6**	**2,973**	**4**
Hosts English proficiency	**4,779**	**5**	**3,326**	**2**	**5,613**	**6**
Cancellation policies and fees	**4,356**	**3**	**4,326**	**3**	**4,373**	**3**
The amount of commissions for the hosts and users	**5,093**	**7**	**5,349**	**8**	**4,947**	**7**
The possibility of errors (lack of control) when setting prices for accommodation (for hosts)	**5,093**	**4**	**4,628**	**5**	**5,36**	**6**
The possibility of hosts last minute cancellation	**4,186**	**1**	**4,395**	**1**	**4,067**	**5**
The difference between verified photos and accommodation in reality	**5,061**	**8**	**5,605**	**4**	**4,760**	**8**

From the data given in table 1. we can see significant differences in the average assessment of individual service elements between the respondents groups of sceptics and pioneers, in the elements: ease of use (F-test[1] = 3.868, p = 0.05), search options (F-test = 2.443, p = 0.05), information about the host (F-test = 4.004, p = 0.05).

From the data given in table 2. we can see significant differences in the average scores of the individual components of the service between subject groups of sceptics and pioneers, and they include the components of the functionality of the service (search speed) (F-test = 1.978, p = 0.05), the possibility of changing accommodation on arrival due to dissatisfaction with accommodation (F-test = 2.255, p = 0.05), the Airbnb ways of doing business as a form of digital economy (F-test = 1.999, p = 0.05).

From the data given in table 3. we can see significant differences in the average assessments of particular shortcomings of service between the respondent groups of sceptics and pioneers, in various

[1] The ANOVA F-test is designed to test if two population variances are equal.

components: the amount of commissions for the hosts and users (F-test = 1.912, p = 0.05), the possibility of hosts last minute cancellation (F-test = 1.696, p = 0.05).

V. CONCLUSION

Airbnb is a fast-growing online service, i.e. an online travel agency for a broad community of clients that offers a worldwide connection of hosts and customers seeking accommodation. The service itself is funded by the commission charged from hosts for the contracted accommodation which amounts to 3% of the amount defined by the host and which is significantly less than the classic tourist agency commission (20-30%), and by the commission of 6% up to 12% charged from guests. Given that today's younger people are good users of information technology, which is evident from the calculated TRI-indexes of respondents, the popularity of the Airbnb service is growing rapidly. This service is best characterized by its ease of use, search options and a lot of good information about the host, the method and the security of payment. Top rated components of this service are the ability to change the reservation date, the functionality of the service and the use of information and communication technologies used for booking.

From previous results we notice that there are no major discrepancies in the average assessments of particular shortcomings, and we see that there are also significant differences in the priority of the greatest shortcomings of the service. We conclude that the biggest drawback is the possibility of the host's last-minute cancellation, followed by the necessity of checking the availability of accommodation with the host and cancellation policies and fees. The group of pioneers who are familiar with the new information technology is not highly dissatisfied with a lack of additional communication with the host and the host's possibility of cancellation at the last moment because they will, with the help of information technology, quickly and easily find alternative accommodation. Because of reduced commission this way of doing business where everything takes place in a virtual world is becoming increasingly common and popular among hosts and users looking for accommodation. This service is significantly growing from year to year and it is expected that more such services, i.e. online travel agencies or companies will become a good example of the digital economy and the companies that generate profits with minimal costs. This research can serve as guidelines about what future online travel agencies should improve or change to make themselves more acceptable to users.

LITERATURE

[1] Panian, Ž. (2013.): Elektroničko poslovanje druge generacije. p.7, Ekonomski fakultet, Zagreb.

[2] Šimović,V. (2009.): Uvod u informacijske sustave. p.221, Golden Marketing, Zagreb.

[3] http://www.airbnb.com/about/ (February 02, 2014)

[4] http://www.econ.ucla.edu/sboard/teaching/tech/Airbnb.pdf (March 05, 2014)

[5] http://research.gigaom.com/2014/03/airbnbs-valuation-reportedly-hits-10-billion/ (March 05, 2014)

[6] http://www.washingtonpost.com/wp-dyn/content/article/2009/07/24/AR2009072401682_2.html?sid=ST2009072402126 (March 11, 2014)

[7] http://tech.fortune.cnn.com/2012/05/03/airbnb-apartments-social-media/ (March 11, 2014)

The digital economy in developing countries - challenges and opportunities

PhD Vujica Lazović
MSc Tamara Đuričković
Faculty of Economics
Jovana Tomasevica 37, Podgorica, Montenegro
e-mail: vujical@t-com.me
tamara.djurickovic@ef.ac.me
Tel: +382 20 241 138

Abstract - The digital economy brings with it a number of opportunities, but also new challenges and rules of the game in the global market. Positioning of the country on the global stage largely depends on its ability to adapt to new conditions. Digital economy brings a new set of benefits, which can make it possible to reduce the differences that exist between rich and poor nations. Developing countries have the opportunity to transform its economy and to contribute to the development of the digital economy. Although these economies are characterized by high added value, faced with numerous obstacles, many developing countries can not adequately respond to the demands of the digital economy. Inadequate access to the latest technology, sophisticated telecommunications infrastructure, low computer literacy as well as numerous cultural and socio-economic factors are just some of the challenges that developing countries have to face. On the other hand, with a clear policy and specific plans and objectives, it is possible to "skip" a few steps and effectively respond to the demands of the global market. This paper aims to analyze the challenges and opportunities that are facing developing countries in the process of creating the digital economy.

Key words: digital economy, ICT, digital divide

I. INTRODUCTION

Concepts 'digital economy', 'information economy', 'information society' and 'knowledge economy/society' are very connected and often interchangeably used. However, the concept of 'information society' substantially starting to be used as a unifying term . In favor of this, some authors suggest that the term 'information society' is presented as a strategic objective in order to overcome existing social stagnation. Central argument is that the information society can create and provide millions of new jobs; it guarantee economic competitiveness; brings environmental benefits; intensify democracy; revolutionize the way of life and work with the help of new communication technologies.

On the other hand, some authors define the information society as a society that is characterized by rapid growth in the use of information - a society in which the quality of life and economic development significantly depend on the information and its use/exploitation, increased use of computers, e-commerce and use of technology for community development. In addition, there is also the explanation that the information societies are the national economy in which more than half of the workforce are employed primarily on activities that are information-oriented and in which the net product of these activities is greater than half the gross national product .

II. CHARACTERISTICS OF THE DIGITAL ECONOMY

The digital economy is characterized by numerous phenomena in the global business environment, including, but not limited to globalization, digitalization, virtuelizaciju, and disintermediation [8]. The phenomenon of globalization has internationalized economy expanding market reach, scope and speed. Borders gave way to electronic constraints that enable millions of businesses through various, mainly online oriented initiatives to carry out (online) transactions in seamless manner. Digitization has eliminated the need for human contact, implying that goods and services are no longer required to be tangible, making activities directed towards the development of eBooks, digital music and the like. On the other hand, visualization has made the physical location and proximity irrelevant, in addition to providing opportunities to business transactions to be carried out in a virtual environment, including communication between multiple entities on both sides of the interaction and also communication between one or more subjects of interaction. Finally, disintermediation has enabled consumers to bypass intermediaries and directly reach to the producers.

Applications of the digital economy (like e-commerce) are mostly driven by information and knowledge sources, and their use is significantly facilitated by widespread access to computers and the Internet. The main resource for production and use in the digital economy is just knowledge, and therefore managing information and knowledge i the key element for organizations that want to improve their business. Centers of information management are focused on several issues, including the creation and maintenance information, search for documents and other data objects as well as reviewing and reuse of data. Further issues include setting information, data extraction, publishing and distribution, as well as providing access to information. As well as information management, effective knowledge management is very important for the development of the digital economy. From the point of knowledge

management, human ingenuity is the most important factor in the growing digital economy. Constant technological innovation make it that a success in the market is largely dependent on the skills of the workforce. In the digital economy, nothing is important as it is human ingenuity, and the future and the success of the digital economy of just relying on this new type of worker. Thus, knowledge management is becoming an important topic in public organizations, although private firms have been already taking into account the management of knowledge as a critical determinant of the competitiveness of the company. In the private sector, the importance of knowledge management has been recognized for more than a decade, and today this sector can boast to have a fairly developed knowledge management systems.

III. DIGITAL ECONOMY DRIVERS

Various factors have given encourage to the transformation from the industrial economy to the new digital economy. Building confidence and security in e-commerce transactions represents one of the most important issues. The development of the digital economy also depends on the elimination of obstacles (such as legal uncertainty regarding the execution of electronic contracts, protection of privacy and personal data, copyrights, etc.) that interfere with transactions in the electronic environment. An appropriate regulatory framework that is adjusted and corresponding to the electronic environment is very important because it can increase: the growth and development of SMEs; workforce with appropriate IT skills, sustainability and a significant number of highly-skilled workforce; growth and greater use of the Internet; significant coverage of the broadband Internet, and self-regulation, especially when it comes to using the Internet [8].

Digital economy can also be enhanced by the extensive development of broadband communications in the business sector, government buildings and residences. Further growth of improving will require enabling legislation and government incentives; effective knowledge workers, education and training, development of governmental and private sector programs that promote digital democracy and ensure that all sectors of society realize the benefits of broadband revolution, as well as the expansion of citizen participation in decision-making processes in government through e-governance. Advances in ICT has been created under the pressure for reform and the digitization of both internal activities as well as activities in relation with the customers and the related trading partners. As a result of these changes, businesses today need to accept the network society and install business processes which are based on information and communication technologies for both front and back offices. Measures that can improve e-business include: enabling legislative systems, the diffusion of electronic payments for the business sector, reducing communication costs, and the development of skills necessary for the use of e-commerce business models. Further, given that it is not simple for individual SMEs to require sufficient human resources, know-how, technology and content, SMEs can use of their characteristic mobility and flexibility to collaborate with others. They can also build mechanisms that will implement the overall business activities over the Internet.

IV. DIGITAL ECONOMY CHALLENGES

In the digital economy era, which is largely defined by applications of e-commerce, it is expected that there will be improvement of efficiency due to reduced transaction costs and time, increased competitiveness, and much more streamlined business processes. Some authors assume that the greater efficiency will be manifested in lower prices, much more frequent modifications price, and a lower price dispersion for identical products. Lower costs of research may also lead to Internet consumers become more sensitive to price changes. Also, the development of the Internet and e-commerce influence on changing the way in which payment is made for goods and services, which as a result may have some implications in the establishment and functioning of monetary policy. The characteristics of the digital economy can change the extent to which e-commerce facilitates much more efficient inventory management, which leads to lower inventory as a ratio of sales. In addition, increased competition in the prices of products on the market may allow the economy to sustain many more jobs without increasing inflation over time, and may also create greater pressure on companies to restrain income growth and change processes that generate inflation.

There are several challenges that the business sector is also met in the environment of the digital economy. These challenges include the small number of trained workers in jobs that use ICT, the dependence of most of the business transactions that require a low level of knowledge intensity, broad and growing social and economic disparities, as well as the challenges of a complete migration services online to make them accessible to everyone. Some authors emphasize the following challenges of the digital economy: problems of lack of information/ computer literacy, slow understanding of importance of broadband, low proportions of people in the "e-workplaces", social and economic inequalities. On the other hand, it is possible to identify the challenges such as productivity in the economy, protection of intellectual property rights, invasion of privacy of the individual, the disappearance of certain types of work.

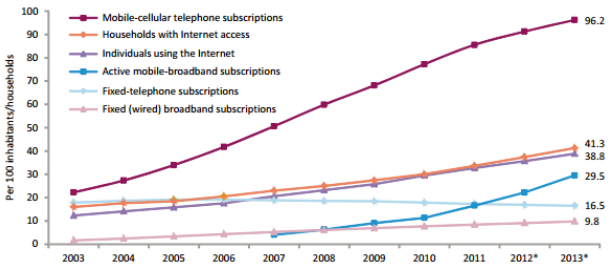


Figure 1: Global ICT developments, 2003-2013 (Note: *estimate) [14]

Some general trends indicate that the digital economy continues its rapid expansion. It is present in every country, in every sector, in most companies, and used over two billion people. By some estimates, already in 2016. year will be 3 billion Internet users. That same year, the digital economy will reach a value of about 4200 billion dollars within the G20. The digital economy in the 2010th took part with an average of 4.1% of GDP in group of the twenty most developed countries, representing about 2300 billion dollars. In some countries such as the United Kingdom, the digital economy contributes to GDP with 8.3%, which in other words presents a great web-dependence. The rapid growth of the digital economy will mostly depend on two factors, namely new users and faster Internet access. Thus the digital economy in the near future will be a very important factor for economic growth and job creation.[4]

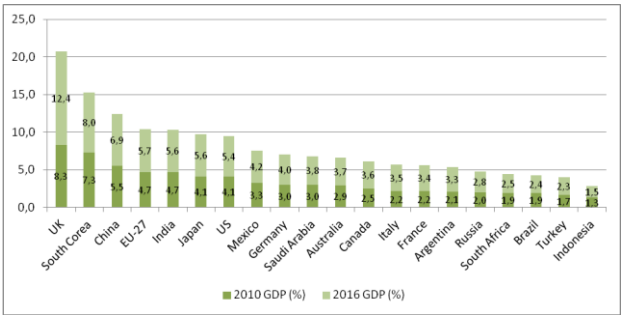


Figure 1. Internet economy as a percentage of 2010 and 2016 GDP [4]

As already stated, the digital economy is becoming a key element of growth and development. Almost all macroeconomic and statistical analysis show that the digital economy participated with 10 % in GDP growth of observed countries in the last fifteen years, and this influence is growing every day. [3]

The influence of the digital economy is evident in regard to a work force worldwide. The general trend is that certain traditional jobs disappear, while on the other hand new jobs are created with different qualifications. According to some estimation in developed countries like France, in the last 15 years the Internet has destroyed 500,000 jobs , while created 1.2 million new, which is a net difference of 700 000 new jobs. In other words , for every lost job elsewhere creates 2.6 new jobs.[3] The greatest significance of the digital economy is reflected in the modernization of traditional activities in the economy. The result is reflected in a major change in value between the sectors in the global economy, and thanks to that all industries have benefits. Creating an attractive business environment is following process of the digital economy and thus investment in the digital economy is the key to the prosperity of any country.

V. DIGITAL ECONOMY IN DEVELOPING COUNTRIES

An important question to be posed to the creators of economic policy is whether the new opportunities that are provided in the digital economy can improve the position of developing countries or they even further deepen the gap that exists between them and the developed countries. According to many authors, globalization and all its consequences make it possible to reduce the differences between rich and poor nations, where a key role is played by information and communication technologies. They provide to a developing countries a wide range of options to transform their economies into information economies with high added value that can compete with advanced economies on a global level. Simply put, developing countries can not "afford the luxury" to miss the opportunities that these technologies create. Some developing countries are increasingly turning to IT sector to accelerate its economic growth. However, this does not apply to all countries, because there is the other side of the coin: faced with numerous obstacles, many countries have not adequately responded to all the demands of the digital economy. Many problems are faced by developing countries, and which concern the implementation of modern information technologies, threatens to make those countries even more marginalized.

One of the main obstacles to the development of information technology in developing countries is the general use of computers by the population, i.e., the possibility of access to modern technologies. As one of the indicators, it can be seen the number of Internet users in these countries. While the world has become home to over 7 billion inhabitants, only one-third use the Internet. The data are worrisome from the standpoint of developing countries since in 2013. year only 30.7% of the population (on average) uses the Internet, which is unfavorable information in the age of the Internet revolution. However, important information that can be gained by analyzing the data is that the number of Internet users in these countries is growing at a higher rate than in developed countries. That's compared to 2005. year, the number of users in developing countries is increased by about 145%, while the indicator for developed countries is at the level of about 50%. So in the future we can expect a gradual "catching up" of developed countries on this indicator. [14]

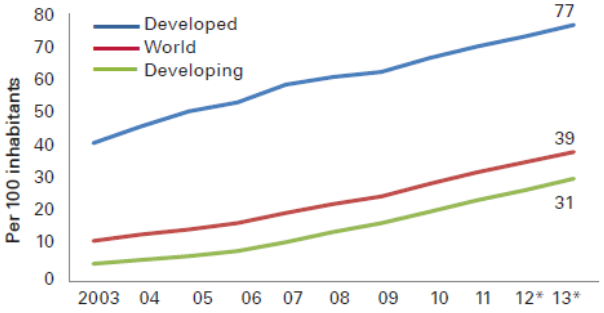


Figure 2: Internet users by development level 2003-2013 (Note: *estimate) [14]

Speaking of the digital economy it is the inevitable question of the use of broadband Internet. This information is important because this type of internet provides high speed data transfer, and thus it is the basis for the development of modern information technology in society. As for the number of Internet users, here is also evident a big gap in developing countries compared to developed countries. This indicator is for less developed countries by as much as 280% less compared to those that have been developed. In addition, the pace of growth of this indicator in these countries was insufficient for a period of 2007th to 2013th year.[14] It is for these reasons the slow growth of broadband has been cited as one of the main challenges of the information economy in developing countries. It is cheaper, faster and more secure access to the Internet, with the implementation of Wi -Fi basic prerequisite for the development of information economics, and thus a factor for the development of developing countries. As one of the recommendations for developing countries in terms of faster infrastructure development in the field of information and communication networks, it is stated the correct level of taxation in the field of ICT, in order to encourage domestic and especially to attract foreign investors. It also highlights the growing cooperation between the private and public sectors, as well as increased investment in research and development of both of these sectors. Telecommunications in developing countries are very weak in comparison with the current international standards.[1] In the 2013 the number of mobile phone subscribers was almost equal to the population of the planet. However, while in developed countries the number of mobile phones exceeds the number of inhabitants (that is 128%) in developing countries this figure is 89%. For example, mobile phone penetration rate in Montenegro is over 180% in many years: in 2008 it was 193%, 2009 - 197% and in 2010 - 197% while living standard i.e. GDP per capita is approximately 5.000 EUR. On the other hand, when compared to developed countries, mobile telephone penetration for example in Denmark is 153% and GDP per capita over 41.000 EUR, or in Netherlands where mobile phone penetration is 121% and GDP per capita over 35.000 EUR. This might be considered individual case instead of common rule for developing countries. However, the fact is that aforementioned level of information and communication technology use is much bigger than expected with regards to living standard of population.[9]

In addition to inadequate access to the latest technology, sophisticated telecommunications systems and infrastructure, also human factors in developing countries occurs as limiting. More specifically, one of the biggest challenges that this countries must face is the problem of low computer literacy as well as the low percentage of people who are dealing with an "e-businesses". Qualification of people to use information technology is often at a very low level. Most developing countries are characterized by low earning opportunities of this kind of education, especially when it comes to practical learning. Some of the measures that have been recommended to address these issues are: reducing cost of computer equipment and Internet access, free computer courses for the population; construction of computer classrooms in schools; constant training and upgrading of the teaching staff of the IT discipline; the introduction of specialized studies in the field of informatics science at universities; consumer protection, their safety and security while using the Internet and the like.[8]

Talking about these recommendations, it should particularly put focus on children and youth as a target group which most easily accept change, and already has a certain knowledge of using the computers. The importance of continuing education and learning is also increased, and in fact it is yet more proof that investing in knowledge is a basic precondition of using all the advantages of information and communication technologies.

The importance of computer literacy can be viewed also through the perspective of business enterprises. Modern enterprises around the world are increasingly turning to electronic commerce. The advantages that this type of business brings are various: starting from cost reductions on various grounds, via possibilities to respond faster to customer needs and attract the same all the way to global market presence.[10] To take full advantage of these benefits, i.e. that the company could start with the implementation of electronic commerce, the basic condition is that the company has a web site, i.e. that is being advertised on the Internet. Unfortunately, many companies in developing countries do not meet the basic criteria, mostly because of computer illiteracy, and therefore it is difficult for them to be competitive in the global market.

Besides the already mentioned problems of computer literacy, on the difficulties in the implementation of information technologies have the influence political and cultural factors. In fact, while in democratic societies, information and communication technologies are more than welcome as it will guarantee transparency, equal access to information and competitiveness, that is not the case in autocratic societies[7]. Different cultural and

sociological frameworks provide different significance to the information. In those environments that are "friendly oriented" when it comes to information, the education system has been created in the way to encourage curiosity and the search for information and answers to various questions. In other areas, people are accustomed to receiving information from the authority, someone who is in the hierarchy above them and therefore do not feel the need to search for them yourself . In these countries educational systems are often created not to encourage curiosity: there is only one right answer to the question, and what is the "right" answer is defined by the educational authorities. Regarding cultural factors, as one of the main problems emphasizes language barrier (about 55 % of web content is in English, 29 % of the content is in Russian, German, Japanese, Spanish, Chinese and French, and 15 % of the content is written on one of the 25 languages of Europe and Asia . Almost no content exists on any of the other 6000 languages in the world that are in use today[11]) .

In many countries, information technologies are hard to accept because they encourage "creative destruction", i.e. changing lifestyles that are traditionally used in these countries for many years. [5] This is especially present in countries with strong cultures, which hardly give up their traditions. Some countries even restrict access to the Internet due to concerns that it badly affects some religious or cultural beliefs. Another type of problem that comes from political authority in the country is the policy of maintaining the monopoly of national companies, which actually the state bureaucracy represents brake for the development. How will the country successfully position itself in the global market in the era of digital economy largely depends on the political process in the country. There are very often cases in which for the sake of achieving certain political interest groups, adversely affects the technological position of the country. Inefficient state monopoly enterprises in the ICT sector can with keeping high prices prevent large numbers of citizens to participate in courses that should characterize a information society.

As one of the key challenges of the digital economy in developing countries is referred to social and economic inequality. Generally speaking, the use of mobile phones and Internet access, namely the application of information technology should reduce the gap between rich and poor nations, because it provides equal opportunities to all. However, one should take into account that not all have the same starting position, or equal access to new technologies, which often affects even greater stratification between developed and developing nations. Also, the Internet in one country, people from the provinces, can provide capabilities that were previously reserved only for residents of the metropolis. But despite this, in some countries the gap between urban and rural areas is becoming increasingly higher. Developed cities are often surrounded by many villages that have almost no access to modern technologies. Due to the low population density in rural areas, they are often confronted with very high costs when it comes to investing in ICT infrastructure.

VI. CONCLUSION

Growth and development of the digital economy will certainly accelerate in the coming years through all areas of society because the number of people accessing the Internet multiplies as its commercial use is growing. Growth will be directed by four types of economic activities: using the Internet, a significant increase of computers, software, services and communications investments, e-commerce in the business sector, with improvements in productivity in the use of electronic networks to create, purchase, distribution, sale and supply of products and services, digital delivery of goods and services, such as software programs, music, etc., as well as retail sales of material goods and services that are produced, stored and delivered physically, such as computers, software, cars, books, flowers, etc.

It has been said that the digital economy brings a number of benefits to developing countries. Use of the Internet leads to greater transparency and strengthening democratic society. Also, it becomes essential to learning and career development. Information technologies increase productivity and affect the economic growth of the country. For this reason, developing countries need all its efforts to focus to overcome all the mentioned obstacles.

Clearly, the information-technology revolution as a new economic paradigm tends to reduce the differences in development between the rich and the poor, i.e. developed countries and developing countries. The application of new technologies, led by computers and the Internet has increased transparency and strengthened democratic society. However, as noted, there are many challenges facing the economies, societies, governments and individuals. All the challenges and obstacles can be overcome, but with the constant growth of productivity and competitiveness by investing in the knowledge that stands as a maxim in the new economic order. The pace of change is accelerating, and to avoid marginalization on the world stage over time it will require increasing investments in information and communication technologies and development of the information society as a whole to help developing countries achieve the objectives of the strategy of rapid, sustainable and inclusive development.

REFERENCES

[1.] Castells M., “Uspon umreženog društva“, Izdavač „Golden marketing“ , Zagreb, 2000.

[2.] Charlene L., Bernoff J., “Veliki val:kako pobijediti u svijetu novih društvenih tehnologija”

[3.] Matthieu Pelissie du Rausas, James Manyika, Eric Hazan, Jacques Bughin, Michael Chui, Remi Said, “Internet matters:The Net's sweeping impact on growth, jobs, and prosperity“, McKinsey Global Institute, 2011.

[4.] David Dean, Sebastian Digrande, Dominic Field, Andreas Lundmark, James O'day, John Pineda, Paul Zwillenberg, „The connected world - The internet economy in the G-20 - The $4.2 trillion opportunity, the Internet Economy in the G-20“, The Boston Consulting Group, 2012.

[5.] Dieter Ernst and Bengt-Åke Lundvall, *“Information technology in the learning economy: challenges for developing countries”,* Danska istraživačka jedinica za industrijsku dinamiku (DRUID), radna studija, 1997.

[6.] Ogunsola L. A., *“Information and Communication Technologies and the Effects of Globalization: Twenty-First Century "Digital Slavery" for Developing Countries--Myth or Reality?”,* Electronic Journal of Academic and Special Librarianship, v.6 no.1-2 (2005);

[7.] Rajesh M., *“A Study of the problems associated with ICT adaptability in Developing Countries in the context of Distance Education”,* Turkish Online Journal of Distance Education-TOJDE, April 2003.

[8.] Stephen, M.M., „Digital Economies:SMEs and E-Readiness“, Izdavač „Business science reference“, New York, 2010

[9.] Vujica Lazovic, Prospects of developing countries in digital economy case study of Montenegro, Acta Economica, Volume XII, No. 20 / February 2014

[10.] http://bryntegict.co.uk/resources/computing/theteacher/ict/ict1_29.htm

[11.] http://carbon.ucdenver.edu/~mryder/dig_div_este.html

[12.] http://ec.europa.eu/europe2020/index_en.htm,

[13.] http://software.benchmark.rs/softver_internet_vesti/google_microsoft_intel_pristup_jeftinijem_internetu

[14.] http://www.itu.int/en/ITU-D/Statistics/Pages/stat/default.aspx

[15.] http://www.microsoft.com/africa/4afrika/images/infographic.gif

[16.] http://ict-adv-disadv.blogspot.com/

SECURE DATA STORAGE

Pančo Ristov[1], Ante Mrvica[2], Toni Mišković[3]
[1]The Faculty of Maritime Studies in Split
[2]Jadrolinija d.d. Rijeka
[3]Student of the Faculty of Maritime Studies in Split

Abstract - Secure storage of data and the current availability of data and information are the most important aspects of any ICT (Information and Communications Technology) system. Data storage systems are mandatory components of modern information systems. The term backup refers to creating a backup or copy of data with the aim of restoring the data in case the original data become corrupted and inaccessible. Reliable and secure automated data storage is nowadays of great importance for business based on the smooth progress of information in an enterprise. Companies have to implement appropriate systems for securing the data storage. Some shipping companies use systems for saving data and applications in the so-called cloud. Cloud computing enables efficient and reliable fleet management. This technology reduces the cost of managing data and information resources, regardless of the size of the fleet.

I. INTRODUCTION

The application of information technology in business processes of the companies generate large amounts of data to be stored, regardless of whether these include output data and information from information systems, ship agents and agencies or data downloaded from the Internet. It is known that data and information represent valuable corporate assets, therefore the loss of data and information may cause great damage, regardless of the company size. For this reason, managements have to consider and decide on the safe storage of their data and information.

Each domestic and foreign shipping company has a specific need to share data, information and patterns among their workers (seafarers), maritime agents, agencies and business associates. Also, individual departments within a company have a need to share some information among their members, while keeping these data unavailable to other employees (e.g. actual cost of the ship in operation, fixed and variable costs of the ship, insurance and insurance premiums, value of the vessel...). Furthermore, each employee such as a ship's master has a need to store data in a way that only one person or a certain number of authorized staff can see them (e.g. technical condition of the vessel, fuel and lubricants...). Therefore, it is necessary to secure locations. These locations include local drives, network drives and servers that are established within the company or at geographically diverse remote locations. Each company should decide on a single, comprehensive and reliable solution for secure data storage. Such a solution should have all or as many of the following characteristics as possible:

- The possibility of saving data in multiple media
- Ability to deploy recovery after failure
- Unique system of the management and monitoring of the data saving process
- Simple interface of management and control system of the data saving process
- Deduplication features and deduplication at the source and the destination (magnetic tape or disk)
- Options for archiving data from file systems, databases, electronic mail, etc.
- Simple upgrade of the management and supervision system of the data saving process
- Deployment of storage systems without burdening the LAN

All companies, especially small and medium-sized ones, have to define their backup policy for securing relevant data. A backup policy is a set of rules within an organization which defines the activities related to data storage. Prior to defining the backup policy the following facts should be taken into consideration:

- Backup frequency - daily, weekly, monthly?
- Point of recovery - depending on the speed of data changes and their importance
- Types of backups - it is possible to implement various types of data backups
- Devices for backup - magnetic tapes, DAS (Direct Attached Storage), NAS (Network Attached Storage), SAN (Storage Area Network) devices...?
- Secure storage at another location - backup data stored at another location. The question is how often and where to perform the backup?
- Backup in the cloud - whether a data backup in the cloud is required?
- Safety - whether the encryption during storage is preferred?
- Items to be stored - applications, business documents, emails, databases?

Backup creation has two different purposes. The primary purpose is to retrieve data in case of data loss, while the secondary purpose of backup is to restore the data from historical moments within the constraints set by the user for defining backup.

The implementation of the backup policy most often depends on the financial resources required for this

investment, on the legal regulations of the country and on the corporate policy.

II. STORAGE MEDIA

Regardless of the system used for secure data storage, the data have to be stored in certain media. A backup device is one of the most important factors when defining the backup policy. Workload during data storage will depend on the selected backup devices. In practice, the following media are in use:

- Magnetic tape is the most widely used media for storing large amounts of data, backup, archiving and sharing. Tape is the media that requires sequential access data. Data access time is longer compared to other media, while periods of continuous reading or writing can be very fast. Newer magnetic drives are faster than modern hard disks. The advantage of this medium is the reliability and availability because the stored content can seamlessly be used for 10, 20 years and over in comparison with other media. Another advantage of the magnetic tape is a better capacity/price ratio compared to other media.
- Hard disk as a storage medium is increasingly preferred for mass storage of data because the ratio between capacity and price has been significantly improved compared to magnetic tape. Computer technology offers high-capacity drives, low access times, availability and easy connection to the computer (portable hard drives). On the other hand, there are a number of downsides: the hard drive may be physically damaged, especially in transport, or burned out by unprofessional handling; moreover, the stability of data stored on hard disks is questionable in the long run (decades). As a data storing medium, the hard disk has been increasingly used in marine information systems such as propulsion information system. The implemented configurations of propulsion information systems use only one hard drive (industrial design). In order to increase the reliability and availability of the information propulsion system and thus increase the security and safety of the ship, it is necessary to implement redundant disks or configure RAID (Redundant Array of Independent Disks) array.
- NAS (Network Attached Storage) devices consist of multiple hard drives. Depending on the number of disks and the possibilities of a computer platform, NAS devices range from small household appliances to the enterprise solutions fitted with ten or more drives in various RAID redundant disk arrays, featuring high capacity and very high reliability. NAS drives are perhaps the best solution for medium and small businesses.
- Optical media include media such as blue-ray, CD (Compact Disc) or DVD (Digital Versatile Disc) discs. Blue-ray is the successor to the DVD. There are two variants of the Professional Disc for Data & Professional Disc for Broadcast. Writing data is based on the "phase change" technique with the possibility of multiple data logger. This technique allows the storage of massive amounts of data and is price competitive, compared with hard drives or tape drives. It is possible to use a CD-R/RW or DVD-R/RW media that are cheaper and more universal than blue-ray discs; however, it is not recommended to use these media as the results of technical and scientific studies indicate that this method of storing data is not sufficiently reliable. If a user is forced to use this media type, the DVD-RAM (Digital Versatile Disc-Random Access Memory) would be a better choice.
- Media with no moving parts include flash memory, USB sticks, SSD (Solid State Drive) drives, etc. Flash memory or flash EEPROM (Electrically - erasable Programmable Read-Only memory) does not require additional drivers to operate and is used where it is essential to store data on a medium that is physically as small as possible. SSD disks have no moving parts, which solves the problem of vibration and noise. In addition, they are characterized by low power consumption and, consequently, low heating rate. Two memory types, NAND flash and DRAM, are used for writing data. Magnetic radiation has no impact on SSD drives. Media with no moving parts are quite expensive compared to offered capacity (poor price/capacity ratio).
- Remote Backup is data storage on remote servers (storage in the clouds). The development of the internet technologies (broadband transmission) and computer technology has enabled the creation of data centers that provide secure availability of critical data, applications, office functions, etc. The services offered by data centers are becoming increasingly popular as a backup on remote servers protects from the worst scenarios such as floods, earthquakes, fires, war and other disasters that might destroy all local backups or those in close proximity. Besides the positive aspects of this method of storing data, there are drawbacks such as speed transmission on the Internet. It is known that the upload speed of the Internet is essentially slower than the speed of transfer to a variety of physical media, especially as upload speed is significantly lower than download, which makes backup extremely slow. Another negative feature of this data storage method is related to the trust in the people working in data centers. This problem can be solved by encrypting data before sending them to the data center.

III. DATA STORAGE TECHNOLOGIES

There are three basic types of data storage technologies: DAS (Direct Attached Storage), NAS (Network Attached Storage) and SAN (Storage Area Network) systems. DAS technology refers primarily to digital storage of data which is directly connected to a workstation or server where a computer network between computers does not exist. Data sharing is performed through computers. Figure 1 shows DAS architecture connecting the computer to a series of hard disk drives (it can be a single

disk) and the storage system. Interfaces such as SCSI, SATA, ATA or FC (Fibre Channel) are used to connect the computer and disks. The core concept of DAS architecture is the inexistence of a computer network for connecting storage devices and client computers. In other words, there is no direct access to storage devices, but the sharing of data is performed through computers. Storage devices are managed by the client operating system and the application accesses the data through the system calls.

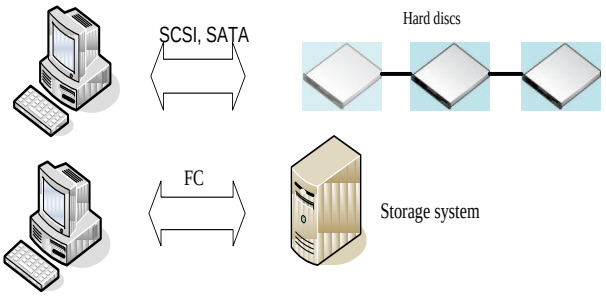


Figure 1. DAS systems

The downside of DAS architecture is that the storage capacity is limited by the number of discs. Adding or removal of one disc results in impaired access to all disks.

NAS technology involves storing data in a computer data storage (NAS device) that is connected to the LAN. Due to its network characteristics, NAS technology allows different client computers that are connected to the network with different operating systems to access the NAS device as shown in Figure 2.

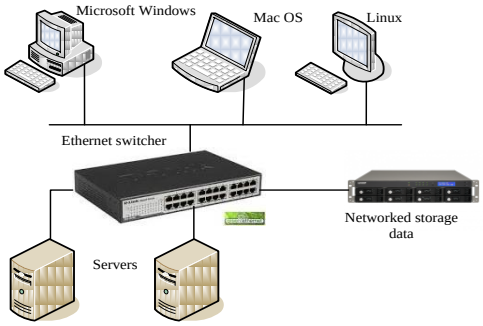


Figure 2. NAS systems

It can be noticed that the networked information storage is connected to the LAN on an equal footing with other computers so that other clients within the network can have access. Also, Figure 2 shows that the NAS device is directly connected to the switch, thus enabling all the computers within the network to share data. NAS architecture can be easily managed (internet or intranet) and the capacity can be increased to store a variety of data types (data, video, image data...). The NAS technology uses protocols such as NFS (Network File System), SMB / CIFS (Server Message Block / Common Internet File System), AFP (Apple Filing Protocol), FTP (File Transfer Protocol), HTTP (Hypertext Transfer Protocol), UPnP (Universal Plug and Play), SSH (Secure Shell), iSCSI (Internet Small Computer System Interface), AoE (ATA over Ethernet), FCoE (Fiber Channel over Ethernet), Rsync. In comparison with the other two technologies, the advantages of NAS technology include easier access to the administration, easier installation of services, data decentralization, easy set-up of RAID arrays and faster data access [1, 2]. Because of its advantages NAS technology prevails in medium and small business environments.

In the IT (Information Technology) market, there are open source software packages that provide services to NAS technology [2]:

- Openfiler is an operating system that provides a repository of data and is based on rPath Linux distribution. It is a free software package that is licensed under the GNU General Public License.
- NASLIite is a free Linux distribution (operating system) designed with the aim of converting computer-based x 86 environments into easily networked information storage. It is available for use on operating systems Windows, Linus, Mac OS K and other.
- Sun Open Storage is an open-source platform for data storage developed by Sun Microsystems organization.
- FreeNAS is a free NAS server which has support for protocols CIFS, FTP, NFS, AFP and others, and is released under the BSD (Berkeley Software Distribution Eng.) license.

SAN technology is similar to NAS technology. Data stores are separate from the LAN and implemented in their SANs networks. SANs commonly use SCSI and FC protocols. Storage of data implies a number of storage devices that are connected to each other sharing their data. A block schematic presentation of the SAN architecture is shown in Figure 3.

It can be noticed that SAN devices are connected to their own network and are separated from the LAN, using the server and thereby not affecting the traffic in the LAN network.

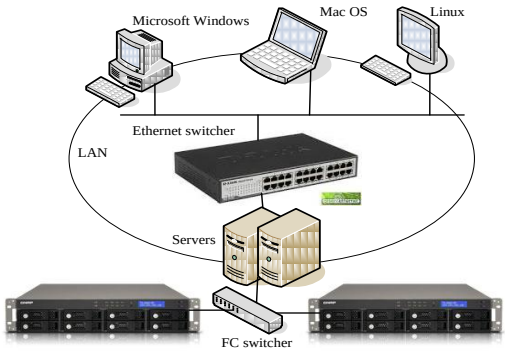


Figure 3. SAN systems

All displayed elements are connected as nodes. Each SAN system works in conjunction with one or more respective servers. With high reliability, they boast a very high configurability and adaptability. The most sophisticated systems can have up to 512 discs in a SAN system. One of the advantages of the SAN option is the quick and easy replacement of servers as SAN network "sees" resources for storage as SCSI devices. SAN architecture is a complex and expensive solution and is usually used only in large business sectors [2].

In practice, a combination of NAS and SAN technologies and SAN-NAS hybrid systems are often used. NAS system is suitable for heterogeneous environments, while SAN system is suitable for environments where large amounts of data are exchanged.

IV. THE TYPES OF DATA BACKUP

The company management shall decide on the type of data and information backup. Modern information technology offers several levels to be used depending on the size, organization, financial options, amount of data, reliability, availability, speed and availability of data, etc.

Level 1 is a way of storing data on a medium that is connected to the system which is always active and available. This level allows restoring the data in the event of a systemic failure of equipment or in the event of data corruption and mistakenly changed or deleted data. It is recommended to run in multiple layers so that individual files from multiple time points can be restored. This level does not protect data from fire, mechanical damage, theft, earthquake or flood and provides but a limited protection from electrical faults.

Level 2 or off-site backup is a way of storing data outside the company's computer center. This level of data storage protects against fire, mechanical damage, earthquakes, all kinds of power failures, floods, earthquakes, deliberate destruction of data, theft, etc. At this level the interaction and care of user (administrator) is necessary. This represents a bottleneck because administrators often do not have time to think about the defined procedures of data storage. The procedure of storing data defines: the storage time, storage media (tape, disk) and their removal to a safe location/position within the company (safe deposit box) or the outside premises of the company.

Level 3 or storing data in the cloud is an automatic back-up via the internet to secure data center servers and to a number of geographically remote locations. This method of secure data storage achieves 100% reliability because the backup procedures are programmed to be carried out automatically and without the direct influence of the administrator.

Some providers of data storage also define the state level 0 (systematic redundancy) which includes measures to be taken to maximize the security and availability of data. Practically this means that the hard drives are configured as RAID (Redundant Array of Independent Disks) arrays [4]. The concept of RAID is to draw most of the disc to the user, whether the user needs speed (RAID 0) security (RAID 1), or any combination (RAID 5, 5e, 6, 6e, 10).

V. STORAGE DATA IN THE CLOUDS

Cloud computing is an information technology that allows three models of service usage, including: software as a service – SaaS, platform as a service – PaaS, and infrastructure as a service – IaaS [4, 5]. Cloud computing represents a combination of many different existing technologies comprising the virtualization, data storage, connectivity and processing power to create a new environment. Storing data in the cloud is an easy and automated backup mechanism for real-time storage of data to a remote location. This mechanism provides a reliable solution for a centrally managed backup that facilitates easy storage of data and enables company managers and owners to protect their data as the most important corporate assets. Furthermore, this mechanism provides a very efficient compression, global data deduplication, significantly reducing the time frame for storage, energy storage capacity and loading on the network infrastructure. Data deduplication eliminates duplicate data at the corporate level, detached offices and mobile users. In addition, the mechanism verifies the integrity (Data Integrity Verification) with automatic recovery and the possibility of advanced rescue (Advance Recovery) that provides immediate availability and eliminates the need for manual data recovery activities.

Data storage system is an ideal way to effectively protect important data on a remote location against unexpected events such as theft, fires, earthquakes, floods… Data storage systems are software solutions that perform storage of data at several remote locations at certain time intervals. In a secure manner the software collects, compresses, encrypts and archives data on the servers of the service provider, which are located at several different remote locations.

Data storage systems in the clouds provide comprehensive support for a variety of user applications and databases. Moreover, the systems provide data archiving and archive of e-mail messages.

Before any piece of data is transferred, the data storage system in the clouds encrypts it with the most secure encryption mechanism. The data remain encrypted during storage and during all transfers. The system has a special module enabling that all data storage and recovery processes are registered, therefore allowing a clear overview of data operations. In addition, such a system features a multiple factor authentication of identity. In other words, the data storage system in the cloud is a set of regulations for digital storage and access to confidential data.

Data storage systems in the clouds are professionally maintained servers and disk systems, providing support 24 hours a day, seven days a week. Data are stored in three or more geographically remote data centers, thus providing unmatched availability and reliability. If there is a failure in the system, the data can be accessed in a secure manner until the re-establishment of the system's full availability.

Systems of data storage in the clouds will be increasingly present within the integrated information and communication systems because they are cost effective in financial and material terms. There is no need for a database administrator and there are no new or additional costs regarding the acquisition and maintenance of the software and hardware... The only cost is related to

paying for the service which eliminates the need for any subsequent significant investment.

Systems for storing data in the clouds provide the following basic functionalities:

- Protection of business assets - data are automatically compressed, encrypted, and transmitted immediately to the determined location or locations designated by the company management. The transfer to a safe location does not require human intervention, hence the risk of human error is reduced and the availability of information systems is increased.
- High-speed data recovery – the system ensures that all lost data can be returned to the current location or the location specified by the user. Any data can be recovered from any storage location and transferred to any specified location.
- The central point of management – the implementation of storage systems in the clouds eliminates the need for the local staff (database administrator) in charge of maintenance and the staff in charge of performing the backup process. The entire process of storing data is supervised from one location: defining the data to be stored, defining storage schedules, initiating retrieval, updating user data, access to data from any place any time. Furthermore, it is possible to monitor the status of servers and delegate users with certain rights from the central location.
- Data archiving – the system offers a simple module for data archiving. All files can be archived, including video recordings and e-mail messages. The system provides modules for searching and restoring individual email messages and attachments.
- Continuous and default schedule storage - in line with the corporate policy, the backup system performs safe storage of data in a well-defined arrangement of storage. Continuous storage allows for selecting the correct version for recovery, and each version indicates the version of the data from the period of storage.
- Secure data removal - when a piece of data reaches the end of the life cycle, the system has a module for removing the data in a secure manner, forbidding the access to unauthorized persons.
- Trace manipulation (handling) data – the system allows comprehensive monitoring of all activities and operations for internal or legally prescribed supervision. This module collects and stores all information in its original form, and provides access only to the legally authorized persons.
- Effective management of network traffic - through hardware and software support it is possible to have a flexible control of the network traffic. If necessary, it is possible to set limits of network traffic for given periods of time.

The advantages of storing data in the cloud include: fixed costs of space rental, expandability (scalability), saving to another location, speed of deployment, ease of administration, low energy consumption, reduction of electronic waste and decreased environmental impact.

The disadvantages of storing data in the cloud include data security, control, openness, scalability, backup duration, making contract with the service provider, etc.

The application of the mechanism of storing data in the cloud enables the companies to focus on their business processes and at the same time allows the technical departments (IT services) to reduce the risk of having difficulties due to the available memory capacity and the increased demand for storage space. The result is the increased efficiency and speed of carrying out the storing data process.

A. The application of cloud computing in the ship's organization

Recognizing the positive aspects of cloud computing, shipping companies such as Gardline have been increasingly implementing and using some or all of the services in cloud computing. A simplified block schematic representation of the use of cloud computing is given in Figure 4.

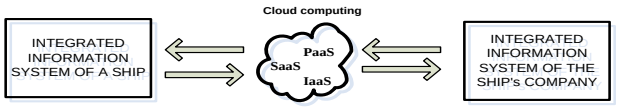


Figure 4. Use of cloud computing in a shipping company

By using cloud computing services the costs for IT management are reduced, regardless of the size of the shipping company or the growth rate. The implementation of cloud computing uses standard IT resources that are in the cloud. Secure and reliable internet connection is required for the end user.

IT companies closely associated with the maritime industry offer software applications for using services from the cloud or software applications for the management of the shipping companies and their fleets. For example, MASPAS AG company from Switzerland has produced technical software for fleet management, software as a service (Technical fleet management software as Software as a Service) [8, 9]. The name of the application is Maspas R5. It is a software application that covers the major functions in terms of the technical management of a fleet of ships including the management of assets, procurement and crew. The principle of operation of the application Maspas R5 is shown in Figure 5.

The staff in the company have access to the central database in real time via a secure internet connection no matter where they are located. Also, the crew members have access to the central database in the offline mode.

The database contains all details regarding the fleet and it is located on a central server in the cloud. Regular synchronization of data and information enables the crew and office staff to operate and use the same data and information. The flow of data between the company's offices, ships and suppliers takes place over Maspas R5.

The crew and staff of shipping companies have access to all data regarding the suppliers that are located in the central database.

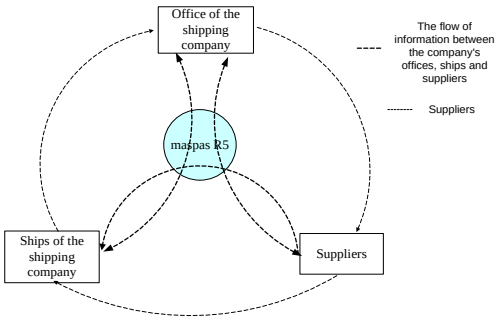


Figure 5. The operation principle of Maspas R5 [10]

Maspas R5 application contains the information system - Planned Maintenance System - PMS, which provides preventive and corrective maintenance. During ship's operation any equipment malfunction is reported via standardized interfaces (*Defect Report*). PMS includes a module for ordering and control of equipment and spare parts that are in the hold. In this way the whole process from the ordering to the receipt of the required equipment or spare parts can be monitored.

The benefits that Maspas R5 application provides include: lower operating costs (no need to purchase, manage and maintain IT resources, automatic updating of hardware and software and an improved quality of maintenance), transparency (overview in real time), improved efficiency of execution of marine processes (clear and precise instructions to the crew of the ship), fast implementation and updating of software with a small amount of data (always working with the latest version of the software), central data repository (automatic data synchronization ship-shore), accessibility to all relevant people and simple scalability (no limit growth) [8, 9].

Minimum requirements for using Maspas R5 [9]:

- Onboard the ship – ship-shore connection and server (industrial design) which is connected to the ship's computer network
- In the office of the shipping company - internet connection (> 500 kbit/s)

Satellite communication or internet connections via satellite are used to establish a communication channel when the ship is in the area outside the range of radio stations. It is known that Inmarst (International Maritime Satellite Organization) is still a leader in the field of maritime satellite communications. The latest Inmarsat's breakthrough is Broadband Global Xpress (Inmarst GX). Inmarsat GX is a global network that provides data transfer at speeds of up to 50Mbps in the incoming direction, while the outgoing speeds are up to 5Mbps. Inmarsat GX provides secure, reliable connections in real time between the ship and the shipping company [14, 16].

VI. CONCLUSION

The system for secure data storage is an integral part of any integrated maritime information system and should be treated accordingly – i.e. seriously.

The companies should have a defined policy for secure data storage practice and observe and enforce the policy of secure data storage because data loss may represent a huge financial cost and may result in problems in the business process and in loss of reputation on the global maritime market.

Combining storage technologies can boost the performance of an integrated information system on board the ship and across the shipping company. It is expected that the increased application of information technology in marine systems will result in a wider use of NAS technology because of data storage requirements.

Cloud computing is a new information technology that will increasingly expand and find application in information systems, owing to the advanced technology and the speed of the connection to the internet (e.g. Inmarsat GX).

The implementation of storing data in the cloud requires lower initial costs and shorter time to achieve productivity compared to traditional ways of storing data. Storing data in the cloud will reduce costs, increase availability and reliability of computer resources and increase profits of maritime companies. The core concept of data storage in the cloud is that the customer pays only for the service of data storage in the cloud. Regardless of the advantages, users are still concerned about the security of their own data stored in the cloud. In the near future the security problems as well as ship-shore communication difficulties will be solved and the servers of cloud computing will enable the use of almost all IT services in the cloud.

REFERENCES

[1]http://security.foi.hr/wiki/index.php/Sigurnosno_spremanje_podataka
[2]Umrežena spremišta podataka, CCERT-PUBDOC-2009-07-271
[3]http://www.callidus.hr/online_backup.php
[4]Beard H., "Cloud Computing Best Practices for Managing and Measuring Processes for On-demand Computing, Applications and Data centers in the Cloud with SLAs", London, UK, 2008, p. 9
[5]NCERT, "Cloud computing", PUBDOC-2010-03-293
[6]http://www.vidi.hr/Non-Tech/Hrvatska/5-GB-besplatne-pohrane-za-sve-korisnike-Vipa
[7]http://www.vecernji.hr/kompanije-i-trzista/usluge-cloud
[8]MESPAS AG Mespas, "The technical management software based on cloud computing", Technical Management Software, Switzerland
[9]MESPAS AG Mespas, "System requirements", Technical Management Software, Switzerland
[10]S. Fiore, G. Aloisio, "Grid and Cloud Database Management", Springer-Verlag Berlin Heidelberg, 2011.
[11]http://www.pmf.unizg.hr/download/repository/Sigurnosna_pohra
[12]https://www.openfiler.com/
[13]http://www.oracle.com/us/products/servers-storage/043711.html
[14]http://www.inmarsat.com/service-collection/fleetbroadband/
[15]R. Buyya, J .Broberg, A. Goscinski, "Cloud Computing: Principles and Paradigms", John Wiley & Sons, Inc., 2011.
[16] S., Krile, "Elektroničke komunikacije u pomorstvu, Mobilne satelitske veze", University of Dubrovnik, 2004.

IMPLEMENTATION OF GREEN ICT FOR SUSTAINABLE ECONOMIC DEVELOPMENT

Marijana Vidas-Bubanja, PhD[1]
Belgrade Business School and Faculty for Trade and Banking, Alpha University, Belgrade
Serbia
e-mail: marijana.bubanja@bbs.edu.rs

Abstract - In the past few years, a new aspect of ICT use has emerged and become a focal point of attention in order to realize development goals respecting in the same time ecological standards and norms. This "green" potential of ICT is oriented to use of ICT in transforming activities to be more energy-efficient and, as a result, to reduce greenhouse gas emissions. There is much hope that ICT can be a major part of the solution in tackling climate change and related environmental challenges.

I. INTRODUCTION

Modern, global world based on the richnes of information and ever more oriented on information society and e-economy put sustainable development in the focus of its existance. Sustainable development must incorporate social and environmental considerations, respecting human and natural resources along with economic factors.

World Commission on Environment and Development defined the sustainable development as „development that meets the needs of the present, without compromising the ability of future generations to meet their own needs" [1]. It implies considering the desired quality of life and the attainable speed of social development, as well as the need to balance different social values. In accordance with theoretic explanations such dvelopment relies on three pillars: the economy and improved standard of living; social equity through the reduction of poverty and enhanced quality of life; and the environment enabling the preservation of natural resources for the future generations. Policy-makers are an important factor in establishing synergies and complementarities among the three pillars.

In 2008, the world was confronted with multiple crises (fuel, food and financial). As a response at negative effects of economic development on environment, but also at the financial crisis, the international community seeks solutions for creating a sustainable economy and society. In this context the concepts of "green economy", "green growth", „and green society" became more present on the international level and in the developed countries. The 'green' economy can be considered synonymous to a 'sustainable' economy. However, the Green Economy concept often carries a more distinctive meaning, one that focuses specifically on the fundamental changes that are required to ensure that economic systems are made more sustainable. That is, green economy is seen as a framework for restoring economic growth meanwhile responding to the climate change and other subjects of environmental sustainability [2].

One of the key roles in this framework has information-communication technologies (ICT) through creation of digital economy. Miller and Wilsdon [3] sustain that digital economy modifies the relation of human being with environment, changing the business models. They propose the concept of „sustainable digital economy" as a solution for the environmental issues and consider the possibility to use the creativity and dynamism of digital economy for the good of economy, environment and society. Green digital economy is based on sustainable business practices as opposed to those which function based on non-renewable resources and inefficient methods. It seeks to promote financial, business and other economic transactions without depending on or damaging the environment.

II. GREEN ICT POTENTIALS

As a "general purpose technology" or an "enabling technology" ICT are widely used in different domains, and have a profound effect on the entire economy. ICT use in business is not only becoming more pervasive, it is also progressing towards more advanced and sophisticated applications.

The role of ICT in economic development is in particular recognized for innovation, for competitiveness and for sustainability. The three domains are strongly

[1] This paper is the result of the research realized under the project 179038 –„ Modelling of the development and integration of Serbia into world trends in the light of the economic, social and political commotions", Ministry of science and education of the Rep. of Serbia.

interrelated. For example, if a company increases its energy efficiency ("sustainability"), it reduces costs and becomes more competitive. However, there can also be conflicts of interests between these domains: investments in innovative green technology may increase the environmental sustainability, but not pay off for the company, at least in the short term [4].

All three domains of ICT impact – sustainability, competitiveness and innovation – are relevant at different levels: at the micro level of the individual company, the meso level of the industry and the macro level of the economy as a whole. Because of the meso and macro effects, industrial and innovation policy in advanced economies has for many years been paying close attention to ICT.

In the past few years, a new aspect of ICT use has emerged and become a focal point of attention: the "green" potential of ICT, in particular the use of ICT to become more energy-efficient and, as a result, reduce greenhouse gas emissions. There is much hope that ICT can be a major part of the solution in tackling climate change and related environmental challenges [4].

Over the past 5–10 years a consensus has emerged that ICTs can support the development of the green economy in three principal ways [5]:

1. by decreasing *direct effects* on the environment of the production, distribution, operation and disposal of ICTs through improved energy and materials efficiency, increased use of renewable energy sources, reduced use of toxic materials and improved recycling and end-of life disposal of ICTs;
2. by increasing the *enabling effects* of ICTs on the development of the green economy through improvements in the efficiency of production, distribution and consumption of goods and services throughout the economy and society; by reducing demand for energy and materials through the whole or partial substitution of virtual products and services for their physical equivalents; and through the dematerialization of human activities and interactions;
3. By supporting *systemic effects* that result in the transformation of behavior, attitudes and values of individuals as citizens and consumers; economic and social structures; and governance processes.

The ICT industry, in partnership with other sectors, has a key role to play in helping to make society's impact visible and to demonstrate the demand for new ways of reducing that impact.

III. GREENING ICT - REDUCING THE ENERGY AND CARBON FOOTPRINTS OF ICT SECTOR

The use and application of information and communication technologies in economic and social life is rapidly increasing throughout the world. For example, it is estimated that there are now some 6 billion mobile phone users and over 2.4 billion Internet users. ICT products and systems are resource-intensive in manufacturing and distribution, consuming ever-greater amounts of energy while in use, and creating escalating volumes of solid and toxic waste. ICT products may also have negative effects on human and social health as they are produced, used, and discarded [6 and 7].

Statistical data show the following [6 and 7]:

a) Energy consumption

- It is estimated that the 1.5 billion computers in the world consume about 90.000 MW of electric power, which is about 10% of global consumption.
- Energy consumption for servers worldwide has doubled over the period from 2000 to 2005. Recent estimates proclaimed that worldwide Internet usage, via the necessary computers and servers, needs the equivalent of fourteen power stations and produces the same amount of carbon dioxide as the entire airline industry.
- Estimates confirm that total electricity consumption by ICT range from 3 percent to 10 percent in the United States and Europe.

b) Carbon emissions

- Global carbon emissions for ICT as an industry are estimated at 2 percent to 2.5 percent of world totals, but would be much higher (14 percent) if indirect energy use were included. IT now causes the release of as much carbon dioxide into the atmosphere as nearly 320 million cars. According to some calculations, worldwide IT now generates CO^2 emissions of about 600 million metric tons a year. And if the sector continues growing at the current rate, emissions in Germany alone, for example, are expected to increase by another 60% by the year 2020.
- In Europe - the use of ICT equipment in the delivery of services represents about 1.75% of carbon emissions and further 0.25% of carbon emissions come from the production of ICT and consumer electronic equipment.

The ICT industry has a major challenge to adopt and implement additional standards for increased energy efficiency of networks and services. In fact, the ICT sector needs to focus on **s**ystemic improvements and further reduce the energy consumption of its own processes (including operations, manufacturing, service delivery and the supply chain). If the sector were to adopt a more systematic approach to monitoring and measuring energy consumption at every step in a process, it could generate verifiable and comparable data, allowing it to identify opportunities for improvements and reduce energy required to deliver ICT product and services [8].

Concerning the carbon emission, the environmental footprint of ICT is large and growing rapidly. The 'carbon footprint' of an organization is the volume of greenhouse gas emissions it generates. It is calculated by assessing energy usage, business travel, and all components of the organization's operations that consume power or generate waste and by-products. An organization is 'carbon neutral' when there is balance between the amount of carbon released and the amount of carbon sequestered (for example, by planting trees). The ICT industry is in a unique position to demonstrate leadership in reducing its footprint, through *structural change* and *innovation* as well as by leading the way in identifying and creating efficient solutions for other socio-economic sectors to follow.

Good example is announcement made by the Microsoft Corporation at Rio+20 Conference. Microsoft promised to become carbon neutral in just a single year—by the start of its fiscal year beginning July 1, 2013. Registered with the UN Global Compact, an Action Network of commitment-makers from business and industry, and under the SE4ALL initiative, Microsoft's promise was to achieve zero net carbon missions (CO_2e) for it's all data centers, software development labs, offices, and employee air travel. Microsoft, one of the most reputable computer technology companies in the world

with almost 100,000 employees in more than 100 countries, generated more than 1.6 million metric tons of CO_2 in 2011. Its pledge to become carbon neutral is significant given the size of its global operations, but it is also an important demonstration of private sector leadership in the global drive to a low-carbon economy [9].

The ICT sector should be called upon to engage collectively in a process of self-improvement by agreeing on common methodologies and metrology tools to generate data about its energy performance, set realistic targets and benchmark progress. Such efforts should give due consideration to the full life cycle and related environmental impact. The ICT sector should lead by example and will be encouraged to commit such innovative practices that can be replicated in other industries.

TABLE I. ICT – VOLUNTARY ICT SECTOR COMMITMENTS TO DECREAS GHG EMISSIONS, AND RISE ENERGY EFFICIENCY

Companies	Target reduction %	Baseline *	Target date	Comment
Bell Canada	15	Not given	2012	GHG emissions
British Tele-Comm.Plc	80	1996	2020	CO 2 emissions per unit of contribution to GDP
Cisco Systems	25	2007	2012	GHG emissions
Dell	Additional 15	Not given	2012	Operational carbon intensity
Deutsche Telecom AG	20	2006	2020	CO 2 emissions
France Telecom	20	2006	2020	CO 2 emissions
IBM	7	2005	2012	GHG emissions
Intel	20 30	2007 2004	2012 2010	Carbon footprint GHG emissions per production unit
Motorola	6	2000	2010	CO 2 emissions
Nokia	6	2006	2012	Energy consumption of offices and sites
Sun Microsystems Inc.	20	2007	2015	GHG emissions
Vodafone Plc	50	2006/2007	2020	CO 2 emissions
European Union (all sectors)	**20** **20**	**1990** **Projected energy use in 2020**	**2020** **2020**	CO 2 emissions Energy savings/efficiency

Source: SMART 2020: Appendix 4, p.75

IV. ICT FOR MONITORING AND CONTROLING PROCESSES IN THE WHOLE ECONOMY

In order to realize "green sustainable economy" it is **crucial to encourage structural changes** aimed at realizing the potential of ICT to enable energy efficiency across the economy. 'Structural change' means re-engineering the way an organization operates. This can be done for instance by replacing products with on-line services ('dematerialization'), by moving business to the internet (e.g. costumer's support), by adopting new ways of working (tele-working and flexi-work enhanced by video-conferencing and tele-presence tools) and by exploring the viability of using green suppliers and energy from renewable resources.

ICTs can be an opportunity for a sustainable and inclusive future by revolutionizing business models and behavior, by making management of energy distribution and transportation more efficient, by enabling new ways of working, learning, e-health and providing public services, and by helping countries adapt to climate change and reducing disaster risk. Green ICT can improve business processes that are not directly involved in ICT, and dramatically lower CO^2 emissions throughout the process chain helping in this way to significantly reduce the other 98 % of all CO^2 emissions not caused by ICT.

With ICTs now an intrinsic element of every economic sector, the need to integrate ICT-based solutions into the environmental actions of each industry sector has become essential.

Company-level - On the company level it is increasingly important to use ICT solutions to support business processes. These solutions can streamline processes and have enormous potential for reducing CO^2 emissions, because they improve the utilization of resources and allow them to be shared. ICT also permits in-depth monitoring of energy consumption and CO^2 emissions throughout the value chain, so processes and organizational structures can be enhanced accordingly.

Fore modern companies "Green IT" is what analysts, manufacturers and providers call all IT solutions that save energy at business organizations. These include hardware, software and services [6].

Where hardware is concerned, energy-efficient desktop PCs, thin-client architectures and data-center hardware offer answers, and so do energy supply and cooling systems.

In the software and service area, there is significant potential in virtualization, in solutions for dynamic capacity management and data-center planning, and in storage-system off shoring.

And with the increasing convergence of IT and telecommunications, it possible to take the Green concept a step further. Green ICT extends to other, less commonly considered aspects of the entire life cycle. This includes eco-friendly procurement, employee behavior, running data centers on sustainable generated energy, environmentally sound disposal of used electrical equipment, and as much recycling as possible.

The most obvious benefits of Green ICT for companies are reduced environmental impact and cost savings. Besides cost reduction a comprehensive Green ICT strategy offers many other benefits – appealing to all stakeholder groups: employee satisfaction increases, the company's standing with capital markets and with society as a whole improves, and the business can attract new customer groups.

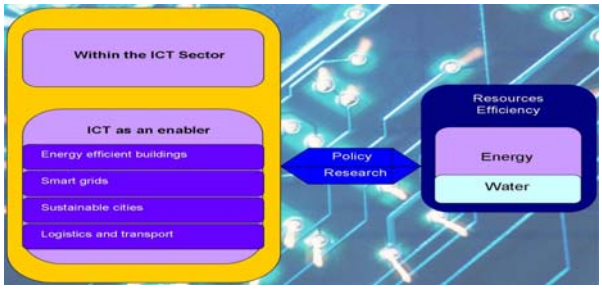
A survey carried out by Forrester in 2009 to identify the three most important reasons for Green IT activities revealed that the principal motivation was to cut spending on energy (69%). The second was to cut down on IT costs (40%) and the third to improve image. The fourth

reason cited was legal compliance (15%). It is remarkable that cost considerations play an above-average role in the United States, while image and legal compliance are more important in Europe [11].

Industry level - The potential savings are particularly great in industries with extensive, complex supply chains. Examples are involving automotive components, and the development and manufacture of cars. ICT can also have a major impact on supporting processes, i.e. via resource utilization in offices. In EU special focus on the industry level is given to buildings and construction, and to transport and logistics [8].

Buildings account for approximately 40% of energy end-use in the EU, of which more than 50% is electrical power. The sector has significant untapped potential for cost-effective energy savings which, if realized, would mean an 11% reduction in total energy consumption in the EU by 2020. There is scope for ICT to contribute to energy and environmental performance, through the application of building and energy management systems, smart metering technologies, solid-state lighting and lighting control systems, intelligent sensors and optimization software.

Transport systems represent about 26% of energy end-use in the EU. Many opportunities exist for improvements in energy efficiency and rationalization, notably through logistics. The focus is on the deployment of ICT to promote modal shift, notably on transport corridors for freight, and through the provision of multi-modal journey planners for passengers to enable significant reductions in congestion.



Picture 1. Resource Efficiency through ICT

Country level - Studies in various countries, generally suggest broad potential for ICT to cut energy use and carbon emissions. One U.S. study projects that changes enabled by broadband networks could save the equivalent of 11 percent of U.S. oil imports through transportation substitution and "dematerialization." A recent study supported by the U.S. Department of Energy demonstrated significant results from "Smart Grid" technologies to monitor and adjust home energy consumption, reducing average household consumption by 10 percent. Proactive management of home energy use by consumers in this way could reduce peak loads on utility grids up to 15 percent annually. The UK RAC Foundation has estimated that "through the use of ICT, there is potential within 10 years for a 15 percent reduction in commuter travel, 18 percent reduction in heavy freight, and a 10 percent reduction in shopping journeys by car" [7].

Partnerships between the ICT sector and other industry sectors are encouraged to identify where and how ICTs can play a role in improving efficiency and reducing emissions in those sectors and thus accelerate the delivery of tools to assess and to optimize energy performance on a comparable basis.

V. ICT IN ENERGY PRODUCTION AND DISTRIBUTION

There are two aspects of the links between energy and ICT [12]. On the one hand, ICT plays a daily part in the production and use of electricity, from the operation of power plants to electricity transport. ICT is used to model, simulate and optimize these processes in order to improve energy efficiency and reduce costs. On the other hand, ICT itself uses a lot of electricity, and work has begun on reducing the consequent economic and ecological costs.

The Information and Communication Technologies are an enabling factor in developing innovation in the whole energy sector. The main initiatives implemented by countries to reach energy efficiency targets are: increased efficiency in generation plants and reduction of network losses (technical and non technical losses). Also challenges deriving from the integration in the network of: large scale renewable sources, electric vehicles, systems for multidirectional flows management, new equipments and power electronics and real time information devices can be faced only with a strong, reliable and fast ICT infrastructure.

The energy transformation sector, dominated by electricity generation, uses around one-third of all primary energy. Given the potential for improvement in electricity generation (estimated at 30-40%) and the considerable losses in transport (2%) and distribution (8%), it is critical to improve transformation efficiency, address losses and identify any potential problems before they compromise supply. ICTs have a major role to play not only in reducing losses and increasing efficiency but also in managing and controlling the whole processes in energy production to ensure stability and reinforce security as well as in supporting the establishment of a well functioning electricity retail market [13].

Smart metering is an activity which introduces ICT potentials in energy sector. Smart metering, introduces ICTs into traditional energy-metering, can be deployed in many ways, taking advantage of ICTs to a greater or a lesser extent. Information can flow between the meter and the energy provider and between the meter and the consumer. The so-called "one-way" implementation, that only serves the energy supplier, is often favored over the 'two-way' implementation which serves both supplier and consumer.

According to the French regulator CRE, with the implementation of smart metering, they estimate an increase in supplier switch capability by factor of 10, decrease residential consumption by 5% and decrease CO2 emissions by 5% [13].

Smart metering is just a first step on the path to smart electricity grids. Smart Distribution network solutions embrace the changing structure of generation, market and use of electricity, and improve the efficiency, reliability, flexibility, accessibility and cost-effectiveness of the end to-end system. Ultimately, smart grids should facilitate, not just a better management of energy consumption, but the integration of alternative and renewable energy sources on a far greater scale than is possible today, with positive impacts for energy security and for the environment. The

liberalization of market for energy, the multiplication of local energy networks, the integration of renewable energy sources (RES), the spread of co- and micro-generation (micro-grids, virtual power plants), and new user demands require the use of the most advanced technologies for monitoring and control as well as for electronic trading of electricity enabled by the smart grids implementation [8].

The grid becomes "smart" because it will not only transport electricity but also information that will become an active part of the electricity supply system. This shall be considered as a paradigm shift in the way the electricity distribution grids are today and how they will become in the future: user and customer centric, service oriented, accommodating all needs and providing adequate solutions, supporting the migration towards and shaping of the low-carbon economy and society.

The energy efficiency targets can only be achieved with the contribution from many different sources, coming from all the energy system stakeholders, mainly customers, utilities, manufacturers, industries, governmental actors and regulators. In fact, the road towards a society without energy waste can be made easier thanks to the availability and the management of large amount of information on energy consumption for all the actors.

VI. GREEN ICT IN EU

Energy security and climate change are becoming more and more prominent on political agendas worldwide and across all sectors of the economy, so that it ranks highest among the EU's priorities. The resolve of the European Council to transform Europe into a low-carbon, high energy efficiency economy means that the continued growth of the European economy, essential to achieve full employment and inclusion, needs to be decoupled from energy consumption. The current trends are unsustainable. Indeed, if nothing were to change, final energy consumption in the EU is predicted to increase up to 25% by 2012, with a substantial rise in greenhouse gas emissions [14].

At the 2007 Spring European Council, the Heads of State and Government highlighted the development of a sustainable integrated European climate and energy policy as a top priority and adopted an energy and climate package to guide the EU towards a competitive and secure energy economy while promoting energy savings and climate-friendly energy sources. That is, Europe has set itself ambitious targets for 2020: saving 20% of primary energy consumption, reducing greenhouse gas emissions by 20% and raising the share of renewable energy to 20%. Fulfillment of these targets faces Europe following main challenges — tackling climate change, guaranteeing secure, sustainable and competitive energy, and making the European economy a model for sustainable development in the 21st century.

Information and Communication Technologies have an important role to play in reducing the energy intensity and increasing the energy efficiency of the European economy, in other words, in reducing emissions and contributing to sustainable growth. In order to achieve the ambitious targets set and meet the challenges ahead, Europe needs to ensure that ICT-enabled solutions are available and fully deployed.

The European Union policy answer to above mentioned challenges can be described as some strong combination of the Knowledge Economy and a Green New Deal. It means putting together economic recovery, by the creation of new ICT markets that address the environmental crisis, with the need to move up the value scale, to a knowledge economy. Essentially the vision is of a *Green Knowledge Society* – a Europe in which citizens and enterprises are empowered through ICT in an inclusive, innovative, secure and sustainable knowledge society [15].

In order to create the *Green* Knowledge Society the following policy issues should be included in an ICT policy agenda for Europe to 2015 [14]:

- A stimulation programme for application of ICTs in energy saving roles across all relevant industry sectors;
- Ensuring ICT is used in more sustainable behavior patterns by citizens and business;
- EU production of a next generation of green ICTs with new technologies and usage patterns – as an opportunity for Europe.

VII. GREEN ICT IN SERBIA

Serbian economy consumes 2.5 times more energy per unit of GDP than the world average and four times more than the average in the EU and can be considered as energy inefficient country. Negative statistics data further confirm need for economic structural change and transition to more energy efficient and environmental responsible economy [16]:

1. Serbian households spend 2.5 times more energy per square meter of living space from some EU countries;
2. Public lighting consume 50 percent more energy than the EU countries;
3. Serbia is dependent on energy imports - even 33.6 percent of the whole import;
4. Emissions of greenhouse gases from the energy sector in Serbia amounts to 76 % of total emissions.

Investing in the use of modern technology and IT solutions is one of the ways to increase energy efficiency and competitiveness of Serbian industry and economy as the whole. Some calculations confirm that the use of advanced ICTs in Serbian enterprises would provide savings of up to 40 percent of the energy. Current data confirm that for the most companies in Serbia energy spending usually accounts up to 30 percent of the operating costs.

Implementation of the advanced IC technologies for the increase of energy efficiency and renewable use, would directly contribute to better position and greater competitiveness of domestic companies, communities and the entire country.

Improving energy efficiency is also one of the key initiatives of the energy policy of the Republic of Serbia. New Law on efficient energy use in Republic of Serbia was adopted by the Serbian Parliament on 15 March 2013. This law establishes a system of organized energy management (energy management) for large energy consumers and the public sector (responsible for about 70 % of primary energy consumption). This Act also introduces minimum energy efficiency requirements for new and revitalized plant in the manufacturing, transmission and distribution, introduce

charges to metered energy consumption and labeling of products that affect energy consumption, and introduce eco-design requirements. The final aim is to open the market for energy services and to implement energy audits in order to identify potential energy efficiency improvements [16].

This Act is also in line with the international obligations of Serbia to save at least 9 % of gross final energy consumption to 2018, and to reduce carbon dioxide emissions by about 9 % to 2030 [17].

VIII. CONCLUSION

Nowdays world is facing multiple crisis and challenges that are mutually interlinked. We have unique opportunity to rethink current perception of growth consumption, inclusion, and how we utilize our limited resources. The transition to a resource-efficient economy and society requires new levels of innovation, breaking out of locked-in systems and technologies. It will be led by forward-looking countries, companies, researchers, inventors and investors.

Such innovations and solutions can be realized only by increasingly closer cooperation among all stakeholders with the aim of unlocking the potential of ICTs to improve energy efficiency, thereby promoting the competitiveness of industry, creating a wealth of opportunities, jobs and services, and building an everyone-wins momentum for industry, users, and society at large.

The private sector can and should have a key role, but it needs the right framework conditions and well-functioning markets. Good regulation and its effective implementation are there to provide a level playing-field and the right incentives and rewards.

Only in such well defined framework the role of ICTs can be fully explored and exploited in the following respects: a) as an enabler of energy efficiency across the economy including fostering the change in citizen's behavior, b) in improving efficiency in the use of natural resources while reducing pollution and dangerous waste.

To put ICTs at the core of the energy efficiency effort and to enable them to reach their full potential, it is necessary to foster research into novel ICT-based solutions and strengthen their take-up — so that the energy intensity of the economy can be further reduced by adding intelligence to components, equipment and services. On the other side, efforts should be made so that ICT leads by example and reduces the energy it uses**.**

The transformation to a Green Economy asks not just business (including producers and retailers), but also consumers, to change behavior. Consumers have the right to be informed about products and they are increasingly asking for more information. However, recent surveys show some worrying trends: half of EU citizens do not trust environmental claims, while 1/3 of them encountered misleading environmental claims. Consumers need to be better informed on how they can contribute to reduce their footprint on the planet and its resources [13].

Increasing social awareness through policies and communications to explain the urgency of the issue and urge conservation is also needed.

The above efforts require that quantifiable benchmarks and metrics be developed to gauge their effectiveness and ensure progress towards established goals.

Final conclusion is that ICT products, systems, and networks are the essential driver of productivity improvements and innovation for the 21st century. They will be the enabler of sustainability solutions in all networks: energy production and use, mobility, water and sewage, open spaces, education, and public health and safety. ICT innovation is also the catalyst for changes in personal, work, and community life that will be a fundamental requirement for sustainable economic development.

REFERENCES

[1] WCED, *Our common future*: The report of the World Commission on Environment and Development, 1987, Retrieved June, 2010, from http://www.un-documents.net/wced-ocf.htm.

[2] Ciocoiu, C.N., Integrating Digital Economy and Green Economy: Opportunities for Sustainable Development, *Theoretical and Emiprical Researches in Urban Managemen*t, Volume 6/ Issue1, February, 2011, pg. 33-43.

[3] Miller, P. and Wilsdon, J., Digital futures: An agenda for sustainable digital economy. *Corporate Environmental Strategy,* 8 (3), 2001, pp. 275-280.

[4] European Commission, ICT and E-Business for an Innovative and Sustainable Economy, 7th Synthesis Report of the Sectoral e-Business Wacth, Brussels, 2010.

[5] IISD, *The Digital Economy and the Green Economy: Opportunities for strategic synergies*, International Institute for Sustainable Development, 2010. Retrieved October, 2010, from: http://www.iisd.org/pdf/2010/com_digital_economy.pdf.

[6] T-Systems, *White Paper Green ICT*. The Greening of Business, Germany, October, 2009.

[7] Cisco Internet Business Solutions Group, *Green ICT: Sustainability for the 21st Century*, Written specially for Connected Development Conference 2008, Connected Urban Developmen, Discussion Brief, 2008

[8] European Commission, Communication on Mobilising Information and Communication Technologies to faciliate the transition to an energy-efficinet, low-carbon economy, Brussels, COM (2009) 111 final, March, 2009.

[9] NRDC (Natural Resource Defence Council), Fulfilling the Rio+20 Promises: Reviewing progress since the UN Conference on Sustainable Developemt, September, Report, 2012..

[10] SMART 2020: *'Enabling the low carbon economy in the information age*

[11] ITU, Goverment of Canada and Prompt, *Why Green ICT: Challange or Opportunity?*, 7th ITU Symposium on ICTs, the environmnet and climate change, Montreal, Canada, 29-31 May 2012.

[12] European Research Consortium for Informatis and Mathematics (2009), „Special Theme: Towardes Green ICT“, in *ERCTIM News*, No.79, October.

[13] European Commission, *ICT for Low Carbon Economy Smart Electricity Distribution Networks*, Brussels, July, 2009.

[14] European Commission, Addressing the challenge of energy efficiency through Information and Communication Technologies, Brussels, COM (2008) 241 final, May, 2008.

[15] Forge, S., Blackman, C., Bohlin, E. and Cave, M., A Green Knowledge Society. *An ICT policy agenda to 2015 for Europe's future knowledge society*, A study for the Ministry of Enterprise, Energy and Communications, Government Offices of Sweden, published by SCF Associates Ltd, September, 2009.

[16] http://www.blic.rs/Vesti/Drustvo/365685/Samit-menadzera-Srbija-izuzetno-energetski-neefikasna, feb 2013

[17] Vidas-Bubanja, M., “Značaj primene ICT za povećanje energetske efikasnosti i zaštitu životne sredine”, magazine *Ecologica*, No 72, Belgrade, 2013, pg. 725-730.

SP

International Conference on

MIPRO Junior - STUDENT PAPERS

Steering Committee

Chairs: Goran Cetušić, KSET – University of Zagreb, Croatia
Andrej Sokolić, KSET – University of Zagreb, Croatia

PAPERS

Complex Networks Measures for Differentiation between Normal and Shuffled Croatian Texts

Domagoj Margan, Ana Meštrović, Sanda Martinčić-Ipšić
Department of Informatics,
University of Rijeka,
Radmile Matejčić 2, 51000 Rijeka, Croatia
Email: {dmargan, amestrovic, smarti}@uniri.hr

***Abstract*—This paper studies the properties of the Croatian texts via complex networks. We present network properties of normal and shuffled Croatian texts for different shuffling principles: on the sentence level and on the text level. In both experiments we preserved the vocabulary size, word and sentence frequency distributions. Additionally, in the first shuffling approach we preserved the sentence structure of the text and the number of words per sentence. Obtained results showed that degree rank distributions exhibit no substantial deviation in shuffled networks, and strength rank distributions are preserved due to the same word frequencies. Therefore, standard approach to study the structure of linguistic co-occurrence networks showed no clear difference among the topologies of normal and shuffled texts. Finally, we showed that the in- and out- selectivity values from shuffled texts are constantly below selectivity values calculated from normal texts. Our results corroborate that the node selectivity measure can capture structural differences between original and shuffled Croatian texts.**

I. Introduction

The complex networks sub-discipline tasked with the analysis of language has been recently associated with the term of linguistic's network analysis. The linguistic network can be based on various language constraints: structure, semantics, syntax dependencies, etc.

In the linguistic co-occurrence complex networks properties are derived from the word order in texts. The open question is how the word order itself is reflected in topological properties of the linguistic network. One approach to address this question is to compare networks constructed from normal texts with the networks from randomized or shuffled texts. Since the majority of linguistic network studies have been performed for English, it is important to test whether the same properties hold for Croatian language as well. So far, there have been only sporadic efforts to model the phenomena of the Croatian language through complex networks [7] [11].

In this paper we address the problem of Croatian text complexity by constructing the linguistic co-occurrence networks form: a) normal texts, b) sentence-level shuffled texts, and c) text-level shuffled texts. This work extends our previous research [11] with additional sentence-level shuffling procedure and by introducing a node selectivity as a new complex network measure. Our experiment tests whether selectivity can differentiate between normal and meaningless texts.

Section 2 presents an overview of related work on complex network analysis of randomized texts. In Section 3 we define measures for the network structure analysis. In Section 4 we present shuffling procedures and the construction of co-occurrence networks. The network measures are in Section 5. In the final Section, we elaborate the obtained results and make conclusions regarding future work.

II. Related work

Some of the early work related to the analysis of random texts dates to 1992, when Li [5] showed that the distribution of word frequencies for randomly generated texts is very similar to Zipf's law observed in natural languages such as in English. Thus, the feature of being a scale-free network does not depend on the syntactic structure of the language. Watts and Strogatz [10] showed that the network formed by the same amount of nodes and links but only establishing links by choosing pairs of nodes at random has a similar small network distance measures as in the original one. Caldeira *et al.* [2] analyzed the role played by the word frequency and sentence length distributions to the undirected co-occurrence network structure based on shuffling. Each sentence is added to the network as a complete subgraph. Shuffling procedures were conducted either on the texts or on the links. Liu and Hu [6] discussed whether syntax plays a role in the complexity measures of a linguistic network. They built up two random linguistic networks based on syntax dependencies and compared the complexity of non-syntactic and syntactic language networks. Krishna *et al.* [4] studied the effect of linguistic constraints on the large scale organization of language. They described the properties of linguistic co-occurrence networks with the randomized words. These properties were compared to those obtained for a network built over the original text. It is observed that the networks from randomized texts also exhibit small-world and scale-free characteristics. Masucci and Rodgers showed [8] [9] that the power law distribution holds when they shuffled the words in the text. Thus, they showed that degree distribution is not the best measure of the self-organizing nature of weighted linguistic networks. Due to the equivalence between frequency and strength of a node, shuffled texts obtain the same degree distribution, but lose all the syntactic structure. They have analyzed the differences between the statistical properties of a real and a shuffled weighted network and showed that the scale-free degree distribution and the scale-free weight distribution are induced by the scale-free strength distribution. They proposed a measure, the node selectivity, that can distinguish a real network from a shuffled network. Selectivity is defined as the average weight distribution on the links of the single node.

Preliminary results of our research on Croatian co-occurrence networks presented in [7] point out that the increase of the co-occurrence window size is followed by a decrease in diameter, average path shortening and, expectedly, the condensing of the average clustering coefficient. The stopwords removal causes the same effect. When comparing Croatian literature networks to networks from other languages such as English and Italian [1] some expected universalities such as small-world properties are shown, but there are still some differences. The Croatian language exhibits a higher path length than English and Italian language which can be caused by the mostly free word order nature of Croatian.

Initial attempt [11] to analyse network properties of normal and shuffled Croatian texts show that the text shuffling causes the decrease of the network diameter, due to the establishment of previously non-existing links. Furthermore, the results indicate a slight difference in the average clustering coefficient which is higher for the networks based on the shuffled text. Also, obtained results showed that node degree distributions are preserved in text-level shuffled networks, due to the same word frequencies (e.g. Zipf's law).

III. The network structure analysis

In the network, N is the number of nodes and K is the number of links. In weighted networks every link connecting two nodes has an associated weight $w \in R_0^+$. The co-occurrence window m_n of size n is defined as n subsequent words from a text. The number of network components is denoted by ω.

For every two connected nodes i and j the number of links lying on the shortest path between them is denoted as d_{ij}, therefore the average distance of a node i from all other nodes is:

$$d_i = \frac{\sum_j d_{ij}}{N}. \tag{1}$$

And the average path length between every two nodes i, j is:

$$L = \sum_{i,j} \frac{d_{ij}}{N(N-1)}. \tag{2}$$

The maximum distance results in the network diameter:

$$D = max_i d_i. \tag{3}$$

For weighted networks the clustering coefficient of a node i is defined as the geometric average of the subgraph link weights:

$$c_i = \frac{1}{k_i(k_i-1)} \sum_{ij} (\hat{w}_{ij}\hat{w}_{ik}\hat{w}_{jk})^{1/3}, \tag{4}$$

where k_i is the degree of the node i, and the link weights $\hat{w}_{ij}$ are normalized by the maximum weight in the network $\hat{w}_{ij} = w_{ij}/\max(w)$. The value of c_i is assigned to 0 if $k_i < 2$.

The average clustering of a network is defined as the average value of the clustering coefficients of all nodes in a network:

$$C = \frac{1}{N} \sum_i c_i. \tag{5}$$

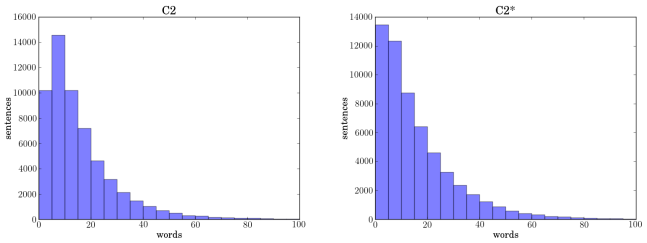


Fig. 1. Histograms of sentence length frequencies for C2 and C2*

If $\omega > 1$, C is computed for the largest network component.

The out-degree and in-degree $k_i^{out/in}$ of node i is defined as the number of its out and in nearest neighbors.

The out-strength and the in-strength $s_i^{out/in}$ of the node i is defined as the number of its outgoing and incoming links, that is:

$$s_i^{out/in} = \sum_j w_{ij/ji}. \tag{6}$$

We then define for the node i the out and in selectivity as

$$e_i^{out/in} = \frac{s_i^{out/in}}{k_i^{out/in}}. \tag{7}$$

IV. Methodology

A. Data

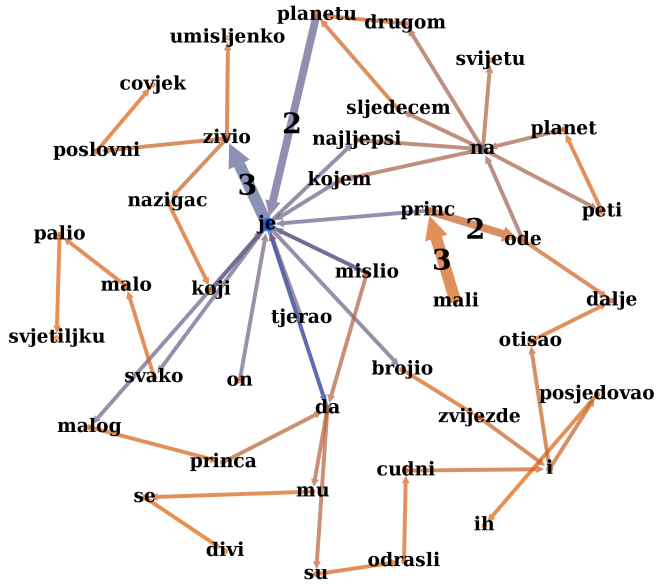
For the construction and analysis of co-occurrence networks, we used the corpora of 10 books written in or translated into the Croatian language. We divided the corpora into two parts: the first - C1 includes one book, and the second - C2 includes all books. The C1 contains: 191941 words (27453 unique) in 17045 sentences; and C2: 888293 words (91420 unique) in 57179 sentences.

B. The shuffling procedure

Commonly, the shuffling procedure randomizes the words in the text, transforming the text into the meaningless form. We shuffled the C1 and C2 corpora in two ways: a) shuffling on the sentence level, and b) shuffling on the text level. In both ways we preserved the vocabulary size, the word and sentence frequency distributions.

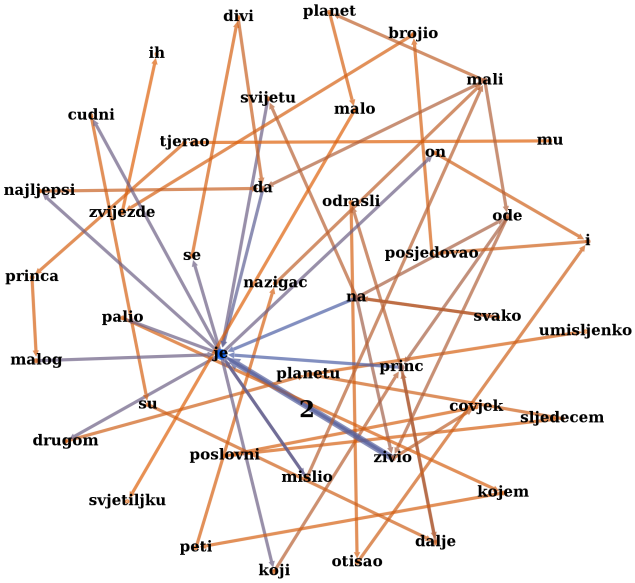
In the sentence-level shuffling we preserved the sentence length (the number of words per sentence) and sentence order. Versions of C1 and C2 shuffled on the sentence level are noted as C1' and C2'. In the text-level shuffling, the original text is randomized by shuffling the words and punctuation marks over the whole text. This approach preserves the number of sentences but changes the number of words per sentence. Versions of C1 and C2 shuffled on the text level are noted as C1* and C2*. Fig. 1 shows the histograms of sentence length frequencies for C2 and C2* corpora.

Our shuffling procedures destroy the sentence and text organization in a way that the standard word-order and syntax of the text is eradicated. As expected, the typical word collocations and phrases of Croatian language are completely lost,



mali princ je mislio da su odrasli cudni i otisao dalje. na drugom planetu je zivio umisljenko. mislio je da je najljepsi na svijetu. tjerao je malog princa da mu se divi. mali princ ode dalje. na sljedecem planetu je zivio poslovni covjek. on je brojio zvijezde i posjedovao ih. mali princ ode na peti planet na kojem je zivio nazigac koji je svako malo palio svjetiljku.

Fig. 2. Example network constructed from original text



mislio mali da je cudni su dalje princ odrasli otisao i. na zivio je drugom planetu umisljenko. na svijetu je mislio je najljepsi da. mu tjerao princa malog je se divi da. mali ode princ dalje. na je zivio covjek poslovni sljedecem planetu. je on i posjedovao brojio zvijezde ih. na svako na ode zivio je koji princ je palio kojem peti nazigac mali planet malo svjetiljku.

Fig. 3. Example network constructed from sentence-level shuffled text

as well as the forms of the morphological structures and local structures of words' neighborhood.

C. The construction of co-occurrence networks

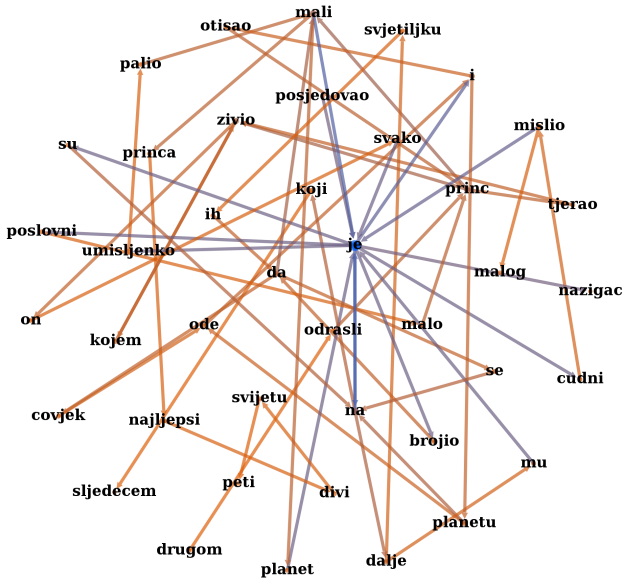
Text can be represented as a complex network of linked words: each individual word is a node and interactions amongst words are links. We constructed six different co-occurrence networks (C1, C1', C1*, C2, C2', C2*) all weighted and directed. Words are nodes linked if they are co-occurring as neighbors to each other in a sentence. The weight of the link is proportional to the overall co-occurrence frequencies of the corresponding word pairs within a corpus.

Network construction and analysis was implemented with the Python programming language using the NetworkX software package developed for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks [3].

Figures 2, 3 and 4 show simple examples of networks constructed from original and shuffled text samples. Networks are directed and weighted; the numbers on the networks' edges represent weights larger than 1. At the bottom of each figure is the text from which the visualized network was constructed. The original sample text (Fig. 2) was shuffled on sentence (Fig. 3) and text (Fig. 4) level in the same way as the corpora C1 and C2.

V. Results

The measures of the networks constructed from C2 corpus, sentence-level shuffled C2', and text-level shuffled C2* are presented in Table 1. The results show that the shuffled networks have decreased values of the average path length L



dalje svjetiljku ih da se na. posjedovao je nazigac. planetu ode i otisao princ tjerao zivio kojem zivio on svako je je su na je na dalje mu je umisljenko palio mali planet je brojio da mali je i planetu na. je cudni mislio je poslovni malo princ mali princa najljepsi divi svijetu peti. zvijezde. mislio malog. da covjek ode. na koji sljedecem. drugom odrasli princ zivio

Fig. 4. Example network constructed from text-level shuffled text

and network diameter D, and have the value of the average clustering coefficient C increased.

In the networks constructed from the corpora shuffled on the text-level the number of nodes N ($N_{C1*} < N_{C1}, N_{C2*} < N_{C2}$) is smaller than the number of words in the original corpora. During the network construction process, the sen-

TABLE I. NETWORK MEASURES FOR ORIGINAL C2, SENTENCE-LEVEL SHUFFLED C2' AND TEXT-LEVEL SHUFFLED C2*

	C2	C2'	C2*
N	91328	91328	91204
K	465196	586110	599335
L	3.097	3.037	2.997
D	23	17	10
C	0.317	0.343	0.354
ω	22	22	10

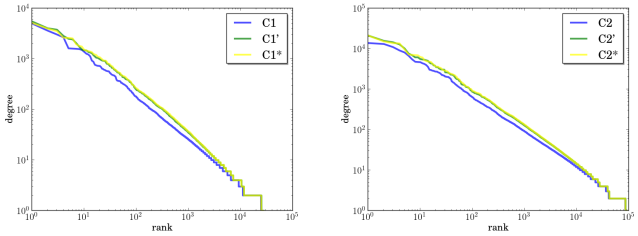


Fig. 5. Degree rank distributions for the C1, C1', C1*, C2, C2', C2*

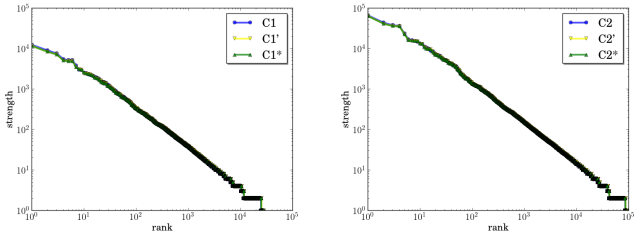


Fig. 6. Strength rank distributions for the C1, C1', C1*, C2, C2', C2*

tences containing just one word are disregarded, because our approach limits the word linkage to the sentence delimiters. This causes sentences with exactly one word to be isolated from the network, which reduces the number of nodes N in C1* and C2*.

One of the standard approaches to examine properties of co-occurrence networks is node degree distribution for unweighted and node strength distribution for weighted. Shuffling preserves the degree distribution due to the word frequencies [9] [11]. In this work we use rank distribution to compare different texts. Rank and frequency distributions are related through Zipf's law: the word frequency is inversely proportional to it's rank.

Initially, we computed the degree and strength rank distributions only for C2 and corresponding shuffled versions. As expected, the degree rank distributions, as well as the strength rank distributions are preserved during the shuffling procedure. In order to explore whether the same holds for substantially smaller corpus we checked degree and strength rank distributions for C1 as well. Fig. 5 and Fig. 6 present degree and strength rank distributions for the C1, C1', C1*, C2, C2' and C2*. Degree rank distributions exhibit no substantial deviation in shuffled networks, and strength distribution is preserved due to the same word frequencies. At this point all network measures and distributions showed no substantial differences between the structure of the original and meaningless texts.

The global network measures: average shortest path length, diameter, clustering coefficient and degree and strength distribution may not be well-suited to discriminate between original and meaningless texts. Therefore, it is necessary to include local (node level) network measures. Motivated by the reported results form [9], which introduced the local measure of node selectivity, we applied the same principle on our data. In weighted and directed co-occurrence networks nodes have ingoing and outgoing links, therefore it is necessary to consider the in- and out- selectivity.

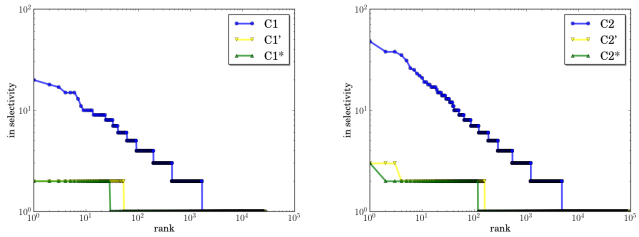


Fig. 7. In-selectivity rank distributions for the C1, C1', C1*, C2, C2', C2*

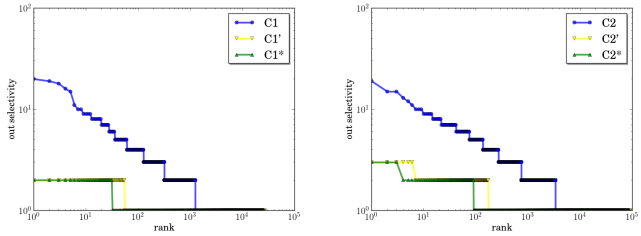


Fig. 8. Out-selectivity rank distributions for the C1, C1', C1*, C2, C2', C2*

Fig. 7 and Fig. 8 present obtained in- and out- selectivity ranks for C1 and C2, and their shuffled counterparts. Regardless of the links direction (in or out), selectivity values of networks from the shuffled texts are constantly below selectivity values calculated from the normal texts. This can be explained by the well known properties found in text - collocations. Collocations are typical word pairs, triples,...etc. which are recognized as standard phrases or names (e.g. New York). Shuffling procedures rearranged the word order, which disassembled collocations. This effect is reflected in the selectivity, since the high weights of common phrases and collocations are lost, which flattens and abates the selectivity ranks.

VI. CONCLUSION

We studied the structure of the linguistic networks constructed from normal and shuffled texts in Croatian. As expected, the text shuffling causes the decrease of the average path length L and the network diameter D, and the increase of the average clustering coefficient C.

We showed that the degree rank distributions exhibit no substantial deviation in shuffled networks, and the strength rank distributions are preserved due to the same word frequencies. The standard approach to study the structure of linguistic co-occurrence networks showed no substantial differences among the topologies of the original and shuffled texts. Therefore we utilized the node selectivity measure proposed by Masucci and Rodgers in [9].

Our results showed that the in- and out- selectivity values from shuffled texts are constantly below selectivity values calculated from normal texts. It seems that selectivity captures typical word phrases and collocations in Croatian which are lost during the shuffling procedure. The same holds for English where Masucci and Rodgers found that selectivity somehow captures the specialized local structures in nodes' neighborhood and forms of the morphological structures in text. Based on this findings, the measure of selectivity can be useful to discriminate between different text types, which will be the part of our future work.

We have shown that the Croatian language networks have similar properties as language networks from English and other languages. Firstly, Croatian text shuffling has no influence on the degree and strength distributions, which has already been shown for English [8], [9], English and Portuguese [2] and English, French, Spanish and Chinese [4]. Furthermore, distance measures (average shortest path length and diameter) show that networks based on normal texts have a greater L and D value than the corresponding network based on shuffled text. The same relations for average clustering coefficient, average shortest path length and diameter are shown in [4] for all studied languages (English, French, Spanish and Chinese). Similar results are shown for English and Portuguese in [2], although the authors used different shuffling procedures.

Our results imply that the node selectivity is a measure suitable for fine-grained differentiation between an original and meaningless text. Furthermore, selectivity can potentially be considered as a measure for discrimination between different text types, capturing the aspect of quality of texts. This should be thoroughly examined in the future work, which will cover: a) differentiation between text types from linguistic network structure, b) finding which set of measures reflects the quality of the texts, c) the analysis of the Croatian linguistic networks using syntax dependencies instead of pure co-occurrence.

REFERENCES

[1] K. Ban, S. Martinčić-Ipšić, and A. Meštrović. "Initial comparison of linguistic networks measures for parallel texts". *5th International Conference on Information Technologies and Information Society (ITIS)*, 97-104, 2013.

[2] S. Caldeira, P. Lobao, R. Andrade, A. Neme, and V. Miranda. "The network of concepts in written texts". *The European Physical Journal B-Condensed Matter and Complex Systems*, 49(4):523-529, 2006.

[3] A. Hagberg, P. Swart, and D. Chult. "Exploring network structure, dynamics, and function using networkx". *Technical report*, Los Alamos National Laboratory (LANL), 2008.

[4] M. Krishna, A. Hassan, Y. Liu, and D. Radev. "The effect of linguistic constraints on the large scale organization of language". arXiv preprint arXiv:1102.2831, 2011.

[5] W. Li. "Random texts exhibit zipf's-law-like word frequency distribution". *Information Theory, IEEE Transactions on*, 38(6):1842-1845, 1992.

[6] H. Liu and F. Hu. "What role does syntax play in a language network?". *Europhysics Letters*, 83(1):18002, 2008.

[7] D. Margan, S. Martinčić-Ipšić, and A. Meštrović. "Preliminary report on the structure of Croatian linguistic co-occurrence networks". *5th International Conference on Information Technologies and Information Society (ITIS)*, 89-96, 2013.

[8] A. Masucci and G. Rodgers. "Network properties of written human language". *Physical Review E*, 74(2):026102, 2006.

[9] A. Masucci and G. Rodgers. "Differences between normal and shuffled texts: structural properties of weighted networks". *Advances in Complex Systems*, 12(01):113-129, 2009.

[10] D. Watts and S. Strogatz. "Collective dynamics of small-world networks". *Nature*, 393(6684):440-442, 1998.

[11] D. Margan, S. Martinčić-Ipšić and A. Meštrović. "Network Differences Between Normal and Shuffled Texts: Case of Croatian". *Springer, Studies in Computational Intelligence, Complex Networks V*, 549:275-283, 2014.

Comparison of the language networks from literature and blogs

Sabina Šišović, Sanda Martinčić-Ipšić, Ana Meštrović
Department of Informatics University of Rijeka, Rijeka, Croatia
ssisovic@uniri.hr, smarti@uniri.hr, amestrovic@inf.uniri.hr

Abstract - In this paper we present the comparison of the linguistic networks from literature and blog texts. The linguistic networks are constructed from texts as directed and weighted co-occurrence networks of words. Words are nodes and links are established between two nodes if they are directly co-occurring within the sentence. The comparison of the networks structure is performed at global level (network) in terms of: average node degree, average shortest path length, diameter, clustering coefficient, density and number of components. Furthermore, we perform analysis on the local level (node) by comparing the rank plots of in and out degree, strength and selectivity. The selectivity-based results point out that there are differences between the structure of the networks constructed from literature and blogs.

I. INTRODUCTION

The representation and analysis of written texts in terms of graphs and complex networks offers an alternative approach for studying the language with different applications in the domain of natural language processing (NLP). Various types of linguistic networks have already been studied: syntax networks [1,2], semantic networks [3], phonological networks [4], syllable networks [5,6], word co-occurrence networks [7-19]. In [3,20,21] a systematic methodological overview of linguistic complex networks principles is presented. Recently, linguistic co-occurrence networks have been intensively studied in order to analyse the structure of the language [7-19].

As the networks incorporate associations between words and concepts, their structure, quantified by global and local network measures [22], such as clustering coefficient, shortest path, diameter, density, node degree, can provide information on some properties of the text. The motivation of our research was to find which network measures are sensitive on different texts categories.

In our previous research [8, 9] we showed the advantages of using a directed and weighted co-occurrence network as the model to capture the structure of a text. In this work we study global and local network measures for the networks constructed from different categories of texts. In particular, at the local level, we applied the node selectivity measure in order to examine if it is sensitive on different styles of texts. Node selectivity is defined as the average weight distribution on the links of the single node [10]. Therefore, in our approach we constructed directed and weighted co-occurrence networks from different texts: 4 books and 4 blogs. We compare global and local network measures for book-blog network pairs.

In the second section we present the overview of related work. In the third section we present key measures of complex networks. In the fourth section the data and network construction techniques are presented. In the fifth section we present the results. In the last section we elaborate on the obtained data and provide concluding remarks.

II. RELATED WORK

Ferrer i Cancho and Solé in [11] first showed that the co-occurrence networks have a small average path length, a high clustering coefficient, and a two-regime power law degree distribution; the network exhibits small-world and scale-free properties. Drogotsev and Mendes [12] used co-occurrence networks to study language as a self-organising network of interacting words. Masucci and Rodgers in [13] investigated the co-occurrence network topology of Orwell's '1984' focusing on the local properties: nearest neighbours and the clustering coefficient. Furthermore, in [10] they introduced the node selectivity measure that can distinguish the difference between normal and randomised text. Liu and Cong [14] constructed co-occurrence networks from text in different languages and used complex network parameters for the classification (hierarchical clustering) of 14 languages, where Croatian was amongst 12 Slavic.

Different applications of linguistic network analysis in NLP includes: evaluation of language complexity [15], automatic summarisation [16] and evaluation of machine translation [17], authorship attribution [18] and text quality analysis [19].

Costa et al. [15] studied the relationship between the topology of network and complexity of the text. They studied texts with different levels of simplification in co-occurrence networks and found that topological regularity correlated negatively with textual complexity. Furthermore, they showed that strength, shortest path, diversity and hierarchical measures can make a distinction between normal text and simplified text. In [16] the authors describe a method that uses complex networks concepts for the summarisation task. In [17] several metrics from complex networks are exploited in order to evaluate the quality of translations. The best distinctions were obtained with the out-degree, in-degree, minimum

path and cluster coefficient. In [18] authors investigate the correlation between the properties of networks and author characteristics. It is shown that the networks produced for each author are sensitive to specific features, which indicates that complex networks can capture author characteristics and, therefore, could be used for the authorship identification. In [19] authors investigate the possibility of automated evaluation of text quality using topological measurements extracted from the corresponding complex networks. All the measures are correlated with grades assigned by human experts. It is shown that out-degree, clustering coefficient and deviation from linear dynamics in the network growth are correlated with the text quality.

However, there are no comprehensive studies focused on finding network measures that are sensitive to different text categories. Our work is the first attempt to analyse which network measures can differentiate between literature and blog networks.

III. THE NETWORK STRUCTURE ANALYSIS

This section contains explanations of network-based measures [22, 10] that are used in our approach for the comparison of two categories of language networks.

Every network has an *N* number of nodes and *K* number of links. Considering the fact that our networks are weighted every link connecting two nodes has an associated weight.The degree of a node *i* is the number of links with which the node is connected, k_i. In the case of the directed network, there are two kinds of the degree: the in-degree, k_i^{in} corresponding to the number of incoming links and the out-degree, k_i^{out} equal to the number of outgoing links. The average degree of the network is:

$$< k >= \frac{2K}{N}. \tag{1}$$

For every two connected nodes *i* and *j*, the number of connections lying on the path between them is represented as d_{ij}, and so d_i is the average distance of a node *i* from all other nodes, and it's obtained by:

$$d_i = \frac{\sum_j d_{ij}}{N}. \tag{2}$$

For the next two measures, if a network contains more than one component, we consider the largest component. The average shortest path length between every two nodes in network is:

$$L = \frac{1}{N(N-1)} \sum_{i \neq j} d_{ij}, \tag{3}$$

and the maximum distance results in the network diameter, D:

$$D = max_i d_i. \tag{4}$$

The clustering coefficient is a measure which defines the presence of connections between the nearest neighbours of a node. And so, ci (clustering coefficient) of a node is a fraction between the number of edges Ei that exist between that ki and the total possible number:

$$c_i = \frac{2E_i}{k(k-1)}. \tag{5}$$

The average clustering of a network is defined as the average value of the clustering coefficients of all nodes in a network:

$$C = \frac{1}{N} \sum_i c_i. \tag{6}$$

Density of network is a measure of network cohesion defined as the number of observed relationships divided by the number of possible relationships:

$$d = \frac{K}{N(N-1)}. \tag{7}$$

Strength of the node i is the number of its outgoing and incoming links (sum of its weights). For the directed networks the in-strength and the out-strength are defined:

$$s_i^{out/in} = \sum_j w_{ij/ji}. \tag{8}$$

The node selectivity measure can capture the effective distribution of numbers in the weighted adjacency matrix, and it is obtained as a ratio of (out/in-) node strength and its (out/in-) degree:

$$e_i^{out/in} = \frac{s_i^{out/in}}{k_i^{out/in}}. \tag{9}$$

All presented measures are standard network measures usually used for network structure analysis, except the node selectivity measure which is introduced in [10] as the measure that can differentiate between networks based on normal and randomized texts. According to these results, we expected that node selectivity may be potentially important for the text categories differentiation and include it in the set of standard network measures.

In order to illustrate the relationships between node degree, node strength (which are the standard local network measures) and node selectivity, we constructed a small network of seven nodes presented in Figure 1.

Additionally, Table 1 contains values of in-degree, out-degree, in-strength, out-strength and in-selectivity and out-selectivity for all seven nodes in the network presented in Figure 1.

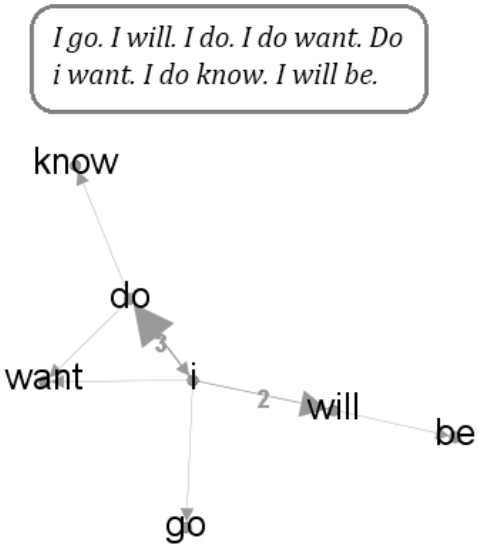


Figure 1 Weighted and directed co-occurrence network of seven nodes, created from the short text below

TABLE 1 - VALUES OF IN/OUT - DEGREE, STRENGTH AND SELECTIVITY

NODE	k^{in}	k^{out}	w^{in}	w^{out}	s^{in}	s^{out}
i	1	4	1	7	1	1,75
do	1	2	3	2	3	1
will	1	1	2	1	2	1
want	2	0	2	0	1	0
know	1	0	1	0	1	0
go	1	0	1	0	1	0
be	1	0	1	0	1	0

IV. NETWORK CONSTRUCTION

A. Data

Our corpus contains 4 books written or translated into the Croatian language, and 4 blog texts written in Croatian language. The books are: *Picture of Dorian Gray*, *Bones,The Return of Philip Latinowicz* and *Mama Leone*. The blogs are: *Index.hr*, *Slobodna Dalmacija*, *Narodne novine and Gospodarski list* (daily newspaper portal, or business portal). The feature which prompted us to do the comparison is the linguistic distinction between book and blog. Books are written in formal language, standard expressions and phrases are used, whilst blogs are mostly written in a casual mode, with the use of slang, the shortenings of the words or mistakes in syntax. Books come in different sizes and so we compared them with the approximately same sized blog (with the same number of different words), which means we had 4 book-blog pairs for comparison. The sizes of books and blogs in number of total words are shown in the first row of Table 2, while the numbers of different words are presented in the second row (as the number of nodes).

B. The construction of co-occurrence networks

We used Python and the NetworkX software package developed for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks [23].

The first step in creating networks was text "cleaning": normalising symbols for Croatian diacritics (č, ć, ž, đ, and š), removing special symbols and removing punctuation which does not mark the end of a sentence. We created 8 networks, weighted and directed. Nodes are words that are linked if they are direct neighbours in a sentence. The next step was creating the networks as weighted edgelists, which contain all the pairs of connected words and their weights (the number of connections between two same words).

V. RESULTS

In this section we present the results of our measuring described in section 2, such as average degree $<k>$, average path distance L, diameter D, and the average clustering coefficient C, density d, node strength s_i and node selectivity e_i.

In Table 2 we present the estimated global network measures. There are certain differences between measures for book-blog network pairs, but there is no uniform rule that may be used to differentiate these two styles of writing.

TABLE 2 - THE COMPARISON OF NETWORK MEASURES FOR BOOK-BLOG NETWORK PAIRS

Measure	**Text**							
	Bones	***Gospodarski list***	***Mama Leone***	***Narodne novine***	***Picture of Dorian Gray***	***Index.hr***	***Return of Phillip Latinowicz***	***Slobodna Dalmacija***
Number of different words	191 986	199 417	85 347	146 731	75 099	118 548	28 137	44 367
Number of nodes (*N*)	27396	27727	13067	13036	15631	15065	9531	9553
Number of edges (*K*)	102052	105171	49383	55661	46201	28972	21760	25155
Average degree (<*k*>)	7,45	7,58	7,56	8,54	3,88	3,85	4,57	5,27
Avg. shortest path (*L*)	3,21	3,28	3,11	3,17	3,45	3,45	3,59	3,56
Diameter (*D*)	10	21	10	12	14	22	16	13
Average clustering coefficient (*C*)	0,25	0,22	0,29	0,22	0,18	0,016	0,15	0,17
Density (*d*)	0,0002	-	0,00056	0,00066	0,0004	0,0002	0,00048	0,00055
No. of connected components	15	7	1	2	1	45	5	3

Furthermore, we compare networks on the node-level using degree, strength and selectivity measures. For the purpose of comparison we used rank plots. The in/out-degree rank function represents the relationship function between the rank and the in/out-degree of the degree sequence of all nodes sorted in decreasing order.

Similarly, the in/out-strength rank plot and the in/out-selectivity rank plot are defined.

Figure 2 represents the results of the comparisons of the in-degree rank plot and the out-degree rank plot for one book-blog network pair.

The plots do not show significant difference for the in-degree nor for the out-degree rank plots between book-blog network pair. We also experimented with additional three book-blog pairs and we obtained similar results (not reported here due to limited space).

The results of the comparisons of the in/out-strength rank plot for the same book-blog network pair are shown in Figure 3. Again, there is no difference except some small deviation that can be noticed in the plot, but we cannot conclude that in/out-strength distinguish books from blogs.

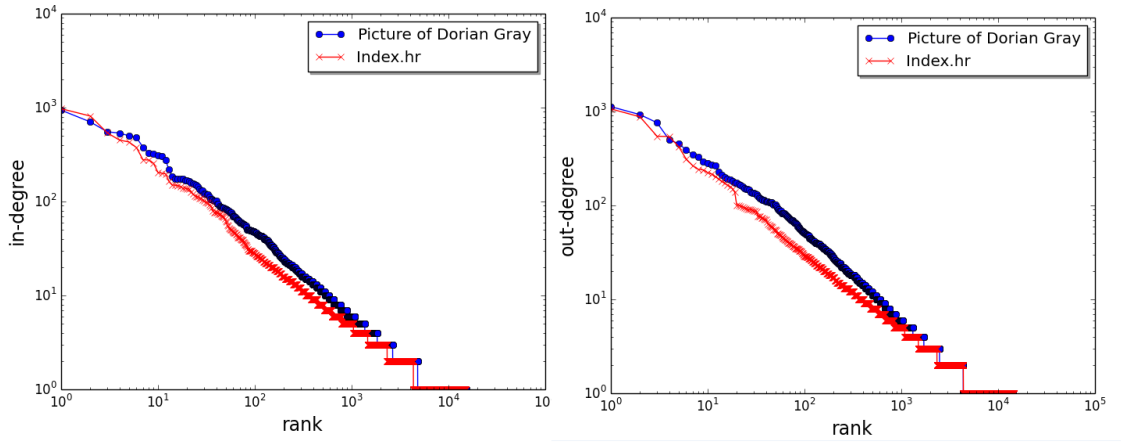


Figure 2 in-degree and out-degree rank plots

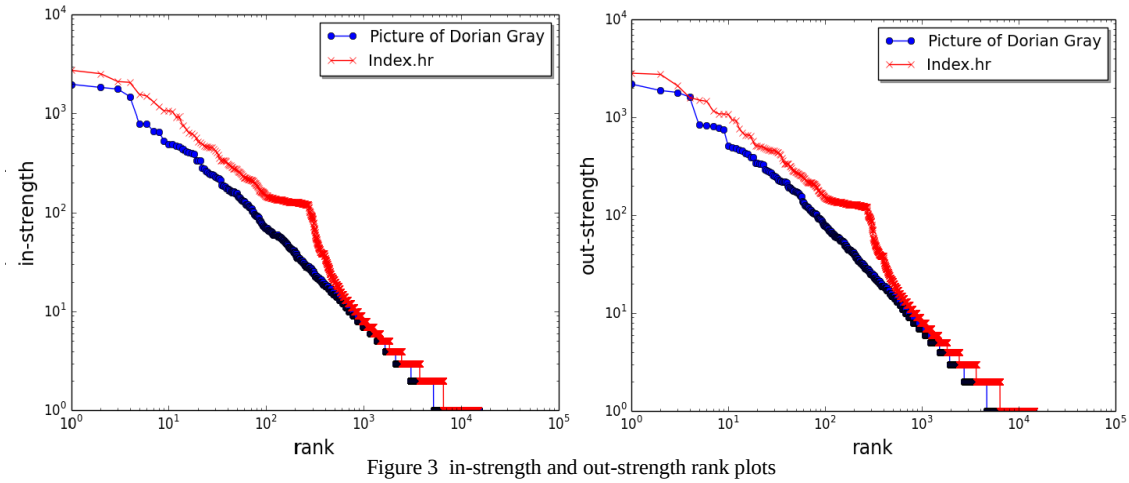


Figure 3 in-strength and out-strength rank plots

The selectivity rank plots are shown in Figure 4 (in-selectivity) and in Figure 5 (out-selectivity). The results show that there are differences in selectivity rank plots between networks constructed from books and networks constructed from blogs for all 4 book-blog network pairs.

In general, all in/out-selectivity values are lower for books than for blogs. We disregarded nodes with zero values of degree because it causes the division by zero (in total 4% of nodes).

VI. CONCLUSION

In this work we analysed which complex network measures can distinguish between different texts categories: literature and blogs.

Our results indicate that global network measures are not precise enough to capture the structural differences between networks constructed from different text categories. Even the compared in/out- degree rank plots and in/out- strength rank plots on the local level do not clearly show the differences. However, in-selectivity and out-selectivity rank plots indicate that there are structural

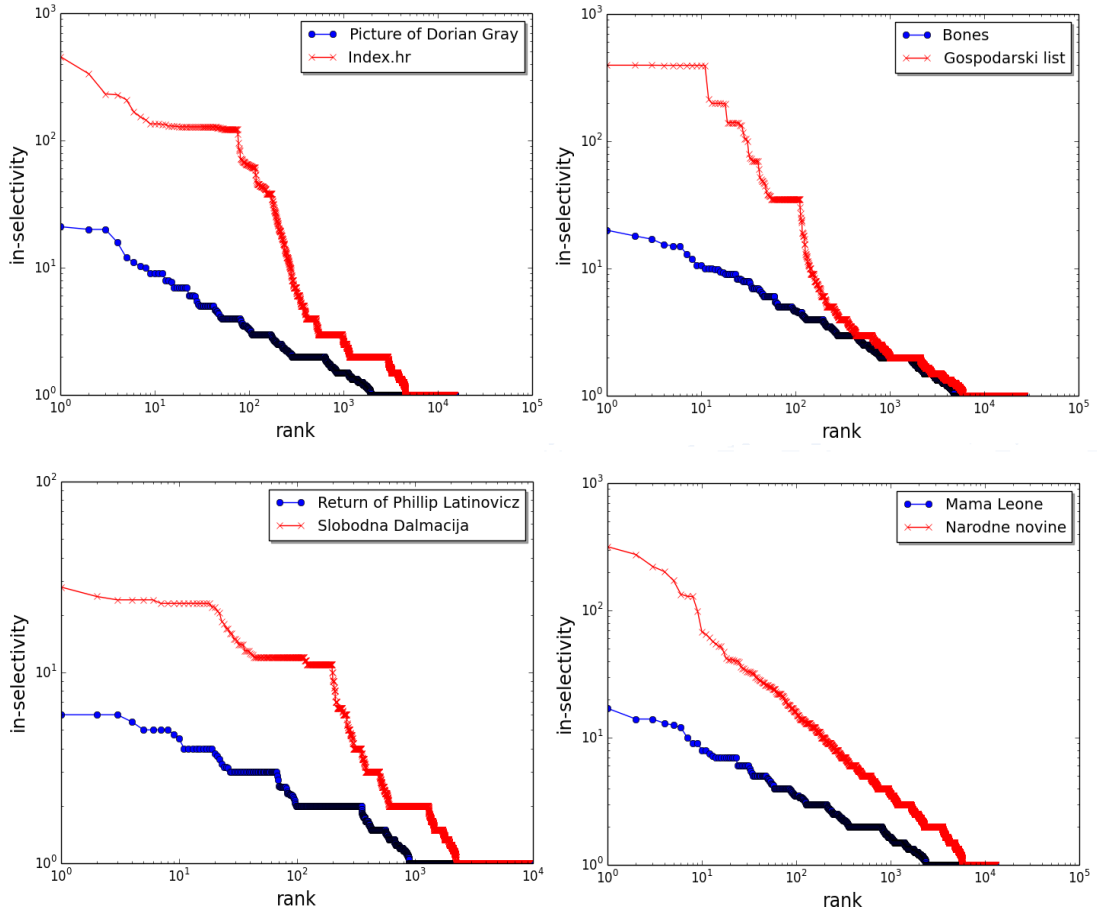


Figure 4 in-selectivity rank plots for 4 book-blog network pairs

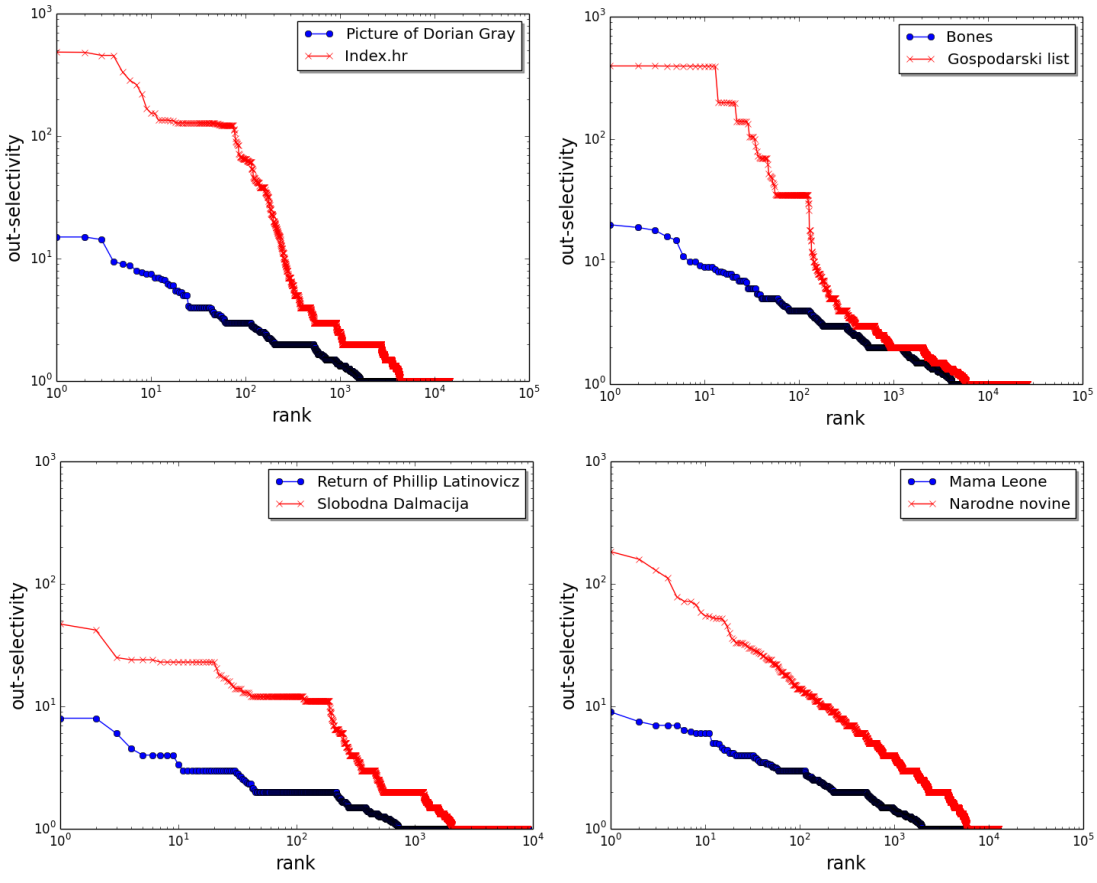


Figure 5 out-selectivity rank plots for 4 book-blog network pairs

and networks constructed from blogs. The values of in- and out- node selectivity measure are higher for networks generated from blogs than for networks generated from books. The node selectivity based measures are the only measures from the set of tested measures that can capture the structural differences between two classes of networks.

The presented approach should be extended with examination of other network measures that depend on the strength, degree and link direction in combination with other measures such as clustering coefficient. For future work we will test the broader set of measures on various categories of texts. We plan to include forum comments, texts from Wikipedia, poems and other genres of literature in our corpus. Obtained results encourage us to investigate the complex network properties for text classification, text evaluation or even text quality assessment.

VII. REFERENCES

[1] R.F. Cancho, et al., "Patterns in syntactic dependency networks", Physical Review, E 69, 051915, 2004.

[2] H. Liu and C. Hu, "Can syntactic networks indicate morphological complexity of a language", *EPL* 93 28005, 2001.

[3] J. Borge-Holthoefer and A. Arenas, "Semantic Networks: Structure and Dynamics", Entropy, 12, 1264-1302, 2010.

[4] S. Arbesman et al., "Comparative Analysis of Networks of Phonologically Similar Words in English and Spanish", Entropy, 12, pages 327-337, 2010.

[5] M. Medeiros Soares et al., "The Network of syllables in Portuguese", Physica A, 355(2-4): 678-684, 2005.

[6] K. Ban et al., "A preliminary study of Croatian Language Syllable Networks", Mipro SP, pp 1697-1701, 2013.

[7] K. Ban et al., "Initial comparison of linguistic networks measures for parallel texts", 5th International Conference on Information Technologies and Information Society (ITIS), pp. 97-104, 2013.

[8] D. Margan, S. Martinčić-Ipšić and A. Meštrović, "Preliminary report on the structure of Croatian linguistic co-occurrence networks", 5th International Conference on Information Technologies and Information Society, Slovenia, pp. 89-96, 2013.

[9] D. Margan, S. Martinčić-Ipšić and A. Meštrović, "Network Differences Between Normal and Shuffled Texts: Case of Croatian", Accepted at CompleNet 2014 Workshop, 2014.

[10] A. P. Masucci and G. J. Rodgers, "Differences between normal and shuffled texts: structural properties of weighted networks", Advances in Complex Systems, 12(01):113-129, 2009.

[11] R. F. Cancho and Richard V Solé. "The small world of human language", Proceedings of the Royal Society of London, 268(1482):2261–2265, 2001.

[12] S. N. Dorogovtsev and J. F. Mendes. "Language as an evolving word web", Proceedings of the Royal Society of London. Series B: Biological Sciences, 268(1485):2603–2606, 2001.

[13] A. P. Masucci and G. J. Rodgers. "Network properties of written human language", Physical Review E, 74(2):026102, 2006.

[14] H. T. Liu and J. Cong. Language clustering with word co-occurrence networks based on parallel texts. Chinese Science Bulletin, 58(10):1139–1144, 2013.

[15] D. R. Amancio et al. "Complex networks analysis of language complexity", *arXiv preprint arXiv:1302.4490* (2013).

[16] L. Antiqueira et al. "A complex network approach to text summarization", Information Sciences, 179(5), 2009, pp. 584-599

[17] D. R. Amancio et al. "Using metrics from complex networks to evaluate machine translation", *Physica A: Statistical Mechanics and its Applications, 390*(1), pp. 131-142. 2009.

[18] L. Antiqueira et al. "Some issues on complex networks for author characterization." Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial 11.36 (2007): 51-58.

[19] L. Antiqueira et al. "Strong correlations between text quality and complex networks features." *Physica A: Statistical Mechanics and its Applications* 373 (2007): 811-820.

[20] R.V. Solé, B.C. Murtra, S. Valverde, L. Steels. "Language Networks: their structure, function and evolution", Trends in Cognitive Sciences, 2005.

[21] M. Choudhury et al. "Global topology of word co-occurrence networks: Beyond the two-regime power-law", In Proceedings of International Conference on Computational Linguistics, pp 162–170, 2010.

[22] L.F. Costa, et. al. Boas, "Characterization of complex networks: a survey of measurements", cond-mat/ 0505185, 2005.

[23] Hagberget. Et al. "Exploring network structure, dynamics, and function using networkx". Technical report, (LANL), 2008.

Performance Analysis of Parallel Master-Slave Evolutionary Strategies (μ,λ) Model Python Implementation for CPU and GPGPU

D. Zubanović, A. Hidić, A. Hajdarević, N. Nosović and S. Konjicija
Faculty of Electrical Engineering, University of Sarajevo
Sarajevo, Bosnia and Herzegovina
{dzubanovic1,ahidic2,ahajdarevi2,nnosovic, skonjicija}@etf.unsa.ba

Abstract – Evolutionary strategies is a heuristic, guided-search based evolutionary algorithm, widely used as optimization technique for computationally intensive problems. Python is a high-level programming language known for code readability, reusability and the ease of use, making it preferable choice for quick and robust software development, although it is lacking in performance and concurrency area. Emerging technologies such as Anaconda Accelerate Python compiler attempt to combine Python's ease of use with both declarative and explicit parallelization and high performance in computationally intensive problems. In this paper an example of master – slave parallel Evolutionary strategy ES(μ,λ) implementation in Python is given, and its performance on CPU and GPU are analyzed.

Keywords – Parallel computing; Python-Anaconda; GPU; Multithreading; Heuristic algorithms; Test functions for optimization;

I. INTRODUCTION

Heuristic algorithms, especially Evolutionary algorithms are widely used as optimization techniques for computationally difficult problems. Due to the nature of such problems, numerous attempts of acceleration of heuristic algorithms have been made. Heuristic algorithms are not embarrassingly parallel, so there is not one simple and obvious approach to their acceleration. Additionaly, good performance and the code's ease of use, readability and usability seldom go with one another, e.g. good performance can be achieved in C++ but the cost is prolonged development time, longer and hard-to-maintain source code. High-level, just-in-time compiled and interpreted languages can alleviate some of the development and maintenance troubles, but underperform in computing intensive situations. Evolutionary strategies (ES) is one of the most widely used heuristic algorithms and thus theoretical models for parallelization have been developed. In this paper, Master-Slave ES model is implemented using special Python compiler that supports declarative parallelizaton of Python code. Performance analysis is based on parallelized code executed both on CPU as multiple slave threads, and on GPU using CUDA integration as multiple GPGPU kernels. For comparison, sequential algorithm's performance is measured as well. Algorithm's performance is measured using standard benchmark functions for heuristic algorithms.

A. Parallelism of cpu

Central Processing Units (CPUs) are devices made to be primary general purpose processing unit in computers. As a result of ever growing demands, CPU architectures evolved and became faster in terms of FLOPS (Floating Point Operations Per Second). Numerous attempts of paralellization have been made, such as clusters and distributed computers, but not on single chip up until Intel HyperThreading concept, and multi core concepts emerged. Modern CPUs consist of certain number of cores ranging from 2 up to 32 or more for todays' most advanced CPUs.

B. Parallelism of gpu

Graphics Processing Units (GPUs) are devices primarily made for processing/rendering and displaying 2D and 3D graphic objects. In an attempt to meet high requirements set primarily by computer games and their graphics engines, GPU architecture radically changed, and it became what is known today as “modern GPU”[2]. New GPU architectures made it possible and feasible to do computations other than processing graphics on GPUs. Modern GPUs are massively parallel structures. Heart of modern GPUs is a set of Multithreaded Streaming Processors. Each of them is made of a certain number of smaller Scalar Processors and shared memory. Architecture is organized so that each stream processor has access to the global video memory. Each scalar processor executes the same code, while the data used as input or output is determined by respective scalar processor's unique identification. This arhitecture is named SIMD (Single Instruction [on] Multiple Data)[3][4].

C. Evolutionary strategies

Evolutionary strategies (ES) represent one of the most commonly used types of evolutionary algorithms, alongside with Genetic Algorithm (GA). It is used for optimization, or, finding adequatly good (best if possible) solution in the designated problem space. Multiple different types of ES have been developed, but the ES(1+1), ES(μ,λ) and ES(μ+λ) are the ones most widely used. Because the ES(1+1) is inherently sequential, only one of the population ES's - ES(μ,λ) is considered [1]. Its search model is considered to be guided because it does not consider all possible solutions (either randomly or in a predefined order). It rather relies on the fact that by creating offsprings of pairs of solutions contained within current population and by randomly mutating the offspring by offsetting it's coordinates by random, normally-distributed value with variable standard deviation, the population's overall fitness will keep increasing until most of the solutions converge to the determined optimum. Initial population can either be fixed, randomly generated or determined by some other algorithm. Unlike the GA, ES

solutions that make the population are tuples of real values and thus the ES eliminates the need for mapping between the algorithm's and problem's domains. However, the quality of the found optimae is generally considered to be lower than it is the case with the GA - operators of crossover used for creating the offspring and mutation have lower impact when used on real numbers than when they are used on tuples of binary values. Therefore, the self-adaptation of the variable standard deviation is the key for obtaining overall adequate results.

II. RELATED WORK

Two major concepts in parallel evolutionary strategies are master – slave model and island model [5].

Master-slave model stores whole population within master and executes mutations and recombinations there while it delegates slaves to execute fitness function evaluation[6]. This approach to parallelization is used when function evaluation is the most expensive operation of the algorithm, as it is the case with multidimensional functions that contain sums or products dependent on the dimension, or that rely heavily on floating point arithmetics [6] .

Island model partitions the population among the slave nodes (threads or seperate nodes in a cluster), and each of them runs algorithm separately on its own population[7]. In island model, master node or master thread is used as a synchronizer between the slaves. More advanced models include the redistribution of the best solutions among the slaves as for to prevent premature convergence.

Numerous attempts have been made to compare performance of GPU and CPU in various scientific computations. Performance comparisons between computations made on CPUs and GPUs are mainly biased or published by GPU or CPU manufacturers or their business partners[4]. In this instance, Evolutionary Strategies (μ,λ) parallellization in Python has been chosen to demonstrate that the difference in performance is not always clear or applicable in problems with different dimensionality. Break-Even analysis is performed as a means to indicate the dimensionality of the problem at which the performance of the CPU and GPU implementations start to diverge. Also, difference in the ease of development using Python with specialized compiler is documented so that it can be factored in when considering the tradeoffs between the different platforms and technologies.

III. MASTER-SLAVE EVOLUTIONARY STRATEGY AND BENCHMARKS

A. *Evolutionary strategy (μ,λ)*

There are various types of evolutionary strategies, roughly they can be divided in population algorithms and algorithm with single individual. As those with single individual are not suitable for parallelization, as there is no parallelizable parts, only population algorithms can be considered for parallelization. Most suitable evolutionary algortihms for parallelization are ES(μ,λ) and ES(μ+λ). Evolutionary strategy ES(μ,λ) is chosen as it converges slower than ES(μ+λ), and for that reason is more suitable when measuring performance of algorithms. In fact, implementation differences are minor as one algorithm can easily be can converted into another.

Generalized procedure of ES(μ,λ) is given below [1]:

1. Population P of μ random solutions according to gaussian distrubtion is instantiated
2. Perform mutation for each solution of population P
3. Perform recombination for 2(λ - μ) selected solutions from P
4. Create new population P' from solutions obtained from steps 3 and 4
5. Evaluate every solution form population P'
6. Sort population P' according to results of evaluation obtained from step 5
7. μ best solutions from P' become P
8. Adapt Ps – parameter of „success“
9. Repeat steps 2-8 if condition of stopping not met
10. Stop

Pseudocode of Mutation operator is defined as:

```
for each solution s in P
    new=s+gauss(0,sigma)
    add new to P'
end for
```

Annotation „gauss“ presents gauss distribution whose parameters are 0 – as mean of distribution, and „sigma“ – as standard deviation.

Pseudocode of Recombination operator is defined as:

```
select 2(λ - μ) solutions form P by uniform
distribution
for each s1,s2 in selected
    new = (s1+s2)/2
    add new to P'
end for
```

In step 8 of generalized procedure of ES(μ,λ) parameter of success or Ps is mentioned. Parameter of succes in algorithm is intended to prevent early convergence of algorithm, or to prevent algorithm from returning false optimum. Parameter of success is used to adjust parameter sigma used as standard deviation parameter for gaussian distribution. Pseudocode of adapting Ps is defined as:

```
if iteration mod 20 == 0
    Ps = nS/iteration
    if Ps>0.2
        sigma = sigma x 1.4
    end if
    if Ps<0.2
        sigma = sigma/1.4
    end if
end if
```

In pseudocode given above nS represents number of successful iterations, where successful iteration is iteration in which mean value of the fitness of the population is better than the one from the previous iteration.

Stopping conditions for Evolutionary strategies may vary between different implementations and problems. Most commonly used, and are also used in this implementation, are fixed number of iterations, and improvement in max value of population, or improvement in mean value of the fitness of the population over n iterations. All of listed criteria are not to be used alone, because in case of fixed number of iterations, algorithm may do half of iterations for no reason if it already had converged to (sub)optimum. Or in case of other two stopping criteria algorithm may never stop.

B. *Master-slave model*

Evaluation of problem solutions is taking most time when executing evolutionary strategies. For that reason evaluation of solutions is parallelized according to master-slave model. Master-slave model of evolutionary strategy is implemented in Python programming language using Anaconda Accelerate module and its features. In this specific case master represents main loop, and delegates slaves to work parallel on evaluation of problem solutions. There is no explicit delegation of slaves, and all work is done implicitly using Anaconda Accelerate features which reduces complexity of the code and makes debugging and maintenance easier but limits the programmer to the predesignated means of utilizing the GPGPU or multithreading capabilites - for obvious reasons it cannot be as flexible as the native API.

Prerequisite for parallelization using Anaconda Accelerate is to vectorize code which is to be parallelized. Criteria functions are vectorized, or customized to work with vectors. This customization is done with „vectorize“ feature of Anaconda Accelerate module.

A simple example of Sphere function is given below.

```
def value(x,y):
    z =[]
    for i in range(0,len(x)):
        z.append(x[i]**2+y[i]**2)
    return z
```

Then vectorization is of function given above can be implemented as follows:

```
from numbapro import vectorize

@vectorize(["float32(float32, float32)"],
target="cpu")
def value(x,y):
    return x**2+y**2
```

As it can be seen vectorization is done by putting „@vectorize“ annotation before definition of function, and iti is necessary to specifiy number and types of input and output arrays. Annotation „target“ can have three different parameters, where parameter „cpu“ defines that function is to be run on single thread, „parallel“ defines that function is to be run on multiple threads, depending on how many are available, and finally “gpu” definest that function is to be run on CUDA platform if available. Vectorized function is than divided to slave components automatically by the compiler.

Master calls the vectorized function just like any other function:

```
import numpy as np
…
y = np.zeros(self.Lambda, dtype=np.float32)
y = value(x1,x2)
```

Parameters x1 and x2 in example above are arrays of problem solutions obtained before by mutation and recombination. Prerequisite for using Anaconda Accelerate, and its vectorize feature is using numpy module arrays.

C. *Benchmarks*

Master-slave evolutionary strategy has been implemented to find the global minimum of Rastrigin and Styblinsky-Tang functions in their five variables versions.
Rastrigin function is defined as:

$$f(x) = 10n + \sum_{i=1}^{n} (x_i - 10\cos 2\pi x_i)$$

Rastrigin function is very difficult because it has numerous local minuma and poorly implemented algorithm can easily return local minimum and converge to it. Rastrigin function has global minimum in f(x) = 0 for x = 0.

Styblinski-Tang function is defined as:

$$f(x) = \frac{\sum_{i=1}^{n} x_i^4 - 16x_i^2 + 5x_i}{2}$$

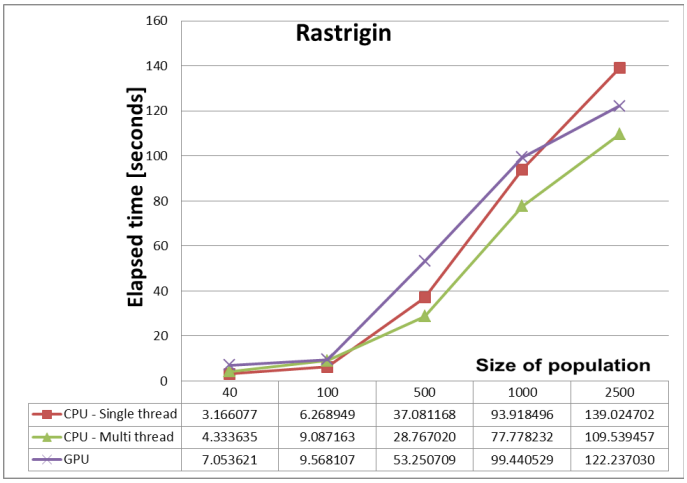
Styblinski – Tang function has 4 minima, one of which is global minimum. All of minima have relatively close values, and function gradient between them is not high so the algorithm can easily converge to one of local minima. Styblinski – Tang is somewhat heavy to evaluate due to having a sum dependent on the dimension of the problem and floating point division in it. One of advantages of Styblinski – Tang function over Rastrigin function as means of testing and benchmarking optimization algorithms is that it has a global minimum in f(-2.903534,…n… -2.903534) = -39.16599n where n is number of problem variables, thus the chances of hitting the minimum by initializing one of the initial solutions directly to or nearby the coordinates (0,…,0) is greatly reduce. That guarantees a certain level of randomness while testing the algorithm's behavior.

IV. PERFORMANCE MEASUREMENT AND ANALYSIS

All performance measurements were done on a machine running Windows 7 on an Intel Core i7 3770 processor with core frequency of 3,4 GHz, with 8GB of DDR3 RAM at 1600 MHz, nVidia GTX 660 GPU with 960 CUDA processors at 980MHz and 2048 MB of GDDR5 graphics memory. Algorithms were implemented in Python v2.7.6 and compiled using Anaconda v1.8.0 MKL and Anaconda v1.8.0 Accelerate features for speedup.

A. *Master slave model results*

Implemented evolutionary strategies algorithm was ran at max 3000 iterations as it was experimentally determined that algorithm converges to one value no sooner than iteration 1500, and no later than iteration 2500. Algorithm was ran with start value of standard deviation sigma set at 1.2. Parameters μ and λ were arranged so that λ is 1.5 times larger than μ. Three types of measurements for each function were conducted. Measurements were conducted on population ranging from (40,60) to (2500,3750). First the measurement was performed on single threaded CPU, than on the multithreaded CPU and lastly on the GPU Python implementation. Every measurement for each population was run 30 times, and mean value was taken in account, so that the standard deviation of measurement tends to zero.



	40	100	500	1000	2500
CPU - Single thread	3.166077	6.268949	37.081168	93.918496	139.024702
CPU - Multi thread	4.333635	9.087163	28.767020	77.778232	109.539457
GPU	7.053621	9.568107	53.250709	99.440529	122.237030

Figure 1 Time representation of Evolutionary strategy run on three platforms when evaluating Rastrigin function

As it can be seen from Figure 1. for small populations sequential implementation elapsed time is considerably smaller due to large overhead for preparation of parallelization. Obviously, for larger population there is more job to be done in parallel, so the benefits from parallellization come to the fore. Multithreaded implementation has better results than GPU implementation due to not enough large data set for GPU to excell. In terms of results every measurement returned results considered as adequate and all results were as high as $1x10^{-6}$ which is a small deviation from the actual minimum.

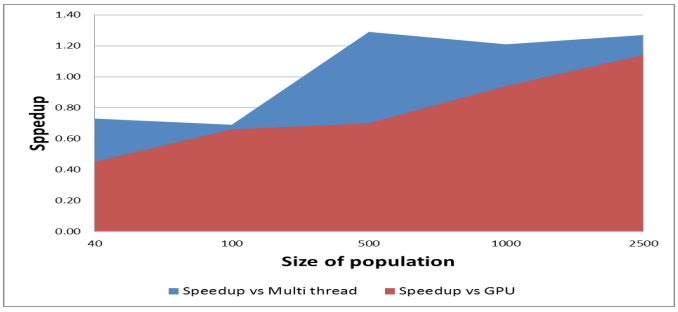
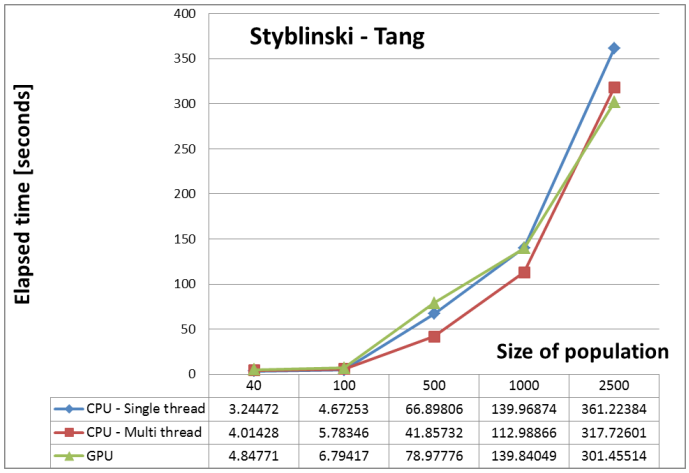


Figure 2 Speedup gained/lost using Multithreading and GPU when evaluating Rastrigin function

All the aforesaid is clearer presented in Figure 2. where it can easily be seen that for smaller populations parallelized implementations slow down the algorithm. It can easily be seen that with increase of population speedup increases as well. In case of Speedup gained by using GPU, speedup can be easily aproximated with linear function

$$y = \frac{0{,}69}{2460}x + 1{,}128$$



	40	100	500	1000	2500
CPU - Single thread	3.24472	4.67253	66.89806	139.96874	361.22384
CPU - Multi thread	4.01428	5.78346	41.85732	112.98866	317.72601
GPU	4.84771	6.79417	78.97776	139.84049	301.45514

Figure 3 Time representation of Evolutionary strategy run on three platforms when evaluating Styblinski - Tang function

As it can be seen from Figure 3. sequential implementation is considerably faster than parallel implementations for smaller populations. Also, the execution times for Styblinski – Tang function are significantly higher than those from Rastrigin function due to Styblinski – Tang being more complicated to evaluate, and also evolutionary strategies algorithm converges slower for Styblinski – Tang, relative to Rastrigin function. Also, for largest population is faster relative to same multithreaded implementation, probably due to Styblinski – Tang being harder to evaluate, and having an arithmetic operation division in it.

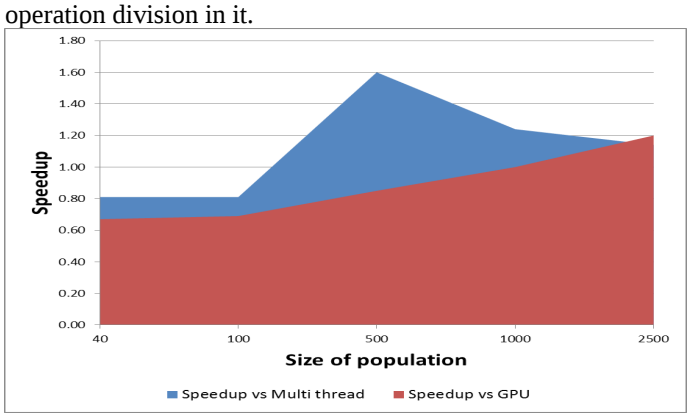


Figure 4 Speedup gained/lost using Multithreading and GPU when evaluating Styblinski - Tang function

Figure 4. gives better preview of all afore mentioned, there is no speedup for smaller populations, but with increase of population amount of speedup gradually increases. Speedup against GPU can be easily aproximated with linear function just as is the case with previous Rastrigin function.

In terms of results of evaluation, smaller population implementations do not always return correct value, or value that is very close to optimum. While larger population implementations, μ higher than 500, always return optimum value. It happens due to large problem space, and Styblinski – Tang complexity.

V. CONCLUSION AND FUTURE WORK

In this paper two kind of performance measurements on GPU, CPU using single thread, and on CPU using multithreading were presented. Results indicate that for smaller problems there is no speedup at all, and for larger problems the

speedup achieved by using parallelized implementations is not as large as it may seem at a first glance.

Presented results in no way indicate that the parallelized implementations should be discarded, or that its performance is not better than sequential one in evolutionary strategies. Performance measurements on Styblinski – Tang function show that there is some progress when problem space becomes larger, evaluation of criteria itself becomes harder, and population becomes even larger. Sequential and parellelized implementations of evolutionary strategies can coexist for different types of problems, depending on their size. Python language's ease of use and inherent reusability can help with maintenance of the code. With the increase of problem dimensionality speedup of parallelized implementation should excell even more relative to sequential implementation.

Sequential implementations of evolutionary strategies should be used for solving small or medium sized problems, while parallelized implementations are to be used for solving of enormous sized problems, where they can gain speedup over sequential implementations.

Travelling salesman problem (TSP), vehicle routing problem (VSP), and similar NP compelete problems higher dimensionality should be consider as candidates to be solved with parallelized implementations of evolutionary strategies algorithm.

Python, when used with adequate technologies, libraries and compilers proved to be a prospective platform for parallelization and high performance computing - parallelization is achieved with low costs and can achieve considerable performance boosts.

REFERENCES

[1] Hans-Georg Beyer, Hans-Paul Schwefel, "Evolution strategies: A comprehensive introduction", Naturac Computing 1: 3-52, Kluwer Academic Publishers, 2002

[2] David Luebke, "GPU Architecture: Implications & Trends", SIGGRAPH 2008: Beyond Programmable Shading course materials, 2008

[3] John Owens, "GPU Architecture Overview", In SIGGRAPH ’07: ACM SIGGRAPH 2007 courses, page 2, New York, NY, USA, 2007

[4] NVIDIA CUDA C Programming Guide, v4.2, April 2012, Available at: http://developer.download.nvidia.com/compute/DevZone/docs/html/C/doc/CUDA_C_Programming_Guide.pdf

[5] T. Hiroyasu, M. Miki, and Y. Tanimura. The differences of parallel efficiency between the two models of parallel genetic algorithms on pc cluster systems. In The Fourth International Conference on High-Performance Computing in the Asia-Pacific Re-gion, volume Volume 2, pages 945–948, Beijing, China, May 2000.

[6] M. Oussaidene, B. Chopard, O. V. Pictet, and M. Tomassini. Parallel genetic programming and its application to trading model induction. Parallel Computing, 23(8):1183–1198, 1997

[7] Ajay Gupta, Garrison W. Greenwood, R. Munnangi, Sanjay Ahire, "Parallel Implementations of Evolutionary Strategies", Proceedings of the Int'l Conf. on High Performance Computing pages 469-474, 1995

Simulating of query processing on multiprocessor database systems with modern coprocessors

Konstantin Y. Besedin*, Pavel S. Kostenetskiy**
* South Ural State University, Chelyabinsk, Russia
besedin.k@gmail.com
** South Ural State University, Chelyabinsk, Russia
kostenetskiy@gmail.com

Abstract - The modern manycore coprocessors and GPUs demonstrate very high performance on certain problems. Recent research has shown that these coprocessors can be used to accelerate database operations. But, to the best of our knowledge, there is only a little prior work on using coprocessors in multiprocessor database systems. This paper focuses on evaluation of database multiprocessor architectures with manycore coprocessors and GPUs. We implemented the emulator of parallel DBMS that uses computing cluster with NVIDIA GPUs or Intel Xeon Phi coprocessors for relational query processing. It allows to simulate simple SELECT and JOIN queries. A number of experiments have been done using this emulator. These experiments have shown that coprocessors are less efficient than CPUs for processing modeled SELECT queries, but more efficient than CPUs for processing INNER JOIN queries.

I. INTRODUCTION

The modern coprocessors like Intel Xeon Phi and NVIDIA Tesla demonstrate very high performance on certain problems. Database community shows a growing interest in applying this power to database systems [1].Recent research has shown that these coprocessors can be used to accelerate database operations. For example, paper [2] covers relational join algorithms. Architecture-sensitive search tree is presented in [3], [4]. Papers [5], [6] focus on sorting algorithms. Paper [7] describes speculative processing of database operations. To the best of our knowledge, there is only a little prior work on using coprocessors to accelerate database operations in multiprocessor database systems. Modern scientific and engineering problems can imply storing and processing terabytes of data what requires use of high performance database multiprocessors. It is important to determine the appropriateness of using different types of coprocessors when constructing such systems. This paper focuses on evaluation of database multiprocessor architectures with manycore coprocessors and GPUs.

II. IMPLEMENATION

To evaluate performance of different coprocessors in database multiprocessor systems we implemented the emulator of parallel DBMS. Emulator uses computing cluster with NVIDIA GPUs or Intel Xeon Phi coprocessors for relational query processing. It allows to simulate simple SELECT and JOIN queries. At this stage, for JOIN, only nested loops algorithm is implemented. Implementation was done with OpenMP, MPI and NVIDIA CUDA technologies. To achieve high level of parallelism, we use the principle of partitioned parallelism [8]. Each relation is partitioned into fragments, which are distributed over different worker processes (parallel agents). Worker processes exchange data using MPI technology.

III. EXPERIMENTS

Experiments with CPUs and Intel Xeon Phi were done on Tornado Supercomputer in South Ural State University, Chelyabinsk, Russia [9]. Experiments with GPUs were done on computing cluster in Lobachevsky State University, Nizhniy Novgorod, Russia. Experiments have used two CPUs or one coprocessor on each computing node. Characteristics of hardware are summarized in table I. To simulate SELECT and JOIN queries processing, we ran our emulator on computing clusters. The following simple queries were simulated:

1. SELECT * FROM Relation WHERE Attribute = Value;
2. SELECT * FROM Inner INNER JOIN Outer ON (Inner.Attribute1 = OuterAttribute2);

In this paper we consider a select on relation with cardinality of 370M that has size 5.5 GB. Join is done on following two relations: inner relation with cardinality of 33500 and size of 0.5 MB and outer relation that has size of 500MB and cardinality of 33.5M. Each relation consists of two integer (int64_t) attributes. Experiments include two steps. The first step is to evaluate query processing performance on a single computing node and determine settings that give maximum performance. In the first step we have run emulator on single computing node and varied its settings. For CPU version of algorithms, we have varied number of MPI processes. For GPU version, we have varied number of CUDA threads. For Manycore

TABLE I. TYPES OF COMPUTING NODES

	Type 1	Type 2
Processor	Intel Xeon X5680 3.33 GHz	Intel Xeon L5630 2.13 GHz
Number of processors/cores	2/12	2/8
RAM	24GB	24GB
Coprocessor	Intel Xeon Phi SE10X	NVIDIA Tesla X2070
Coprocessor memory	8GB	6GB

version, we have varied number of OpenMP threads. Results of the first step are summarized in table II. The second step is to evaluate query processing performance on a computing cluster, using settings from previous step In the second step we have run emulator on computing cluster and varied it's the number of nodes from 1 to 8. On each computing node, emulator has been run with settings from previous step. Results of second step are summarized in figure 1 and figure 2. As can be seen from figures, coprocessors are less efficient than CPUs for processing simulated SELECT query. Small speedup is explained by overhead costs for transferring data over the network and over PCI Express bus. In the case of INNER JOIN query, all versions of algorithms have similar speedup. However, coprocessors can process these queries faster, than CPUs. Thus, one can conclude that coprocessors can be effectively used for processing JOIN queries in multiprocessor database systems.

IV. CONLUSION

This work focuses on evaluating performance of manycore coprocessors and GPUs in multiprocessor database systems. We have implemented the emulator of parallel DBMS. This emulator uses computing cluster with NVIDIA GPUs or Intel Xeon Phi coprocessors for relational query processing. All algorithms have been implemented for CPUs, NVIDIA GPUs and Intel Xeon Phi coprocessors. A number of experiments have been done using this emulator. These experiments have shown that coprocessors are less efficient than CPUs for processing modeled SELECT queries, but more efficient than CPUs for processing INNER JOIN queries. Results of experiments will be used for extending DMM (Database Multiprocessor Model) [8]. Future work directions include:

1. Performance evaluation of heterogeneous computing systems in database applications.
2. Implementation and evaluation of more complicated JOIN algorithms, such as Hash Join for different types of coprocessors.
3. Design and implementation of algorithms that can use CPUs and coprocessor simultaneously for processing database queries.

ACKNOWLEDGMENT

This work was supported in part by the Russian Foundation for Basic Research, project no. 14-07-00420 and the Russian Federation president grant no. MK-3711.2013.9.

TABLE II. BEST EMULATOR SETTINGS FOR A SINGLE COMPUTING NODE

Algorithm	Time(s)	Settings
CPU JOIN	258	12 MPI processes
GPU JOIN	90	20736 CUDA threads
Manycore JOIN	213	240 OpenMP threads
CPU SELECT	22.96	12 MPI processes
GPU SELECT	55.73	5120 CUDA threads
Manycore SELECT	69.84	140 MPI threads

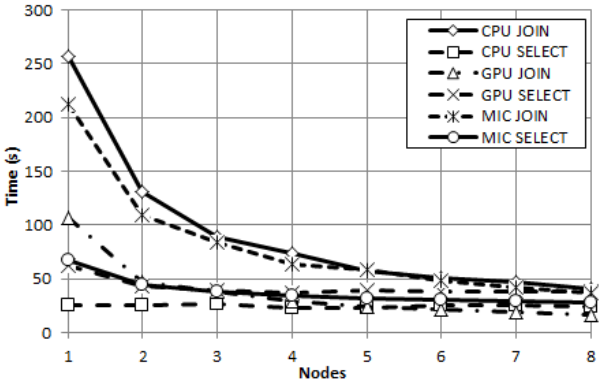


Figure 1. Query processing time

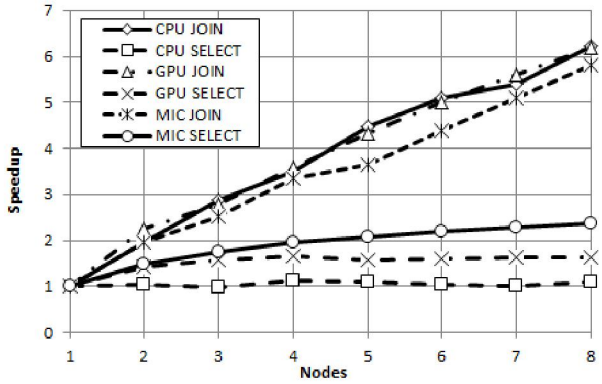


Figure 2. Query processing time

REFERENCES

[1] A. Di Blas and T. Kaldewey, "Data monster," IEEE Spectr., vol. 46, no. 9, pp. 46–51, Sep. 2009.

[2] B. He, K. Yang, R. Fang, M. Lu, N. Govindaraju, Q. Luo, and P. Sander, "Relational joins on graphics processors," in Proceedings of the 2008 ACM SIGMOD international conference on Management of data, ser. SIGMOD '08. New York, NY, USA: ACM, 2008, pp. 511–524

[3] C. Kim, J. Chhugani, N. Satish, E. Sedlar, A. D. Nguyen, T. Kaldewey, V. W. Lee, S. A. Brandt, and P. Dubey, "Fast: fast architecture sensitive tree search on modern cpus and gpus," in Proceedings of the 2010 ACM SIGMOD International Conference on Management of data, ser. SIGMOD '10. New York, NY, USA: ACM, 2010, pp. 339–350.

[4] C. Kim, J. Chhugani, N. Satish, E. Sedlar, A. D. Nguyen, T. Kaldewey, V. W. Lee, S. A. Brandt, and P. Dubey, "Designing fast architecture-sensitive tree search on modern multicore/many-core processors," ACM Trans. Database Syst., vol. 36, no. 4, pp. 22:1–22:34, Dec. 2011.

[5] N. Satish, C. Kim, J. Chhugani, A. D. Nguyen, V. W. Lee, D. Kim, and P. Dubey, "Fast sort on cpus, gpus and intel mic architectures," Intel Labs, Technical Report, 2010.

[6] D. G. Merrill and A. S. Grimshaw, "Revisiting sorting for gpgpu stream architectures," in Proceedings of the 19th international conference on Parallel architectures and compilation techniques, ser. PACT '10. New York, NY, USA: ACM, 2010, pp. 545–546.

[7] L. W. Volk P .B., Habich D., "Gpu-based speculative query processing for database operations," in First International workshop on accelerating data managment systems using modern processor and storage architectures in conjunction with VLDB, Sep. 2010.

[8] P. S. Kostenetskii, A. V. Lepikhov, and L. V. Sokolinskii, "Technologies of parallel database systems for hierarchical

multiprocessor environments," Autom. Remote Control, vol. 68, no. 5, pp. 847–859, May 2007.

[9] The Top 500, "Supercomputer Tornado SUSU of South Ural State University," http://www.top500.org/system/177983.

Simulacija dotoka autoregresivnim modelom prvog reda

S. Bošnjak, E. Sutlović
Sveučilište u Splitu, Fakultet elektrotehnike, strojarstva i brodogradnje, Split, R. Hrvatska
Stipo.Bosnjak@fesb.hr, Elis.Sutlovic@fesb.hr

Sažetak - U elektroenergetskim sustavima sa znatnim udjelom hidroenergije planiranje rada hidroelektrana je od izuzetnog značaja. Nepoznavanje i stohastički karakter budućih dotoka je glavni problem. Cilj planiranja je naći „optimalan“ način korištenja raspoložive količine vode nastojeći zadovoljiti brojna ograničenja. U optimizacijskom proračunu prirodni dotoci na određenoj lokaciji mogu se predstaviti deterministički, u kojem se dotoci tretiraju kao poznate veličine ili stohastički, u kojem se već u modelu ulaznih podataka u proračun nastoje ugraditi stohastičke karakteristike. Cilj rada je vrednovanje prikladnog stohastičkog modela za generiranje dotoka na nekom mjestu. Testirana su dva stohastička modela koji dotoke u različitim vremenskim periodima tretiraju kao međusobno ovisne slučajne varijable opisane autoregresivnim Markovljevim modelom prvog reda. Jedan model se koristi ako se dosadašnji dotoci na tom mjestu podvrgavaju normalnoj, a drugi ako se dosadašnji dotoci na tom mjestu podvrgavaju log-normalnoj razdiobi. Nakon izrade programa i simuliranja niza dotoka po oba modela, generiranim dotocima ispitivana su razna statistička obilježja: srednja vrijednost, standardna devijacija, faktori korelacije dotoka u uzastopnim vremenskim periodima itd. Generirani dotoci zadržali su sva ispitivana obilježja kao stvarni dotoci mjereni na tom mjestu u nizu godina. Oba ispitivana modela podataka su ocjenjeni kao pogodni za potrebe srednjoročnih i dugoročnih simuliranja i optimiranja rada hidroelektrana.

I. UVOD

U elektroenergetskim sustavima sa znatnim udjelom hidroenergije planiranje rada hidroelektrana, naročito hidroelektrana sa velikim akumulacijskim bazenima je od izuzetnog značaja. Kada vodu čuvati, a kada je i koliko ispustiti iz akumulacijskog bazena, utječe na moguću proizvodnju nizvodnih hidroelektrana, a posredno na potrebnu proizvodnju iz ostalih elektrana, na troškove pogona elektroenergetskog sustava, na profit itd. Cilj planiranja je naći „optimalan“ način korištenja raspoložive količine vode nastojeći zadovoljiti brojna ograničenja [1].

Proces planiranja obično se dijeli na kratkoročno (do sedam dana), srednjoročno (nekoliko mjeseci do godinu dana) i dugoročno (godina dana i duže). Dok se kod kratkoročnog planiranja redovito polazi od pretpostavke da su nam budući dotoci u riječne profile i akumulacijske bazene poznati, kod srednjoročnog i dugoročnog planiranja nepoznavanje i stohastički karakter budućih dotoka je glavni problem. Pristup problemu neizvjesnosti budućih dotoka određuje i karakter cjelokupnog optimizacijskog postupka, utječe na izbor optimizacijske metode, definiranje funkcije cilja i ograničenja itd.

Za rješavanje problema planiranja rada akumulacijskih bazena razvijeni su brojni optimizacijski proračuni koji koriste različite kriterije pri definiranju funkcije cilja (npr. maksimalna proizvodnja hidroelektrana, minimalni troškovi proizvodnje u EES-u, maksimalni profit proizvođača električne energije itd.) i različite matematičke optimizacijske metode (gradijentna metoda, linearno programiranje, dinamičko programiranje, itd.). U svim tim analizama cjelokupni period promatranja se dijeli u manje, jedinične vremenske periode odnosno diskretne vremenske korake unutar kojih se različite varijable i ulazni parametri ne mijenjaju. Obično se za osnovni vremenski period kod srednjoročnog i dugoročnog planiranja rada akumulacijskih bazena uzima jedan tjedan ili jedan mjesec [2] - [8].

U optimizacijskom proračunu prirodni dotoci na određenoj lokaciji mogu se predstaviti na više načina. Najjednostavniji način je deterministički model dotoka u kojem se svi dotoci (na svim lokacijama i u svim vremenskim periodima) tretiraju kao poznate veličine. Prednost determinističkog pristupa je što jednostavan model ulaznih podataka dozvoljava korištenje jednostavnijih i pouzdanijih optimizacijskih metoda koje sigurno vode ka „optimalnom“ rješenju. Jednostavnije optimizacijske metode mogu uzeti u obzir niz dodatanih ograničenja i utjecajnih parametara što kod složenijih optimizacijskih metoda može dodatno komplicirati proračun. Međutim, deterministički model ulaznih podataka zahtijeva čitav niz proračuna sa različitim mogućim scenarijima. Svaka kombinacija ulaznih podataka dat će različite rezultate stoga je nužna dodatna analiza i obrada odgovarajuće velikog broja izlaznih rezultata, čto često nije jednostavno. Kvaliteta dobivenih rezultata ovisi o broju analiziranih scenarija.

Za razliku od determinističkog pristupa, stohastički pristup pri simulaciji ili optimizaciji nastoji već u modelu ulaznih podataka i varijabli u proračun ugraditi stohastičke karakteristike. U stohastičkom modelu izlaz je po prirodi slučajan. Deterministički model je samo specijalan (pojednostavljen) slučaj stohastičkog modela. Za stohastičko modeliranje često se koristi tzv. nezavisni model koji dotoke (na svim lokacijama i u svim vremenskim periodima) tretira kao nezavisne slučajne varijable opisane statističkim parametrima ili pak zavisni

model podataka koji protoke tretira kao međusobno ovisne slučajne varijable opisane Markovljevim lancima

Cilj ovog rada je vrednovanje prikladnog zavisnog stohastičkog modela za generiranje dotoka na nekom mjestu predstavljenog autoregresivnim Markovljevim modelom prvog reda. Ispitivani model uzima u obzir stohastički karakter dotoka vodeći računa i o sezonskim varijacijama tijekom godine.

Izrađen je program za simulaciju budućih dotoka te je analizirano da li simulirani „budući“ dotoci zadržavaju karakteristike do sada zabilježenih dotoka na tom mjernom mjestu.

II. MODELIRANJE DOTOKA

Ako se zabilježeni dotoci u riječna korita ili akumulacijske bazene u razumno dugačkom periodu promatranja mogu opisati stacionarnim stohastičkim procesom (procesom čiji se parametri ne mijenjaju tijekom vremena), tada se može kreirati stohastički model dotoka prilagođen mjerenim podacima. Taj stohastički model može potom generirati proizvoljno dugačak niz pogodan za različite proračune, simulacije, optimiranja itd. Prvi korak u kreiranju prikladnog stohastičkog modela je ekstrakcija osnovnih statističkih karakteristika dotoka u različitim vremenskim periodima.

A. Model dotoka

Model dotoka kojeg ćemo ispitivati izmjerene dotoke (srednje dnevne, srednje mjesečne, ...) definira kao jednodimenzionalni vremenski niz opažanja bilježen u pravilnim vremenskim intervalima $t_0+n\tau$, n= 1, 2, ..., N. Za razliku od nezavisnih slučajnih uzoraka, kod vremenskog niza red opažanja je važan. To znači da opažanja nisu nezavisna. Slučajna varijabla čija se vrijednost mijenja u vremenu prema zakonu vjerojatnosti je stohastički proces. Određen vremenski niz se interpretira kao realizacija slučajnog procesa.

Značajno pojednostavljenje pri modeliranju čine slučajevi kada se budući razvoj procesa može predvidjeti samo pomoću trenutnog stanja, bez pozivanja na prošlost. To je Markovljevo svojstvo. Specijalni slučaj Markovljevog procesa je onaj kod kojeg stanja mogu biti poprimiti samo diskretne vrijednosti. Takav se proces zove Markovljev lanac. Radi pojednostavljenja modela, kontinuirani stohastički procesi se često aproksimiraju Markovljevim lancem.

Autoregresivni proces je niz slučajnih varijabli uzastopno definiranih određenim pravilom. Kod autoregresivnog modela izlazna varijabla je linearno ovisna o vlastitoj prethodnoj vrijednosti. Oznaka AR(*p*) označava autoegresivni proces reda p. AR(*p*) model se općenito definira kao

$$X_t = c + \sum_{i=1}^{p} \varphi_i X_{t-i} + \varepsilon_t \,, \tag{1}$$

gdje su φ_1, ..., φ_n parametri modela, c je konstanta a ε_t je niz nezavisnih normalnih slučajnih varijabli s očekivanjem nula i varijancom σ^2 (bijeli šum). Autoregresivni proces prvog reda AR(1) ovisan o jednom parametru se definira kao

$$X_t = c + \varphi X_{t-1} + \varepsilon_t \,. \tag{2}$$

Uobičajena i razumna pretpostavka je da se godišnji dotoci mogu tretirati kao Markovljev proces prvog reda. Izbor modela ovisi o raspoloživim podacima, karakteristikama dotoka i namjeni modela. Pretpostavit ćemo da imamo relevantne podatke o izmjerenim dotocima na nekom mjestu, sa poznatim statističkim parametrima. Fokusirati ćemo se na dva modela koji reproduciraju dotoke u kraćim vremenskim periodima (mjesecima ili tjednima) prikladna za potrebe srednjoročnog i dugoročnog planiranja rada hidroelektrana.

Ako se mjesečni dotoci na nekom mjestu podvrgavaju normalnoj distribuciji sa poznatim parametrima budući se dotoci mogu simulirati koristeći autoregresivni Markovljev model sa prvog reda [6 ,7]

$$\begin{aligned} i_{j,m} = \mu_m + r_m(\sigma_m / \sigma_{m-1})\left(i_{j,m-1} - \mu_{m-1}\right) + \\ + z_m \sigma_m \sqrt{1 - r_m^2} \,. \end{aligned} \tag{3}$$

Dotoci na nekom mjestu $i_{j,m}$ u godini j (j=1, 2, ,n) i mjesecu m (m=1, 2, ...,12) mogu se izračunati ako se poznaje prethodni dotok $i_{j,m-1}$, parametri normalne razdiobe (μ_m i σ_m) za svaki mjesec te koeficijenti korelacije r_m između dva uzastopna mjeseca. Ovi se parametri mogu izračunati na temelju višegodišnjeg mjerenja dotoka koristeći izraze

$$\mu_m = \frac{1}{n} \sum_{j=1}^{n} i_{j,m} \,, \tag{4}$$

$$\sigma_m = \sqrt{\frac{1}{n-1} \sum_{j=1}^{n} \left(i_{j,m} - \mu_m\right)^2} \,, \tag{5}$$

$$r_m = \frac{\sum_{j=1}^{n} (i_{j,m} - \mu_m)(i_{j,m-1} - \mu_{m-1})}{\sigma_m \sigma_{m-1}} \,. \tag{6}$$

Varijabla z_m u (1) je standardna normalna slučajna varijabla sa očekivanjem 0 i varijancom jednakoj 1. Uvjetna vjerojatnost dotoka u mjesecu m uz zadani dotok u prethodnom mjesecu

$$\mu_{m|m-1} = \mu_m + r_m(\sigma_m / \sigma_{m-1})\left(i_{j,m-1} - \mu_{m-1}\right) \tag{7}$$

je deterministička komponenta u izrazu (1) dok je uvjetna standardna devijacija

$$\sigma_{m|m-1} = \sigma_m \sqrt{1 - r_m^2} \tag{8}$$

pomnožena sa z_m slučajna komponenta koja predstavlja slučajno odstupanje dotoka neovisno o dotoku u prethodnom mjesecu. Za r_m=0 jednadžba (1) bi postala nezavisni model mjesečnih dotoka.

Ako se dotoci podvrgavaju log-normalnoj razdiobi, izraz (1) se može primjeniti na logaritme dotoka koji se podvrgavaju normalnoj razdiobi.

$$\ln(i_{j,m}) = \mu'_m + r_m'(\sigma'_m / \sigma'_{m-1})\left(\ln(i_{j,m-1}) - \mu'_{m-1}\right) + \\ + z_m \sigma_m' \sqrt{1 - r'^2_m}\ , \tag{9}$$

gdje su μ_m', σ_m' te μ_{m-1}', σ_{m-1}' srednja vrijednost i standardna devijacija logaritama protoka $\ln(i_{j,m})$ and $\ln(i_{j,m-1})$. Srednja vrijednost i standardna devijacija log-normalne razdiobe za sve jedinične periode može se izraziti pomoću srednje vrijednost i standardne devijacije normalne razdiobe [10]

$$\mu'_i = \ln \frac{\mu_i^2}{\sqrt{\sigma_i^2 + \mu_i^2}}\ , \tag{10}$$

$$\sigma'^2_i = \ln\left(1 + \frac{\sigma_i^2}{\mu_i^2}\right). \tag{11}$$

Koeficijent korelacije r_m' između $\ln(i_{j,m})$ and $\ln(i_{j,m-1})$ također se može izraziti preko koeficijenta korelacije r_m no precizniji rezultati se dobiju ako direktno mjerimo korelacije logaritama protoka [7].

Varijabla z_m u (7) je također standardna normalna slučajna varijabla sa očekivanjem 0 i varijancom jednakoj 1.

Ako normaln varijabla $\ln(i_{j,m})$ slijedi autoregresivni Markov model (7), tada odgovarajući dotok $i_{j,m}$ slijedi model

$$i_{j,m} = (i_{j,m-1})^{r_m' \frac{\sigma'_m}{\sigma'_{m-1}}} \cdot \exp\left[\mu'_m - r_m' \frac{\sigma'_m}{\sigma'_{m-1}} \mu'_{m-1}\right] \cdot \\ \cdot \exp\left[z_m \sigma_m' \sqrt{1 - r'^2_m}\right] \tag{12}$$

Autoregresivni modeli dotok (3) i (12) mogu se primjeniti i na tjedne dotoke, ako postavimo da je m index tjedna (m=1, 2, ..., 52), te da su poznate srednje vrijednosti, standardne devijacija i faktori korelacije između uzastopnih mjeseci.

B. Generiranje dotoka

Za simulaciju budućih dotoka napravljena su dva programa, jedan koji dotoke generira prema jednadžbi (3), a drugi koji dotoke generira prema jednadžbi (12). Rezultati rada programa su svaki put različiti jer se pri svakom pokretanju programa generira drugi skup slučajnih brojeva z_m za svaki vremenski period m. Programi se mogu koristiti za generiranje tjednih i mjesečnih protoka za niz godina jer se broj vremenskih perioda m (12 ili 52) i broj godina definiraju na početku.

III. PRIMJER

Za testiranje rada programa mogu se uzeti bilo kakvi početnii podaci, no mi smo ipak uzeli stvarne podatke o dotocima u jedan veliki akumulacijski bazen mjerenz dugi niz godina. Cilj testiranja je ispitati da li dotoci koji se generiraju pomoću programa za niz godina zadržavaju početne karakteristike.

A. Generiranje i analiza dotoka koji se podvrgavaju normalnoj razdiobi

U tablici 1 u prva tri retka su prikazani zadani ulazni podaci μ_m, σ_m i r_m za svih 12 mjeseci te prosječna vrijdnost ovih veličina tijekom godine. Sve vrijednosti dotoka prikazane u tablici, te srednje vrijednosti i varijance date su m^3/s. Npr. srednji godišnji protok iznosi 36,98 m^3/s. Jedino je faktor korelacije r veličina bez dimenzije.

Nakon toga su, u slijedećih 30 redaka prikazani dotoci generirani u programu za 30 godina u nizu. Također su u predzadnjoj kolone prikazane prosječne vrijednosti dotoka generirani za svaku godinu te razlika („pogreška“) između prosječnog godišnjeg protoka dobivenog simulacijom i zadanog prosječnog godišnjeg dotoka u postotcima.

U slijedeća tri retka prikazani su izračunati μ_m, σ_m i r_m za svih 12 mjeseci koji su dobiveni simulacijom, prosječni μ_m, σ_m i r_m dobiveni simulacijom te razlika prosječnih simuliranih i prosječnih zadanih vrijednosti ove tri veličine. U posljednja 3 retka prikazana je razlika za μ_m, σ_m i r_m u postotcima između zadanih i simuliranih vrijednosi dotoka za sve mjesece.

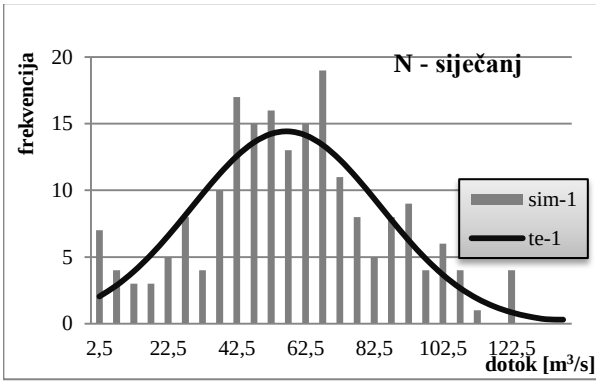
TABLICA 1 DOTOCI GENERIRANI PREMA (1) ZA NIZ OD 30 GODINA

MJE.	**1**	**2**	**3**	**4**	**5**	6	7	8	9	10	11	12	**avg**	**g%**
μ	56,94	43,20	47,23	58,90	41,98	24,69	15,55	12,34	16,97	20,02	42,29	63,62	**36,98**	
σ	27,52	17,24	18,07	22,19	13,56	6,18	4,18	5,69	13,94	11,14	22,03	31,58	**16,11**	
r	0,005	0,575	0,487	0,596	0,664	0,604	0,809	0,411	0,487	0,721	0,341	0,32	**0,5016667**	
1	88,08	62,90	84,54	83,70	56,21	25,26	14,02	12,50	27,84	24,27	40,33	57,83	**48,12**	30,14
2	46,94	43,31	41,32	48,42	57,92	30,46	17,80	17,95	32,38	37,95	34,67	56,96	**38,84**	5,03
3	76,53	47,19	44,97	54,88	9,12	17,63	11,60	-1,17	-6,71	13,47	12,47	105,49	**32,12**	-13,12
4	60,58	38,61	14,24	-8,91	6,10	11,54	7,95	8,75	15,53	26,29	57,46	34,79	**22,75**	-38,49
5	76,94	57,97	57,01	59,10	63,83	24,74	11,86	13,06	35,69	30,93	76,92	59,74	**47,32**	27,97
6	106,94	62,49	64,72	77,20	40,38	28,04	16,80	13,51	3,32	1,01	25,27	111,63	**45,94**	24,24
7	31,60	25,99	37,65	47,04	45,15	19,97	20,06	12,33	14,05	24,79	57,36	68,78	**33,73**	-8,78
8	58,50	36,66	49,27	71,03	35,33	20,86	18,17	13,67	19,05	11,06	24,19	58,39	**34,68**	-6,21
9	96,12	56,55	72,17	110,09	60,86	31,79	19,12	13,72	9,03	27,83	56,51	94,26	**54,00**	46,05
10	127,54	48,77	47,40	17,44	32,85	15,71	12,63	14,26	33,63	18,88	34,62	27,40	**35,93**	-2,84
11	62,33	38,77	53,54	29,52	24,99	18,41	12,91	18,78	19,94	26,15	49,06	88,08	**36,87**	-0,28
12	77,41	48,07	35,66	71,68	39,40	27,15	20,53	23,44	41,27	39,05	62,48	107,03	**49,43**	33,68
13	51,28	27,51	47,27	55,79	31,47	16,71	14,34	17,67	22,69	25,86	47,30	87,27	**37,10**	0,33
14	59,23	40,77	44,65	26,94	22,65	22,47	12,98	9,04	9,25	13,85	21,55	64,95	**29,03**	-21,50
15	68,72	44,66	64,60	59,35	46,39	26,12	15,87	10,59	15,75	13,14	8,06	59,27	**36,04**	-2,52
16	30,14	20,16	36,02	62,09	37,90	13,88	11,63	17,17	35,18	21,07	88,45	33,62	**33,94**	-8,21
17	51,27	49,05	35,75	29,51	34,84	22,43	16,86	12,76	35,82	41,56	46,81	44,90	**35,13**	-4,99
18	90,81	80,88	63,08	89,60	63,58	26,49	18,09	15,97	10,33	11,49	29,07	59,48	**46,57**	25,95
19	7,38	17,82	9,33	45,17	32,40	19,67	17,09	14,66	17,52	19,97	-10,78	4,72	**16,25**	-56,06
20	51,77	27,49	28,76	65,30	53,07	22,31	14,47	11,16	15,70	7,22	3,54	74,90	**31,31**	-15,33
21	80,59	47,38	42,27	56,35	48,37	26,42	14,26	6,71	21,93	17,46	25,33	76,18	**38,60**	4,40
22	36,59	29,90	17,52	22,74	28,94	20,05	13,69	9,11	12,31	17,86	15,26	40,25	**22,02**	-40,45
23	115,62	68,61	48,23	54,47	15,26	23,12	16,23	11,92	1,55	13,01	34,04	55,97	**38,17**	3,23
24	46,45	48,95	46,77	59,40	35,48	9,62	9,55	13,18	25,39	32,32	50,97	28,50	**33,88**	-8,37
25	91,86	42,42	24,99	37,24	35,19	23,61	16,13	15,86	40,29	27,18	53,82	34,24	**36,90**	-0,20
26	62,78	19,79	43,37	51,35	46,85	27,35	12,88	16,74	8,78	27,32	46,89	62,89	**35,58**	-3,77
27	62,25	36,81	30,02	30,59	12,96	12,16	10,16	10,40	27,47	25,09	18,96	108,12	**32,08**	-13,24
28	45,78	40,27	46,71	72,39	52,68	35,93	20,63	19,53	40,15	27,41	72,82	88,50	**46,90**	26,84
29	49,33	55,23	44,20	47,39	68,34	30,47	18,28	13,38	11,92	17,24	81,64	119,85	**46,44**	25,59
30	17,33	42,96	33,61	52,73	32,27	19,53	12,66	10,23	12,96	33,33	38,52	30,39	**28,04**	-24,16
avg μ	64,29	43,60	43,65	52,65	39,03	22,33	14,98	13,23	20,33	22,47	40,12	64,81	**36,79**	-0,50
avg σ	27,84	14,80	16,54	23,67	16,47	6,29	3,33	4,54	12,53	9,63	23,80	29,30	**15,73**	-2,37
r	0,390	0,780	0,680	0,659	0,691	0,741	0,670	0,439	0,659	0,694	0,518	0,106	**0,585**	16,69
g_μ%	12,91	0,91	-8,18	-11,85	-7,58	-10,57	-3,82	6,74	16,55	10,90	-5,42	1,85		
g_σ%	1,16	-16,52	-9,27	6,26	17,66	1,77	-25,50	-25,30	-11,26	-15,66	7,43	-7,78		
g_r%	7700,43	35,60	39,60	10,50	4,00	22,64	-17,14	6,84	35,24	-3,78	51,79	-66,94		

Analizirajući rezultate dane u tablici 1, ali i niza sličnih tablica koje se dobiju pri svakom novom pokretanju programa, možemo uočiti slijedeće:

- Razlika srednjih godišnjih protoka generiranih u programu i zadanog srednjeg godišnjeg protoka u nizu godina je obično unutar ±40%. Kod stvarnih podataka u zadnjih 30 godina razlike prema prosjeku su varirale od – 31% 2004. godine do +40% 2011. godine.
- Razlika simuliranog srednjeg godišnjeg protoka za svih 30 godina i izmjerenog srednjeg godišnjeg protoka je uvijek puno manja (unutar ±5%). U ovom primjeru +3,45 %.
- Srednje vrijednosti generiranih dotoka po mjesecima μ_m su također blizu polaznih vrijednosti. Razlike su obično unutar ±15%. Slično se ponašaju varijance generiranih dotoka po mjesecima σ_m.
- Nešto veće razlika između početnih i simuliranih vrijednosti se javljaju kod faktora korelacije r_m, odnosno velika razlika se javlja kod faktora r_m jer je zadani faktor korelacije izuzetno nizak, dok su razlike za ostale mjesece obično unutar ±25%.

Dotoci simulirani u programu moraju zadržati karakteristike raspodjele prema kojoj se raspodjela vrši. Rezultati simulacije dani u tablici 1 obuhvatili su samo 30 godina što je malo za točnije utvrđivanje pripadne raspodjele. Stoga smo napravili novi pokus sa istim početnim parametrima, ali smo generirali protoke za 200 godina. Time smo dobili 200 protoka za svaki mjesec koje smo grupirali u razrede i grafički prikazali. Na slici 2 stupčastim grafom je prikazana raspodjela simuliranih protoka u 1. Mjesecu, a glatkom krivuljom toeretska normalna razdioba za zadane parametre μ_1, σ_1. Standardnim χ^2 testom sa pragom od 0,05 ispitivana je hipoteza o „prilagođenosti“ simulirane i teoretske razdiobe za sve mjesece. U većini slučajeva hipoteza se mogla prihvatiti. Za slučaj na slici vrijednost χ^2 testa je 0,034343.

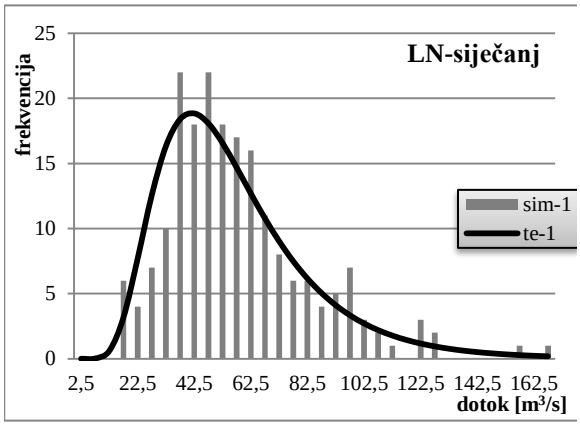


Slika 1. Toretska normalna raspodjela i raspodijela generiranih dotoka u prvom mjesecu za niz od 200 godina

B Generiranje i analiza dotoka koji se podvrgavaju log-normalnoj razdiobi

Kao što smo analizirali dotoke generirane prema normalnoj razdiobi, napravili smo analizu dotoka generiranih prema log-normalnoj razdiobi. Razlike statističkih parametar dotoka μ_m', σ_m' i r_m' dobivene simulacijom u odnosu na početne vrijednosti su u sličnim graničnim vrijednosti kao kod normalne razdiobe.

Također su simulirani dotoci zadržali željeni oblik log-normalne razdiobe što je provjeravano grafički i preko χ^2 testa. Na slici 2 stupčastim grafom je prikazana raspodjela simuliranih protoka u 1. mjesecu na uzorku od 200 dotoka, a glatkom krivuljom toeretska log-normalna razdioba za zadane parametre μ_1' i σ_1'. Vrijednost χ^2 testa u ovom prikazanom primjeru je 0,034343.



Slika 2. Toreska log-normalna raspodjela i raspodijela generiranih dotoka u prvom mjesecu za niz od 200 godina

IV. ZAKLJUČAK

Stohastički model podatka je onaj koji prepoznaje slučajnu prirodu ulaznih komponenti. U radu su testirana dva stohastička modela za simulaciju budućih dotoka na nekom mjestu koji dotoke u različitim vremenskim periodima tretiraju kao međusobno ovisne slučajne varijable opisane autoregresivnim Markovljevim lancima. Jedan model se koristi ako se dosadašnji dotoci na tom mjestu podvrgavaju normalnoj, a drugi ako se dosadašnji dotoci na tom mjestu podvrgavaju log-normalnoj razdiobi.

Nakon simuliranja niza dotoka po oba modela, generiranim dotocima ispitivana su razna statistička obilježja: srednja vrijednost, standardna devijacija, faktori korelacije dotoka u uzastopnim vremenskim periodima itd. Generirani dotoci zadržali su sva ispitivana obilježja kao dotoci mjereni na tom mjestu u nizu godina. Oba ispitivana modela podataka za simulaciju dotoka su relativno jednostavna linearna regresijska modela, robusna, primjenjiva za potrebe srednjoročnih i dugoročnih simuliranja i optimizacijskih proračuna rada hidroelektrana. Primjenu testiranih modela može otežat potreba za poznavanjem što točnijih karakteristika dotoka na nekom mjestu koje se mogu dobiti samo dugogodišnjim mjerenjem.

LITERATURA

[1] B. Udovičić, „Elektroenergetika“, Školska knjiga Zagreb, 1983.

[2] E. Buffet i ostali, „Stohastičko modeliranje“, Faculty and Institute of Actuaries , S interneta

[3] T. G. Siqueira, M. Zambelli, M. Cicogna, M. Andrade, S. Soares, „Stochastic dynamic programming for long-term hydrothermal scheduling considering different streamflow models“, *Proceedings of 9th International Conference on Probabilistic Methods Applied to Power Systems*, Stockholm, Sweden, June, 2006.

[4] M. Pereira, N. Campodónico and R. Kelman, „Long-term hydro scheduling based on stochastic models“, *Proceedings of EPSOM'98*, Zurich, September 23-25, 1998.

[5] A Gjelsvik, B. Mo. and A Haugstad, „Long- and medium-term operations planning and stochastic modelling in hydro-dominated power systems based on stochastic dual dynamic programming“, *Handbook of Power Systems I*, Energy Systems, pp 33-55, Springer 2010.

[6] D. P. Loucks and E. van Beek, „Water resources systems planning and management - An introduction to methods, models and applications“ – ISBN 92-3-103998-9 – © UNESCO 2005

[7] R. A. Wurbs, „Modeling and analysis of reservoir system operations,“ Prentice Hall PTR, USA, 1996.

[8] E. Sutlović, S. Čujić Čoko, I. Medic, “Characteristics of basin inflows - A statistical analysis for long-term/mid-term hydrothermal scheduling”, *Proceedings of SDEWES 2013*, September 2013, Dubrovnik, Croatia

[9] R. Žugaj, „Hidrologija“,Udžbenik Sveučilišta u Zagrebu, Zagreb, 2000.

[10] I. Pavlić, „Statistička teorija i primjena“, Tehnička knjiga, Zagreb, 1988.

Korelacijska analiza dotoka

K. Zoričić, E. Sutlović,
Sveučilište u Splitu, Fakultet elektrotehnike, strojarstva i brodogradnje, Split, R. Hrvatska
Kristian.Zoricic@fesb.hr, Elis.Sutlovic@fesb.hr

Sažetak – U okviru korelacijske analize ustanovljava se postojanje veze između slučajnih varijabli, njen oblik, jačina i smjer. Cilj rada je na osnovu relevantnih hidroloških podataka mjerenih u nizu godina na jednom mjernom mjestu napraviti korelacijsku analizu prirodnih dotoka grupiranih u manje vremenske periode. U radu su izračunate simetrične matrice Pearsonovih koeficijenata korelacije za srednje mjesečne dotoke i srednje tjedne dotoke. Analizirane su prosječne vrijednosti koeficijenata korelacije srednjih mjesečnih i srednjih tjednih dotoka za uzastopne periode, ali i za veće vremenske korake. Na kraju su računate korelacije među prirodnim logaritmima srednjih mjesečnih i srednjih tjednih protoka potrebne pri modeliranju dotoka koji se podvrgavaju log-normalnoj razdiobi. Uočene su razlike u prosječnoj jakosti veza obzirom na tip korelacijske veze i period promatranja. Dobiveni rezultati su značajni za razvoj modela podataka prirodnih dotoka za potrebe optimizacijskih proračuna.

I. UVOD

Optimalno planiranje i korištenje voda na nekom hidroenergetskom slivu temelji se na sustavnom višegodišnjem prikupljanju hidroloških podataka te njihovoj analizi i obradi. Međusobne veze u pojavi određenih dotoka značajno utječu na modeliranje dotoka u proračunima i složenost proračuna. U okviru korelacijske analize ustanovljava se postojanje veze, njen oblik, jačinu i smjer ne ulazeći u to što je uzrok a što posljedica.

Cilj rada je na osnovu relevantnih hidroloških podataka mjerenih u nizu godina napraviti korelacijsku analizu prirodnih dotoka. Treba istražiti i statistički definirati korelacijske veze među dotocima grupiranih u manje vremenske periode. Prema jakosti veze ocijeniti da li je opravdano zanemariti neke korelacijske veze u određenom tipu proračuna.

Izračunate su simetrične matrice Pearsonovih koeficijenata korelacije za srednje mjesečne dotoke i srednje tjedne dotoke. Analizirane su prosječne vrijednosti koeficijenata korelacije srednjih mjesečnih i srednjih tjednih dotoka za uzastopne periode, ali i za veće vremenske korake. Na kraju su računate korelacije među prirodnim logaritmima srednjih mjesečnih i srednjih tjednih protoka potrebne pri modeliranju dotoka koji se podvrgavaju log-normalnoj razdiobi. Uočene su razlike u prosječnoj jakosti veza obzirom na tip korelacijske veze i period promatranja.

II. KORELACIJSKA ANALIZA

Pojave u prirodi mogu biti povezane različitim vezama. Neka neki događaj ima dva obilježja x i y, gdje je x mjerena vrijednost jednog, a y mjerena vrijednost drugog obilježja. Nakon n mjerenja dobit ćemo skup od n parova (x_i,y_i), i=1,2, ..., n. Budući da je riječ o uređenim parovima, nakon n mjerenja rezultat ćemo moći predstaviti sa n točaka u koordinatnom sustavu. Skup tih točaka se zove dijagram rasprostiranja. Oblik dijagrama rasprostiranja otkriva zavisnost između obilježja x i y:

- kod funkcionalne veze točke (x_i,y_i) leže na jednoj krivulji, a ako je ta krivulja pravac govorimo o linearnoj funkcionalnoj vezi;
- kad obilježja x i y nisu vezana funkcionalnom vezom, ali dijagram ipak pokazuje tendenciju izduživanja u jednom smjeru govorimo o stohastičkoj vezi među slučajnim varijablama x i y;
- ako točke (x_i,y_i) ne pokazuju tendenciju gomilanja u nekom smjeru, obilježja x i y su nepovezana.

Kod stroge funkcionalne veze (ili determinističkog odnosa) za svaku vrijednost x se točno zna vrijednost varijable y. Kada su slučajne varijable x i y stohastički povezane, za jednu vrijednost od x dolazi više vrijednosti od y koje su raspoređene po nekom zakonu vjerojatnosti. Stoga se pri stohastičkoj vezi za određeni x može samo odrediti interval kojem se ta vrijednost nalazi i to uz izvjesnu vjerojatnost.

Jakost stohastičke veze mjeri se koeficijentom korelacije ρ. Koeficijent linearne korelacije osnovnog skupa (populacije), ili Pearsonov koeficijent definiramo kao

$$\rho = \frac{ko\,\mathrm{var}\,ijanca(x,y)}{\mathrm{var}\,ijanca(x)\cdot \mathrm{var}\,ijanca(y)} = \frac{\sigma_{xy}}{\sigma_x^2\sigma_y^2} \qquad (1)$$

Vrijednost koeficijenta korelacije je $-1 \geq \rho \geq +1$. Za $\rho=\pm 1$ govorimo o funkcionalnoj vez dviju varijablii, a za $\rho=0$ varijable su nezavisne. U ostalim slučajevima imamo stohastičku vezu među varijablama. Općenito labavijoj vezi odgovara manji $|\rho|$, a čvršćoj veći $|\rho|$.

Nepristranu procjenu koeficijenta korelacije ρ, označimo sa r. Tada je koeficijent linearne korelacije uzorka [9]

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}} = \frac{s_{xy}}{s_x^2 s_y^2}. \qquad (2)$$

Nepristrana procjena kovarijanca varijabli x i y je

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) . \quad (3)$$

Nepristrana procjena varijance varijable x je

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 . \quad (4)$$

Nepristrana procjena varijance varijable y je

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2 . \quad (5)$$

U primjeni se stupanj jakosti korelacije r obično definira na slijedeći način

| r |	**Jakost korelacije**
0	nema korelacije
0 - 0,5	slaba korelacija
0,5 - 0,8	srednje jaka korelacija
0,8 - 1	jaka korelacija
1	potpuna korelacija

Kada imamo više varijabli i zanima nas veza među svim kombinacijama varijabli možemo kreirati matricu korelacija. Redovi i kolone matrice predstavljaju promatrane varijable, a članovi matrice su koeficijent korelacije. Dobivena matrica je simetrična ($r_{ij}=r_{ji}$), pa je dovoljno promatrati gornji ili donji trokut matrice. Svi dijagonalni članovi su jedinice jer je svaka varijabla sama sa sobom u potpunoj korelaciji.

Ako su varijable članovi nekog vremenskog niza, stupanj k označava razmak između osnovnih perioda, a r(k) koeficijent korelacije stupnja k. Za k=1 imamo koeficijent korelacije susjednih vremenskih perioda itd.

Međutim, Pearsonov koeficijent korelacije r može se računati na gornji način kod linearne povezanosti varijabli koje imaju simetričnu raspodjelu. Prirodni dotoci u pravilu nemaju simetričnu distribuciju jer dotoci ne mogu biti negativni. Distribucija prirodnih dotoka u pravilu ima pozitivni koeficijent asimetrije, donja granica protoka je obično blizu nule, a gornja je gotovo neograničena. Distribucija prirodnih dotoka obično više naliči log-normalnoj nego normalnoj distribuciji.

U takvim slučajevima općenito je preporučena metoda transformacije slučajnih varijabli u varijable koja se podvrgava normalnoj (simetričnoj) distribuciji i korištenje koeficijenata korelacije među simetričnim distribucijama definirane na prethodni način [2].

Ako se npr. slučajna varijabla Q, koja predstavlja protok na nekom mjestu, podvrgava dvoparametarskoj logaritamsko-normalnoj distribucijom, tada se ln(Q) podvrgava normalnoj distribuciji. Očekivanje, varijanca, koeficijent asimetrije itd. log-normalne razdiobe mogu se jednostavno izračunati iz očekivanja, varijanca i koeficijenta asimetrije normalne razdiobe koristeći dobro poznate izraze [8].

Kod kasnijeg modeliranja protoka Q koji se podvrgavaju log-normalnoj razdiobi, veza stupnja k među protocima u periodu u i v označena sa $\rho_{LN}(k,u,v)$ može se odrediti iz veze stupnja k među transformiranim varijablama ln(Q) koje se podvrgavaju normalnoj razdiobi označena sa $\rho_N(k,u,v)$ i njihovih varijanci $\sigma^2_N(u)$ i $\sigma^2_N(v)$ [2] koristeći izraz

$$\rho_{LN}(k,u,v) = \frac{\exp\left[\rho_N(k,u,v)\cdot\sigma_N(u)\cdot\sigma_N(v)\right]-1}{\sqrt{\exp(\sigma_N^2(u)-1)}\cdot\sqrt{\exp(\sigma_N^2(v)-1)}} \quad (6)$$

Mnogo efikasnija metoda estimacije stvarne zavisnosti među protocima koji se podvrgavaju log-normalnoj razdiobi je transformacija zabilježenih protoka u ekvivalente ln(Q) koji se podvrgavaju normalnoj razdiobi te korištenje zabilježenih korelacija između transformiranih varijabli kao estimacija za $\rho_N(k,u,v)$.

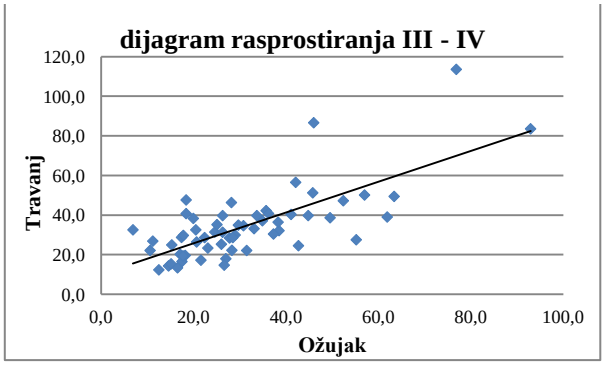
III. PRORAČUN MATRICA KORELACIJE

Korelacijsku analizu dotoka proveli smo nad mjerenim podacima dotoka u rijeku Krku. Imali smo dva niza mjerenja zabilježenih na istom mjernom mjestu. To mjerno mjesto se nalazi uzvodno od svih vodnih zahvata na slivu pa predstavljaju prirodne dotoke. Prvi niz su zabilježene vrijednosti srednjih mjesečnih dotoka u periodu od 59 godina (1950 – 2013 s prekidom 1991-1995). Drugi niz je kontinuirani niz srednjih dnevnih dotoka zabilježen u posljednjih 17 godina (1997-2013).

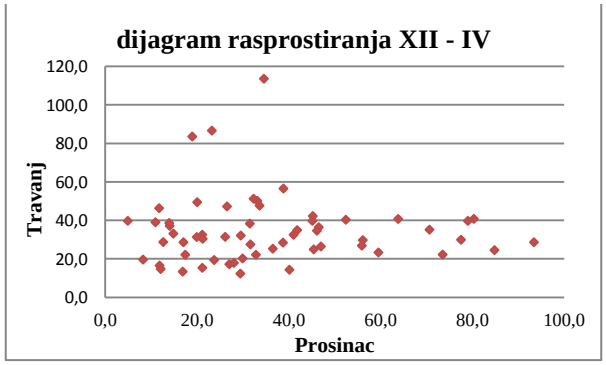
Na uzorku srednjih mjesečnih dotoka u periodu od 59 godina izračunata je matrica 12*12 Pearsonovih koeficijenata linearne korelacije (tablica1). Matrica je simetrična pa je naglašen samo gornji trokut. Prema očekivanju koeficijenti korelacije između susjednih mjeseci (k=1) su značajno veći nego koeficijenti korelacije sa k>1. Na slici 1 i slici 2 prikazani su dijagrami rasprostiranja i to u prvom slučaju kad dijagram pokazuje tendenciju izduživanja (r=0,728) i kada točke ne pokazuju tendenciju gomilanja (r=0,029). U drugom stupcu tablici 2 sažeto su prikazani prosječni iznosi koeficijenta korelacije srednjih mjesečnih dotoka u periodu od 59 godina iz tablice 1 za vremenski korak 1 do 4.

TABLICA 1. MATRICA KOEFICIJENATA KORELACIJE SREDNJIH MJESEČNIH DOTOKA U PERIODU OD 59 GODINA

	1	2	3	4	5	6	7	8	9	10	11	12
1	1	0,497	0,376	0,284	0,126	0,197	0,144	0,058	0,020	0,011	0,142	0,023
2	0,497	1	0,643	0,458	0,350	0,288	0,253	0,256	0,178	0,012	0,147	0,114
3	0,376	0,643	1	0,728	0,476	0,315	0,147	0,121	0,012	0,077	0,019	0,136
4	0,284	0,458	0,728	1	0,505	0,187	0,068	0,001	0,068	0,168	0,025	0,029
5	0,126	0,350	0,476	0,505	1	0,440	0,384	0,262	0,020	0,040	0,124	0,051
6	0,197	0,288	0,315	0,187	0,440	1	0,676	0,432	0,232	0,190	0,103	0,201
7	0,144	0,253	0,147	0,068	0,384	0,676	1	0,600	0,154	0,244	0,152	0,107
8	0,058	0,256	0,121	0,001	0,262	0,432	0,600	1	0,464	0,204	0,283	0,073
9	0,020	0,178	0,012	0,068	0,020	0,232	0,154	0,464	1	0,090	0,078	0,089
10	0,011	0,012	0,077	0,168	0,040	0,190	0,244	0,204	0,090	1	0,449	0,114
11	0,142	0,147	0,019	0,025	0,124	0,103	0,152	0,283	0,078	0,449	1	0,495
12	0,023	0,114	0,136	0,029	0,051	0,201	0,107	0,073	0,089	0,114	0,495	1



Slika 1. Dijagram rasprostiranja između protoka u ožujku i travnju na uzorku od 59 godina



Slika 2. Dijagram rasprostiranja između protoka u prosincu i travnju na uzorku od 59 godina

Drugi niz podataka o srednjim dnevnim protocima grupirali smo u manje vremenske jedinice, mjesece i tjedne te dobili niz srednjih mjesečnih protoka za svih 12 mjeseci i niz srednjih tjednih protoka za sva 52 tjedna u nizu od 17 godina. Potom smo još jednom izračunali koeficijente korelacije srednjih mjesečnih dotoka, ali ovaj put na uzorku od posljednjih 17 godina. Također smo dobili matricu 12*12 Pearsonovih koeficijenata linearne korelacije. U trećem stupcu tablice 2 prikazani su prosječni iznosi koeficijenta korelacije srednjih mjesečnih dotoka u periodu od 17 godina za vremenski korak 1 do 4. Nešto veći iznosi prosječnih koeficijenata korelacije u drugom nizu podataka su vjerojatno posljedica kratkog perioda mjerenja 2. niza te što je u drugi niz kontinuiran. Cilj ove analize nije točan iznos koeficijenata korelacije na određenom mjernom mjestu, za što je potreban što duži niz mjerenja, već usporedba srednjih mjesečnih i srednjih tjednih koeficijenata korelacije, a to je najbolje raditi na istom uzorku.

Obzirom da je za očekivat kako se srednji mjesečni dotoci ne podvrgavaju simetričnoj normalnoj već nekoj asimetričnoj razdiobi tipa log-normalne, izvršena je transformacija prirodnih srednjih mjesečnih dotoka Q na uzorku od 17 godina u logaritme tih dotoka te ponovno izračunata matrica koeficijenata linearne korelacije među varijablama koje se sada sigurno podvrgavaju normalnoj simetričnoj razdiobi. Sažeti prikaz prosječnih koeficijenata korelacije između dva mjeseca u ovom slučaju je prikazan u drugoj koloni tablice 3.

Pearsonovi koeficijenata linearne korelacije srednjih tjednih dotoka računati su u nizu od posljednjih 17 godina. Rezultat analize je simetrična matrica veličine 52*52. Sažeti prikaz prosječnih koeficijenata korelacije između dva tjedna na uzorku od 17 godina je prikazan u četvrtoj koloni tablice 2.

Na kraju smo, slično kao kod srednjih mjesečnih dotoka, izvršili transformacija srednjih tjednih dotoka Q na uzorku od 17 godina u logaritme tih dotoka te ponovno izračunali matricu veličine 52*52 koeficijenata linearne korelacije logaritama protoka. U trećoj koloni tablice 3 dan je sažeti prikaz prosječnih koeficijenata korelacije u ovom slučaju.

TABLICA 2. PROSIJEČNA VRIJEDNOST KOEFICIJENTA KORAELACIJE PROTOKA Q ZA RAZLIČITE VREMENSKE KORAKE

Korak	Prosječni koef. korelacije između dva mj. na uzorku od 59 godina	Prosječni koef. korelacije između dva mj. na uzorku od 17 godina	Prosječni koef. korelacije između dva tj. na uzorku od 17 godina
1	0,508	0,599	0,743
2	0,273	0,400	0,572
3	0,208	0,302	0,491
4	0,095	0,213	0,440

TABLICA 3. PROSIJEČNA VRIJEDNOST KOEFICIJENTA KORAELACIJE LOGARITAMA PROTOKA LN(Q) ZA RAZLIČITE VREMENSKE KORAKE

Korak	Prosječni koeficijent korelacije između dva mjeseca na uzorku od 17 godina	Prosječni koeficijent korelacije između dva tjedna na uzorku od 17 godina
1	0,656	0,838
2	0,402	0,678
3	0,292	0,589
4	0,236	0,516

Za srednje mjesečne protoke napravili smo proračun koeficijenata korelacije prema (6), na osnovu koeficijenata korelacije izračunatih za logaritme protoka. Rezultati su prikazani u tablici 4. Iznosi koeficijenata stvarne korelacije među protocima r_{LN} su neznatno niži neko r_N iz kojih se računaju. Proračune koeficijenata korelacije r_{LN} za srednje mjesečne protoke nismo računali.

TABLICA 4. PROSIJEČNA VRIJEDNOST KOEFICIJENTA KORALACIJE STVARNIH PROTOKA r_{LN} RAČUNATA IZ KORELACIJE LOGARITAMA PROTOKA r_N ZA RAZLIČITE VREMENSKE KORAKE

Korak	Prosječni koef. korelacije između dva mjeseca na uzorku od 17 godina
1	0,619
2	0,367
3	0,265
4	0,209

Uspoređujući rezultate u tablicama 2 i 3 vidimo da su korelacijske veze između tjedana uvijek jače nego korelacijske veze između mjeseci, što je bilo očekivano. Dotoci su općenito jače korelirani za kraće vremenske intervale.

Analizirajući iznose koeficijenata korelacijskih veza možemo uočiti da se koeficijenti korelacije između uzastopnih mjeseci kreću oko 0,6, što predstavlja srednje jaku korelacijsku vezu. Iznosi koeficijenata korelacije za k>1 su uvijek manji od 0,5, što označava slabe korelacijske veze. Prema tome, kod modeliranja srednji mjesečnih dotoka, precizniji model treba uzeti u obzir veličinu dotoka u prethodnom periodu, a ostale može zanemariti. Jednostavniji i manje precizan model može zanemariti utjecaj svih prethodnih dotoka.

Koeficijenti korelacije srednjih tjednih dotoka za k=1 od približno 0,8 govore o srednje jakoj i jakoj korelacijskoj vezi susjednih tjedana. Čak i koeficijenti korelacije srednjih tjednih dotoka za k=2 nisu beznačajni. Dakle, kod modeliranja srednji mjesečnih dotoka, manje precizan model treba uzeti u obzir veličinu dotoka u prethodnom periodu, a ostale može zanemariti. Precizniji model može uzetu u obzir i korelacijske veze sa korakom k>1, ili na neki drugi način uzeti u obzir utjecaje ostalih korelacijskih veza. Jedan od načina je npr. računanje parcijalnog koeficijenta korelacije između dviju varijabli uz „uklanjanje“ utjecaja drugih nezavisnih varijabli.

Za kraj uočimo da se svi naši računati koeficijenti korelacije približno pokoravaju uobičajenoj formuli koja se koristi pri modeliranju koeficijenata korelacije višeg reda [2].

$$r(k) = [r(1)]^k = r^k . \quad (7)$$

IV. ZAKLJUČAK

Cilj rada je napraviti korelacijsku analizu prirodnih dotoka na osnovu relevantnih hidroloških podataka mjerenih u nizu godina. Prema jakosti veze ocijeniti da li je opravdano zanemariti neke korelacijske veze. Izračunate su simetrične matrice Pearsonovih koeficijenata korelacije za srednje mjesečne dotoke i srednje tjedne dotoke. Analizirane su prosječne vrijednosti koeficijenata korelacije srednjih mjesečnih i srednjih tjednih dotoka za uzastopne periode, ali i za veće vremenske korake.

Korelacijske veze između tjedana uvijek su jače nego korelacijske veze između mjeseci, što je bilo očekivano. Uzastopne mjesece karakterizira srednje jaka korelacijska veza, a ostale korelacijske veze među mjesecima su slabe i zanemarive. Uzastopne tjedne karakterizira srednje jaka i jaka korelacijska veza. Korelacijske veze koraka 2 između tjedana također nisu zanemarive. O ovim činjenicama treba voditi računa prilikom modeliranja mjesečnih i tjednih protoka za različite proračune

LITERATURA

[1] R. A. Wurbs, „Modeling and analysis of reservoir system operations,“ Prentice Hall PTR, USA, 1996.

[2] D. P. Loucks and E. van Beek, „Water resources systems planning and management - An introduction to methods, models and applications“ – ISBN 92-3-103998-9 – © UNESCO 2005

[3] T. G. Siqueira, M. Zambelli, M. Cicogna, M. Andrade, S. Soares, „Stochastic dynamic programming for long-term hydrothermal scheduling considering different streamflow models“, *Proceedings of 9th International Conference on Probabilistic Methods Applied to Power Systems*, Stockholm, Sweden, June, 2006.

[4] M. Pereira, N. Campodónico and R. Kelman, „Long-term hydro scheduling based on stochastic models“, *Proceedings of EPSOM'98*, Zurich, September 23-25, 1998.

[5] A Gjelsvik, B. Mo. and A Haugstad, „Long- and medium-term operations planning and stochastic modelling in hydro-dominated power systems based on stochastic dual dynamic programming“, *Handbook of Power Systems I*, Energy Systems, pp 33-55, Springer 2010.

[6] E. Sutlović, S. Čujić Čoko, I. Medic, "Characteristics of basin inflows - A statistical analysis for long-term/mid-term hydrothermal scheduling", *Proceedings of SDEWES 2013*, September 2013, Dubrovnik, Croatia

[7] R. Žugaj, „Hidrologija“,Udžbenik Sveučilišta u Zagrebu, Zagreb, 2000.

[8] I. Pavlić, „Statistička teorija i primjena“, Tehnička knjiga, Zagreb, 1988.

[9] N. Šakić, „Korelacijska i regresijska analiza“, Sveučilišta u Zagrebu, Fakultet strojarstva i brodogradnje, Zagreb, 2010.

Statistička analiza funkcije vjerojatnosti prirodnih dotoka

I. Perišić, E. Sutlović
Sveučilište u Splitu, Fakultet elektrotehnike, strojarstva i brodogradnje, Split, R. Hrvatska
Ivan.Perisic.Joso@fesb.hr, Elis.Sutlovic@fesb.hr

Sažetak - Za srednjoročno i dugoročno planiranje rada hidroelektrana obično se ciljani period promatranja od nekoliko mjeseci do godine dana dijeli u manje vremenske intervale, dugačke jedan tjedan ili jedan mjesec. U radu su analizirani višegodišnji rezultati mjerenja prirodnih dotoka na jednom mjernom mjestu grupirani u mjesečne i tjedne intervale. Ispitano je da li teorijske funkcije vjerojatnosti kontinuiranih razdioba, koje se uobičajeno koriste u hidrologiji za modeliranje dotoka, dobro aproksimiraju empirijske podatke prirodnih dotoka koje smo imali na raspolaganju. Ispitivanje kompatibilnosti serija srednjih mjesečnih i serija srednjih tjednih empirijskih podataka s teorijskim funkcijama raspodjele vjerojatnosti je provedena sa χ^2 (hi-kvadrat) testom. Prema veličini χ^2 testa rangirali smo distribucije u svakom periodu promatranja. Cilj rada je naći tip distribucije koja se dobro prilagođava mjerenim podacima u velikom broju promatranih osnovnih vremenskih perioda, a također je relativno jednostavna za korištenje - po mogućnosti dvoparametarska distribucija i da se parametri distribucije određuju jednostavnim analitičkim postupkom. Na osnovu tog tipa distribucije koji najbolje opisuje učestalost pojavljivanja protoka po mjesecima ili tjednima kasnije želimo graditi „univerzalni" model dotoka u nekom osnovnom vremenskom periodu.

I. UVOD

Optimalno korištenje vodnih potencijala temelji se na dugogodišnjim hidrološkim i meteorološkim mjerenjima te statističkim obradama izmjerenih podataka. Pouzdani podaci prikupljeni na odgovarajući način podloga su prvo za planiranje razvoja i izgradnje hidroenergetskih sustava, a potom za planiranje optimalnog korištenja raspoloživih količina vode na nekom slivu. Statistička analiza prikupljenih podataka i određivanje prikladnih funkcija vjerojatnosti je znanstveni pristup problemu neizvjesnosti prirodnih dotoka [1]-[6].

Srednjoročno i dugoročno planiranje rada hidroelektrana izučava moguće scenarije u budućem ciljanom periodu od nekoliko mjeseci do godine dana, ponekad i duže. Neizvjesnost i stohastički karakter budućih dotoka je glavni uzrok složenosti ovakvih analiza. Ukupni ciljani period planiranja se obično dijeli u manje, osnovne ili jedinične vremenske intervale, dugačke jedan tjedan ili jedan mjesec. Pretpostavka je da se u osnovnim vremenskim intervalima ne mijenjaju vrijednosti varijabli u proračunu. Isto tako možemo pretpostaviti da će budući dotoci, koji su uvijek dio ulaznih podataka raznih analiza i optimizacijskih proračuna, zadržati statistička obilježja kakva su imali do sada.

Korištenje prethodno zabilježenih vrijednosti neke slučajne varijable, npr. prirodnog dotoka, za simuliranje mogućih vrijednosti koju ta varijabla može poprimiti u budućnosti je prilično nepraktično. Nepoznata je matematička forma koja bi opisala taj događaj. Jednostavnije je tim empirijskim podacima prilagoditi neku teorijsku distribuciju, a potom koristiti matematičku formu teorijske distribucije za predstavljanje empirijske raspodjele učestalosti, generiranje mogućih vrijednosti, modeliranje u optimizacijskim proračunima itd.

Poznat je veliki broj funkcija vjerojatnosti koje se mogu koristiti u različitim područjima za statistički opis podataka i više statističkih metoda koje pomažu pri odabiru prikladne funkcije učestalosti. Obično treba prilagoditi nekoliko teorijskih distribucija, koje po svojim karakteristikama odgovaraju prirodi problema kojim se bavimo, usporediti rezultate i izabrati najbolju. Nakon toga potrebno je odrediti kako dobro se odabrana teorijska distribucija prilagođava mjerenim podacima koristeći prikladne testove. Prema rezultatima testa, odnosno stupnju prilagodbe, bira se model koji najbolje opisuje stvarne podatke.

U radu smo analizirati višegodišnje rezultate mjerenja dotoka na jednom mjernom mjestu i tražili prikladnu teorijsku funkcija vjerojatnosti za opis izmjerene učestalosti pojavljivanja tih dotoka u jediničnim vremenskim intervalima. Svi prikupljeni podaci su mjereni u jednoj mjernoj stanici na isti način kako bi se smanjila moguća greška pri mjerenju. Nakon sortiranja i grupiranja izmjerenih podataka ispitano je da li teorijske funkcije vjerojatnosti kontinuiranih razdioba koje se uobičajeno koriste u hidrologiji za modeliranje dotoka (normalna, log-normalna, gama, log-gama, Pearson tip 3, log-Pearson tip 3, Weibull, Gumbel Max, …) dobro aproksimiraju empirijske podatke dotoka koje smo imali na raspolaganju. Ispitivanje kompatibilnosti serija srednjih mjesečnih i serija srednjih tjednih empirijskih podataka s teorijskim vjerojatnosti raspodjele funkcija je provedena sa χ^2 (hi-kvadrat) testom. Prema veličini χ^2 testa rangirali smo distribucije u svakom periodu promatranja.

Cilj rada je naći tip distribucije koja se dobro prilagođava mjerenim podacima u velikom broju promatranih osnovnih vremenskih perioda, a također je

relativno jednostavna za korištenje. Jednostavna za korištenje znači da je po mogućnosti dvoparametarska distribucija i da se parametri distribucije određuju jednostavnim analitičkim postupkom. Na osnovu tog tipa distribucije koji najbolje opisuje učestalost pojavljivanja protoka po mjesecima ili tjednima kasnije želimo graditi „univerzalni“ model dotoka u nekom osnovnom vremenskom periodu.

Ispitivanja smo proveli koristeći vlastita programska rješenja i jedan dostupan računalni program sa s gotovim algoritmima obrade.

II. STATISTIČKA ANALIZA PRIRODNIH DOTOKA

A. Osnovna statistička obrada nizova podataka

Kada distribucija slučajne varijable nije poznata, ali je raspoloživ niz od N nezavisnih mjerenja (x_1, ..., x_N), numeričke karakteristike nepoznate distribucije slučajne varijable x mogu se procijeniti na osnovu empiričkih mjerenja na uzorku veličine N koristeći slijedeće izraze [1].

Nepristrana procjena očekivanja μ varijable x je aritmetička sredina slučajnog uzorka

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \,. \tag{1}$$

Nepristrana procjena varijance σ^2 varijable x jest veličina definirana izrazom

$$\hat{\sigma}_x^2 = s_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2 \,. \tag{2}$$

Nepristrana procjena koeficijent asimetrije α_3 varijable x jest veličina definirana izrazom

$$\hat{\alpha}_3 = Cs_x = \frac{N}{(N-1)(N-2)}\sum_{i=1}^{N}\frac{(x_i - \bar{x})^3}{{s_x}^3} \,. \tag{3}$$

Prirodni dotoci su uvijek pozitivne veličine s asimetričnim oblikom distribucije čiji je koeficijent asimetrije u pravilu pozitivan. Donja granica protoka je obično blizu nule, a gornja je gotovo neograničena.

B. Teorijske razdibe kontinuirane slučajne varijable

Za modeliranje učestalosti srednjih tjednih i srednjih mjesečnih dotoka razmatrano je nekoliko jednostavnih kontinuiranih distribucija koje imaju slijedeće karakteristike:

- Definirane su na području vrijednosti varijable x (0,+∞) kao i empirički podaci. Izuzetak je normalna razdioba koje je, kao najčešća kontinuirana razdioba, prikazana radi usporedbe s drugim razdiobama;
- Oblik funkcije vjerojatnosti razmatrane teorijske razdiobe i pridruženi statistički parametri (koeficijent simetričnosti itd.) približno su jednaki karakteristikama histograma zabilježenih vrijednosti dotoka;
- razdiobe se uobičajeno koriste u hidrologiji, a parametri razdiobe se daju lako estimirati bez posebnih statističkih programa za estimiranje.

Teorijske funkcije vjerojatnosti i izrazi za estimiranje parametara razmatranih raspodjela dani su u nastavku.

Normalna (ili Gussova) razdioba slučajne varijable x definirana je u rasponu (-∞, +∞), simetrična je (α_3=0) te ima karakterističan zvonoliki oblik opisan funkcijom vjerojatnosti [7]

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \,. \tag{4}$$

Normalna razdioba je jednoznačno određena očekivanjem μ i varijancom σ^2 pa se uvodi i kraća oznaka za tu razdiobu $N\{\mu, \sigma^2\}$.

Da bismo empiričkoj distribuciji prilagodili normalnu razdiobu, aproksimirat ćemo očekivanje μ aritmetičkom sredinom $\bar{x}$ empiričkih podataka (1), a varijancu σ^2 opet varijancom empiričkih podataka s_x^2 (2).

Logaritamsko-normalna (ili Galtonova) razdioba varijable x je razdioba kod koje je lnx distribuirano po zakonu normalne razdiobe, a područje vrijednosti varijable x je (0, +∞). Dakle, ako y = lnx pripada normalnoj raspodjeli $N\{\mu_y, \sigma_y^2\}$, funkcija vjerojatnosti varijable x glasi [7]

$$f(x) = \frac{1}{x\sigma_y\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\ln x-\mu_y}{\sigma_y}\right)^2} \,. \tag{5}$$

Očekivanje, varijanca i koeficijent asimetrije varijable x glase:

$$\mu_x = e^{\mu_y + \sigma_y^2/2} \,, \tag{6}$$

$$\sigma_x^2 = e^{2\mu_y+\sigma_y^2}(e^{\sigma_y^2} - 1) \,, \tag{7}$$

$$\alpha_{3x} = (e^{\sigma_y^2} + 2)\sqrt{e^{\sigma_y^2} - 1} \,. \tag{8}$$

Logaritamsko-normalna razdioba je potpuno određena sa očekivanjem μ_x i varijancom σ_x^2 varijable x, što kraće pišemo $N_L\{\mu_x, \sigma_x^2\}$. Očekivanje μ_y i varijancu σ_y^2 varijable y možemo izraziti očekivanjem i varijancom varijable x:

$$\mu_y = \ln \frac{\mu_x^2}{\sqrt{\sigma_x^2 + \mu_x^2}}, \quad (9)$$

$$\sigma_y^2 = \ln\left(1 + \frac{\sigma_x^2}{\mu_x^2}\right). \quad (10)$$

Prilikom prilagođavanja log-normalne razdiobe empiričkim podacima, aritmetičku sredinu $\overline{X}$ i varijancu s_x^2 empiričkih podataka uvrstit ćemo na mjesto μ_x i σ_x^2 u gornje izraze i dobiti aproksimativne vrijednosti za μ_y i σ_y^2.

Gama razdioba je familija dvo-parametarskih funkcija raspodjele za kontinuiranu slučajnu varijablu. Za varijablu x kažemo da je distribuirana po zakonu gama razdiobe ako je područje njenih vrijednosti interval (0, +∞), a funkcija vjerojatnosti

$$f(x) = \frac{1}{\Theta^k \Gamma(k)} x^{k-1} e^{-x/\Theta}. \quad (11)$$

U gornjem izrazu k je parametar oblika, a Θ parametar mjerila (k, Θ > 0), pa gama razdiobu kraće označavamo kao Γ(k, Θ). Različite vrijednosti parametra k omogućavaju gama funkciji da poprimi različite oblike pa zato kažemo da je gama razdioba familija dvo-parametarskih funkcija raspodjele.

Za gama razdiobu se očekivanje, varijanca i koeficijent asimetrije mogu definirati kao:

$$\mu = k\Theta, \quad (12)$$

$$\sigma^2 = k\Theta^2, \quad (13)$$

$$\alpha_3 = 2\sqrt{k}. \quad (14)$$

Vidimo da je gama razdioba uvijek pozitivno asimetrična ($\alpha_3 > 0$), ali kada k teži u beskonačno gama funkcija teži u simetrični oblik ($\alpha_3 \to 0$).

Gama raspodjela ima nekoliko generalizacija. Ako se uvede i treći parametar, tzv. parametar položaja x_0 tako da se $(x - x_0)$ podvrgava gama razdiobi, distribucija varijable x je troparametarska gama distribucija ili **Pearson 3** distribucija čija je funkcija raspodjele za $x \geq x_0$ i k, Θ > 0 definirana slijedećim izrazom

$$f(x) = \frac{1}{\Theta^k \Gamma(k)} (x - x_0)^{k-1} \exp(-\frac{x - x_0}{\Theta}). \quad (15)$$

Parametri gama distribucije **(**11**)** mogu se estimirati kosisteći slijedeće izraze [1], [8]

$$\hat{k} = \frac{(\bar{x})^2}{s_x^2}. \quad (16)$$

$$\hat{\Theta} = \frac{s_x^2}{\bar{x}}, \quad (17)$$

dok se parametri Pearson 3 distribucije (15) mogu estimirati koristeći izraze [1]

$$\hat{k} = \frac{4}{(Cs_x)^2}, \quad (18)$$

$$\hat{\Theta} = \frac{s_x Cs_x}{2}, \quad (19)$$

$$\hat{x}_0 = \bar{x} - 2\left(\frac{s_x}{Cs_x}\right), \quad (20)$$

pri čemu su estimirane vrijednosti aritmetičke sredine $\bar{x}$, varijance s_x^2 i koeficijenta asimetričnosti Cs_x dane sa (1), (2) i (3).

Ako pak logaritmi slučajne promjenjive X slijede Pearson 3 razdiobu, tada varijabla X slijedi **log-Pearson 3 (**ili kraće LP3) razdiobu čija je funkcija vjerojatnosti dana izrazom

$$f(x) = \frac{1}{x\Theta^k \Gamma(k)} (\ln(x) - y_0)^{k-1} \exp(-\frac{\ln(x) - y_0}{\Theta}) \quad ,(21)$$

a definirana je za $y_0 \leq \ln(x) < \infty$ i k > 0. Srednja vrijednost i varijanca za LP3 distribuciju se definiraju sa [1]

$$\mu_x = \left(\frac{1}{1-\Theta}\right)^k \exp\left(\frac{y_0}{\Theta}\right), \quad (22)$$

$$\sigma_x^2 = \left[\left(\frac{1}{1-2\Theta}\right)^k - \left(\frac{1}{1-\Theta}\right)^{2k}\right] \exp\left(-2\frac{y_0}{\Theta}\right). \quad (23)$$

Kada logaritmi slučajne promjenjive X slijede dvo-parametarsku gama razdiobu, varijabla X slijedi **log-gama** razdiobu čija je funkcija vjerojatnosti može napisati kao (19) uz izostavljanje člana y_0.

LP3 distribuciju možemo prilagoditi empiričkim podacima tako da za prirodne logaritme zabilježenih protoka odredimo estimirane vrijednosti aritmetičke

sredine $\bar{x}$, varijance s_x^2 i koeficijenta asimetričnosti Cs_x dane sa (1), (2) i (3), a nakon toga pomoću izraza (18), (19) i (20) odredimo parametre P3 distribucije od ln (x). Log-gama distribuciju prilagođavamo empiričkim podacima na analogan način uz korištenje izraza (19) i (20) za određivanje parametara gama distribucije od ln(x).

Gumbel_Max razdioba (ili „Maksimum Extreme Value Type 1“ razdioba) je nesimetrična dvoparametarska razdioba čija je funkcija vjerojatnosti definirana sa

$$f(x) = \frac{1}{\beta}\exp(-\frac{x-\mu}{\beta})\exp(-\exp(-\frac{x-\mu}{\beta})) \qquad (24)$$

gdje je β parameter mjerila (β> 0), a μ parametar mjesta (pomaka). Ovu razdiobu karakterizira:

Očekivanje	μ+0.5772β
varijanca	$\frac{\pi^2}{6}\beta^2$
koeficijent asimetrije	≈1,14

Parametri razdiobe mogu se estimirati koristeći slijedeće izraze [8]:

$$\hat{\beta} = \frac{s\sqrt{6}}{\pi}, \qquad (25)$$

$$\hat{\mu} = \bar{x} - 0{,}5772\hat{\beta} = \bar{x} - 0{,}45006s, \qquad (26)$$

gdje su $\bar{x}$ i s aritmetička sredina (1) i standardna devijacija uzorka (2).

Weibull razdioba je kontinuirana distribucija definairana na području varijable x (0, +∞) čija je funkcija vjerojatnosti definirana sa

$$f(x) = \frac{\alpha}{\beta}\left(\frac{x}{\beta}\right)^{\alpha-1}\exp\left(-\left(\frac{x}{\beta}\right)^{\alpha}\right), \qquad (27)$$

gdje je α parametar oblika (α> 0), a β parametar mjerila (β > 0). Moguće je uvesti i treći parametar, parametar pomaka γ ako umjest x uvrstimo (x- γ) pa tada dobivamo troparametarsku weibull razdiobu.

Estimiranje parametara Weibull razdiobe zahtjeva provedbu iterativnog proračuna na računalu. Mi smo koristilo komercijalne programske pakete za proračun parametara α i β , a potom računali teorijske vrijednosti raspodjele prema (27) i dobrotu prilagodbe .

C. Testiranje dobrote prilagodbe funkcije raspodjele

Analizu dobrote prilagođavanja teorijskih raspodjela empiričkim podacima proveli smo uobičajenim HI – kvadrat (χ^2) testom. Obzirom da smo imali uzorak većeg opsega ovaj test je ocijenjen kao najpodesniji [8[,[13].

Za uzorak od N elemenata (x_1, ..., x_N) koji se grupiraju u k razreda neka f_{ei} označava broj elemenata i-tog razreda. Izraz

$$\chi_0^2 = \sum_{i=1}^{k}\frac{(f_{ei} - f_{ti})^2}{f_{ti}} \qquad (28)$$

predstavlja mjeru odstupanja empirijskih frekvencija f_{ei} od teorijskih f_{ti}.

Hipotezu da empirički podaci slijede neku teorijsku raspodjelu prihvaćali smo uz prag značajnosti od 0,05. Pri tome smo vodili računa o broju parametara koji definiraju neku teorijsku razdiobu. Prema rezultatima testa, odnosno stupnju prilagodbe, rangirali smo testirane razdiobe.

III. NUMERIČKI PRIMJER

Empiričkim podacima prirodnih dotoka u rijeku Krku pokušali smo prilagoditi različite teoretske distribucije opisane u prethodnom poglavlju i tražili razdiobu koja se najbolje prilagođava. Na raspolaganju smo imali niz srednjih mjesečnih dotoka zabilježenih na jednom mjernom mjestu u periodu od 59 godina (1950 – 2013 s prekidom 1991-1995). Također smo imali kontinuirani i niz srednjih dnevnih dotoka mjerene na istom mjernom mjestu u neprekidnom nizu u posljednjih 17 godina (1997-2013). Mjereni dotoci su „sirovi“ podaci, odnosno nikakve uzvodne aktivnosti nemaju utjecaja na mjerene dotoke. Svi dotoci su mjereni na jednak način, kako bi se smanjila mogućnost greške.

Zabilježene vrijednosti srednjih dnevnih dotoka grupirani su u manje vremenske jedinice, mjesece i tjedne. Zabilježene vrijednosti srednjih mjesečnih dotoka mogli smo grupirati samo po mjesecima. Prema (1) – (3) izračunate su nepristrane procjene očekivanja, standardne devijacije (i varijance) te koeficijenta asimetrije za pojedine vremenske intervale. Osnovne statističke parametre računali smo za svih 12 mjeseci i desetak tjedana.

Distribucija srednjih mjesečnih dotoka je asimetrična. Na uzorku od posljednjih 17 godina koeficijenti asimetričnosti za pojedine mjesece variraju između 0,5 i 2, prosječno 1,42. Na većem uzorku od 59 godina prosječni koeficijent asimetrije je 1,56. Niz srednjih tjednih dotoka, u analiziranim tjednima na uzorku od posljednjih 17 godina, karakterizira nešto veća asimetričnost, oko 1,8. Distribucije srednjih dnevnih dotoka, bez obzira da li su grupirani po tjednima mjesecima ili godinama, su izrazito asimetrične. U našem primjeru vrijednosti koeficijenata asimetrije su uvijek veće od 1, prosječno oko 3.

Empirički podaci srednjih mjesečnih dotoka na uzorku od 59 godina mogu se, različito po mjesecima, opisati sa nekoliko testiranih razdioba. Za mjesece sa koeficijentom asimetrije oko prosječnog ili većim od prosječnog 1,56 bolje se daju prilagoditi logaritamske razdiobe (LN, LP3, log-gama), dok su u mjesecima sa

koeficijentom asimetrije manjim od prosječnog dobrima pokazuju i neke druge razdiobe (Weibull, Pearson 3, …). U pravilu je LN razdioba u svim mjesecima bila dobra. Nešto malo lošija se u našem primjeru pokazala LP3 razdioba. Obzirom da χ^2 test ovisi o broju razreda u koje su podaci grupirani, varirali smo širinu razreda. Promjena širine razreda ponekad bi promijenila rank distribucija, ali LN i LP3 su se u pravilu prilagođavale empirijski podacima dovoljno dobro za prihvaćanje hipoteze da podaci slijede danu distribuciju.

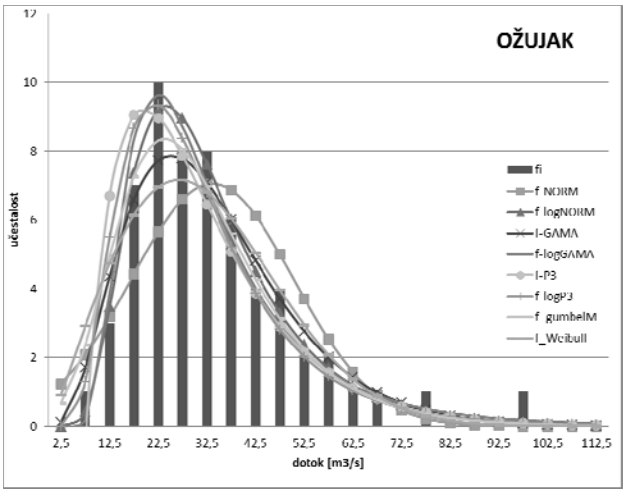
U tablici I dati su izračunati parametri razdioba koje smo prilagođavali empiričkim podacima za mjesec ožujak. Prema izračunatim parametrima izračunate su teorijske frekvencije pojedinih razdiba, prikazane tablicom II. Na slici 1 histogramom su prikazani empirički podaci za mjesec ožujak, a krivuljama različite teorijske distribucije sa parametrima dobivenim u postupku prilagođavanja.

TABLICA I. PARAMETRI RAZDIBA ZA MJESEC OŽUJAK

RAZDIOBE	**PARAMETRI**
Normalna	μ=31,334, σ=16,661
Log-normalna	μ_y=3,404, σ_y=0,469
Gama	k=4, Θ=8,400
log-gama	k=47, Θ=0,0723
Pearson 3	k=2, Θ=11,792, x_0=7,792
LP3	k=1031, Θ=-0,015925, y_0=19,733
Gumbel Max	μ=12,991, β=23,8
Weibull	α =2,092, β=36,404

TABLICA II. ZA MJESEC OŽUJAK

xi	fi	N	LN	G	LG	P3	LP3	gMax	Wbull
2,5	0	1,239	0,000	0,293	0,000	0,185	0,003	0,669	0,907
7,5	1	2,073	0,416	2,703	0,230	4,223	1,301	2,372	2,911
12,5	3	3,169	3,473	5,618	3,869	7,525	5,494	4,964	4,741
17,5	7	4,428	7,349	7,503	8,319	8,745	8,668	7,256	6,138
22,5	10	5,655	9,198	8,072	9,630	8,440	9,336	8,308	6,955
27,5	8	6,601	8,954	7,638	8,703	7,338	8,367	8,056	7,156
32,5	8	7,043	7,616	6,636	7,026	5,967	6,794	6,976	6,806
37,5	6	6,868	5,994	5,426	5,374	4,631	5,213	5,594	6,042
42,5	4	6,121	4,504	4,239	4,007	3,472	3,868	4,256	5,038
47,5	4	4,987	3,291	3,197	2,957	2,536	2,815	3,125	3,960
52,5	2	3,713	2,366	2,343	2,177	1,813	2,025	2,239	2,942
57,5	2	2,527	1,684	1,679	1,607	1,275	1,448	1,578	2,070
62,5	1	1,572	1,194	1,180	1,191	0,884	1,033	1,100	1,381
67,5	1	0,894	0,845	0,816	0,889	0,606	0,737	0,761	0,874
72,5	0	0,464	0,598	0,556	0,669	0,411	0,526	0,524	0,526
77,5	1	0,221	0,424	0,374	0,506	0,276	0,377	0,359	0,300
82,5	0	0,096	0,302	0,250	0,386	0,184	0,271	0,246	0,163
87,5	0	0,038	0,216	0,165	0,297	0,122	0,195	0,168	0,084
92,5	0	0,014	0,155	0,108	0,230	0,081	0,141	0,114	0,041
97,5	1	0,005	0,112	0,070	0,179	0,053	0,103	0,078	0,019
102,5	0	0,001	0,081	0,045	0,141	0,034	0,075	0,053	0,009
107,5	0	0,000	0,059	0,029	0,111	0,022	0,055	0,036	0,004
112,5	0	0,000	0,043	0,019	0,089	0,019	0,041	0,025	0,001
SUM	59	57,729	58,872	58,958	58,588	58,843	58,887	58,857	59,071



Slika 1. Histogram srednjih mjesečnih dotoka za mjesec ožujak i prilagodba različitih teorijskih funkcija distribucije

Izmjereni dotoci u mjesecu ožujku imaju $\overline{X}$ =31,33, s_x=16,66 i Cs_x=1,42. Slične dijagrame dobili smo i za ostale mjesece, a mjesec ožujak smo prikazali obzirom da je procijenjeni koeficijent asimetrije Cs_x najbliži prosječnom. Dobrotu prilagođavanja pojedinih teorijskih razdioba empiričkim podacima u mjesecu ožujku sažeto smo prikazali u tablici III. Vrijednost χ^2 za pojedine razdiobe je računata prema (28) grupirane u k razreda uz stupanj slobode s. U zadnjem stupacu data je vjerojatnost testa. Vidimo da se sve razdiobe, s izuzetkom Pearson 3, dobro prilagođavaju empiričkim podacima u ovom primjeru. Također se vidi da se LN i LG izuzetno dobro prilagođavaju empiričkim podacima.

Teorijskim funkcijama distribucije pokušali smo opisati i izmjerene srednje tjedne dotoke međutim uzorak od 19 godina, odnosno skup od 19 srednjih tjednih dotoka, je prekratak period za relevantnu analizu. Izabrana širina razreda ima značajan utjecaj na oblik distribucije i rezultat χ^2 testa. Ipak, za analizirane tjedne ponovno su se LN i LP3 razdiobe dobro prilagođavale empiričkim podacima u većem broju slučajeva, što ne čudi obzirom na slične koeficijente asimetrije srednjih tjednih i srednjih mjesečnih dotoka.

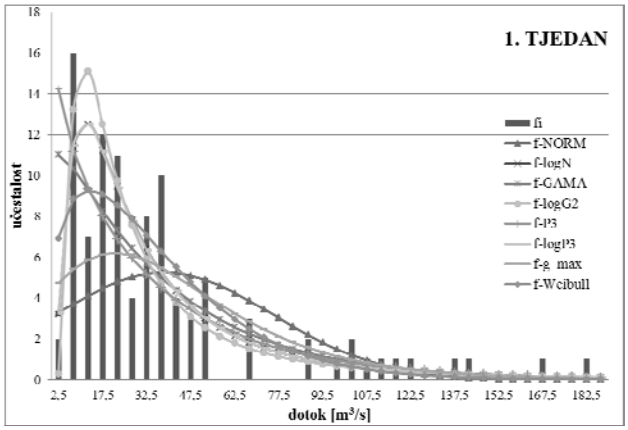
Također smo, radi usporedbe, analizirali empiričke podatke srednjih dnevnih dotoka (u nekom tjednu i mjesecu). Obzirom da su distribucije srednjih dnevnih dotoka izrazito asimetrične, uvijek su se mjereni podaci bolje prilagođavali logaritamskim razdiobama (LN, LP3, log-gama) nego ostalima koje smo razmatrali. Neke od razmatranih razdioba samo u nekim vremenskim intervalima bi uspješno opisale empiričke podatke (gama, Pearson 3, Weibull, Gumbel max), dok se normalna razdioba nije dala prilagoditi empiričkim podacima u niti jednom slučaju. Dakle, ponovno su LN i LP3 bile dominantne. Slično kao na slici 1, na slici 2 su histogramom prikazani empirički podaci za 1. tjedan u godini, a krivuljama različite teoretske distribucije sa parametrima dobivenim u postupku prilagođavanja. Izmjereni dotoci u 1. tjednu imaju $\overline{X}$ =38,41, s_x=37,22 i Cs_x =1,91. Dobrotu prilagođavanja pojedinih teorijskih razdioba empiričkim podacima sažeto smo prikazali u tablici IV. Sada je samo LN razdioba prešla prag od 0,05.

TABLICA III. DOBROTA PRILAGOĐAVANJA TEORIJSKIH RASPODJELA EMPIRIČKIM PODACIMA ZA MJESEC OŽUJAK

	N	LN	G	LG	P3	LP3	gMax	Wbull
χ_0^2	7,3671	0,2118	3,3817	0,5111	8,5723	2,2519	2,6589	4,3059
k	7	7	7	7	7	7	7	7
s	2	2	2	2	3	3	2	2
k-s-1	4	4	4	4	3	3	4	4
P	0,1177	0,9948	0,4961	0,9724	0,0356	0,5218	0,6164	0,3662

TABLICA IV. DOBROTA PRILAGOĐAVANJA TEORIJSKIH RASPODJELA EMPIRIČKIM PODACIMA ZA 1. TJEDAN

	N	LN	G	LG	P3	LP3	gMax	Wbull
χ_0^2	77,622	11,620	22,752	23,511	26,983	11,682	40,751	16,835
k	9	9	9	9	9	9	9	9
s	2	2	2	2	3	3	2	2
k-s-1	6	6	6	6	5	5	6	6
P	1,1E-14	0,0710	0,0009	0,0006	5,7E-05	0,03941	3,2E-07	0,0099



Slika 2. Histogram srednjih dnevnih dotoka u 1. tjednu i prilagodba različitih teorijskih funkcija distribucije

IV. ZAKLJUČAK

Pronalaženje odgovarajuće teorijske funkcije vjerojatnosti koja dobro opisuje neki stohastički događaj je od izuzetne važnosti pri modeliranju tih podataka za različite potrebe. U radu smo analizirati višegodišnje rezultate mjerenja dotoka na jednom mjernom mjestu i tražili prikladnu teorijsku funkcija vjerojatnosti za opis izmjerene učestalosti pojavljivanja tih dotoka u jediničnim vremenskim intervalima. Cilj rada je naći tip distribucije koja se dobro prilagođava mjerenim podacima u velikom broju slučajeva. Distribucija mora biti prikladna za korištenje, sa parametrima koji se estimiraju jednostavnim analitičkim postupkom.

Vrijednosti koeficijenta asimetrije izmjerenih podataka definiraju tip distribucije prikladne za opis izmjerenih vrijednosti. Distribuciju srednjih mjesečnih dotoka i srednjih tjednih dotoka karakterizira srednja ili velika asimetričnost (koeficijent asimetrije između 0,5 i 2) dok je kod distribucije srednjih dnevnih dotoka asimetričnost izrazito velika (prosječni koeficijent asimetrije oko 3). Što je osnovni vremenski period kraći koeficijent asimetrije je veći.

Srednji mjesečni i srednji tjedni dotoci, u većini analiziranih slučajeva su se dobro prilagođavali log-normalnoj i log-Pearson 3 razdiobi. Promjena širine razreda ponekad bi promijenila rank distribucija određen χ^2 testom, ali LN i LP3 su se u pravilu prilagođavale empirijski podacima dovoljno dobro za prihvaćanje hipoteze da podaci slijede danu distribuciju.

LITERATURA

[1] D. P. Loucks and E. van Beek, „Water resources systems planning and management - An introduction to methods, models and applications“ – ISBN 92-3-103998-9 – © UNESCO 2005

[2] R. A. Wurbs, „Modeling and analysis of reservoir system operations,“ Prentice Hall PTR, USA, 1996.

[3] E. Sutlović, S. Čujić Čoko, I. Medic, “Characteristics of basin inflows - A statistical analysis for long-term/mid-term hydrothermal scheduling”, *Proceedings of SDEWES 2013,* September 2013, Dubrovnik, Croatia

[4] T. G. Siqueira, M. Zambelli, M. Cicogna, M. Andrade, S. Soares, „Stochastic dynamic programming for long-term hydrothermal scheduling considering different streamflow models“, *Proceedings of 9th International Conference on Probabilistic Methods Applied to Power Systems*, Stockholm, Sweden, June, 2006.

[5] Z.Yu, F. T. Sparrow, B. H. Bowen, “A new long-term hydro production scheduling method for maximizing the profit of hydroelectric systems”, *IEEE Transactions on Power Systems*, Vol. 13, No. 1, pp 66-71, 1998.

[6] Y. Ikura, and G. Gross, “Efficient large-scale hydro system scheduling with forced spill conditions”, *IEEE Transactions on Power Apparatus and Systems*, Vol. PAS-103, No. 12, pp 3502-3520, 1984.

[7] I. Pavlić, „Statistička teorija i primjena“, Tehnička knjiga, Zagreb, 1988.

[8] Engineering Statistics, *NIST/SEMATECH e-Handbook of Statistical Methods*, http://www.itl.nist.gov/div898/handbook/, 7/01/2014.

[9] “Probability distribution”, s Interneta, http://en.wikipedia.org/wiki/Probability_distribution, 16.12.2013.

[10] H. Aksoy, "Use of Gamma Distribution in Hydrological Analysis", *Turk J. Engin Environ Sci*, 24 (2000), pp. 419 – 428

[11] M. A. Al-Fawzan, “Methods for Estimating the Parameters of the Weibull Distribution”, InterStat, May 2000.

[12] R. Žugaj, „Hidrologija“,Udžbenik Sveučilišta u Zagrebu, Zagreb, 2000.

[13] N. Ožanić, “Statističke obrade velikih voda hidromelioracijskih sustava – seminar Velike i male vode”, Građevinski fakultet Rijeka, 2004.

Cloud dilemma

A.B. Robert Petrunić, bacc.ing.comp. Algebra university college for applied computer engineering, robert@petrunic.com C. Igor Orović, bacc.ing.comp Algebra university college for applied computer engineering, igor.orovic@racunarstvo.hr i D.E. Vjeran Bušelić, Senior Lecturer, MApMa, Polytechnic of Zagreb, vbuselic@tvz.hr

Cloud is changing the way IT thinks and works, it's changing the way administrators do their job, the way businesses offer their services, how consultants create new offerings and solutions, and how management divide capital and operational expenses.

Most of the administrators are reluctant to use cloud services because of the fear that the Cloud will make them obsolete. The rest of them, including the management have a dilemma whether to use it or not, and most important whether they are allowed to do it, based on local and global laws and regulatory compliances.

This paper will deal with dilemmas administrators, consultants and management may encounter during the transition to the cloud solution, from the technical, through business, to the regulatory and legal perspective. It will give an overview on Croatian laws and rules created by Croatian regulatory bodies, which will help to solve this dilemma and make the transition to the cloud solution easier and smoother for most of the businesses that are having abovementioned dilemma.

I. INTRODUCTION

Cloud is the way of service delivery that makes the service more accessible and chipper to the end user, and it changes the usage of computing resources by using the innovation known as utility computing [1]. Utility computing is an approach that packages computing, storage and network resources as a new metered service ready to be delivered to the end user.

During the last decade, more and more companies are trying to cut IT (information technology) expenses by changing the business model from CAPEX (capital expenses) to OPEX (operational expenses). By "leasing" the computing resources they are able to achieve desired state. Cloud is offering the same thing with better TCO (total cost of ownership) and ROI (return on investment) by offering "pay-per use" or "pay as you go" services, where one has to pay only for used resources, with "easy to grow" options that allows a company to cover greater demand on resources when needed with minimal impact on services and finances. There is also an additional benefit, because cloud is offering shorter "time to market", especially for startups and small to medium companies, with easy to configure services. This is only one part of potential cloud users, because the other companies, the ones that do not care about the price of the service, can, and will use the cloud as long as the service is offering a benefit (financial or any other).

One of the characteristics of the cloud is the easy way to increase the resources needed for the running services and to decrease the resources when the demand has lowered. Together with already mentioned "pay-as-you-go" system it changes the IT perception on computing, storage and network resources, and it's opening the door to replace classical oversized datacenter (CAPEX) into a fixed and manageable system (OPEX).

An example of capacity to demand ratio in classical (static) datacenter and cloud based datacenter is shown on figure 1 and figure 2.[1]

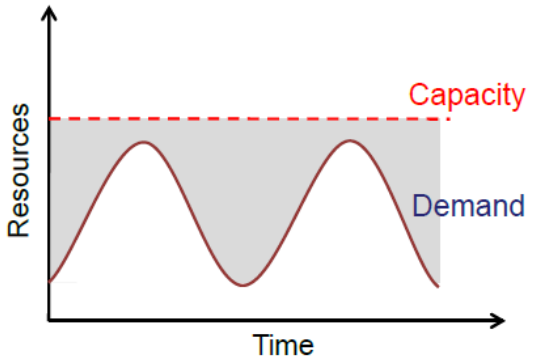


Figure 1: Capacity and demand ratio in cloud based (dynamic) data center

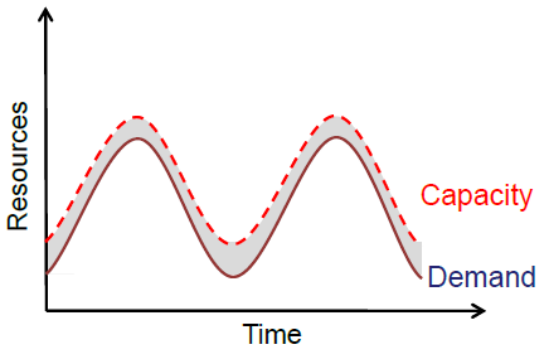


Figure 2: Capacity and demand ratio in cloud based (dynamic) data center

[1] Franković, Vračić: Algebra University College for applied computer engineering, 2014. Retrieved 9th of February 2014.

From the IT management perspective, cloud is categorized as: public, private, hybrid and community cloud.

Public cloud is accessible to anyone, and this is the service offered by a provider to a general public.

Private cloud is, as the word suggests, a cloud created and used by a "private" entity.

Hybrid cloud is a type of cloud used by medium sized companies and it allows for merger of private cloud with public one. This merger can be triggered on a whole or only for parts of the private cloud and it can be dynamic.

Community cloud is a special cloud type made by group of companies offering their services for closed group of consumers.

Because cloud offerings have different "flavors" there is another classification used to subdivide and describe the services offered by a cloud provider, and those are: SaaS (software as a service), PaaS (platform as a service) and IaaS (infrastructure as a service).

SaaS is the type of service where the end user is offered fully functional service that can be used with no or minimal configuration from the user side.

PaaS is the type of service where the end user is offered everything except the application. The service provider is offering fully functional OS (operating system) and additional services needed by most of the applications like database, and those parts will be patched and guaranteed to be functional by service provider. The user has to install, configure and patch its own application by using the provided platform services.

IaaS is the type of service where the end user is offered computing, storage and network resources, and service provider guarantees that those resources will be available, while OS, platform modules and applications are user responsibility. User will have to patch all the components used.

More formal definition for cloud types and services is given by NIST [2]

II. CLOUD BENEFITS

Some of the cloud benefits are:
Geopolitical independence, redundant locations, standardization, risk management, manageability and measurability, high availability and scalability, "pay-as-you-go" model, less or ne CAPEX, faster deployment time, shared resources, increased mobility, easy to manage and higher productivity.

Geopolitical independence is cloud characteristic that can play a crucial role in a decision process whether to use cloud or not. Since cloud providers are usually geopolitically independent, they can offer their services even do countries using those services are politically unstable and on the brink of war, or in a situation where natural disaster or some other type of devastation is inevitable or imminent. Being geopolitically independent makes the cloud providers the best solution for startups and medium to large companies, because they can offer some kind of security and continuity.

Redundant locations is at the same time an addition and the part of already explained geopolitical independence. By having redundant locations, cloud providers can offer continuity in any case of political or natural disaster.

High availability is the feature used in abovementioned benefits that offers availability of the system when some parts of this system are malfunctioned or nonfunctional for any other reason.

Scalability on the other hand is a feature that allows the cloud provider and cloud user to grow their resources as needed horizontally (by adding more instances of the application, service or OS) and vertically (by adding more resources - CPU and memory to an existing systems).

Standardization helps cloud provider to lower their expenses because standardization will make a system more reliable, easier to manage and by doing that the system will give better ROI.

Risk management is the next step in lowering the cost of the system and creating it more secure and robust. During the risk management process most of the threats to the service will be treated and corrective measures will be taken. This process will in the long run save money, make system more resilient and customer more satisfied.

Manageability and measurability are the most important parts of the cloud provider services, because this is the only way to make the services useful and chargeable. By having manageable system, cloud provider can manage the system and offer the service to the customer, and by having this service measurable service provider can measure and charge used resources, thus offering cloud services.

From the economic perspective, "pay-as-you-go" and less or no CAPEX are attractive to almost any manager searching to cut expenses and save the money.

From the administrator and IW (information worker) point of view, faster deployment is offering more time and less work to be done, shared resources are allowing for cheaper testing and working environments, and increased mobility makes IW more productive and agile.

III. CLOUD CHALLENGES

There are some factors one has to take in the account when discussing the cloud, and those factors are related to possible problems, disadvantages and challenges of the cloud based products and solutions. Some of those factors are: regulatory compliance, local and global laws, data location, data separation, loss of manageability, forensics and/or incomplete data deletion, vendor lock-in, lack of education, client part of the communication channel, public and private sector dependency, new technologies and business models and more and more complex attacks.[3] [4]

Data location maters in various situations. For instance, most of the countries have laws preventing sensitive data from being stored on a computer systems located outside the country boundaries. Some countries

have policy to spy on the data (US for instance) and by using US based cloud provider with datacenter located in US, data itself will be treated by US laws, not by the laws of the country the user is from. [5]

Data separation is one other sensitive data related problem because it may happen that company A is sharing the resources with company B, and if data separation (multi-tenancy) is not configured and enforced correctly company A might be able to see the data belonging to company B and vice versa.

Manageability loss depends on the cloud service type used (SaaS, PaaS, IaaS), and it's emphasized in the SaaS and PaaS type of service. IaaS type of service is as manageable as any other OS in the internal network if the machine is accessible over the network.

Forensics and/or incomplete data deletion is one of the security question one will have to ask when an incident occurs, and one have to analyze it. The problem with forensic investigation in the cloud environment is that data is scattered across various storage devices, possible across datacenters too. This will most certainly make the forensic investigation difficult if not impossible. The other part of the problem is incomplete data deletion that might have an impact on data privacy and isolation. If the data is not entirely deleted on the source drive, somebody might recover it and gain access to sensitive data. This can also be done if data separation is not correctly implemented.

Vendor lock-in is something one have to deal with in the cloud transition planning phase. The thing is that if one choose a particular cloud provider, and this provider is using particular technology that no other provider is offering, one will have a problem changing providers, thus ending up in a vendor lock-in or being tied to a single vendor. This is again highly dependent on type of service used and it's less problematic with IaaS then with SaaS and PaaS, because IaaS VMs can be easily converted between vendors (VMware to Microsoft and vice versa for instance).[2]

Lack of education is one of the possible problems that's not so obvious at the first glimpse. In order to accept and properly use the cloud services one have to be familiar with the technology and capable of supporting and using it.

Client part of communication channel is not directly related to the cloud provider, but since it's used to access cloud provider services it has to be properly secured. Most of the security breaches are done on the client part of the communication channel, thus in order to have secure and reliable cloud service, client has to be secured and properly used.

Public and private sector dependency is security and accessibility related. Regarding the security, both sectors want to have all three CIA (confidentiality, integrity/authenticity and availability) triangle requirements met. When the accessibility is in question, what will happen should the Internet in Croatia is down for an hour? A day? A week? A month? With cloud in perspective, everything is dependent on Internet connection, so the infrastructure has to be reliable and accessible in order to be able to use the cloud services

Regulatory compliance, local and global laws are one of the biggest issue for most of the government and government related companies and banks. In Croatia, HNB (Hrvatska Narodna Banka – Croatian National Bank) define rules and guidelines that all the banks have to obey.

IV. REGULATORY COMPLIANCE AND LAWS IN CROATIA

If we look at the regulatory compliance and the banks in Croatia, there are some rules defining how the banks can store the data, and where this data can be stored. The regulatory body (HNB) has defined exact rules that has to be followed in two documents: Guidelines for IS management in order to lower the operational risk [6] and Guidelines for adequately managing outsourcing related risks. [7]

The most important part related to cloud services is defined in those two documents with the following rules (translation – just the parts needed for this theses):

… Provider can be resident or nonresident.... direct supervision by HNB should never be (by any means) obstructed or denied … thus, the bank should determine whether the provider local laws are compatible with this guideline … [3]

By reading this clipping, one can understand one of the biggest problem related to cloud computing, and this is the territory, or to be more precise no territorial boundaries in the cloud can lead to problem with the laws of the country where the data is at the particular point in time. For instance if Croatian bank is using the cloud provider with datacenter in US, data will be treated by local US laws, and HNB will not be able to access the data in an unobstructed way. The other problem is that cloud providers are utilizing their own cloud power, so they have datacenters on different continents and the data is being moved from one datacenter to another (in some cases) without user knowledge. So, one day data can be in US, the other in Europe and after that in Asia, and the laws applicable to data being in the cloud are changing without user knowledge. Of course, some cloud providers gives you the ability to choose where the data will be stored, but by reading the "small letters" one can see that they have the right to move the data wherever they want to in order to preserve the ongoing operation of the cloud and possible backup scenarios. From this perspective it is difficult for the bank to find the right public cloud provider, thus leaving them with the option to use private cloud instead.

Regarding the applicable laws in Croatia there are: law on private data protection [8], information security law [9], data secrecy law [10] and law on electronic signature [11].

[2] Starwind software. Retrieved 9th of February 2014. http://www.starwindsoftware.com/converter

[3] Croatian National Bank (October, 2005.) "Guidelines for adequately managing outsourcing related risks". Retrieved 9th of February 2014. http://www.hnb.hr/supervizija/eksternalizacija/h-smjernice-za-adekvatno%20upravljanje-rizikom-eksternalizacije.pdf

Private data protection law has similar parts as abovementioned HNB guidelines regarding personal data collections: (translation – just the parts needed for this theses):

Article 13: Private data collections can be taken out of Croatian territory only if the country hosting the data has appropriate regulations regarding private data protection…[4]

Office of the National Security Council is the roof body responsible to support the abovementioned laws, by creating regulations regarding data secrecy, manipulation etc. In the document called "Regulation on information security measures" [12] there is article 6:1 discussing that classified data can be exchanged only with countries and international organizations if they have signed the document on mutual protection of classified data with Republic of Croatia. There are also rules regarding the physical requirements that have to be met in case the document is classified as Top secret, secret or confidential.

One more regulatory body in Croatia that have his own rules regarding data that leaves the country boundaries is Ministry of Finance. They have rules regarding gambling and technical requirements that has to be met. The Minister of Finance shall prescribe standards on the technical requirements of betting through interactive sales channels online gaming [13], and the registration and identification of players, the terms of software (software) and equipment, certification systems and games, handling transactions of receipts and payments, security principles, conditions of supervision, responsible for preparing the principle, and more. The organizer who wants to organize games betting has to coordinate system for the preparation of the technical standards set by the Minister of Finance. Given the above, we conclude that a taxpayer who has no registered office, permanent establishment, domicile or habitual residence in the country is not obliged to register for PDV purposes in the Republic of Croatia, was to perform a host or computing cloud computing taxpayer undertaking abroad. Also note that the issue of invoices and payment of PDV that the taxpayer is obliged to apply the rules of the Member State of residence or permanent establishment from which the service is performed

V. CLOUD SECURITY

Security in a cloud is on a higher level than in the normal small to medium sized business environment, because higher standards are being met in the process of creating and managing the cloud environment. In spite of that it has some security issues. Top 9 threats to the cloud environment are as follows: [14] data breaches, data loss, account hijacking, insecure APIs, DoS (denial of service), malicious insiders, abuse of cloud services, insufficient due diligence and shared technology issues. Those security concerns are not exclusively cloud based – they are all applicable to normal computing environment, but in the cloud each and every of those implications could be multiplied and much more intense in nature.

Data breaches are taken to a higher level in the cloud environment. If a person breaks into an application running on the shared database system she could have an access to other tenant's data. The other problem lies in the virtualized nature of cloud, and one compromised VM (virtual machine) could lead to all machines on the same hypervisor being compromised should hypervisor have unknown or unpatched vulnerability.

Data loss is administrators nightmare, because by losing data, customer is denied the sole purpose of the IS (information system) and the business will have to deal with the consequences. Data loss can be on the cloud provider side (hardware or software failure, and no backup in place), it can be malicious in nature (hackers, competitors, dissatisfied employees), and customer based (if customer loses encryption keys – cloud provider backup is unusable, the other incident could be accidental data deletion).

Account (or service) hijacking is being more and more implemented in targeted phishing attacks, and it can be triggered by various fraud schemes and software vulnerabilities exploitations. By having account or service hijacked an attacker can use it for its own purposes and to their own willing. Hijacked services become part of the botnets, being run and managed by the attacker allowing her to use the computing, storage and network resources to her own will. This can have implications in various levels regarding law and regulatory compliance, just to mention some of them.

Insecure APIs can lead to service degradation or hijacking, because APIs are used for provisioning, monitoring and management. By having an access to APIs, an attacker can have access to client data and services, and in some cases to all other tenants on the cloud system.

DoS can also have multiple implications in cloud environment. DoS can be used to deny an access to the targeted system or to use a system to attack other systems. Usually botnets are used to create DDoS (distributed denial of service) attack against a target, and in some cases even the largest providers are unable to stop, mitigate or deflect such an attack.

Malicious insiders definition by CERT: (computer emergency response team) is as follows: [5] *"Malicious insider threat to an organization is a current or former employee, contractor, or other business partner who has or had authorized access to an organization's network, system, or data and intentionally exceeded or misused that access in a manner that negatively affected the confidentiality, integrity, or availability of the organization's information or information systems."* Insiders or privileged users have access to secure systems and are capable of creating more damage than outsiders by having the privileges and physical access.

Abuse of cloud services is again interesting field for most of the attackers because it can give them financial or

[1] [4] Croatian parliament (12th of Jun 2003) "Private data protection law". Retrieved 9th of February 2014. http://narodne-novine.nn.hr/clanci/sluzbeni/305952.html

[5] CERT. Retrieved 09th of February 2014. http://www.cert.org/insider-threat/index.cfm

other gains. For instance, cracking the encryption keys can be much easier by having virtually unlimited resources in the cloud, and this is just one of the possible implications. Calculating bit coins is another example, because it will give an abuser direct financial gain.

Insufficient due diligence is entirely on the cloud provider domain, because users of the service cannot directly influence the cloud providers irresponsibility. Since the cloud services are growing each and every day, more and more irresponsible providers are emerging.

Shared technology issues are in the very nature of the cloud. Cloud services are based on virtualization technologies that share CPU, memory and network resources between virtual machines. Those virtual machines can belong to different customers (tenants) and should anyone be able to access the shared resources from another tenant he will be able to access sensitive information and possibly the part of, or the entire cloud environment.

VI. CONCLUSION

By looking at all the benefits, challenges and security related issues the cloud based services have to offer, deal with and are vulnerable to, one has to ask a question whether to use it or not. The answer depends on the company risk management procedures, knowledge about the cloud services and security awareness. It is obvious that some companies will benefit from using public cloud, and yet some will find hybrid or private cloud to be more flexible and appropriate solution.

LITERATURE

[1] Simson Garfinkel (3 October 2011). "The Cloud Imperative". Technology Review (MIT). Retrieved 9th of February 2014.

[2] National Institute of Standards and Technology, The NIST Definition of Cloud Computing, 2011. Retrieved 9th of February 2014

[3] Petrunic, Robert (April 2011.) "Cloudy on the Client side" Windays 2011. Retrieved 09th of February 2014

[4] Petrunic, Robert (April 2011.) "Cloud v2.0" Windays 2012. Retrieved 9th of February 2014

[5] US department of Justice (11th September 2011). Retrieved 9th of February 2014.

[6] Croatian National Bank (March, 2006.) "Guidelines for IS management in order to lower the operational risk". Retrieved 9th of February 2014

[7] Croatian National Bank (October, 2005.) "Guidelines for adequately managing outsourcing related risks". Retrieved 9th of February 2014

[8] Croatian parliament (12th of Jun 2003) "Private data protection law". Retrieved 9th of February 2014

[9] Croatian parliament (13th of July 2007) "Information security law". Retrieved 9th of February 2014

[10] Croatian parliament (13th of July 2007) "Data secrecy law". Retrieved 9th of February 2014

[11] Croatian parliament (17th of January 2002) "Electronic signature law". Retrieved 9th of February 2014

[12] Office of the National Security Council parliament (18th of April 2008) "Regulation on information security measures". Retrieved 9th of February 2014

[13] Croatian parliament (10th of July 2009) "Law on lottery". Retrieved 9th of February 2014

[14] Cloud security alliance (February 2013) "The Notorious Nine Cloud Computing Top Threats in 2013". ". Retrieved 9th of February 2014

Microcontroller managed module for automatic ventilation of vehicle interior

Boris Raić* and Aleksander Radovan**
*Polytechnic of Zagreb, Zagreb, Croatia
**University of Applied Sciences Velika Gorica, Velika Gorica, Croatia
boris.raic@gmail.com
aleksander.radovan@vvg.hr

Abstract – In order to mitigate overheated interior of a vehicle parked in the hot summer sun and thereby to make the entering into the vehicle more comfortable, microcontroller managed module for automatic ventilation of vehicle interior is made. The module is implemented using a microcontroller as a central logical unit and a series of sensors which provide sufficient data to ensure functional, but also efficient, reliable and safe ventilation. The ventilation process is performed by opening vehicle windows slightly, which enables air to circulate. Microcontroller controls the position of the windows autonomously and independently of the driver's presence, following predefined algorithm that uses sensors data obtained from the vehicle's surroundings. Besides temperature, the most important factors to ensure quality implementation of ventilation are detected movements around the vehicle, the presence of precipitation and other. This paper shows the components, their purpose and capabilities, advantages and disadvantages, as well as potential implementations and upgrades. The test results give insight into utilization options of this module and its usefulness.

Key words – automatic ventilation of vehicle interior, microcontroller, sensors, temperature, controlling position of vehicle windows.

I. INTRODUCTION

Closed area without air circulation and directly exposed to sunlight heats up much faster and reaches significantly higher temperatures than the open space where air circulates freely – so-called "greenhouse effect". That is exactly what happens in a vehicle parked in the hot summer sun, which later represents a problem when entering such overheated vehicle.

The basic idea of this project is to prevent or at least mitigate overheating of the vehicle interior in order to facilitate entering into the vehicle.

The problem can be solved by leaving the vehicle windows partially opened so air can circulate, which reduces the heating of the vehicle interior. However, as the ventilation process takes place in a parked vehicle, without human presence, and the vehicle is exposed to external influences, such as precipitation or potential burglary, it is unreliable and unsafe. Consequently, it is necessary to automate the ventilation process.

According to [12], the complete automatic ventilation process is managed by microcontroller default algorithm, considering input signals read using various sensors. The temperature sensor measures the temperature inside the vehicle. If it exceeds predefined level of "comfortable temperature", microcontroller automatically lowers the power windows to enable air circulation which alleviates aforementioned "greenhouse effect".

The automatic ventilation process must be reliable and not affect the safety of the vehicle. For this reason, additional sensors are used, whose task is to annul dangers appeared with this ventilation method. Before all, these are precipitation sensor that makes the ventilation process reliable, and motion sensors that make it safe. There are also other sensors whose implementation gives additional quality, but these are the most important of them. Together with temperature sensor, they meet the minimum requirements to perform automatic function of ventilation, without endangering the vehicle.

Following chapters show components and technologies used, operation modes, advantages and disadvantages, capabilities and potential upgrades of the module, as well as ability of integration with existing automatic systems within the vehicle. Results of testing the module in real conditions are presented and analyzed.

II. COMPONENTS AND MODULE ARCHITECTURE

According to [12], the module's task is to control the ventilation process autonomously in order to facilitate vehicle utilization, i.e. providing conditions to make the entering into the vehicle parked in the sun more comfortable. In order to meet all the criteria of quality ventilation, the module has to collect enough data from the vehicle's immediate environment and, based on that data, it has to decide whether it is necessary to lower the windows and thus start the ventilation process. Accordingly, the module can be divided to three logical units: peripheral unit - collects data, control unit - manages module operations, and switching unit – adjusts signals. Figure 1 shows logic diagram of module.

The switching unit exclusively receives commands, which makes it the output unit. The control unit, or microcontroller, receives, processes and sends signals which makes it the central logical unit, while peripheral

Polytechnic of Zagreb, Zagreb, Croatia
University of Applied Sciences Velika Gorica, V. Gorica, Croatia

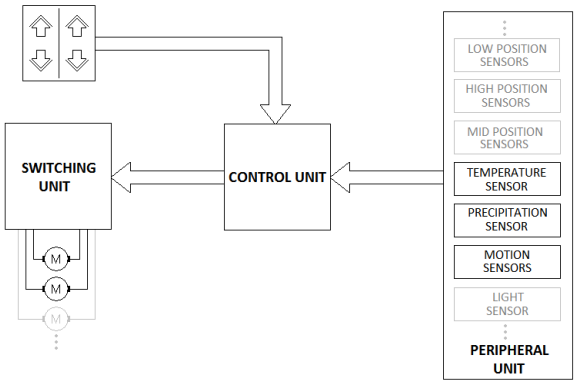


Figure 1. Logic diagram

unit exclusively collects data and sends it to control unit, which makes it the input unit.

A. Peripheral unit

Since the peripheral unit is also the input unit, it is presented first, for better insight into module operations and modes.

By having to collect many different types of data, the peripheral unit is complex enough as it is. The necessity of placing different sensors in separate locations, in order to achieve data collection, makes it even more complex.

- The temperature sensor is the most important sensor of this module because measuring of the vehicle interior temperature enables this method of automatic ventilation. The sensor is located inside the vehicle, protected from direct sunlight in order to measure the actual temperature of the vehicle interior, not the ostensible temperature (which is slightly higher due to the direct expose to the sun rays). The sensor consists of temperature sensitive resistor - NTC (Negative Temperature Coefficient) as a main part of sensor's electronic circuit with other necessary electronic components. In accordance with temperature change, sensor gives different output value. The control unit processes it and thus receives information about the actual temperature inside the vehicle.
- Precipitation sensor [9] is very important because it makes the automatic ventilation reliable. If precipitation occurs during the ventilation process, sensor detects the new circumstances and sends a signal to the control unit which orders to close the windows immediately. The sensor is situated on the windshield inside the vehicle, above the rear-view mirror in order to detect the falling drops of water on the glass. LED (Light Emitting Diode) or IR (Infra-Red) diode is used as a transmitter that emits light at a certain angle onto the windshield. The light of the expected intensity reflects to the photodiode which is used as a receiver. If droplets appear onto the windshield, the light reflects differently which causes a change in the receiving electronic circuit connected to the microcontroller. That is recognized as the presence of precipitation and windows close. Figure 2 shows precipitation sensor.
- Motion sensors are very important for the vehicle's safety while the ventilation process takes place. If there are movements detected within the vehicle's surroundings, it is classified as a potential danger and the system closes the windows. Commercial motion sensors on the principle of infrared technology - "Velleman PIR416" [10] are used for purpose of this module. The infrared motion sensor works on a similar principle as the precipitation sensor; by detecting the infrared radiation reflected from a moving object within the sensor's range. Infrared sensors have the disadvantage of not receiving signal effectively through the glass, so they should be mounted outside the vehicle, which is aesthetically undesirable. For this reason, microwave motion sensor(s) can be used, such is [11], which will surely detect all motions within the vehicle's surroundings, from inside the vehicle.

The basic function of automatic ventilation inside the vehicle is achieved using temperature sensor. Ventilation inside the vehicle is achieved by method of slightly lowering windows for a few centimeters, which makes it functional, discreet enough, but relatively reliable and safe. By adding the precipitation and motion sensors, the ventilation process becomes reliable and safe, but it opens up the possibility of even better ventilation. That is, now when there is a control system and protection from external impact, the windows can be further lowered to increase air circulation, and more air circulation means more quality ventilation. These critical input units are the most important for stable ventilation process. Other input units which contribute to efficiency and vehicle safety are:

- The light sensor; used for economic reasons. Given that the interior can be heated more than the environment only when the vehicle is exposed to direct sunlight, which does not exist at night, whole ventilation process at night does not make sense, and it is automatically suspended until morning. It results in a more efficient and economical usage of the module. Sensor is placed inside the vehicle, protected from direct

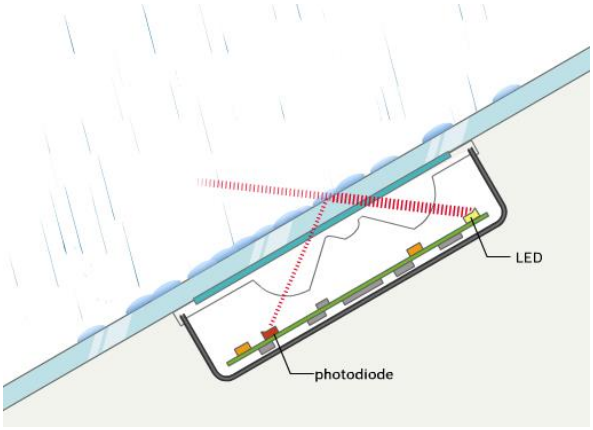


Figure 2. Precipitation sensor

sunlight in order to reduce sudden changes in light intensity readings caused by disruptive factors such as clouds. The sensor itself is actually a photo resistor that has a characteristic of resistance change depending on the intensity of light in the area where it is located. Photo resistor used here has an extremely fast response to changes in light intensity. Resistance changes are large even for the least change of the brightness. Therefore, it is necessary to use additional electronic components that slow down and reduce these rapid and large changes to obtain stable and reliable signal which can then be sent to the control unit for processing. [12]

- Position sensors; contribute to the quality work of module. These are the lower and upper end as well as mid position sensors. They are located in the vehicle doors, that is, the end positions of windows movement range. They signal an entirely opened or entirely closed position of the window. The mid position sensor signals when window reaches position where should be placed in the automatic ventilation mode. Position sensors provide a reliable windows setting in the desired position. This benefits to all aspects of the module quality; provides the reliability and security because the windows will always reach an end position whenever it is necessary, improves efficiency because windows will stop immediately when they reach a target position, preventing unnecessary overload of the electric motors and needles consumption of limited electrical energy, as well as potential danger of blowing electric motors or module is prevented and the vehicle is protected from potential conflagration. Moreover, a very important input information is obtained - direct feedback that informs the microcontroller on the execution of the given commands. All this greatly improves the quality of the ventilation process. Position sensor functionality is achieved using conventional buttons and switches. When reaching the mid or end position, the window carrier press the button/switch that closes the electronic circuit which confirms successfully executed operation to microcontroller. According to [12], various components and methods were considered for this purpose, such are photo and infra-red diodes or instead of using additional components this function is programmable via microcontroller in the form of time circuit. However, classic buttons and switches are the simplest, most reliable and most resistant components that are to be applied for this purpose.

B. Control unit

Control, or central logical unit, is the "brain" of the module. By default algorithm and collected input signals, module independently controls power windows in automatic mode - the automatic ventilation inside the vehicle. In manual mode, when a person is present in the vehicle, module just relays human commands onto the windows.

For this purpose is used AVR ATmega32 microcontroller. Since the microcontroller is available as one component, this is seemingly a very simple logical unit. However, it is a highly complex integrated circuit that contains thousands of electronic components inside [8]. All in all, this microcontroller is too powerful for this module. However, as it has an affordable price, and considering the additional module upgrades and extensions, it would be imprudent to limit its potential immediately at start by implementing a microcontroller with fewer features.

The control unit could be implemented even without microcontroller by using some other logic electronic components, however, the module would not be programmable, nor upgradeable or be adaptable to individual vehicles and customer requirements.

C. Switching unit

Switching unit is also used as the output unit and is responsible for the stable transmission of the output control signals from the microcontroller and to adjust them for window's electric motors. It consists of several components that are responsible for the conversion of weak control signals from the microcontroller to the strong electric signals that directly drive the electric motors. These are various transistor amplifiers and relays which, in addition to amplifying the weak control signals from the microcontroller, also galvanically isolate logic module's electronic circuit from a strong motor's electric circuit. There are various options of transfer and adjustment of control signals from the microcontroller to the electric motors. In this case, a driver L293NE is used. It turns logic signal from the microcontroller into much stronger electric signal that can directly power the electric motor. Considering that in real circumstances current reaches 1 A, it turns out that this driver component overheats and does not deliver enough power for electric motor to move windows quality enough. Therefore, besides the aforementioned galvanic circuit separation, relays are used to prevent overheating of the driver component, achieving a sufficient power for quality work of electric motors and prevent the flow of large currents within the module, which is not advisable because of the strong magnetic fields that are generated by large currents and their negative effects on sensitive electronic components.

D. Other units

Peripheral unit, control unit and switching unit compose a logical unit that is oriented to the specific module's role - automatic ventilation inside the vehicle. In addition to them, the module includes other components that perform a particular role; voltage converter and stabilizer, integrated connector and a set of components that allow the microcontroller to be programmed directly within the module.

III. ALGORITHM AND MICROCONTROLLER PROGRAMMING

It is mentioned that a microcontroller decides what to do and operates independently. The microcontroller basically just follows the algorithm behavior, predefined program of what to do in a certain situation. Simplified flowchart shows the microcontroller's algorithm of behavior (Figure 3).

Method of operation can easily be noticed by the flowchart. At the beginning, initialization of microcontroller and auxiliary variables is performed. After completion of the initialization, program starts checking parameters and thus sets the state of the output. First, it determines whether the vehicle is parked or in use. This is determined by checking whether the key is in the ignition lock, which is not a one hundred percent confirmation, but it is considered satisfactory. Depending on the result, manual or automatic mode is switched on. In manual mode, while a person uses the vehicle, microcontroller only transmits input state to output and thereby has only a minor role in the protection of electric motors from overload. When the automatic mode is on, other preconditions are checked in order of priority; whether it is night, is there precipitation and are there suspicious objects near the vehicle. If the answer to any of these queries is confirmative, it means the ventilation process is not necessary or it could endanger the vehicle and will not even start until the appropriate conditions. If all conditions are satisfied, only then microcontroller checks if the vehicle interior is overheated and, if it is, opens the windows for ventilation. If, at any time during the ventilation process, disturbing factors appear, the windows close, thus suspending the ventilation process until the required conditions are met again. Left and right window can be opened and closed separately in automatic mode. This also provides the efficiency. For example, if in the middle of the ventilation suspicious movements are detected on the right side of the vehicle, only the right window will close which still allows partial ventilation through left window only, which is better than none.

IV. FINAL IMPLEMENTATION

The module is designed and developed for the installation into a one older car without automatic air conditioner and with already existing basic electric mechanism for lowering and raising the windows. The

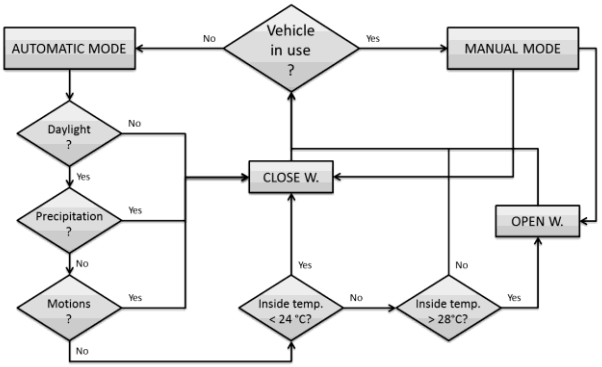


Figure 3. Flowchart diagram

car has two power windows which can be managed. That is why this module prototype manages only two sets of input and output signals for connecting two independent windows. Figure 4 shows the electric scheme of a module prototype. Control unit, switching unit, other units and some of the peripheral units (temperature and light sensors) can be perceived. Since the rest of the sensors cannot be integrated into the module's main electronic circuit, connectors for their signals are located inside the module. However, what is not mentioned so far are a lot of buttons, switches and LEDs that can be seen alongside input and output connectors. With buttons, switches and LEDs, this prototype module becomes the simulator as well – inside the module are simulators of all input and output signals that make this prototype even more convenient for the further development and testing. Although this is just a module prototype, it satisfies all requirements for working in real conditions and can be implemented in any vehicle.

Considering different characteristics of individual microcontroller pins and the various possible upgrades, important criterion was to place and connect components in a logical and calculated order, with respect to all present and possible hardware or software upgrades so it is important to be careful while designing and connecting all electronic circuits, inputs, outputs and electronic components in general. Integration of certain electronic components protects module against different sort of nuisances that exist or could appear in real circumstances. Electronic circuit was developed parallel with the program code to achieve even better congruence between individual components and operations. Besides all the attention that was given in the hardware development, a lot of attention was committed to the software segments that improve quality work of module and protect system against contingent situations. Aesthetic appearance of the assembly was also taken into consideration. Result is a complete product that is ready for real use. Of course, the software can still significantly improve in order to obtain a more "intelligent" and efficient module, but it is efficient enough, functional, reliable and safe to use.

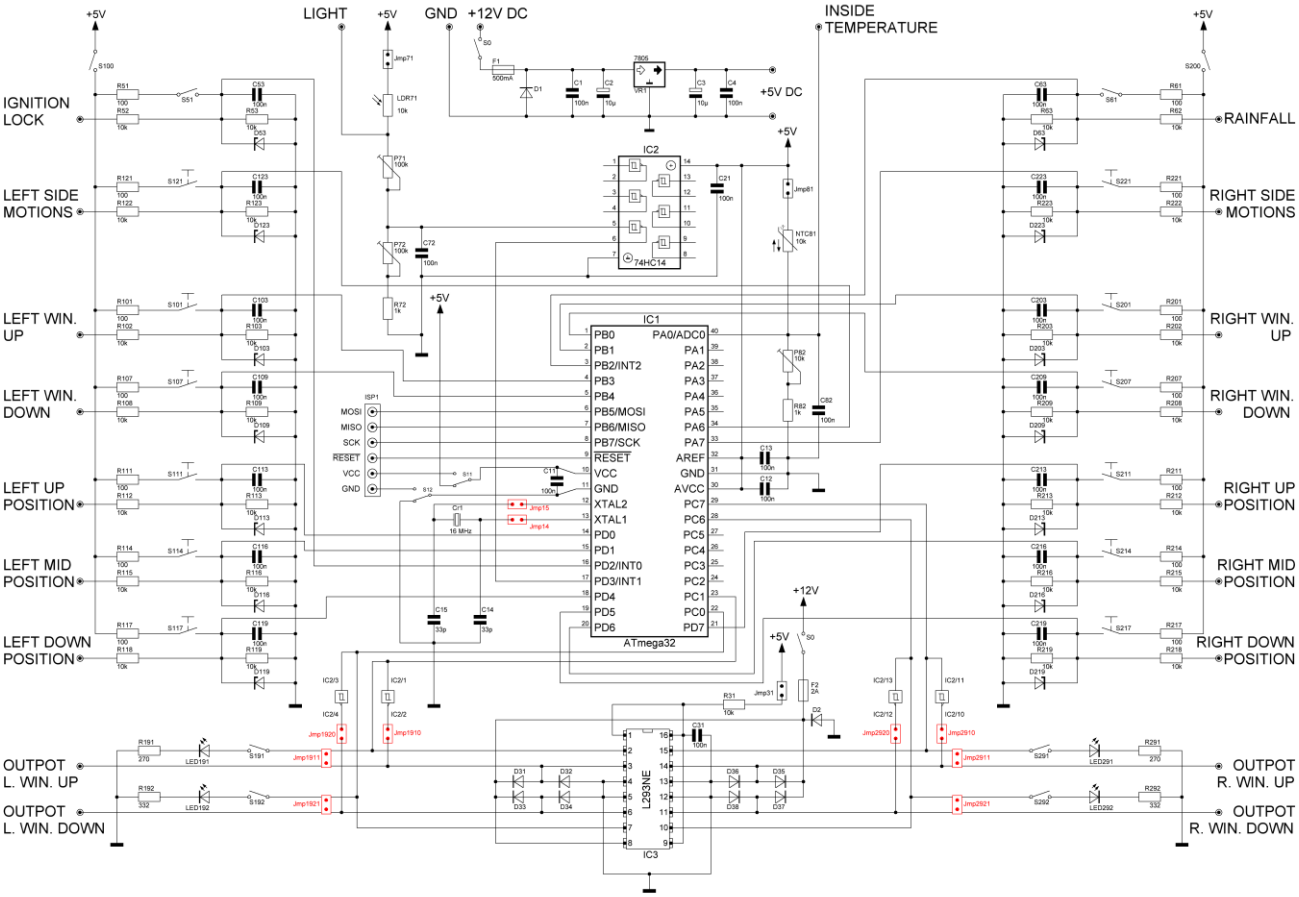


Figure 4. Electric scheme

The module is assembled using DIP (Dual In-line Package) electronic components, primarily due to the ease of making, testing and changing the hardware, which is quite important since it is a prototype that will spend most of the time in simulation instead of operating into the real world. Figure 5 shows the final implementation of printed circuit board.

Necessity of placing sensors around vehicle interior is already presented, but it is good to emphasize it again because it directly disables implementation of this module as one physical unit, which triggers other aggravating effects for easy installation and minimum costs.

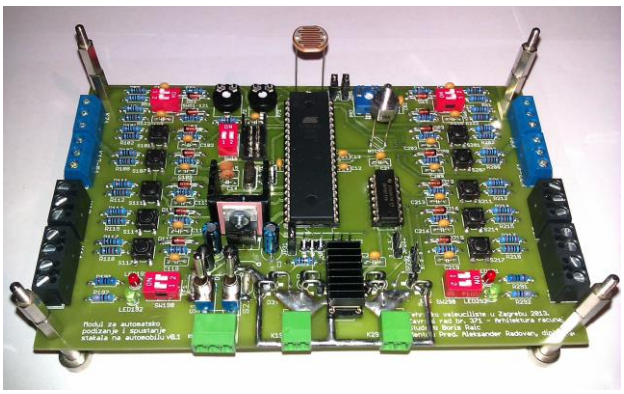


Figure 5. Final implementation

V. TEST RESULTS AND DISCUSSION

Module has been tested under real operating conditions in few vehicles on outside temperatures in range of 21 °C to 27 °C. Mid position sensors were set on few different levels. Also, temperatures inside vehicles while fully opened/closed windows were measured for better comparison. Measured temperatures are presented by Figure 6. The samples were collected during longer period of time. Comparing measured inside temperatures it is concluded that slightly opened windows (up to 5 cm) do not get satisfying results. To achieve quality ventilation it was necessary to lower windows at least 10 to 15 centimeters. Of course, more opened windows, and finally completely opened windows gives the best results, but 15 to 25 centimeters is optimal regarding inside temperature drop, energy consummation and safety concerns.

Presented results should be taken into consideration as informative since measurements were not taken in exactly the same weather conditions for each individual vehicle so there are visible disrupting effects. It can be remarked that many different factors, besides obvious direct sunlight, outside temperature and windows position directly affect ventilation process; weather conditions have significant influence on the ventilation process – e. g., wind noticeably accelerates and enhances ventilation. Further, size of the vehicle, even vehicle body color and vehicle orientation relative to the sun makes noticeable difference.

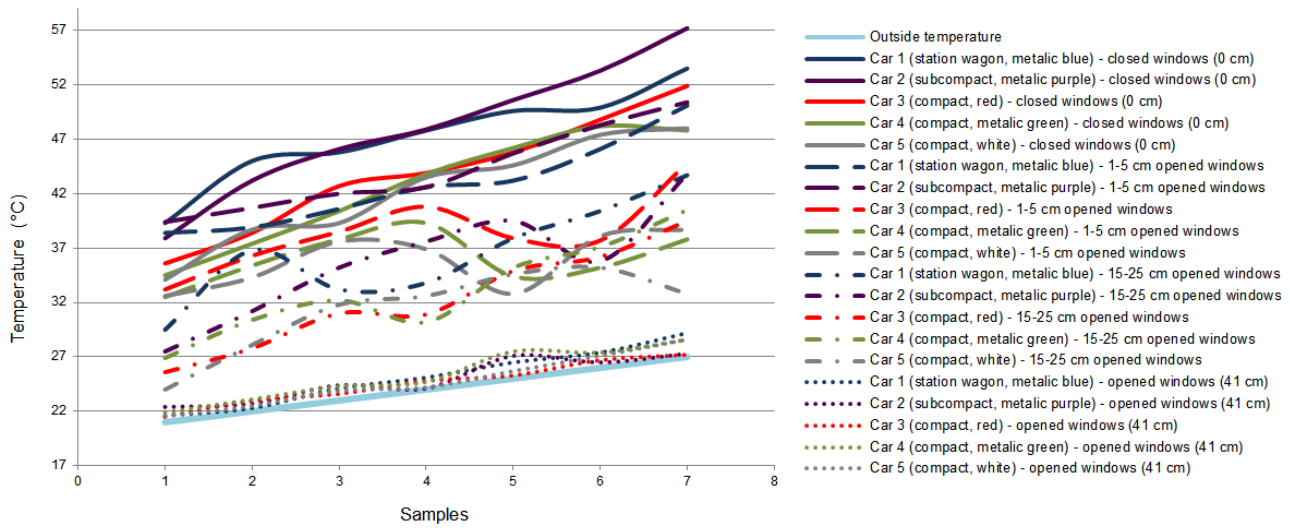


Figure 6. Measured temperatures

Except module proved to be safe, reliable and efficient, satisfactory results of the ventilation function were achieved. Without the module and with closed windows (or slightly lowered windows), temperature inside of the vehicle parked in the sun on light spring temperatures of 25 °C can reach over 50 °C. In summer, inside temperature rises even more drastically. While lowering windows for at least 10 to 15 centimeters to enable quality ventilation process, inside temperature becomes much bearable. Considering the fact that this ventilation process is a passive method of cooling overheated interior, it accomplishes very good results.

VI. CONCLUSION

It can be summarized that this module, although only a prototype, is completely functional device that performs its function safely, reliably and efficiently. Possibility of application is not limited only to the vehicle ventilation. The module is extremely expandable in both ways: it is upgradeable for new tasks as well as compatible for integration into some existing systems. Nearly every component can be used for several other functions, which can be achieved only by software upgrade. Upgrading hardware opens even greater possibility to upgrade the software and there are countless capabilities. Despite all the positive sides of this module, the question remains whether it can find its application in the market. Since the module is impracticable as a physical unit (because of peripheral unit), it is necessary to adapt it to any individual vehicle. Thus its individual installation is not cost-effective. It could find its purpose and economic efficiency if it would be serially mounted in same type of vehicle.

ACKNOWLEDGEMENTS

The authors are grateful to Polytechnic of Zagreb and University of Applied Sciences Velika Gorica for financial support.

REFERENCES

[1] Leo Budin, Microcomputers and microcontrollers, Element, 2004

[2] Slobodan Ribarić, Advanced microprocessor architectures, Element, 1999

[3] Antun Šarčević, Electronic components and analog circuits, Technical school of Ruđer Bošković, 1996

[4] Aleksandar Szabo, Impulse and digital electronics – part 1, Technical school of Ruđer Bošković, 2000

[5] Aleksandar Szabo, Industrial electronics, Technical school of Ruđer Bošković, 1996

[6] Stanko Paunović, Digital electronics – part 1, Školska knjiga, 2005

[7] Hans Bürklin, The whole electronics – Main catalog '92, datasheet, 1991

[8] Atmel Corp., AVR ATmega32(L), datasheet, 2011 www.atmel.com/images/doc2503.pdf

[9] Guangzhou Besyo auto parts Co, Ltd., BY-R606N, brochure http://www.besyoautoparts.com/Products_hb.asp?id=1704

[10] Velleman nv, PIR416, user manual, 2011 http://www.velleman.eu/downloads/3/pir416a705.pdf

[11] SOAN Electronic Technology Co., Ltd., KG001-12, user manual http://www.savebase.com/infobase/downloads/Kris_products/KG001-12V.pdf

[12] Boris Raić, Module for automatic raising and lowering car windows, unpublished

Acquisition Card NI USB-6009 in Automatic Control Laboratory Exercises

Hrvoje Benić, Mato Fruk, Goran Vujisić
Department of Electrical Engineering
The Polytechnic of Zagreb
Konavoska 2, Zagreb, Croatia
hbenic@tvz.hr, mato.fruk@tvz.hr, gvujisic@tvz.hr

ABSTRACT - This paper presents methods for collecting and processing data using the acquisition card NI USB-6009 and MATLAB's Data Acquisition Toolbox. It also enables analysis of control system elements and responses of closed loop systems. Results obtained by series of experiments are used as examples for further use in Automatic Control Laboratory Exercises. A detailed explanation of procedures can be used as a base for the preparation of further laboratory exercises. The paper covers topics of initial installation and various use of the acquisition card.

I. INTRODUCTION

This paper will show the necessary steps how to link acquisition card NI USB-6009 and Matlab. After these procedures various experiments and measurements will be presented how to use this device. There are some problems that occur for this type of acquisition card that can be minimized by certain methods and procedures that will be explained in the paper. Today's technology allows the collection and processing of data from many physical processes using acquisition cards and computers. That makes it easier to process data for analysis of control systems. The initial idea of this paper is to improve and simplify laboratory exercises. It will be further described and done in this paper. The paper will also describe how to use Data Acquisition Toolbox, which is located within Simulink, which is part of the Matlab.

II. DATA ACQUISITION

Data acquisition is the procedure of collecting information from the process. With analog instruments it is difficult to collect data. There is also a lot of problems such as noise, drift, instability and high energy consumption. In digital systems it is easier to control noise and they provide easy transfer and storage of data. Digital signal processing is virtually unlimited. It all makes digital systems better for data acquisition.
Measurement and acquisition system consists of three parts:

1. acquisition
2. analysis
3. data presentation.

Part that relates to the data acquisition is a card that connects to a PC. Here it is realized by:

1. connection
2. A/D conversion
3. signal processing.

The analysis part is performed by calculations such as filtering, spectral analysis, data formatting and statistics. Data formatting is necessary because different devices and instruments may require information in a different format. Third part of the system is a presentation, which provides displaying, saving and printing of results, data transfering between different applications and communications between systems and networks.

III: A/D CONVERSION

The tasks in this paper were processed with analog data and for that there are given basics of analog-digital conversion. A conversion process consists of three steps.

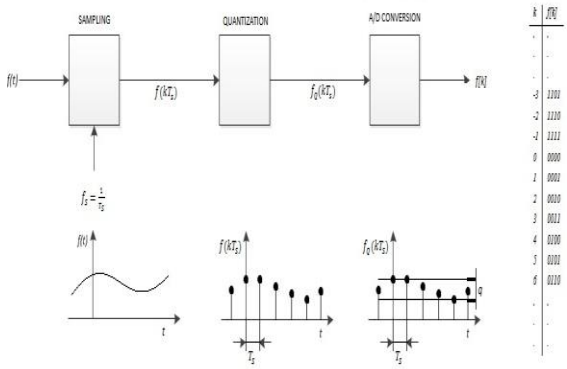


Fig. 1 The basic model of the conversion of the analog signal f (t) into a digital format that is shown a series of f [k]

The first step is carried out by the time discretization. With quantization level signal becomes a discrete value. In the final step there is performed coding of quantization level samples $f_Q(kT_S)$ or analog digital conversion. Each sample is associated with a corresponding binary number *f[k]*. A/D simbol stands for unit for analog to digital signal conversion (coder). That unit makes digital signal in the third step of the conversion. In practice an electronic circuit performs operations from Fig 1. Some of the signals between the two steps don't exist in practice. That is why usually ADC (*Analog Digital Converter*) stands for analog to digital converter as a unit that makes operations from each of three steps of the conversion. Analog signal must be reconstructed after the digital signal processing.

That conversion takes place in two steps. In the first step impulses of samples were reconstructed based on digital data *f[k]*. There is performed digital to analog conversion. This produces quantized level pulses $f_Q(kT_S)$. In the second step the analog form of the signal is beening reconstructed$f_r(t)$. It can be expected deviation of reconstructed analog signal$f_r(t)$. This deviation is due to quantization errors. Corresponding analog signal is reconstructed by filtering samples.

IV. ACQUSITION CARD NI USB-6009

Acquisition card *National Instruments USB-6009* allows data acquisition for mobile measurements, practice and lab measurements etc. It connects to the computer via the USB interface and that is a big advantage. The card is good for some complex measurements that can be seen at lab exercises.

A. SPECIFICATIONS

Acquisition card *NI USB-6009* has 8 analog inputs, 2 analog outputs, 12 digital inputs/outputs and 32-bit counter. Maximum sample speed by each analog input is 48 kS/s. Sample speed on analog outputs is 150 S/s and it can't be changed. Analog inputs have 14-bit resolution and analog outputs have 12-bit resolution. USB interface allows better transferability and easier connecting with the PC. The external appearance of the card is shown in Fig. 2.



Fig. 2 Acquisition card NI USB-6009

Maximum voltage that can be connected to analog inputs is from -20 V to 20 V (card specifications says -10 V to 10 V, but in practice -20 to 20 V can be connected). Analog outputs voltage is from 0 to 5 V and it can't be changed. The card connects with PC via USB interface. At digital inputs it can be connected voltage from 0 to 5 V and at digital outputs it gives 5 V. The card also has a counter that gives voltage from 0 to 5 V and maximum frequency of 5 MHz.

The card also has a green control LED that switches on when the card is connected to a PC via USB interface. If LED is not switched on card is not connected properly or it is malfunctioned. If it flashes that means that the card is in function.

Conductors merge with inputs and outputs with screw fittings connections that are at sides of the card.

Analog inputs and outputs were used in measurements.

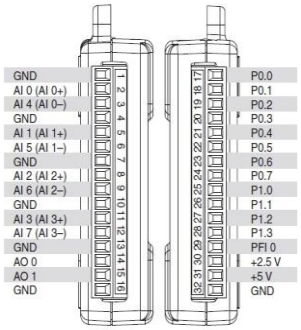


Fig. 3 Colloaction of screw fittings connections

B. CARD INSTALATION

At the beginning of using the card it is necessary to install a driver that is obtained on a CD with the card. Program *Measurement & Automation Explorer* needs to be installed. *Measurement & Automation Explorer* is used to test the card, reset it and adjust its characteristics.

V. FUNCTIONAL TESTING OF NI USB-6009

A. USE OF PROGRAMMING PACKAGE MATLAB

Work with acquisition card *NI USB-6009* is possible with 32-bit programm package *Matlab* and its library *Simulink*. Inside of *Simulink* there is a toolbox called *Data Acquisition Toolbox* which reads the card and makes possible to use its inputs and outputs in schemes inside of *Simulink*. That makes measurement easier because it enables to work with data inside of program.

Data Acquisition Toolbox takes place inside of *Simulink* library. *Simulink* library takes place inside of *Matlab*.

B. THE INITIAL MEASUREMENT

The first task was to observe the signal delay when processing data inside the card. Sinusoidal frequency 1 Hz was selected for the test signal. It was obtained from the function generator and it was brought to analog input and observed with 2-channel digital oscilloscope on analog input and output of the card. The model is made inside *Simulink* with usage of *Data Acquisition Toolbox* and it is very simple. Input and output parameters are set within the model. The window in which the parameters are set opens by double-clicking on the

input block. It is possible to select the input that will be used, sampling frequency, input voltage range, mode and whether there is used one or more cards. Modes are asynchronous and synchronous. The difference is that the asynchronous mode is FIFO (First In, First Out), or the card sends data out in the order that it receives and synchronous mode is such that all data are processed within the card and then it is sent to the output. The difference between these two modes is observed in larger models and affects the duration of the simulation. Asynchronous mode is faster. Output parameters that can be set are: output that we want to use and how many cards we want to use. Output voltage range is from 0 to 5 V and it can't be changed.

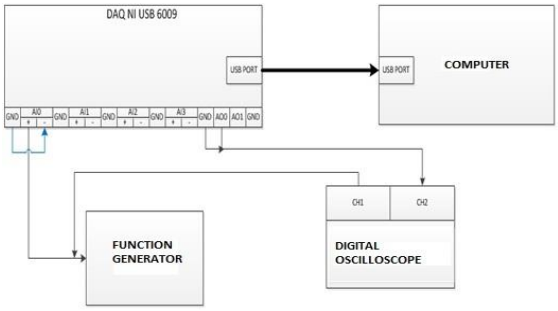


Fig. 4 Linking Up

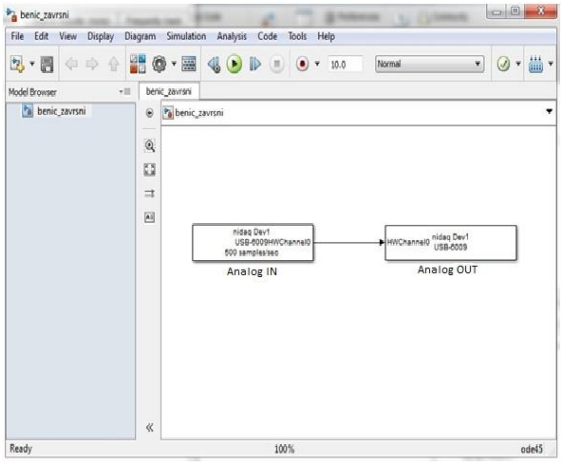


Fig. 5 Simulink model

The measurement signal on the analog input was sampled with the frequency of 500 S/s. Sample frequency is one of the factors that affects on data processing speed and the duration of the simulation. Sampling frequency at analog output is 150 S/s and it can't be changed as mentioned before.

Measured result is that sinusoidal frequency at the output of the card has delay of approximately 136 ms. Reason for that is data processing inside of the card that requires A/D conversion on the input of the card and D/A conversion at the output. Another reason is time that is necessary for acquisition and data processing that is given by a *Simulink* model. The result of this measurement is shown in fig 6.

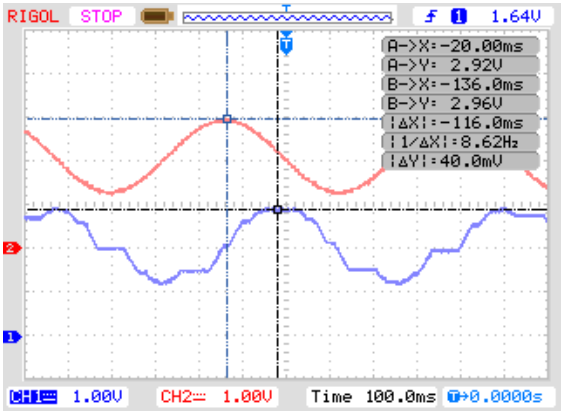


Fig. 6 Initail measurement of card delay

C. TEST RESULTS FROM THE INITIAL MEASUREMENT

Given results were obtained with repeated measurements. The test signal obtained from function generator was now rectangular with the frequency of 1 Hz and 10 Hz. Digital oscilloscope *Velleman PCSU1000* which is connected to the computer was used. Given results show that card has delay in range of 120-140 ms which confirms result that was given in initial measurement.

VI. RECORDING AND PROCESSING TIME RESPONSES OF SECOND ORDER ELEMENTS

A. ESTIMATION OF THE TRANSFER FUNCTION OF PT_2 ELEMENT

Double RC circuit is second order element. It is described by differential equation of the second order. It has two time constants and two energy storges. Energy storges are its two capacitors.

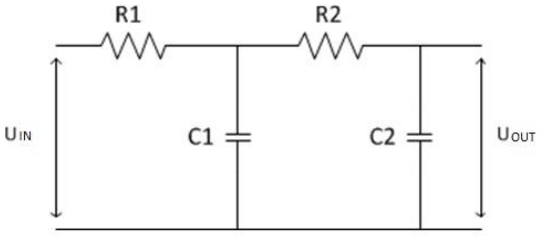


Fig. 7 Double RC circuit

The purpose of this experiment is to examine the possibilities of card. Tool *SIT* (*System Identification Toolbox*) is used in this experiment. *SIT* is used in a way that it compares the given time response with some known response inside *SIT* and it determines in which percentage they match with each other. *SIT* automatically determines system parameters (gain, time constants, dead time and transfer function).

The experiment measurement was made in a way that the double RC circuit was connected with card.

Input voltage of the double RC circuit is program designed step which is connected to the analog output of card. Output voltage of the double RC circuit is brought to the analog input of the card and it is observed on program oscilloscope.

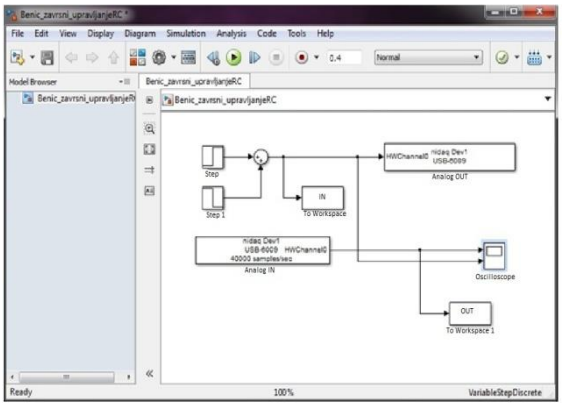


Fig. 8 Simulink model

Sampling frequency at the analog input is 40000 S/s and at the analog output it is 150 S/s. Input voltage of the double RC circuit are two program designed steps which starts in different moments. At Fig. 8 it is evident that given response matches with assumed in 98.68%.

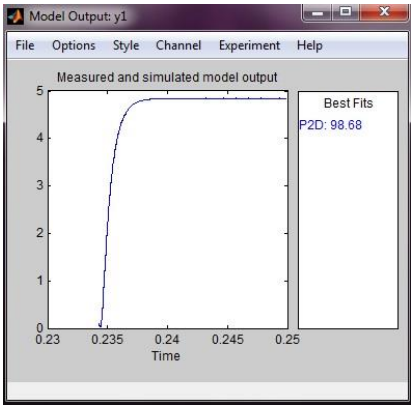


Fig. 9 The resulting accuracy of double RC circuit

Obtained accuracy is satisfying. Obtained time constants are almost equal to the time constants obtained by habitual classic procedure.
That all means that the card is good for measurements like in this experiment and that it can be used for laboratory exercises of this type.

B. ESTIMATION OF THE TRANSFER FUNCTION OF PT_2S ELEMENT

Serial RLC circuit is the element of second order. It has two energy stores and two time constants. It is described by differential equation of second order. Its energy stores are capacitor and coil.

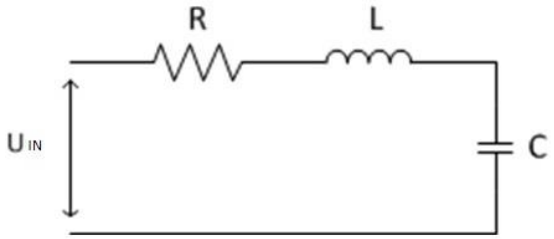


Fig. 10 Serial RLC circuit

The purpose of this experiment also is to examine the possibilities of card. The experiment measurement is made in the same way as last one. *SIT* (*System Identification Toolbox*) is also used. The expected response of the system is damped oscillatory response. The difference compared to the previous experiment is that the frequency of sampling analog input is set on 10000 S/s. Obtained system response is shown on Fig. 11.

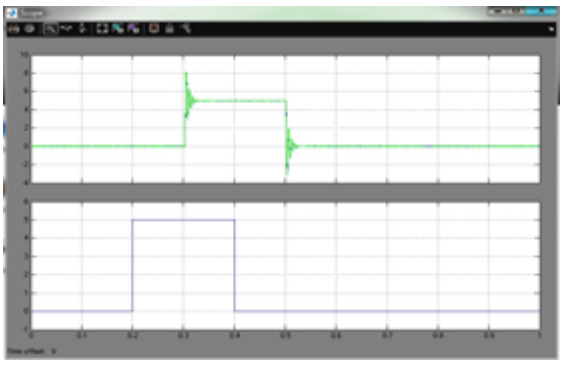


Fig. 11 Response of serial RLC circuit

System response is like it was expected. SIT is used in a further process of identifying the parameters of system as in the previous measurement.

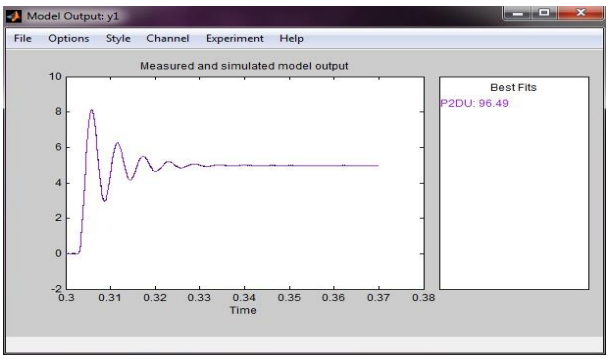


Fig. 12 The resulting accuracy of serial RLC circuit

Obtained accuracy is 96.49%. That is also good result like in previous measurement.
Conclusion of this test is that card is suitable for this type of measurement in laboratory exercises.

VII. CONCLUSION

After various tests of *NI USB-6009* work potentials with use of *Matlab* conclusion is that the card is good for laboratory excersises. Plenty of measurements performed on laboratory exercises in subject *Automatic control* can be simplified by using the card and *Matlab*. It is useful that the students familiarize themselves with the data acquisition and virtual instrumentation because that is present and the future and it is is likely for them to encounter some kind of data acquisition in their profession after they graduate.

While the card was tested some conclusions about possibele improvements were made. Folowing pro and cons:

Disadvantages:

1. Card with use of *Data Acquisition Toolbox* works only on 32-bit *Matlab*
2. The voltage at the analog outputs of the card is limited to the range of 0-5 V, and can not be changed.
3. Samle frequency at analog outputs is 150 S/s and can not be changed.
4. Maximum sample frequency at the analog inputs (48 kS/s) can be achieved using only one analog input. Maximum sampling frequency is reduced if used multiple inputs simultaneously.
5. It often happens that *Matlab* simulation stops working and that *Matlab* blocks if the card is not previously reseted inside *Measurement & Automation Explorer*.
6. The card has a signal latency which is not fixed.
7. Card does not work in real-time mode.
8. Simulation time and real time of process are not the same. Eg. for simulation time 10 s process usually will not last 10 s. The time required for the implementation of the simulation depends on many factors, including the complexity of the model and the step size inside Solver.
9. Card has a problem with noise and drift while receiving and processing analog data.

Benefits:

1. Connecting to a computer via the USB interface.
2. Easy to work in a graphical environment (*Data acquisition toolbox*).
3. Small dimensions.
4. No additional power supply of the card is needed.
5. Using the card reduces the necessary equipment to perform labaratory exercises.
6. The card "remembers" the last voltage state in which it found itself.
7. Simulation time inside of *Simulink* can be set on unlimited in a way that instead of numerical value for end of the simulation enters *inf*
8. Using the *Rate transition* block.



Fig. 13 Rate Transition Block

The *Rate transition* block is inside of S*imulink* library (**Simulink → Signal attributes → Rate transition**) and it is used for changing sample frequency to its divisor. (Eg. 1000 S/s u 100 S/s). The purpose of this is that signal at analog input can be sampled on a given frequency, and within the simulation can be viewed only parts of the signal (Eg. if at input is 1000 S/s it is possible for some part of the simulation later to put 100 S/s and watch every tenth sample of the input signal). It makes possible that in simulation every block needs not to have same sample frequency and it is possible to work with fixed discretization step.

Possible improvements:

1. Work with 64-bit *Matlab*.
2. Possibility of changing range of output voltage at analog outputs.
3. Possibility of changing sample frequency at analog.
4. Troubleshooting blocking simulation.

BIBLIOGRAPHY

1. http://sine.ni.com/nips/cds/view/p/lang/en/nid/201987, 9.6.2013.
2. I. Alfirevićet, "Inženjerski priručnik IP3", Školska knjiga, Zagreb, 2002.
3. User guide and specifications NI USB-6008/6009, http://www.ni.com/pdf/manuals/371303m.pdf, 9.6.2013.
4. J. Božičević, "Automatsko vođenje procesa: Prvi dio", Tehnička knjiga, Zagreb, 1971.
5. G.Vujisić, "Identifikacija parametara linearnih procesa korištenjem programskog paketa Matlab System Identification Toolbox", TVZ, Diplomski rad br. E191, Zagreb, 2012.

Udaljeno praćenje ljudske aktivnosti prikupljanjem podataka s ugrađenih senzora pametnog mobitela

J. Begovac i L. Celić
Sveučilište u Zagrebu Fakultet elektrotehnike i računarstva, Zagreb, Hrvatska
juraj.begovac@fer.hr
luka.celic@fer.hr

Sažetak - Korištenje pametnih mobitela koji sadrže veliki broj senzora, kao što su žiroskop, magnetometar, akcelerometar i sl. raširenо je u populaciji, pa se sve više razmišlja o njihovom korištenju za udaljeno praćenje kroničnih bolesnika i uključivanju u programabilno zdravstvo (m-zdravstvo). Cilj ovoga rada je ispitati vjerodostojnost udaljenog praćenja osoba prikupljanjem podataka izmjerenih troosnim akcelerometrom koji je ugrađen u pametni mobitel. Istraživanje je provedeno na ispitanicima koji redovito i rutinski koriste pametni mobitel. Kao kontrolni parametar o trenutnoj fizičkoj aktivnosti ispitanika, ispitanik je ručno unosio podatak o planiranoj aktivnosti. Fizičke aktivnosti koje su se pratile tijekom istraživanja su hodanje, hodanje uz stepenice, hodanje niz stepenice, brzo hodanje i trčanje, te sjedenje, kao neaktivno stanje. Obrada prikupljenih podataka je obavljena pomoću programskog paketa Matlab. Za signal svake aktivnosti izračuna se brza Fourierova transformacija te se usporedbom spektara signala pojedinih aktivnosti traže razlike radi vrednovanja mogućnosti automatske klasifikacije aktivnosti. Temeljem spektra signala ubrzanja, moguće je dobro klasificirati sjedenje, trčanje i općenito hodanje, ali je teža klasifikacija pojedinih oblika hodanja (normalno, uz stepenice, niz stepenice, brzo hodanje).

I. UVOD

Zahvaljujući inženjerskim rješenjima danas je moguće pratiti zdravstveno stanje osoba oboljelih od kronične bolesti i/ili tijekom rehabilitacije i izvan zdravstvene ustanove [1,2,3]. Ovisno o situaciji (bolesti, ozljedi itd.), osobe kroz dnevni život, terapiju i rehabilitaciju vodi stručni tim uz pomoć informacijsko-komunikacijske tehnologije. Prijedlozima i savjetima tim pomaže praćenoj osobi što djelotvornije provoditi predviđene dnevne aktivnosti i mjere potrebne za oporavak. U mjere prevencije pogoršanja bolesti, uz edukaciju i promociju zdravog načina življenja, uključuje se fizička aktivnost putem vježbanja, rehabilitacija u izvan-zdravstvenim ustanovama te udaljeno praćenje osobe oboljele od kronične bolesti u vlastitom domu [4].

Mobilna tehnologija pruža velike mogućnosti u daljnjem unaprjeđenju zdravstvene njege kroz mobilne zdravstvene aplikacije, ugrađene senzore u mobilne uređaje i povezivanje mobilnih uređaja s medicinskim uređajima [5]. U današnje vrijeme raširenо je korištenje pametnih mobitela (*smartphone*) [6]. Pametni mobiteli sadrže veliki broj senzora, kao što su žiroskop, magnetometar, akcelerometar i sl., pogodnih za praćenje fizičke aktivnosti oboljelih osoba i primjenljivih u njihovoj rehabilitaciji [7]. Cilj ovoga rada je ispitati vjerodostojnost udaljenog praćenja osoba prikupljanjem podataka izmjerenih troosnim akcelerometrom koji je ugrađen u pametni mobitel.

Za istraživanje provedeno ovim radom je korišten sustav koji čine aplikacija za operacijski sustav (OS) Android razvijena za potrebe ovog rada te računalo na kojem su obrađeni prikupljeni podaci (Slika 1.). Funkcionalnost Android aplikacije je prikupljanje podataka s akcelerometra ugrađenog u pametni mobitel i prikupljanje podataka o aktivnosti ispitanika (korisnikov *feedback*).



Slika 1. Sustav za ispitivanje vjerodostojnosti udaljenog praćenja osobe pomoću pametnog mobitela

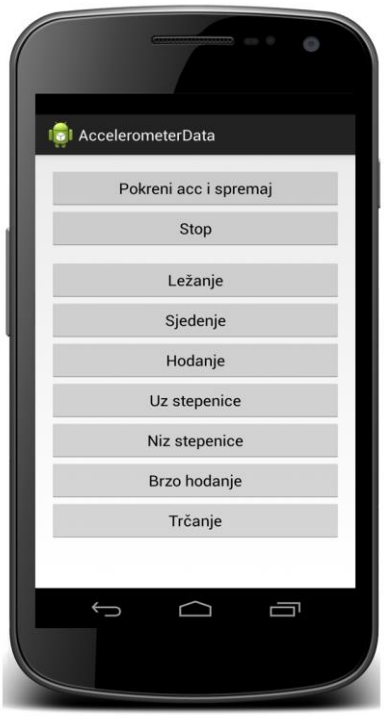
II. MJERENJE I OBRADA PODATAKA

A. *Aplikacija za operacijski sustav Android*

U ovom radu je kao ispitni uređaj korišten Samsung Galaxy Nexus I9250, koji ima ugrađen troosni akcelerometar BMA220 [8].

Razvijena je aplikacija naziva „AccelerometarData“ za operacijski sustav Android (Slika 2.). Primarna funkcionalnost aplikacije je prikupljanje izmjerenih podataka sa senzora pametnog mobitela o fizičkoj aktivnosti korisnika aplikacije (ispitanika). Međutim, kako bi se dobila kontrolna informacija o fizičkoj aktivnosti korisnika, aplikacija ima funkcionalnost prikupljanja informacije korisnika o fizičkoj aktivnosti koju trenutno izvršava (korisnikov *feedback* o njegovoj fizičkoj aktivnosti). Sljedeća funkcionalnost aplikacije je spremanje navedenih podataka i informacija u tekstualnu datoteku u memoriju pametnog mobitela.

Korisničko sučelje aplikacije se sastoji od jednog zaslona koji sadrži tipke. Tipke su podijeljene u dvije sekcije. U prvoj sekciji se nalaze dvije tipke koje imaju funkcionalnost pokretanja i zaustavljanja Android servisa koji radi u pozadini. Funkcionalnost navedenog servisa je prikupljanje mjernih podataka s akcelerometra.
Drugu sekciju čine tipke čija je funkcionalnost prikupljanje korisnikovog unosa o vlastitoj aktivnosti.



Slika 2. Aplikacija za praćenje fizičke aktivnosti

B. *Snimanje signala*

Za potrebe istraživanja izabrane su uobičajene dnevne aktivnosti, a snimanje signala akcelerometra tijekom pojedine fizičke aktivnosti odvijalo se na Fakultetu elektrotehnike i računarstva. Signali su snimani na odabranoj grupi studenata. Snimale su se sljedeće fizičke (ne)aktivnosti: sjedenje, hodanje, hodanje uz stepenice, hodanje niz stepenice, brzo hodanje i trčanje.

U istraživanju je korišten pametni telefon koji je bio smješten u džepu uskih hlača (traperica) ispitanika za vrijeme obavljanja svake navedene fizičke aktivnosti.

Moguće frekvencije očitavanja podataka s akcelerometra (*sampling rate*) ovise o primijenjenom mobilnom uređaju. Za potrebe ovog istraživanja odabrana je frekvencija očitavanja SENSOR_DELAY_GAME, tj. frekvencija vrijednosti 50Hz [7].

C. *Obrada podataka i klasifikacija aktivnosti*

Nakon što su podaci snimljeni i spremljeni u tekstualnu (.txt) datoteku u memoriju pametnog mobitela, moguće ih je dohvatiti na računalu povezivanjem mobitela s računalom.

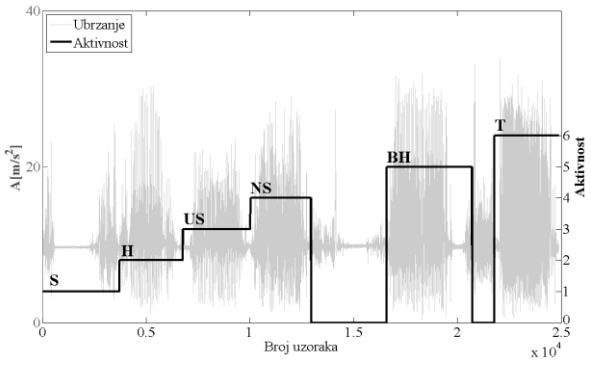
Nakon dohvata podataka na računalu, obavlja se obrada podataka u programskom paketu Matlab. Za potrebe rada su razvijene Matlab skripte (.m skripte) koje sadrže programski kod za obradu podataka.

Budući da akcelerometar mjeri komponente ubrzanja u tri smjera (x, y i z) u ortogonalnom sustavu, u prvom koraku se podaci učitani iz tekstualne datoteke spremaju u četiri vektora (*ax*, *ay*, *az* i *activity*). Prva tri navedena vektora predstavljaju vrijednosti komponenata ubrzanja izmjerene akcelerometrom pametnog mobitela. Dodatni vektor (*activity*) predstavlja zapisani podatak korisnika o aktivnosti koju je obavljao.

Zatim se pomoću (1) izračuna iznos (*magnitude*) linearnog ubrzanja [9].

$$A = (ax^2+ay^2+az^2)^{1/2} \qquad (1)$$

Paralelni prikaz iznosa ubrzanja (A) i unesene informacije o aktivnosti koju korisnik obavlja nalazi se na slici (Slika 3.).



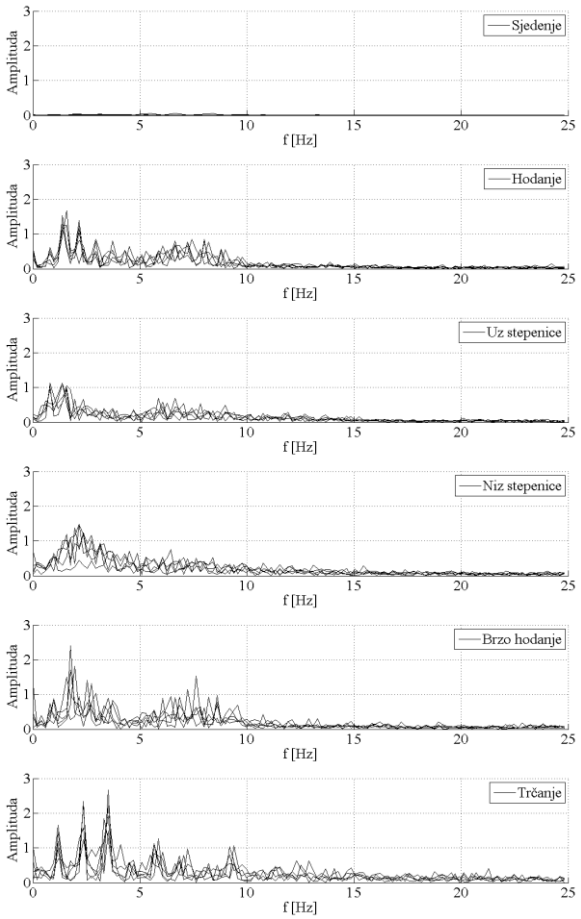
Slika 3. Paralelni prikaz ubrzanja i korisničkog unosa tipa aktivnosti
(S – sjedenje, H – hodanje,
US – hodanje uz stepenice,
NS – hodanje niz stepenice,
BH – brzo hodanje, T – trčanje)

Nakon usporedne analize signala ubrzanja i korisničkog unosa tipa aktivnosti, izdvajaju se uzorci signala ubrzanja za svaku aktivnost i pohranjuju u zasebne datoteke. Za

svaku aktivnost, nad 256 uzoraka izračunava se brza Fourierova transformacija (*Fast Fourier Transform*, FFT) te se usporedbom izračunatih spektara pojedinih aktivnosti traže razlike radi vrednovanja mogućnosti automatske klasifikacije aktivnosti.

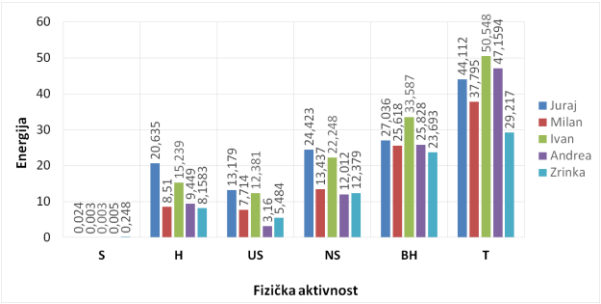
Prikaz spektara svih aktivnosti jednog ispitanika se nalazi na slici (Slika 4.). Slika spektra se dobiva tako što se nad pet različitih dijelova signala od 256 uzoraka izračunava FFT za svaku aktivnost. Na temelju izračunatih spektara, može se zaključiti:

- za vrijeme sjedenja amplitude signala na svim frekvencijama su približno jednake nuli,
- u amplitudnom spektru signala trčanja (u usporedbi sa signalom hodanja) izraženije su amplitude na višim frekvencijama,
- temeljem spektra signala ubrzanja, moguće je dobro klasificirati sjedenje, trčanje i općenito hodanje, bez obzira o kojem obliku hodanja se radi (normalno, uz stepenice, niz stepenice, brzo hodanje),
- temeljem spektra signala ubrzanja teža je klasifikacija pojedinih oblika hodanja.



Slika 4. Spektri aktivnosti jednog ispitanika

Dodatno se može primijetiti da je ukupna energija prilikom trčanja značajno veća od energije za bilo koji oblik hodanja. Ostale amplitude variraju od ispitanika do ispitanika (Slika 5.).



Slika 5. Izmjerena energija pojedinih aktivnosti za različite ispitanike
(S – sjedenje, H – hodanje,
US – hodanje uz stepenice,
NS – hodanje niz stepenice,
BH – brzo hodanje, T - trčanje)

III. ZAKLJUČAK

Na temelju istraživanja može se zaključiti da je analizom spektra signala ubrzanja moguće dobro klasificirati (ne)aktivnosti sjedenje, trčanje i općenito hodanje. Međutim, teže je klasificirati slične aktivnosti kao na primjer klasifikacija između pojedinih oblika hodanja (normalno, uz stepenice, niz stepenice, brzo hodanje). Uključivanjem dodatnih senzora (npr. žiroskopa) koji su ugrađeni u pametni mobitel odluke o klasifikaciji aktivnosti bile bi izraženije, a kombiniranjem signala s različitih senzora dobila bi se bolja klasifikacija aktivnosti.

U daljnjem istraživanju praćenja ljudske aktivnosti prikupljanjem podataka s ugrađenih senzora pametnog mobitela planirano je praćenje višesatne aktivnosti ispitanika. Planirano je provođenje obrade signala i klasifikacija aktivnosti na pametnom mobitelu umjesto na računalu. Također, planirano je ispitivanje automatiziranog praćenja aktivnosti ispitanika smještanjem pametnog mobitela na različita mjesta uobičajena za nošenje pametnog mobitela (npr. džep od jakne, torbica ili džep od komotnije odjeće poput trenirke).

ZAHVALA

Zahvaljujemo se našem mentoru i profesoru Ratku Magjareviću na tome što nas je poticao u izradi ovoga rada te se zahvaljujemo na njegovim savjetima. Također, zahvaljujemo se Kristijanu Jurkoviću koji nam je pomogao sa savjetima pri izradi aplikacije za operacijski sustav Android.

LITERATURA

[1] M. Varga, T. Pozaić, S. Žulj, L. Celić, R. Magjarević, „Raspoznavanje i kvantifikacija tjelesne aktivnosti“, vol. 5, MIPRO, 2011, pp. 426 – 431

[2] T. Pozaić, D. Džaja, M. Varga, I. Matec, L. Celić, I. Lacković, R. Magjarević, „Quantitative and Qualitative Assessment of Assisted Strength Exercising“, vol. 42, IFMBE, 2014, pp. 71-75.

[3] M. Varga, J. Begovac, J. Cvitanović, „Interaktivna Android aplikacija za praćenje vitalnih funkcija inteligentnom majicom“, neobjavljen

[4] H. Viswanathan, C. Baozhi, D. Pompili, "Research challenges in computation, communication, and context awareness for ubiquitous healthcare", vol. 50, IEEE Communication Magazine, 2012, pp. 92-99

[5] D. M. West, „Improving Health Care through Mobile Medical Devices and Sensors“, Center for Technology Innovation at Brookings

[6] Mobile Future: Mobile Year in Review 2013., 25.03.2014. (zadnji pristup), http://mobilefuture.org/resources/2013-mobile-year-in-review/

[7] Google Inc, „Sensors Overview”, 27.02.2014. (zadnji pristup) http://developer.android.com/guide/topics/sensors/sensors_overview.html

[8] BOSCH, „BMA220 Digital, triaxial acceleration sensor”, Datasheet

[9] Q. Li, J. A. Stankovic, M. A. Hanson, A. T. Barth, J. Lach, G. Zhou, „Accurate, Fast Fall Detection Using Gyroscopes and Accelerometer-Derived Posture Information“, BSN, 2009, pp. 138-143

Visualization of Streaming Trajectories

D. Osmanović, E. Mešković
University of Tuzla, Faculty of Electrical Engineering, Franjevačka 2,
75000 Tuzla, Bosnia and Herzegovina
dario.osmanovic90@gmail.com, emir.meskovic@untz.ba

Abstract - One of the basic requirements when managing moving objects in data stream management systems (DSMS) is visualization of continuous queries results on the map. In most cases, the result of a continuous query is data stream whose elements contains information about locations of moving objects so that we are talking about visualization of streaming trajectories. Such visualization involves redrawing the location of moving objects for each new arriving element of data stream that represents the result of a continuous query. In this paper we present an application that allows you to pose continuous queries and visualize their results in OCEANUS DSMS prototype for managing spatio-temporal data streams. The application is based on the Java programming language, swingx-ws library for interacting with web services and OpenStreetMap web service that provides free maps of the world. The application allows users familiar with CSQL to pose queries directly using this query language, while beginners can pose a query through creating the conditions by selecting one or more operations from a set of basic operations on moving objects. In addition to presenting current positions of each moving object which satisfies the condition of continuous query it is also enabled to view the entire trajectory of a moving object.

I. INTRODUCTION

Moving object databases support objects whose position and/or extent changes continuously and enable posing queries not just about history of movements, but also about current and future movements. Such databases are overloaded with frequent updates of moving objects' positions which is the reason for objects' movements to be considered as continuous, time varying and possibly unbounded data streams. Several research projects provided prototype systems for managing such large and possibly unbounded data streams and continuous queries, while other attempt to extend these prototypes with support for spatio-temporal data types and operations in order to efficiently manage moving objects. OCEANUS [1] is one of those prototypes that presents an effort to extend TelegraphCQ DSMS (TCQ) with spatial support providing a platform for spatio-temporal streaming applications.

Besides defining appropriate spatio-temporal data types and algebra that would satisfy all requirements for moving objects management, one of the basic requirements is also a possibility to visualize results of spatio-temporal continuous queries on the map. Such visualization is possible if the result of a continuous query consists of elements containing information about locations of moving objects. In that case we have a data stream that represents trajectory of moving object and requirement to redraw a moving object's position with each new arriving element, since each spatio-temporal continuous query returns incremental results refreshed with the most recently received positional information generated at data source.

Applications for visualization of such streaming trajectories have to offer appropriate method for processing incremental results of continuous queries, as well as map background for their appropriate presentation. Since there are a number of web services with open source visualization libraries offering capability to build an application with appropriate functionality, it is reasonable to select one of them as a basis for developing a prototype for visual presentation of streaming trajectories. Our OCEANUS Visualizer prototype is based on OpenStreetMap web service that provides free maps of the world and swingx-ws library for interacting Java programs with web services. Prototype also enables posing continuous queries directly using CSQL query language or indirectly by choosing query parameters.

II. RELATED WORK

Modeling, analyzing, monitoring and visualizing trajectories of moving objects has been a focus of research in recent years. This research was mostly done separately in the context of conventional spatio-temporal databases in which data is assumed to be less dynamic and in the context of continuous, time varying and possibly unbounded spatio-temporal data streams carrying information about moving object's positions at any time instant.

Visualization of a generic model for moving point objects in spatio-temporal database management system (DBMS) is presented in [2]. It is based on ESRI Tracking Analyst extension on ArcGIS for handling spatio-temporal data in a way that enables displaying point and tracking data, showing the aging of data, symbolizing actions based on attribute or spatial queries and highlighting.

Conceptual model for visual exploration of trajectory data with multivariate attributes by using timemap.js web visualization library is proposed in [3]. Developed prototype has the capability to display overview of each trajectory and the details of each trajectory with different zoom levels through map view and timeline.

An online spatial data analysis and visualization system, TerraFly GeoCloud, which facilitates end users to visualize and analyze spatial data, and to share the

analysis results is presented in [4]. It is built on the TerraFly Geo spatial database and can efficiently support many different visualization functions and spatial data analysis models. It also enables the MapQL technology to customize map visualization using SQL-like statements.

Work presented in [5] introduces design and implementation of a geospatial data stream query engine in SECONDO, together with web interface for dynamic visualization of streaming results. It uses AJAX technology to support asynchronous rendering of streaming contents on a map interface. The GUI can accept SQL queries from the web form and pass the queries to the parser. The streaming results returned are then dynamically visualized.

Method for visualization of trajectory data proposed in [6] utilize XSLT and XPath technologies to convert the original trajectory data to GML format for storage and transmission and then convert GML to KML for visualization of trajectory data in browser. Trajectory data are then visualized rapidly in Google Maps and Google Earth.

Design and implementation of a powerful map application for managing, querying and visualizing evolving locations of moving objects is presented in [7]. This application offers capabilities for specifying typical continuous queries (such as range, distance and nearest neighbor search) and visual display of objects' trajectories and online collection of movement statistics. Interface was implemented with Java SDK 1.4.2, making use of ESRI's MapObjects-Java 2.0 libraries for rendering and manipulating geographic features on map.

Our approach for developing special-purpose application for handling and visualizing streaming trajectories is closed to one presented in [7] since this application is developed on the top of TCQ on which OCEANUS is based on. But we also overcome limitations of this approach originating from PostgreSQL built-in spatial data types and operators in a way that visualization includes spatio-temporal extension in OCEANUS which relies on full standardized spatial support for data types and operations found in PostGIS.

III. OCEANUS VISUALIZER ARCHITECTURAL OVERVIEW

OCEANUS Visualizer represents mapping application that enables querying and visualizing streaming trajectories through a user-friendly graphical interface and OpenStreetMap background. It is possible to pose various spatio-temporal continuous queries using embedded set of operations in OCEANUS DSMS, each one producing incremental results. To ensure such comprehensive functionality, we have designed and implemented OCEANUS Visualizer as a desktop application consisting of three layers illustrated in Fig. 1.

At the first (bottom) layer we have an OCEANUS prototype of spatio-temporal data stream processing engine that extends TCQ with a spatial support for efficient evaluation of continuous queries over moving objects. This support is ensured with PostGIS extension, Proj4 and GEOS libraries, and MovingObject library

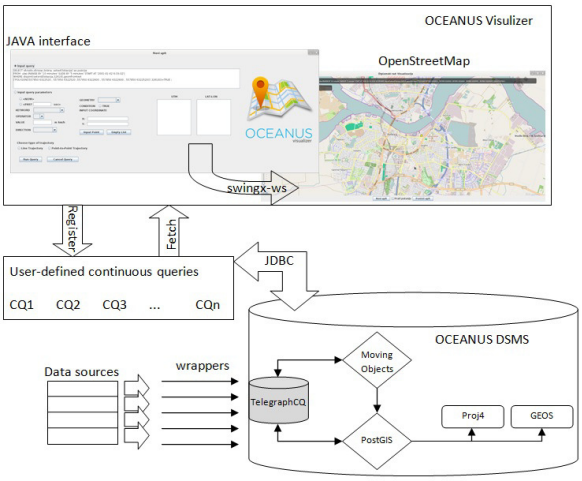


Figure 1. OCEANUS Visualizer architecture

containing data types an functions for moving objects. Special wrappers that read data from external sources representing moving objects and forward them into system with desired rate for further processing have also been designed.

At the second (intermediate) layer we have a Java interface for posing spatio-temporal continuous queries and retrieving their incremental results. Connection with data stream processing engine at the bottom layer is established through standard JDBC (Java Database Connectivity) interface for PostgreSQL 7.3, since TCQ is based on it. Continuous queries for reporting locations of moving objects can be specified using spatial operators inherited from PostGIS (such as *distance*, *disjoint* or *within*) and functions from MovingObject library (such as *passes*, *mdirection*, *trajectory*, etc.). This means, it is also possible to define conditions that combine dynamic spatio-temporal streams with static spatial entities. At running time, user can immediately suspend continuous query execution and start the new one.

At the third (top) layer we have a map offered by OpenStreetMap web service that persistently reads result of a continuous query containing locations of moving objects, and redraws it with each new arriving element in resulting stream. Communication with Java interface at the intermediate layer is accomplished with swingx-ws library for interaction with web services. Static spatial entities involved in condition of continuous query are also presented on map taking into account typical cartographic functionality such as zoom.

Since the visualization of continuous queries is done at the third layer, we will describe its implementation briefly. The map viewer is a common component needed in applications for generating and using geographic data and we use the open source JXMapViewer that enables embedding mapping abilities into Swing-based Java application. It is just a special JPanel whose API enables converting coordinates to pixels, caching tiles, and stitching them together on screen. We use JXMapKit as a wrapper for the JXMapViewer because it includes the most commonly used features as zooming and mini-view and also has many customization options. First thing need

to be done is setting a map center by calling *setAdressLocation()* method with GeoPosition object as a parameter representing coordinate at the globe as a latitude and longitude. WayPoint class is used for implementing moving objects on the map which are customized by implementing the WaypointRenderer interface. These objects are drawn by using WaypointPainter, the JXMapViewer's special painter that can properly draw points on the map, and by calling *paintWaypoint()* method. For drawing static spatial objects, trajectory of a moving object and text of a posed query we use painters (objects of a Painter class) which can customize drawing of a JXMapViewer by setting them with *setOverlayPainter()* method.

IV. TRAJECTORY VISUALIZATION IN OCEANUS VISUALIZER

Modular design of OCEANUS Visualizer application presented previously offers a comprehensive framework for managing, querying and visualizing streaming trajectories of moving objects. User-friendly graphical interface for posing various spatio-temporal continuous queries and advanced capabilities for visual display of moving object's trajectories make it suitable for wide range of users, from beginners to advanced ones. Hence, Java graphical interface is divided into two sections enabling two different methods for definition of a spatio-temporal continuous queries.

A. JAVA grafical user interface

In first section, users can type query directly in text area using CSQL query language. This section is intended for advanced users since there is no limitations in using PostGIS spatial operators and functions or functions for operations over moving objects embedded in OCEANUS prototype. Application parse posed continuous query just in a way to test weather selection list is suitable for a visualization on map (i.e. does one contains information about location of a moving object). Except this case, entire query parsing is addressed to a data stream processing engine at the bottom layer, and if query results with an exception, application discard it without any visualization and informs user with appropriate message.

In second section of JAVA interface users can pose spatio-temporal query in a way to define a condition over streaming trajectory by choosing certain operation from a offered subset of PostGIS functions and operations on moving objects. According to selection made by user, application offers him to choose just certain options in order to properly define a condition of query. List of all required options and parameters for each available operation is given in Table I and section for parameter query is illustrated in Fig. 2. For each continuous query

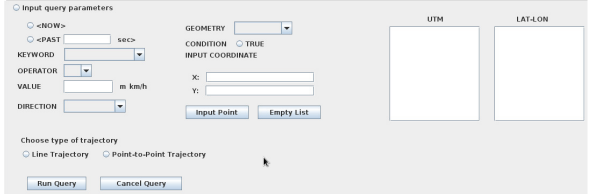


Figure 2. Section of JAVA interface for defining input query parameters

TABLE I. OPERATIONS AND REQUIRED OPTIONS AND PARAMETERS

Operation	Operator and value	Static spatial object	Direction options
within		Yes	
disjoint		Yes	
distance	Yes	Yes	
passes		Yes	
trajectory length	Yes		
{*min* | *max*} *distance*	Yes	Yes	
{*min* | *max*} *speed*	Yes		
direction			Yes

user have to determine time based sliding window and in this interface section *now* or *past* window can be chosen. For *past* window chosen, one must enter a time period in seconds. When user selects an operation that involves interaction with static spatial object, one must be defined with its coordinates. User first chooses a type of spatial object (*point*, *linestring*, *polygon*) and then defines one by one points it consists of by entering their coordinates. By not choosing an option "true", users determine to select and visualize locations of moving objects that don't fulfill the specified condition.

According to subset of operations listed in Table I, spatio-temporal queries that can be defined by choosing query parameters could be classified into three classes:

- positional queries – continuous queries that examine spatial interaction of moving objects with specific static spatial object. This class includes operations *within*, *disjoint* and *passes*.
- distance queries – continuous queries that founds current locations of moving objects within or outside a given distance from a reference object (*distance*), or current location of moving objects in particular relation with distance from a reference object in the past (*minimal distance* and *maximal distance*).
- scalar queries – continuous queries that examine values of particular features of moving objects such as their trajectory length during some period, minimal or maximal speed reached or direction of their movements (south, north-east, west, etc.).

OpenStreetMap uses the WGS84 spatial reference system used by the Global Positioning System (GPS) (geographic coordinates between -180° and 180°

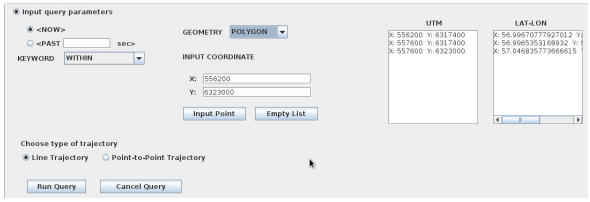


Figure 3. Defining condition of query Q1

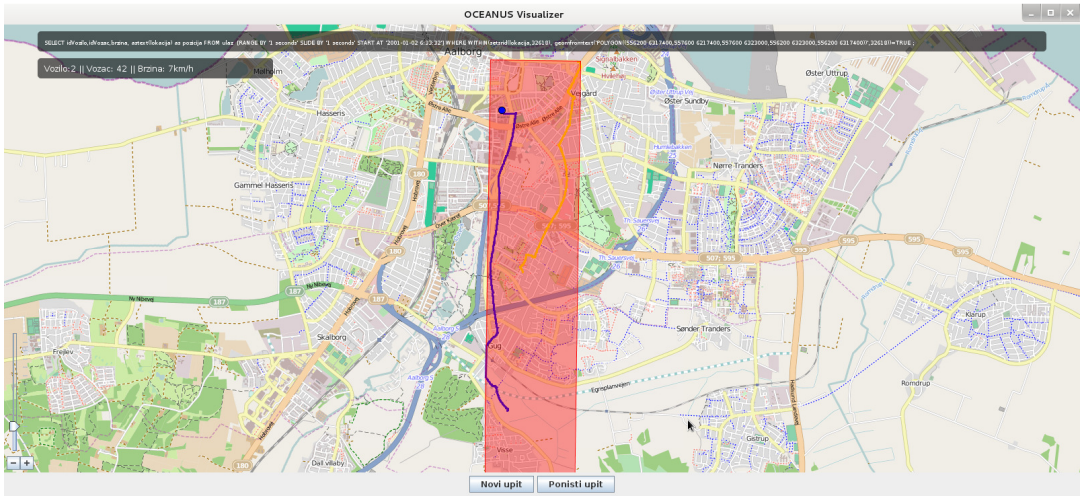


Figure 4. Visualization of query Q1

longitude and -90° and 90° latitude) and application also offers possibility to convert coordinates from Universal Transverse Mercator (UTM) projected coordinate system to WGS84. Hence, users can define spatial objects involved in query execution by using any of these two spatial reference systems.

B. Query examples and visualization

Following the rules for defining query condition explained in previous subsection, users can pose various spatio-temporal queries in second section of Java interface. Query examples in this subsection are based on the following schema[1] of a data stream receiving data from moving objects:

```
carStream <
   id:        integer,
   carType:   varchar(8),
   driver:    smallint,
   pos:       intime(point),
   speed:     real,
   tcqtime:   timestamp timestampcolumn
>
```

For example, if user wants to pose the following query:

Q1: *Continuously report current locations (or trajectories) of all vehicles within particular region of interests.*

he will need to: choose a window *now*, choose an operation *within*, choose polygon type of geometry and define it with its coordinates, as it is illustrated in Fig. 3. This will result with the following CSQL query generated by the application:

```
SELECT id, val(pos), speed
FROM carStream <Now>
WHERE within(val(pos),geomfromtext
('POLYGON((557850 6322520, 557950 6322520,
557950 6322600, 557850 6322600, 557850
6322520))'));
```

[1] Schema is derived from INFATI data set [8], used for generating *archived* streams

After defining a query, Java interface forwards it to the OCEANUS prototype at the bottom layer for a registration as a continuous query and displays an OpenStreet map for a visualization. Map continuously listens for the results of the query and when first element of data stream that satisfies query condition arrives it starts displaying results. There are two ways for visualizing results, either as a continuously "moving" marker on the map or as a continuously "growing" trajectory of the moving object with marker representing last location at the top. Second way of visualization of a query Q1 is illustrated in Fig. 4. Notice that static spatial object defined for evaluating a condition of continuous query is also displayed on map.

In query Q1 we have used PostGIS function *within* in combination with window *now*. In next example, we will illustrate usage of operation on moving objects in combination with window *past* since it considers recent history of movement. For query such as:

Q2: *Continuously report current locations (or trajectories) of all vehicles that were heading north-east 40 seconds ago.*

user just needs to select an operation *direction* and appropriate value for it (north-east), as it is illustrated in Fig. 5. Notice that for this type of operation user can not define any other parameter values except the one for direction. In this way, users don't have to be familiar with operation signatures (interface) because application takes care about appropriate usage of each operation. Choice made by user given in Fig. 5 will result with the following CSQL query generated by the application:



Figure 5. Defining condition of query Q2

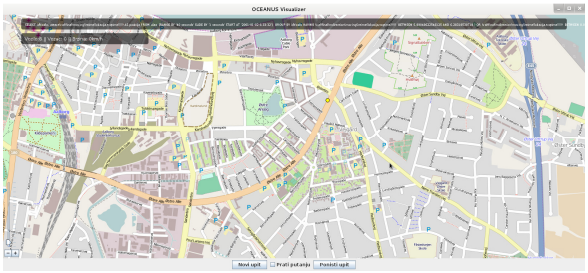


Figure 6. Visualization of query Q2

```
SELECT id, val(final(moving(pos)))
FROM carStream <Past 40 seconds>
GROUP BY id
HAVING
val(initial(mdirection(moving(pos))))
BETWEEN (3.14159265)/8 AND
(3.14159265*3)/8;
```

Visualization of query Q2 is illustrated in Fig. 6 presenting moving objects as continuously "moving" markers on the map.

V. CONCLUSIONS AND FUTURE WORK

In this paper we have presented OCEANUS Visualizer, an application for querying and visualizing streaming trajectories of moving objects. It is developed on the top of OCEANUS prototype engine for processing spatio-temporal data streams and is based on Java programming language and OpenStreetMap web service. Communication between application modules is accomplished through standard JDBC interface and singx-ws library for interaction with web services. Possibility to define continuous query directly using CSQL language enables posing wide range of spatio-temporal queries that include PostGIS operators and functions and operations on moving objects as well. Through query examples presented in the last section, we have illustrated advantages of defining continuous query condition by a simple mouse clicks and definitions of parameter values. With this approach, application is found appropriate for beginners and advanced users as well.

Focus of our future work will be improvement of existing and adding new functionalities. We will examine a possibility of adding new operations for defining conditions of continuous query that would enable users not familiar with CSQL to take advantage of operations on moving objects such as *atinstant*, *present*, *at*, *intersect*, *inside*, etc. Since specifying static spatial objects for query condition is not an easy task (user must know the exact coordinates of an object), our future work will also focus on specifying these objects interactively with the map. An open issue is also to find a possibility to express results of continuous queries that don't represent streaming trajectories, i.e. data streams whose elements except location of moving objects contain other information as well.

REFERENCES

[1] Z. Galić, E. Mešković, K. Križanović, M. Baranović, "OCEANUS: A Spatio-Temporal Data Stream System Prototype," in Proc. IWGS 2012 3rd ACM SIGSPATIAL Int. Workshop on GeoStreaming, Redondo Beach, CA, USA, 2012, pp. 109-115

[2] M. Baars, P. van Oosterom, E. Verbree, B. Gorte, "Structuring, Indexing, Querying and Visualizing Moving Objects in a DBMS," AGILE, Heraklion, 2004, pp. 251-260

[3] H.N. Long, "Web visualization of trajectory data using web open source visualization libraries", Msc thesis, March, 2010

[4] Y. Lu, M. Zhang, T. Li, Y. Guang, N. Rishe, "Online Spatial Data Analysis and Visualization System," in Proc. IDEA '13 Proceedings of the ACM SIGKDD Workshop on Interactive Data Exploration and Analytics, New York, USA, 2013, pp. 71-78

[5] C. Zhang, Y. Huang, T. Griffin, "Querying Geospatial Data Streams in SECONDO," in Proc. 17th ACM SIGSPATIAL Int. Conf. on Advances in Geographic Information Systems, New York, USA, 2009, pp. 544-545

[6] J. Li, J. Wang, L. Yu, R. Qi, J. Zhang, "Visualization Method of Trajectory Data Based on GML, KML," Advances in Computer Science and Information Engineering, Vol. 2, Springer Berlin Heidelberg, 2012, pp 479-484

[7] K. Patroumpas, E. Kefallinou, T. Sellis, "Monitoring Continuous Queries over Streaming Locations," in Proc. 16th ACM SIGSPATIAL Int. Conf. on Advances in geographic information systems, New York, USA, 2008, Article No. 81

[8] C. S. Jensen, H. Lahrmann, S. Pakalnis, J. Runge "The INFATI Data", *TimeCenter Technical Report*, 2006.

SPAM trends and techniques used on Exchange Online system used for racunarstvo.hr domain

Igor Pavleković[1] and Vjekoslav Gašparević [1]
[1]University College Algebra, Ilica 242, Zagreb, Croatia
[1] igor.pavlekovic@racunarstvo.hr

Abstract - This paper will give an overview of the technologies and techniques employed in SPAM protection on University College Algebra (UCA). UCA uses Exchange Online system included in Office 365 service that is free of charge for every educational institution covering majority of countries around the world. In today's world of electronic communication, e-mail is still primary system for exchanging messages and data. Globally, average mail user daily sends and receive around 115 e-mail messages and from that number around 19% or 13 of received e-mail messages are considered as SPAM. Exchange Online system is based on the Microsoft Exchange 2013 server and it is located in the multi-tenancy datacentres around the world, serving corporate, governmental and educational users. That system has built in anti-SPAM protection and reporting services that will be used for conducting results in this paper. Those results will be than compared with other publicly available researches to achieve comparison reports that is presented in this paper.

Keywords **– SPAM; Exchange Online; Office 365**

I. INTRODUCTION

SPAM [1] is defined as irrelevant or unsolicited messages sent over the Internet with the purpose of advertising, phishing, spreading malware etc.

Majority of world's SPAM messages are sent by using infected computers with worms or viruses, which are centrally controlled. Those kind of networks are called bot-nets or zombie-networks. History shows that when those kind of networks are turned off, volume of SPAM messages drops dramatically by 60% or more [2].

One of the primary task of messaging administration includes fighting SPAM messages. This job is not an easy one, because e-mail systems need to resolve messages that are SPAM and treat them by company rules and on other hand to resolve messages that aren't SPAM not blocking their normal delivery. Today's anti-spam systems use different techniques to achieve this goal. Some of them includes investigating originating e-mail's Internet Protocol (IP) address, sender's server and body and the content of the e-mail message itself.

A. Investigation of sender's IP adress

When e-mail message enters e-mail server, the server will do few checks on the sender server IP address. The first and most basic check is checking on IP allow and IP block list. Those lists are generated by messaging administrator by manually entering IP addresses of allowed of denied sender server IP addresses.

Next checks include checking on Real-time Black Lists (RBL) and Real-time White Lists (RWL). RBLs and RWLs are provided by third party providers to what anti-SPAM systems are subscribed to. They check, in real time whether an IP address is on that real-time list or not. One of the well-known and used provider of RBLs and RWLs is Spamhaus Project [3].

One more common check is Sender ID defined in RFC 4406 [4]. This check presumes that sender's domain has valid Sender Policy Framework (SPF) record in Domain Naming System (DNS) system stating IP addresses of the servers allowed to send e-mail for that domain. If the sender's server IP address is not stated on the list, recipient's server can discard that e-mail message or treat it as a SPAM message. This technique tries to stop phishing e-mail messages.

B. Analisys of the senders' servers characteristics

This usually includes more different tests that are done on the senders' servers. When the receiving server receives a message from the senders' server, it will do several checks to assess it.

Most common check is called open-relay test. It means that receiving server will do open-relay test to see if the senders' server will deliver open relay message or not [5].

Other checks include checking if the sending IP address has its corresponding DNS pointer (PTR) record and if that record matches the hostname stated in the HELO/EHLO part of the Simple Mail Transfer Protocol (SMTP) communication.

Based on servers' characteristics, e-mail messages can receive higher Spam Confidence Level (SCL) value, messages can be rejected, or even IP addresses of the servers can be put on the temporary block list – so called gray listing.

C. Analisys of the content

Last tests that are done on the e-mail message inspect content of it. This includes searching of the keywords, different Hyper Text Markup Language (HTML) elements

in the message, language of the messages, attachment names and extensions and similar tests.

Usually, there are definitions used by the anti-spam software provided by 3rd party vendors.

Depending on the content, message receives SCL value and based on the anti-spam policy, messages can be altered, rerouted, quarantined, rejected or even deleted.

II. ANTI-SPAM IN EXCHANGE ONLINE

Microsoft Exchange Online (EXO) uses Exchange Online Protection (EOP) service as its anti-malware and anti-spam solution. Even dough EOP is available as a standalone service, UCA uses it as a part of its MS Office 365 subscription with EXO services that are provided to all students, faculty and staff on racunarstvo.hr domain.

EOP uses all three anti-spam techniques described in Introduction, as well as anti-malware inspection of every e-mail message that enters or exits EXO.

As it is visible in Fig. 1. every e-mail message that comes to the EOP system in Microsoft datacenter is firstly filtered by connection policy. If the message fails to pass policy rules it will be rejected or deleted. After that

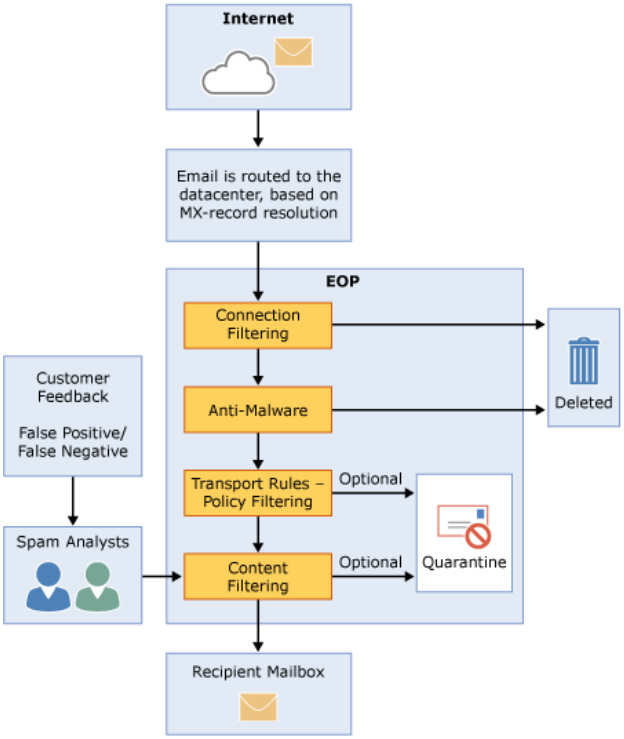


Figure 1. Diagram of inbound message flow through EOP system [6]

message passes through anti-malware filtering and policy. Also if the message fails to pass that settings, it will be deleted automatically. Third filter checks for policies that define content of the message in terms of subject text, body text, attachments or other relevant part of the message. This kind of filtering is called transport-rules filtering. Last filter that checks content of the message in terms of overall content, HTML elements and anti-spam definitions provided by EOP itself.

Interesting thing is how Microsoft builds anti-spam definitions. Majority of the other anti-spam vendors need to collect respectable amount of SPAM messages so they will be able to comply anti-spam definitions. They do that by opening so called honey-pots, ie. E-mail addresses that are subscribed to vast number of newsletters, porn sites and similar web sites and by that way they collect large number or SPAM messages. On the contrary, Microsoft uses their free e-mail service Outlook.com that enables them to rely on their customers classification of messages as if they are junk or not. Based on that, Microsoft will comply that information with the anit-spam definitions that are used in EOP and even in Microsoft Office Outlook product used as Personal Information Management (PIM) application.

III. EOP IMPLEMENTATION IN UNIVERSITY COLLEGE ALGEBRA

To avoid deletion of any e-mail message that could be detected as false-positive and eventually not delivering that message to the recipient's mailbox, UCA doesn't use any custom policy at all. This means that EOP configuration is default, without any custom policies. E-mail message will pass the system and the only place that the message can be deleted or rejected is by means of Connection Filtering component. Otherwise, e-mail message will go through the filtering rules and based on the SCL values set by Content Filter, message will be passed to the recipients Inbox folder where the user itself can choose to create rule to transfer message to the Junk folder or let MS Office Outlook handle message automatically.

By using this approach, every e-mail user is responsible for maintaining hygiene of their mailbox by using default setup or defining their own custom rules for managing messages with higher SCL values. As a best practice, messages with SCL values between 5 and 7 are most likely to be spam and those with values between 8 and 9 are definitely spam messages.

EOP services enables its customer receiving statistical data and reports for the period of last 60 days with detailed reporting for the last 7 days.

For this paper, statistical data was collected in two periods:

- first reporting period (FRP) – from Dec-12-13 to Feb-10-14
- second reporting period (SRP) – from Feb-11-14 to Apr-24-14

Data from FRP is as follows:

- total number of e-mail messages sent and received is 27,414 messages or in average 456.9 messages per day
- total number of received spam messages is 5,268 or in average 87.8 messages per day
- percentage of spam messages in total volume of received messages is 24.30%

Data from SRP is as follows:

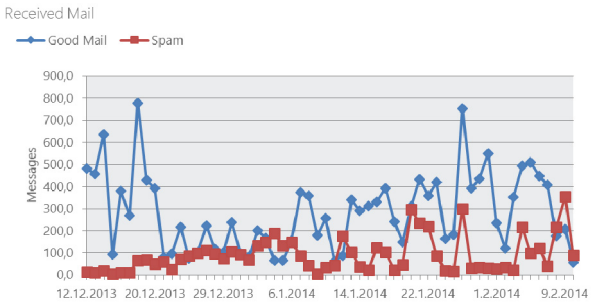


Figure 2. Report on daily distribution of received good and spam messages for FRP

- total number of e-mail messages sent and received is 53,517 or in average 891.95 messages per day
- total number of received spam messages is 22,998 or in average 383.3 messages per day
- percentage of spam messages in total volume of received messages is 42,97%

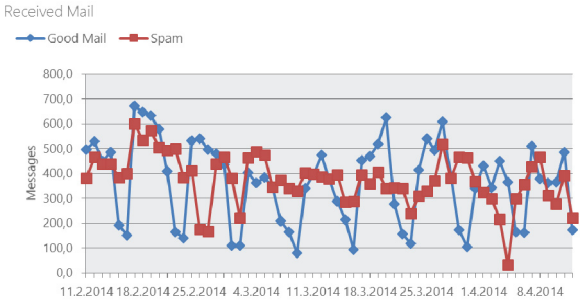


Figure 3. Report on daily distribution of received good and spam messages for SRP

When taken into consideration that "many Internet networks report the figure of spam messages greater than 90% and some as high as 96%" [7] it is obvious that UCA has much lower volume of SPAM messages than average computer network.

A. *First reporting period analisys*

In Figure 2, for FRP it is visible that number of SPAM messages vary from day to day. It can be concluded that only in two days, daily volume of SPAM messages was higher than non-SPAM messages. It was on January 4 and January 5 when it was weekend in Croatia. In addition, two dates are spiked with number of total number of received mail. First of them is December 18, 2013, but number of SPAM messages is not spiked, so it means that users received larger amount of messages, possibly as messages for the upcoming holydays. Second spike occurred on January 28, 2014 when there was intense communication between teachers and supporting staff for preparing upcoming semester in UCA.

In further analysis of the received SPAM messages for the same reporting period, it is visible that:

- 91.66 % of all spam messages is filtered based on IP receive blocking mechanism like RBL
- 0.01% of all spam messages are filtered based on SMTP communication analysis
- 8.31% of all spam messages are marked as spam messages by assigning them SCL value of 8 or 9

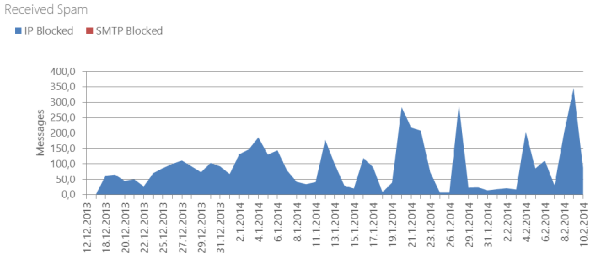


Figure 4. Report on dailiy distribution of spam messages filtered by Connection Filtering system for FRP

B. *Second reporting period analisys*

Figure 3. for SRP shows different e-mail distribution that it was in FRP. In this period is can be differentiated when there is weekend as volume of good e-mail messages drops. There are also many more periods in which number of SPAM messages is higher than good e-mail messages. Also, it is visible, that number of SPAM messages is relatively steady during whole reporting period, with only one drastic drop on Apr-4-14.

Detailed analysis shows following:

- 96,4% of all spam messages is filtered based on IP receive blocking mechanism like RBL
- 3,60% of all spam messages are marked as spam messages by assigning them SCL value of 8 or 9

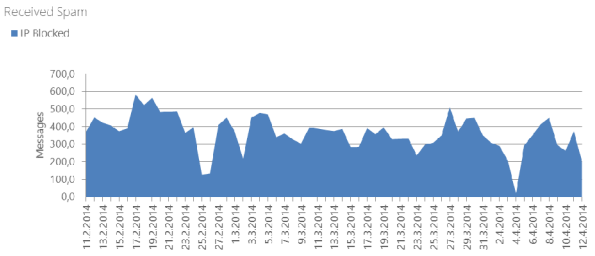


Figure 5. Report on dailiy distribution of spam messages filtered by Connection Filtering system for SRP

IV. CONCLUSION

By comparing global statistics and data about spam messages in global e-mail systems, volume of spam messages in UCA e-mail system is at least 65% lower than in global systems. In addition, after crunching

numbers from available reports, it is visible that only every fifth message that enters system is spam message in FRP, but in SRP every 2nd message is marked as spam. In average 94.05% of the rest spam messages will be filtered and deleted or rejected, and only 5.95% of spam messages will be delivered to the users' mailboxes. This means that in received 1000 e-mail messages to the system, only 59 of them will end up in users' mailboxes.

When taking into consideration that UCA has around 450 mailboxes and average daily number of received messages is 674 for the both reported period, this means that mailboxes will daily receive 14 spam messages, or this means that around 3% of all recipients on domain racunarstvo.hr will have one spam message delivered to their mailbox daily.

In addition, it is imperative to stress out that those messages delivered to the users' mailbox will have SCL value 8 or 9 and the users' PIM application, like Microsoft Outlook will treat that message as a spam and will move them to the Junk folder automatically. Of course, user itself can change this behavior by turning off that functionality so users will control all inbound messages that are not filtered by EOP, by users themselves.

By analyzing report on Figure 3. it is visible that number of SPAM messages vary from day to day. This depends on the global SPAM activity and supporting bot-nets that are active on the internet and that are used for sending SPAM messages. Based on that information, one can conclude, that during FRP period Jan-29-2014 until Feb-3-2014 some global bot-network was taken offline and that is the reason for much lower SPAM activity during that period.

However, it is importing to stress that during two reporting periods mentioned here, average volume of e-mail messages, either legitimate or spam varies. It is visible that during SRP daily e-mail volume almost doubled than in FRP. It can be concluded that during FRP there was less e-mail activity as UCA had winter holidays and there were no classes. Never the less, it is visible that number of spam messages quadrupled in SRP meaning that global spam activity increased, but EOP itself has blocked 5% more messages based on IP receiving mechanism, and number of messages that got to the users' Inbox was decreased by 5%. This means that IP blocking mechanism got better. Microsoft didn't announce any change in the configuration of EOP that could justify that change.

REFERENCES

[1] Oxford Universirty Press, http://www.oxforddictionaries.com/definition/english/spam?q=SPAM

[2] KrebsonSecurity website, http://krebsonsecurity.com/2011/03/rustock-botnet-flatlined-spam-volumes-plummet/

[3] The Spamhaus Project, website, http://www.spamhaus.org

[4] Sender ID, Request for Comments 4406, http://www.ietf.org/rfc/rfc4406.txt

[5] Definition of open mail relay on Wikipedia, http://en.wikipedia.org/wiki/Open_relay

[6] Exchange Online Protection Overview, http://technet.microsoft.com/en-us/library/jj723119(v=exchg.150).aspx

[7] Spamhouse project report, http://www.spamhaus.org/whitepapers/effective_filtering/

An Approach using Simulation Techniques to estimate Quality of Service Parameters in Communication Networks

Hrvoje Kozačinski*, Petar Knežević *
* Faculty of Electrical Engineering and Computing/Department of Fundamentals of Electrical Engineering and Measurements, Zagreb, Croatia
hrvoje.kozacinski@fer.hr

Quality of Service in communication networks is a request for minimal required performance needed for transferring IP packets in a network. The performance can be ensured with a set of mechanisms that guarantee required performance through shaping and policing of packet traffic. Their goal is to prevent congestion, packet delays and packet loss. The goal of this paper is to introduce some of the commonly used Quality of Service mechanisms and analyze their impact on network traffic in order to improve its performance, mainly packet delay and buffer congestion. The mechanisms are analyzed trough a simple packet transfer simulator that offers different combinations of mechanisms to be used in the simulation, allowing different packet transfer scenario setups. After a simulation is complete, the simulator offers results along with relevant packet transfer graphs. The impact of Quality of Service mechanisms is analyzed and interpreted trough those results. It is shown that combining Quality of Service mechanisms in packet transfer generally gives better results. Also, different combinations of mechanisms give different results, allowing more flexibility in improving packet transfer in a desired way, but also making it more predictable and controllable.

I. INTRODUCTION

Basic elements of a computer network include hardware, software and protocols. The interrelationship of these elements constitutes the infrastructure of the network. A network infrastructure is the topology in which the nodes of a network are connected to each other. Connections involve routers, switches, bridges and hubs connected by cables or wireless technologies.

Because of the way the communication networks are built, with many possible transfer routes and with possibility that some routes can be congested or slower than others, IP packets travelling through the network can arrive too late, with errors, in wrong order or be discarded during transfer. This causes transfer errors and degradation of transfer performance.

Quality of Service (QoS) is a set of requirements that guarantee certain performance in packet transfer. A good implementation of QoS usually minimizes packet delay and reduces the number of discarded packets. To ensure such performance, QoS has a set of mechanisms that help regulate the transfer, so that the desired performance requirements are met [1].

There are different approaches to solving inadequate transfer performance, mainly by introducing different QoS mechanisms to communication networks [2, 3, 4]. Also, different approaches and recommendations are given through a number of RFCs, regarding guaranteed QoS [5], defining a information model for QoS network managing policies [6], QoS Mechanism Selection [7] and others [8, 9]. All of these approaches help reduce packet delay and packet loss. The problem this paper analyzes through a simple network simulator is how exactly QoS mechanisms impact a communication network performance in attempt to improve packet transfer, primarily by reducing packet delays and packet loss associated with the buffer mechanism.

II. QUALITY OF SERVICE MECHANISMS

Some of the main QoS mechanisms that are used in communication networks will be explained in this paragraph, along with their desired effects and shortcomings.

Retransmission: resends packets lost in transmission, thus preventing loss of information. It helps with acquiring complete information, but with a cost - increased transfer time and packet delay.

Buffer: container used to temporarily store packets that cannot be processed immediately upon arrival, preventing retransmission to cause bigger time delay and unneeded bandwidth usage. However, increasing the buffer size also extends packet delay, unless other mechanisms are applied [10].

Buffer collecting strategies: various ways of collecting packets from the buffer, involving various criteria like order of arrival, packet priority, etc. Different approaches try to improve on different aspects of QoS. For instance, round robin buffer collecting strategy is used to even buffer delay times of packets in the buffer.

Packet priority: classifies packets by their importance - important packets get higher priority and will be processed first. This helps reducing packet loss and delays of packets with higher priority, but by increasing packet delay of lower priority packets [11].

Policing mechanisms: mechanisms used to discard packets from the buffer to prevent its congestion, allowing newer packets to enter the buffer in favour of discarded packets. The most common policing mechanisms are RED (Random Early Detection; also Discard or Drop)[12] and WRED (Weighted RED, uses packet priority in deciding which packets to discard). If the priority mechanism is used, WRED can discard less significant packets to prevent congestion and allow future higher priority packets to be received [3, 13].

Shaping mechanisms: mechanisms used to shape the sending flow of packets to prevent buffer congestion, packet drop and retransmission, avoiding unnecessary bandwidth usage, possible congestion and packet delays. The most common are the Token Bucket and Leaky Bucket algorithms [14, 15].

III. THE SIMULATOR OF QUALITY OF SERVICE MECHANISMS

In order to better study the impact of different QoS mechanisms on a communication network, a network simulator was developed. Its function is to simulate a simple TCP packet transfer between two network nodes, in a dumbbell topology, with different QoS mechanisms involved and track buffer capacity, packet loss and packet time delay. The mechanisms implemented in the simulator are: retransmission, buffer, buffer collecting strategies (first packet in the buffer or round robin), packet priority with 3 priorities (0-2, 0 being the highest) and the Token Bucket algorithm.

The simulator in its essence is a multithreaded Java application, in which the sender thread generates packets and sends them to the buffer, while the receiver thread collects packets from the buffer. There is also a third thread that collects relevant transfer data. Packet generation is a simple, best effort solution, packet length is fixed and packet priority is generated randomly.

The simulator is controlled through the main screen. The user can set up the desired simulation parameters (buffer size, number of packets generated, sender and receiver speed - slowdown in milliseconds) and choose the desired QoS mechanisms, with the exception of retransmission, as it cannot be disabled.

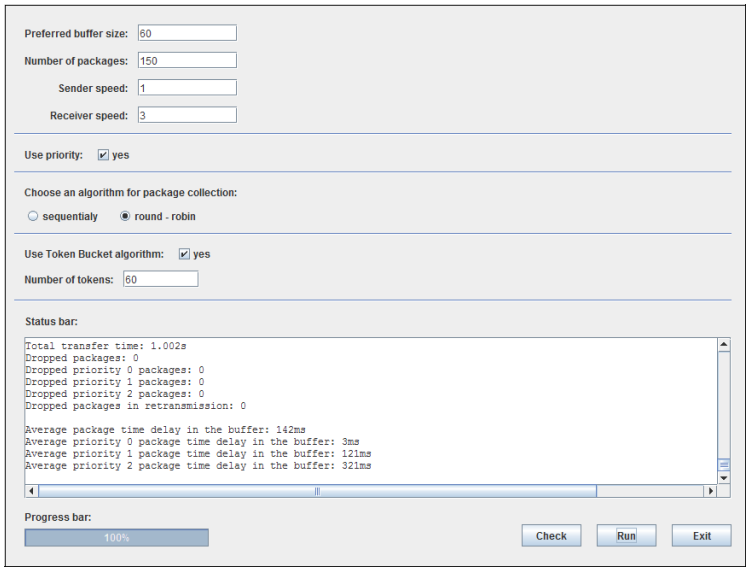


Figure 1. Main simulator screen after a completed simulation

TABLE I. SIMULATION SCENARIOS AND MECHANISMS USED

Scenarios	Priority	Buffer collecting	Token bucket
Scenario 1	No	First available	No
Scenario 2	Yes	First available	Yes
Scenario 3	No	Round robin	No
Scenario 4	Yes	Round robin	Yes

After the simulation is completed, three graphs are generated: two packet transfer delay graphs (with and without marked priorities) that show the time delay of each packet in the buffer and a buffer capacity graph that shows how the buffer capacity changed during the transfer. The simulator also generates log files and displays relevant transfer data: total transfer time, number of discarded packets (total and divided by priorities), number of packets discarded in retransmission and average packet time delay inside the buffer (total and divided by priorities). An example of a finished simulation is shown in Figure 1.

There have been a lot of different studies of QoS mechanisms where a simulation approach was used, but they were mostly used in specific areas like Composite Web Services [16], VoIP [17], Video Traffic [18], Mobile Ad Hoc Networks [19], UMTS Terrestrial Radio Access Networks [20] or evaluating complex systems [21]. Also, there are discrete event simulators targeted at networking research. One example is the ns-2 simulator [22]. The simulator of this paper is constructed to implement only the needed components and its deliberate simplification allows better emphasis and clearer study on how the buffer handles incoming packets with different QoS mechanisms involved and how do those mechanisms impact packet delay and congestion in the buffer.

There is also a matter of packet generation. While using Poisson distribution for packet generation [23] provides a more realistic simulation, it was not necessary for the analysis of observed QoS mechanisms and their impact on buffer capacity and packet time delay.

The impact on packet transfer will be observed and measured in four transfer scenarios. In every scenario, the basic transfer parameters will remain constant:

- Buffer size: 60
- Packets to send: 140
- Sender speed: 1
- Receiver speed: 3
- Tokens: 60 (maximum number of tokens for the Token Bucket algorithm)

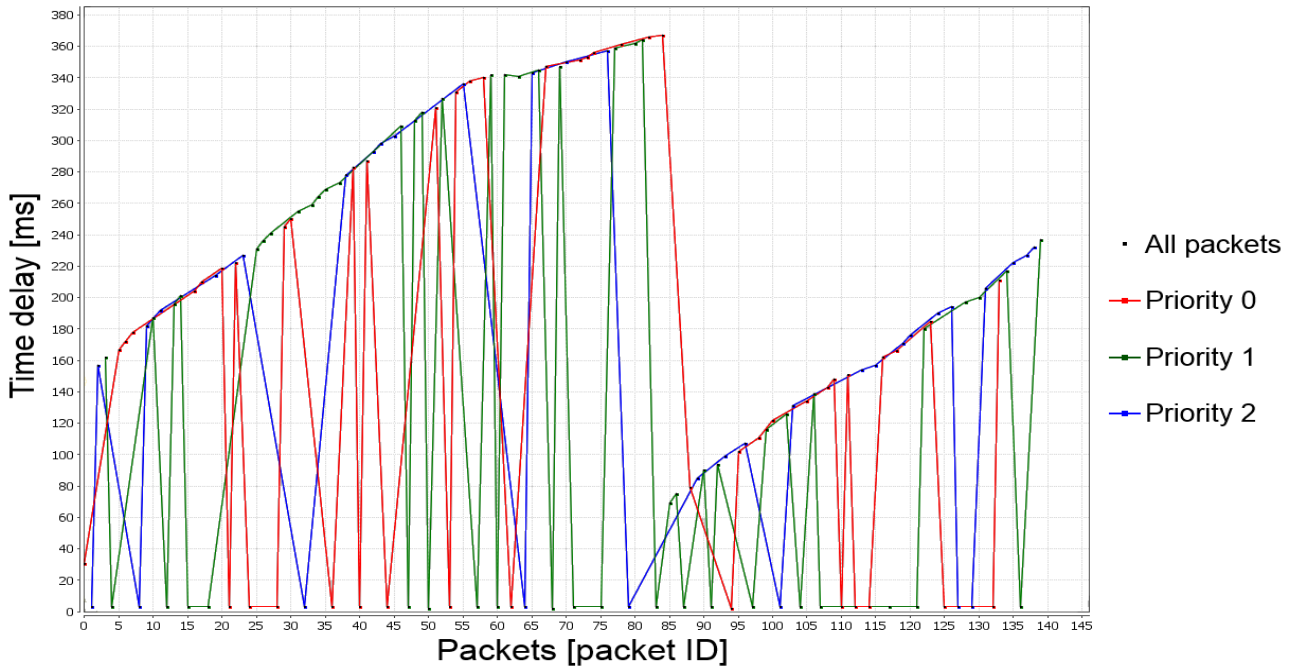


Figure 2. Packet delay during transfer graph, scenario 1 (no mechanisms, basic collecting)

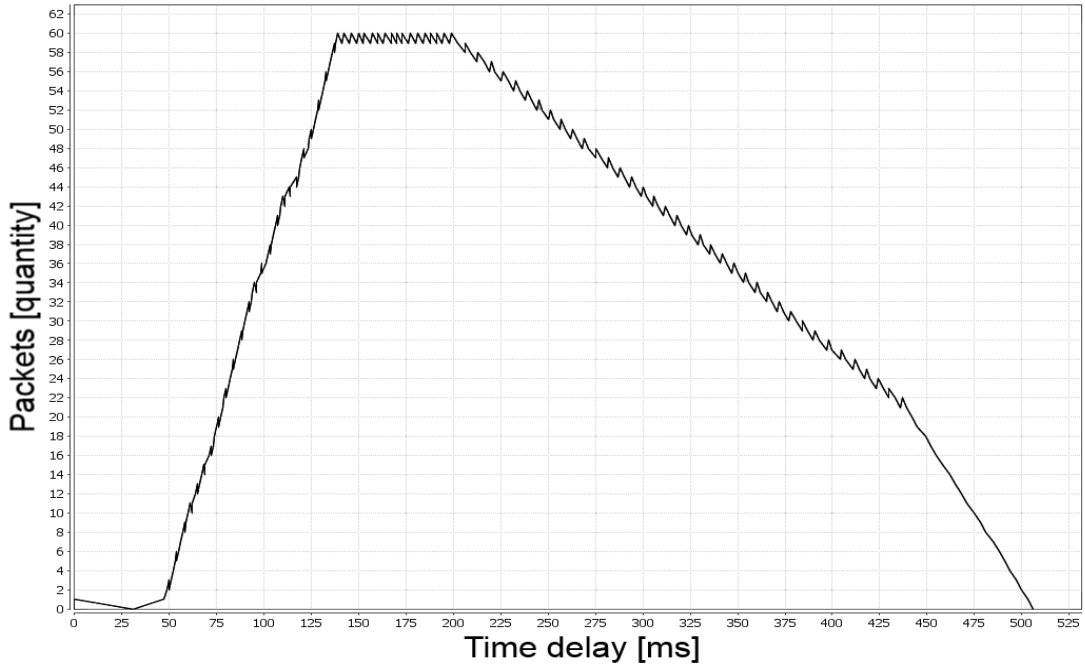


Figure 3. Buffer capacity during transfer graph, scenario 1 (no mechanisms, basic collecting)

Each scenario was run 25 times and average values are presented. This was necessary to reduce measuring errors, because the simulation was run on a personal computer which contained other running processes.

IV. SIMULATION SCENARIOS

The simulation is divided in four different scenarios, depending on the combination of QoS mechanisms used, as seen in Table 1. All scenarios use retransmission and the buffer mechanism with one of the buffer collecting strategies. Optional mechanisms are packet priority, buffer collecting strategy and the token bucket algorithm. Each scenario will be explained in more detail, along with its results, through one typical simulation of each scenario.

A. Scenario 1 - no mechanisms, basic collecting

The first scenario uses only the unavoidable mechanisms (retransmission and buffer) along with the buffer collecting strategy that collects the first available packet in the buffer.

Delays of packets can be seen on Figure 2. The graph marks different packet priorities with different colours, but because the priority mechanism isn't activated, packet priority has no effect on delay times. There are three

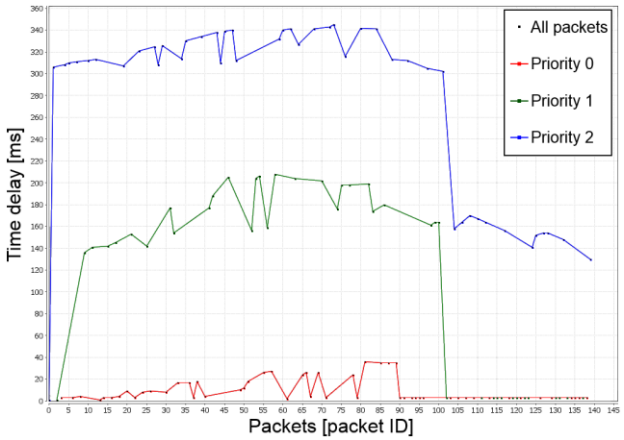


Figure 4. Packet delay during transfer graph, scenario 2 - all mechanisms, basic collecting

distinct lines on the graph, one horizontal line and two lines with linear growth. The horizontal line is caused by the buffer collecting strategy, which collects the first packet in the buffer it comes across, resulting in low time delays. The first linear growth is caused by the delay packets have while waiting in the buffer. The second linear growth is the result of retransmission that started after approximately 85 packets were sent and the buffer was congested. The retransmission time delay appears smaller, but it is in fact larger. This graph shows only time delays of packets while waiting in the buffer. Packets that went into retransmission were sent, rejected, retransmitted again and then waited in the buffer, so their total time delay is larger than that of the packets that immediately went into the buffer.

Another relevant graph is the buffer capacity graph, shown in Figure 3. The buffer capacity was exponentially filled to the point of congestion, at approximately 135 ms. After that, surplus packets were discarded and sent into retransmission. At that point, the receiver collected packets, but the sender filled the empty buffer spot quickly, as is shown on the graph with the horizontal line spanning to 200 ms in transfer. After that the drop in sending rate because of the slow retransmission rate that helped lower the buffer capacity, until all packets were collected from the buffer.

The total transfer time was 507 ms, total of 38 packets were dropped in transfer (12 priority 0, 12 priority 1 and 14 priority 2). Average packet time delay is 161 ms (164 ms priority 0, 154 ms priority 1 and 168 ms priority 2). The results clearly show that without a priority mechanism all packets are treated equally. This is the basic scenario on which all others will try and improve.

B. Scenario 2 - all mechanisms, basic collecting

This scenario adds the priority mechanism to scenario 1, to observe and measure its effects on packet transfer, seen on Figure 4. Packets are clearly distinguished by their priority: priority 0 has the lowest time delay, followed by priority 1 and priority 2 with the biggest time delay.

This scenario also introduces the Token Bucket algorithm. The idea behind the algorithm is to control the rate at which packets are sent. The algorithm periodically fills a container (bucket) with tokens, up to a set capacity. A packet can only be sent if there are sufficient tokens in the bucket to "pay" for sending. At first, with enough tokens at the beginning of transfer, packets can be sent in a large burst, until the container has no more tokens. After that, the periodical appearance of tokens controls the send rate. This helps buffer congestion, reduces the need for retransmission and because of that also reduces packet drop rates and delay times [14].

The total transfer time was 446 ms and no packets were dropped in transfer. Average packet time delay is 127 ms (9 ms priority 0, 115 ms priority 1 and 272 ms priority 2). In this case, combining both mechanisms really paid off, especially when observing the priority 0 delay times. The time delay drop at the end of the graph is the result of the Token Bucket algorithm that reduced the send rate once it ran out of tokens. The shaping mechanism helped with packet drops, but this is an idealized example, where the algorithms tokens are refilled at the same rate the receiver can receive a packet. In a real life situation, the algorithm would require a careful parameter calibration to get as close to the idealized example as possible. Higher sending rate would cause packet drops and lower rate would cause bigger packet time delay.

The buffer capacity graph (Figure 5) looks similar to the graph in scenario 1, only the horizontal line where the sender and the receiver compete while the buffer is at full

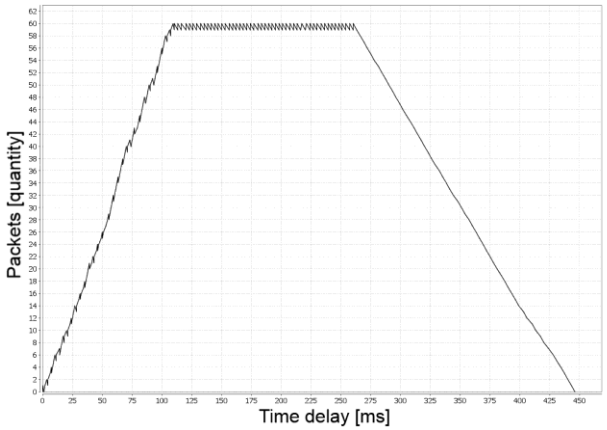


Figure 5. Buffer capacity during transfer graph, scenario 2 - all mechanisms, basic collecting

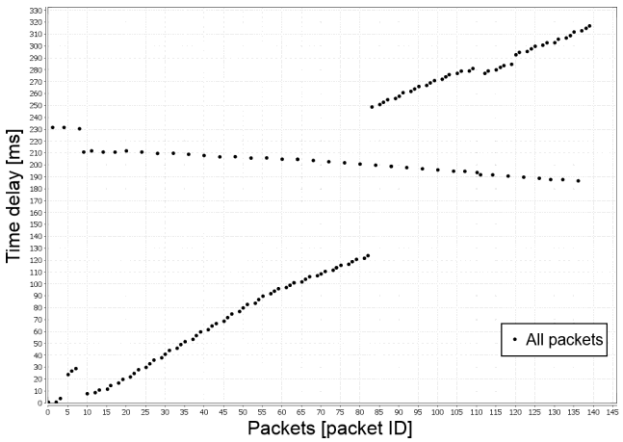


Figure 6. Packet delay during transfer graph, scenario 3 - no mechanisms, round robin

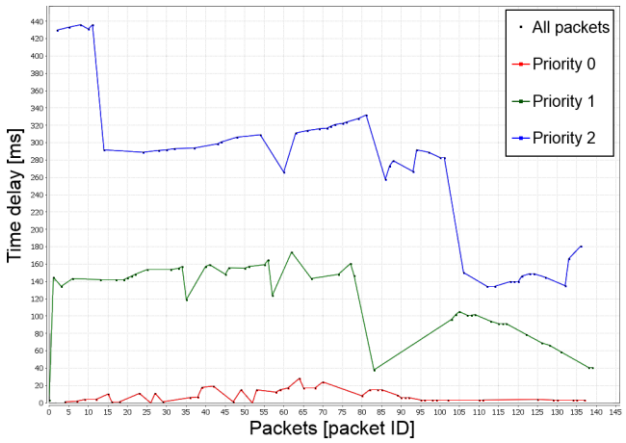


Figure 7. Packet delay during transfer graph, scenario 4 - all mechanisms, round robin

capacity lasts longer and ends with a smooth linear drop.

C. *Scenario 3 - no mechanisms, round robin*

In this and the following scenario, the buffer collecting strategy is changed. Packets will be collected using the round robin strategy. This scenario has no additional mechanisms. The effect this collecting strategy has on packet transfer can be seen in Figure 6.

The idea of the round robin collecting strategy is to improve the basic collecting strategy. Because of the way packets are collected, this strategy evens time delays across all packets that enter the buffer. In the graph this is shown as a horizontal line in the middle. Two linear growths appear because the sender is filling the buffer faster than the receiver can collect packets and because of inevitable retransmission once the buffer is congested.

The total transfer time was 486 ms, total of 41 packets were dropped in transfer (19 priority 0, 12 priority 1 and 10 priority 2). Average packet time delay is 166 ms (182 ms priority 0, 157 ms priority 1 and 158 ms priority 2).

The buffer capacity graph is similar to the same graph in scenario 1.

D. *Scenario 4 - all mechanisms, round robin*

In this scenario both the priority mechanism and the shaping mechanism are combined with the round robin buffer collection strategy. The effect this has on packet transfer is visible on Figure 7.

The total transfer time was 450 ms and no packets were dropped in transfer. Average packet time delay is 133 ms (8 ms priority 0, 120 ms priority 1 and 270 ms priority 2). In this case, combining both mechanisms also paid off, like in scenario 2, especially when observing the priority 0 delay times. The time delay drop at the end of the graph is the result of the Token Bucket algorithm that reduced the send rate once it ran out of tokens.

The buffer capacity graph is similar to the one in scenario 2.

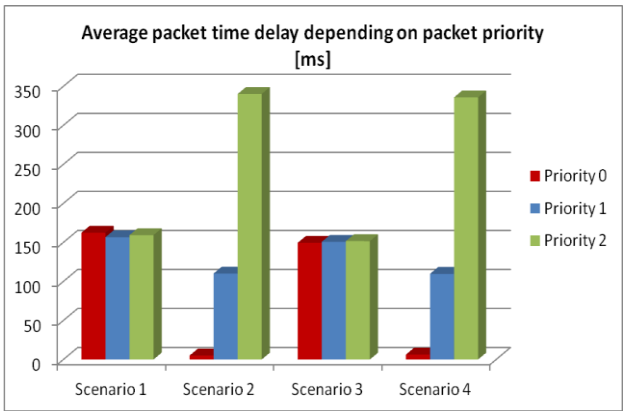


Figure 8. Average packet time delay depending on packet priority

V. SIMULATION RESULTS

After all the scenarios were run 25 times, average values of observed parameters were calculated. The most interesting parameters to analyze are average packet time delays depending on packet priority, given in Figure 8, and the average transfer times, given in Figure 9.

If average transfer times are observed, the conclusion is that scenario 3 (no priority, round robin, no Token Bucket) has the lowest transfer time of 469.72 ms. The highest transfer time can be seen in scenario 4 (round robin, priority, Token Bucket) with 575.8 ms, followed closely by scenario 2 (575.24 ms). The conclusion that can be drawn from this figure is that if lowest possible transfer time is desired in a communication network, in this case, only a buffer and the round robin collecting strategy will suffice.

On the other hand, if average packet time delay depending on packet priority is observed for each scenario, there are two distinct patterns. The even columns of scenarios 1 and 3 are the result of the scenarios not having the priority mechanism, and are not relevant for this observation. Other two scenarios have more interesting results. The lowest time delay for priority 0 can be seen in scenario 2 (5.04 ms), with scenario 4 close by (6.32 ms). They also have the lowest time delay of priority 1 packets (110.16 ms and 109.72 ms respectively). However, their delay times of priority 2 packets is substantially (more than two times) higher than in scenarios 1 and 3.

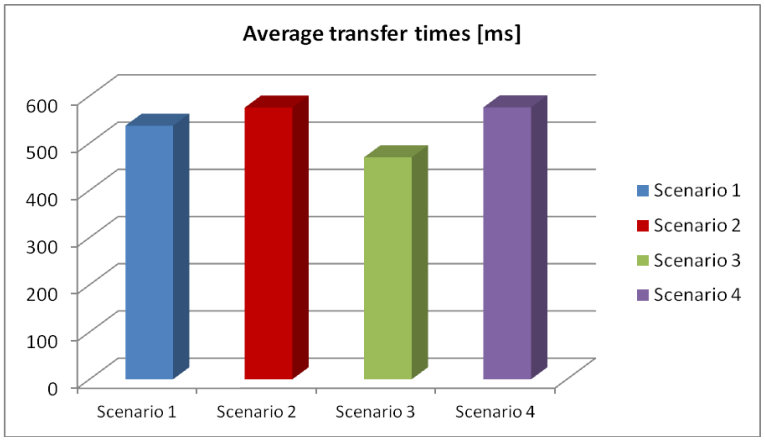


Figure 9. Average transfer times for all scenarios

For best priority 0 and 1 packet time delay, the priority mechanism should be used along with the Token Bucket algorithm, regardless of the buffer collecting strategy. Round robin strategy does give a bit higher priority 0 delay, but has better delays for other priorities, which can also be relevant, depending on the situation.

VI. CONCLUSION

This paper deals with IP packet delay, the impact it has on communication networks and how it can be reduced to improve Quality of Service, with the buffer as the main observation point. A simulator was developed to further study the impact each of the mechanisms has on buffer packet delays and buffer congestion, alone or combined with other mechanisms. The results show that a clear solution to QoS in communication networks does not exist. Instead, the solution greatly depends on the network itself and the individual needs for that particular network. The results have also shown that different combinations of QoS mechanisms improve network traffic in different ways. If there is a need for fast packet transfer, regardless of packet priority, the round robin buffer collection strategy can prove sufficient. If, on the other hand, packets have different priorities or some are more sensitive to delays than others, priority mechanism should be used combined with the Token Bucket shaping mechanism.

Other previously mentioned studies also show that network traffic improves with the introduction of QoS mechanisms. Again, their conclusions are drawn from simulations of specific, complex systems, with real life application, while simulation in this paper was deliberately simplified to emphasize how QoS mechanisms affect packet delay and buffer congestion.

Although an ultimate solution does not exist, a careful selection of mechanisms can provide a good solution, because using the QoS mechanisms makes packet transfer more clear, more predictable and more manageable, eliminating or lessening the undesirable effects in network transfer. This also allows declaring minimal performance guarantees necessary when establishing QoS in a communication network[1, 24, 25]. As a next step, the simulator will implement more advanced QoS mechanisms and will be improved to support larger traffic flows and allow flow customization to further study the impact of QoS mechanisms on different traffic flows.

REFERENCES

[1] Szigeti, T. & Hattingh, C. (2004). *End-to-End QoS Network Design: Quality of Service in LANs, WANs, and VPNs*, Cisco Press, Indianapolis

[2] *System i Networking Quality of Service (Version 6 Release 1)*, IBM Corp., 2009., *Available from:* http://publib.boulder.ibm.com/infocenter/iseries/v6r1m0/topic/rzak8/rzak8.pdf, *Accessed:* 2014.15.1.

[3] *Quality of Service Networking*, Cisco Press, *Available from:* http://docwiki.cisco.com/wiki/Quality_of_Service_Networking, *Accessed:* 2014.17.1.

[4] *Enterprise QoS Solution Reference Network Design Guide*, Cisco Press, San Jose

[5] Shenker, S., Partridge, C., Guerin, R. (1997). *Specification of Guaranteed Quality of Service*, RFC 2212, September 1997.

[6] Snir, Y., Ramberg, Y., Strassner, J., Cohen, R., Moore, B. (2003). *Policy Quality of Service (QoS) Information Model*, RFC 3644, November 2003.

[7] Polk, J., Dhesikan, S., Camarillo, G. (2009). *Quality of Service (QoS) Mechanism Selection in the Session Description Protocol (SDP)*, RFC 5432, March 2009.

[8] Chan, K., Sahita, R., Hahn, S., McCloghrie, K. (2003). *Differentiated Services Quality of Service Policy Information Base*, RFC 3317, March 2003.

[9] Crawley, E., Nair, R., Rajagopalan, B., Sandick, H. (1998). *A Framework for QoS-based Routing in the Internet*, RFC 2386, August 1998.

[10] Gettys, J. & Nichols, K. (2011). *Bufferbloat: Dark Buffers in the Internet*, Queue - Virtualization, Volume 9, Issue 11.

[11] Ke, P., Li, Y., Ni, F. (2013). *Multiple Services Scheduling with Priority Queuing Model*, IJCIT, Volume 3, Issue 1, *Available from:* http://ijcit.org/ijcit_papers/vol3no1/IJCIT-120702.pdf, *Accessed:* 2014.25.1.

[12] Patil, G., McClean, S., Gaurav, R. (2011). *Drop Tail and Red Queue Management with Small Buffers: Stability and HOPF Bifurcation*, ICTACT, Volume 2, Issue 2, *Available from:* http://eprints.ulster.ac.uk/21987/1/jct_Spl_Paper_339_344.pdf, *Accessed:* 2014.25.1.

[13] Francis-Cobley, P. (2004). *Congestion Control*, unpublished, *Available from:* http://ftp.utcluj.ro/pub/users/cemil/prc/CONGESTION%20CONTROL.ppt, *Accessed:* 2014.17.1.

[14] Ali Mantar, H. (2000). *The Token Bucket (Leaky Bucket) Model*, Syracuse University, unpublished, *Available from:* http://qbone.internet2.edu/bb/Bucket.doc, *Accessed:* 2014.15.1.

[15] Domokos Botond, B. (2001). *Traffic Shaping*, unpublished, *Available from:* http://ftp.utcluj.ro/pub/users/tarc/biro.ppt, *Accessed:2014.17.1.*

[16] Silver, G., Maduko, A., Jafri, R., Miller, J.A., Sheth, A.P. (2003). *Modeling and Simulation of Quality of Service for Composite Web Services* in Proceedings of the 7th World Multiconference on Systemics, Cybernetics, and Informatics (SCI'03), Orlando, FL, July 2003, Nagib Callaos, Anna M. Di Sciullo, Toshizumi Ohta, and Te-Kai Liu (Eds.), International Institute of Informatics and Systemics, 2003, pp. 420-425.

[17] Al-Naamany, A., Bourdoucen H., Al-Menthari W. (2008). *Modeling and Simulation of Quality of Service in VoIP Wireless LAN*, Journal of Computing and Information Technology, vol 16, no. 2, pp. 131–142

[18] Chen, B., (2002). *Simulation and Analysis of Quality of Service Parameters in IP Networks with Video Traffic*

[19] Oladosu, J. B., O Alao, A., Adigun M., Emuoyibofarhe, J. O. (2006). *Design and Simulation of a Quality of Service (QoS) Guaranteed Model for Mobile Ad Hoc Networks (MANets)*, SATNAC 2006, paper no. 250

[20] García, A. B., Álvarez-Campana, M., Vázquez, E., Berrocal, J. (2002). *Simulation of Quality of Service Mechanisms in the UMTS Terrestrial Radio Access Network*. Proc. of the 4th IEEE Conference on Mobile and Wireless Communications Networks (MWCN 2002). Stockholm, Sweden, 9-11 September 2002. pp. 174-178.

[21] Kanters, O. (2010). QoS analysis by simulation in Reo

[22] *The ns Manual*, unpublished, *Available from:* http://www.isi.edu/nsnam/ns/ns-documentation.html, *Accessed:* 2014.02.04.

[23] Kleinrock, L. (1976). *Queueing Systems, Vol. 2: Computer Applications*, Wiley Interscience, New York

[24] Kozacinski, H. & Knezevic, P. (2011). *Configuration of Quality of Service Parameters in Communication Networks*, Annals of DAAAM for 2011 & Proceedings of the 22nd International DAAAM Symposium, 23-26th November 2011, Vienna, Austria, Volume 22, No. 1, ISSN 1726-9679, ISBN 978-3-901509-83-4, Katalinic, B. (Ed.), pp. 1369-1370, Published by DAAAM International Vienna, Vienna

[25] Kozacinski, H. & Knezevic, P. (2013). *Configuration of Quality of Service Parameters in Communication Networks*, 24nd International DAAAM Symposium, 23-26th November 2013, Zadar, Croatia, Procedia Engineering, vol 69, pp. 655-664.

Mobile data Wi-Fi offloading

[1]Martin Sraga,[1]Norman Momčilović, [1]Alberto Tekovic
[1]University College Algebra, Ilica 242,10 000 Zagreb, Croatia
norman.momcilovic@racunarstvo.hr

Abstract - The predicted exponential growth of mobile data traffic demands suitable solutions for coping with the high amount of data. Increasing number of Wi-Fi enabled smartphones and Wi-Fi access becoming more widely available is opening the possibility of offloading a large chunk of data traffic from 3G/ 4G mobile networks to Wi-Fi IEEE 802.11. However, the quality of experience must be preserved no matter which network layer is currently engaged. Therefore, the scope of this paper was to determine the usability of the free of charge Wi-Fi available networks and to compare it with the legacy 3G\4G networks.

I. INTRODUCTION

This paper refers to the survey conducted on December 2013 in the city of Zagreb capital of Croatia mainly focusing on the center of the capital.

The goal of the survey was to determine the current performance of available Wi-Fi and 3G\ 4G networks, used by the typical user in the busy hour.

It is expected that the mobile data traffic will reach the following milestones within the next five years [1].

- By 2018, over half of all devices connected to the mobile network will be "smart" devices.
- Over two-thirds of the world's mobile data traffic will be video by 2018.
- The average smartphone will generate 2.7 GB of traffic per month by 2018, a 5-fold increase over the 2013 average of 529 MB per month.
- By 2018, 4G will be 15 percent of connections, but 51 percent of total traffic.
- By 2018, more than half of all traffic from mobile-connected devices (almost 17 exabytes) will be offloaded to the fixed network by means of Wi-Fi devices and femtocells each month.

From the milestones cited above, it is clear that existing 3G and 4G network won't be able to handle the expected traffic load. Beside few other proposed offload techniques, Wi-Fi offload is currently one of the most promising strategies.

II. WI-FI & 3G TECHOLOGY

A. 3G\4G and Wi-Fi

Wi-Fi and 3G\ 4G are both types of wireless data networks, providing high-speed Internet service for laptops, desktop PCs, smartphones and tablets. Although Wi-Fi enjoys an advantage in terms of raw speed, a 3G connection offers greater mobility; devices get a service across a much wider area. Many mobile devices can access either kind of network automatically, depending on which has the stronger signal. But it should be noted that in order to access WLAN network, one has to manual connect to it. In a future scenario it will be possible to automatically connect to the WLAN with the authentication being resolved via the SIM card [2].

3GPP Release 8 introduces a client-server based protocol, Dual Stack Mobile IP (DSMIP), to enable seamless handover between 3G and Wi-Fi. DSMIP is a mobility protocol specified in IETF that provides IP address preservation for IPv4 and IPv6 sessions, allowing the user to roam independently in IPv4 and IPv6 accesses.

The solution does not require any support from WLAN accesses. If the cellular radio access network supports a Home Agent (HA), the only new requirement is that the UE (i.e., client) and the Home Agent (i.e., server) are dual-stack capable [3].

Although the described Wi-Fi offload solution doesn't require any support from WLAN accesses, that doesn't mean that there is no ongoing work on the IEEE side. The standardization bodies for the next generation Wi-Fi (802.11ac) has taken into account increased Internet usage, smartphones, portable and smart TV devices, offload from 3GPP networks, and the fact that video is still growing [4].

B. Range

As user moves farther from a wireless network's antenna, the signal strength fades, limiting ability to get a connection, and its maximum possible speed. Wi-Fi ranges vary from 45 meters for the oldest versions to 70 meters for 802.11n [1]. A single Wi-Fi access point is sufficient for a coffee shop, but larger organizations need multiple access points. 3G covers a much wider area; service providers already have multiple cell towers servicing cities, towns and major highways.

C. Transfer data

The download speed of a Wi-Fi connection varies depending on the type of Internet connection that the Wi-Fi access point is connected to. Download speeds can be as low as 1 megabit per second (Mbps) or as high as 5 to 8 mbps or more. With 3G Internet service, download speeds can vary between 1 Mbps at the cell border, up to 25 Mbps in case of dual carrier capable terminal and network and a good radio conditions. The average upload speed on a 3G connection is approximately 1.5 Mbps, depending on signal strength and network congestion. The 4G speeds are even greater. Typically, a Wi-Fi connection will have upload speeds ranging from approximately 300 to 600 kbps [5].

III. MEASUREMENT METHODOLOGY

In order to evaluate achievable performance on currently available 3G and Wi-Fi networks in the city

center, three main KPI's were measured: FTP download throughput (Mbps), FTP upload throughput (Mbps) and PING (ms).

Measurements were conducted by the usage of a regular commercial smartphone Samsung Galaxy S3 that was downloading the information via the VIP server.The radio network parameters were acquired using measurement application NEMO Handy. Speedtest application has been used to acquire KPI data. Measurement has been conducted in the city of Zagreb, on nine different micro locations. On every measurement spot, thirty minutes test has been performed. Fifteen minutes for 3G\ 4G, and fifteen minutes for Wi-Fi. The final result for each location was taken from the average of conducted ten test samples per technology.



Figure 1 Measurement locations

IV. RESULTS OF THE SURVEY

The accumulated data from the results of the survey can be found below represented graphically

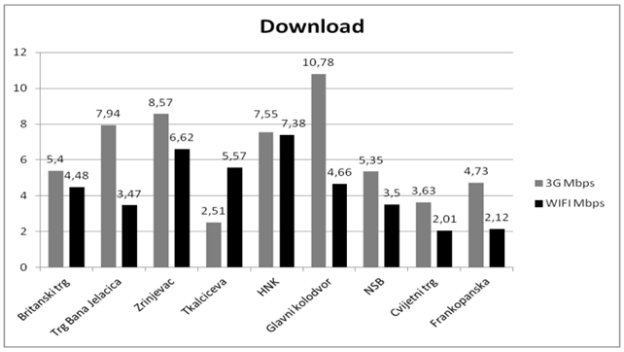


Figure 2 FTP download test

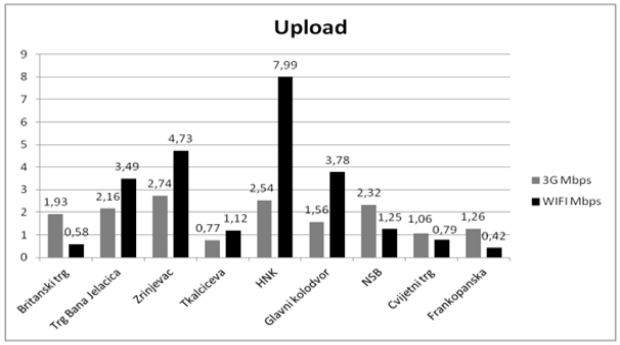


Figure 3 FTP upload test

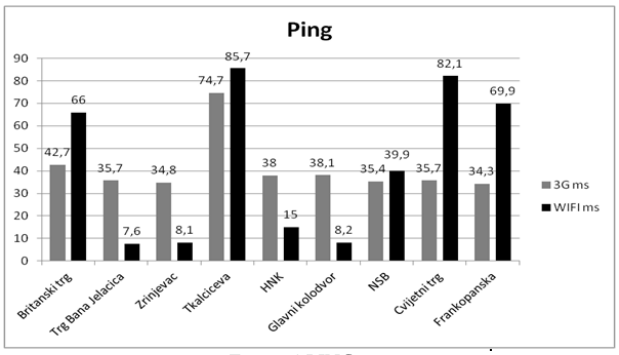


Figure 4 PING test

From the presented graphs it can be concluded that there is a good quality Wi-Fi network currently availability for the regular consumer as a free of charge alternative to 3G\ 4G network. In the city center there are enough open Wi-Fi locations to accommodate the basic needs: reading, sending emails and internet browsing. Average download throughput of 4-5 Mbps is currently enough to partake in those kinds of activities without experiencing any noticeable problems. 3G\ 4G technology provides a good quality of service through the whole city.

V. CONCLUSION

In the center of Zagreb city both network types have a good quality of service and they fulfill the average needs of a wireless internet user, but still 3G\4G networks performs slightly better keeping in mind that may be due to Wi-Fi measurements being performed outdoors while the access points are located indoors and for that reason may have slightly worse results. Taking into account that the two networks are still not "communicating" with each other, it is required to reconnect from one to the other, which from the user perspective can be found annoying. Therefore, it seems that the planned 3G\ 4G and WLAN integration will be really appreciated from the users, since it will enable automatically camping on the best network. That can be accomplished by using the 802.11f standard if all the providers deal with the handover in the same manner that implies not modifying the standard to their needs but following it.

ACKNOWLEDGMENT

We would like to thank Professor Alberto Teković for giving us the opportunity to make this paper and the device required to complete it.

REFERENCES

[1] Cisco,"Cisco Visual Networking Index", available at http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-520862.html, Februrary 2014

[2] Vijay Garg," Interworking between WLANs and 3G", available at http://www.eetimes.com/document.asp?doc_id=1278379

[3] *Qualcomm, "3G/*Wi-Fi *Seamless* Offload", available at http://www.qualcomm.com/media/documents/3gwifi-seamless-offload, July 18, 2013

[4] M. Abbas, "The impact of 5th Generation Wifi" (802.11ac), availabe at http://www.slideshare.net/mazlan1/the-impact-of-80211ac, Jan 16, 2013

[5] J.T. Barett, Demand Media "3G Speed Vs. Wi-Fi", available at http://techchannel.radioshack.com/3g-speed-vs-wifi-1572.html

Testing optical splitters

Ivan Žuti [1], Aleksandar Komparić [1]
[1] University College Algebra, Ilica 242, 10 000 Zagreb, Croatia
izuti@racunarstvo.hr

Abstract – This paper gives an overview of bidirectional optical splitter characteristics. It outlines the basics of passive optical network infrastructure, describes the most common attenuation mechanisms in optical fibers and the testing methodology for measuring optical splitter performance. The paper compares two types of Planar Light wave Circuit (PLC) splitters and provides measurement results for samples of both.

I. INTRODUCTION

Single mode optical fiber is now the transport medium of choice in metropolitan and long-haul passive optical networks (PONs). Fiber optic cables are becoming common even in the last mile – the segment of the network that extends from the central office to the subscriber. In order for the subscribers to share the total available bandwidth, a splitter is commonly used. The splitter is a bidirectional branching device that has one input port and multiple output ports, 1xN [2]. The input optical signal is divided among the output ports. Splitters are passive devices because they require no external energy source other than the incident light beam. Depending on network topology, there may be one splitter or several cascaded splitters. ITU-T recommendation G.983 allows split ratios of up to 32, while Recommendation G.984 allows split ratios of up to 64 splits. The splitter must accommodate the allowed optical budget, regardless of topology [1].

Splitters can be packaged in different shapes and sizes, depending on the basic technology used. The most common type is a planar waveguide (Fig. 1), typically used for high split ratios, and fused biconic taper, usually used for low split counts. As a result, the planar waveguide splitters are more commonly used.

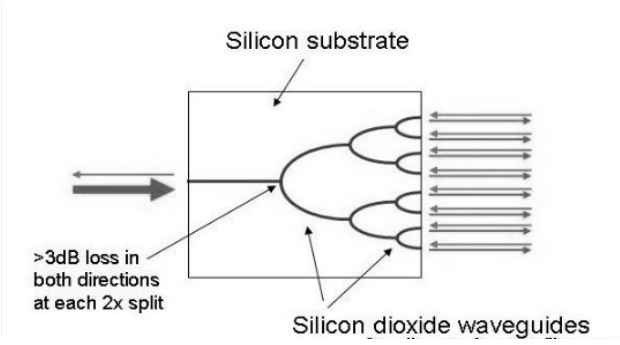


Figure 1 Planar light wave circuit diagram [7]

Planar Light wave Circuit (PLC) is based on semiconductor chip fabrication technology and is cost effective at large splitting ratios. However, since it is not fiber based it suffers from problems with fiber connectivity and can suffer from polarization effects due to its design. Because splitters account for the majority of signal loss in an optical network, they must be thoroughly tested to minimize that loss [2].

Fused Biconic Taper (FBT) is a type of splitter formed by two independent optical fibers. It works as a result of an energy transfer caused by the coupling proximity between optical fiber cores. Using two separate fibers will always result in a 2x2 coupler. To obtain a 1x2 coupler, one of the input fibers must be cut and terminated to avoid back-reflections from that end [6].

II. PASSIVE OPTICAL NETWORK ARCHITECTURE

Passive optical network (PON) is a telecommunications network that uses point-to-multipoint fiber to the premises in which unpowered optical splitters are used to enable a single optical fiber to serve multiple premises. A PON consists of an optical line terminal (OLT) at the service provider's central office and a number of optical network units (ONUs) near end users [2] as shown in Figure 2.

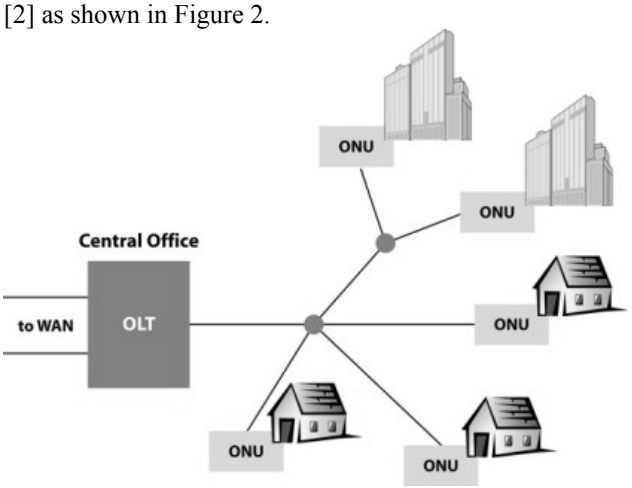


Figure 2 PON example [9]

PON reduces the amount of fiber and central office equipment required compared with point-to-point architectures. A passive optical network is a form of fiber-optic access network. In most cases, downstream signals are broadcast to all premises sharing multiple fibers. Encryption can prevent eavesdropping. Upstream signals are combined using a multiple access protocol, usually time division multiple access (TDMA) [2].

Optical network starts at the headend, where the public-switched telephone network and Internet services are interfaced with the optical distribution network via the optical line terminal. They, along with a TV signal, are converted into an optical signal and sent by the optical line terminal. That means that the same optical fiber simultaneously carries different information in different directions. This feeder fiber cable then connects to a splitter which splits the signal depending on the network design [1].

The signal can be split again as long as the network remains within the allowed optical budget. Each splitter compounds the signal loss and with it, the bandwidth that is available to the subscriber is lesser. The signal then arrives to the subscribers endpoint device, be it an optical network unit which can host a number of subscribers, or an optical network terminal which serves a single subscriber. That device then transforms the optical signal into electronics in order to present the delivered content on its respective devices [1].

Fiber optic cable installation is one of the most costly elements in network deployment, and depends on numerous factors, such as whether the fiber is installed in new premises or in an existing development over active routes. There are several cable installation methods used, the most common being direct burial, in which the cable is placed underground, duct installation, which is more expensive but allows for easy maintenance later, and aerial installation, in which the cable is typically installed on poles or towers [2].

III. FIBER OPTIC ATTENUATION

Attenuation in fiber optics, also known as transmission loss, is the reduction in intensity of the light beam (or signal) with respect to distance travelled through a transmission medium. Attenuation coefficients in fiber optics usually use units of dB/km through the medium due to the relatively high quality of transparency of modern optical transmission media. The medium is typically a fiber of silica glass that confines the incident light beam to the inside. Attenuation is an important factor limiting the transmission of a digital signal across large distances. Thus, much research has gone into both limiting the attenuation and maximizing the amplification of the optical signal. Empirical research has shown that attenuation in optical fiber is caused primarily by both scattering and absorption [5]. This leads to the inherent signal loss along the fiber. With it, there are several other causes for signal loss in an optical network.

Total fiber optic attenuation can be the result of several optical fiber loss mechanisms:

- Rayleigh scattering
- Absorption
- Macroscopic and microscopic bends
- Simulated nonlinear scattering
- Mechanical issues around fiber connections

Attenuation coefficients in fiber optics typically use units of dB/km because of the high quality materials used in modern fiber optic mediums.

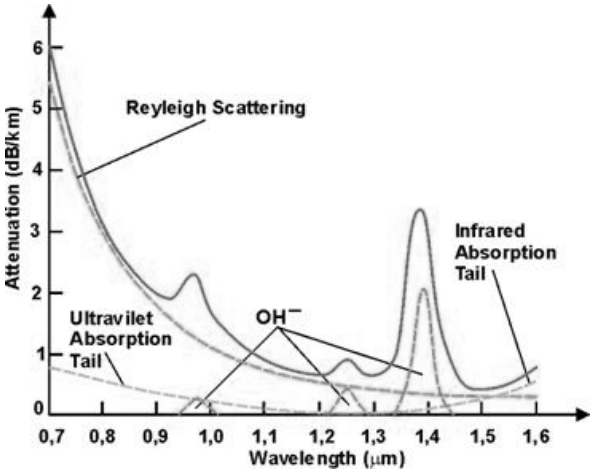


Figure 3 Spectral attenuation of a silica optical fiber [4]

Rayleigh scattering occurs when irregularities exist in the optic fiber, and they are smaller than wavelength. These irregularities can either be uneven internal surfaces or in homogeneities in the material [1]. Rayleigh scattering depends on the wavelength; the losses can be expressed as:

$$\alpha_r[\text{dB/km}] = 1.7(0.85/\lambda[\mu\text{m}])^4. \quad (1)$$

Rayleigh scattering defines the fundamental limit of reachable fiber attenuation, and it is the most significant attenuation mechanism within optical limits [4]. Absorption is a process of conversion of electro-magnetic energy into other forms of energy [4]. In optical fibers, the electro-magnetic energy of the light is converted into lattice vibrations. Each material used as a fiber optic medium has different absorption rates based on the wavelength of the signal. Most impactful impurity comes from water in form of hydroxyl ions (OH^-).

Macroscopic bends occur when the optical fibers are installed. If the band diameter is greater than 50 mm the losses are negligible.

Microscopic bend losses occur as a result of localised bends or bumps in the fiber core which then redirect or scatter the laser signal.

Simulated nonlinear scatterings are typically divided into Stimulated Brillouin backscattering and Stimulated Raman scattering. Stimulated Brillouin backscattering (SBS) occurs when the laser signal travelling through the optic produces vibrations in the medium due to the variations in the electric field of the beam itself. Threshold power for the SBS effect is:

$$P_B = 4.4 * 10^{-3} d^2 \lambda^2 \alpha_{dB} \upsilon. \quad (2)$$

Where d is fiber core diameter, α_{dB} is fiber attenuation, λ is beam wavelength and υ is optical beam bandwidth. Stimulated Raman scattering (SRS) occurs when in elastically scattered photons are forced to stay in the

medium, which causes high frequency photon generation that propagates in both directions in the medium. Threshold power for the SRS effect is:

$$P_R = 5.9 * 10^{-2} d^2 \upsilon \lambda \alpha_{dB}. \quad (3)$$

The threshold optical power of the SRS effect can be up to three orders of magnitude higher than the Brillouin threshold.

The connection points of two fibres are susceptible to varying degrees of mechanical issues. These include discontinuities, gaps, misalignments, cracks, faults etc. All of these affect the signal travelling through the optical fiber.

IV. TESTING OPTICAL SPLITTERS

Splitter loss usually accounts for the majority of the loss in the system. It depends mainly on its number of output ports. The downstream optical signal is divided equally into a cascade of branches, so each path gets the corresponding percentage of the original incident beam. The splitter also adds approximately the same amount of loss for the light travelling in the upstream direction [1].

One of the most important factors in ensuring proper transmission is controlling the power loss in the network against the link loss budget specifications. This is done firstly by establishing a total end-to-end loss budget with a sufficient margin. In addition, back reflections must be reduced to a minimum [1].

Optical return loss (ORL) is defined as the ratio of incident power (P_i) to reflected power (P_r) (4) and is measured at the input of a device under test. ORL is measured in dB and is a positive number.

$$ORL[dB] = 10\log_{10} P_i/P_r \quad (4)$$

The higher the ORL value, the better the system will perform. Optical loss is defined as the difference in power level between the transmitting source and the receiving power meter.

Insertion loss (IL) is the loss of optical energy resulting from the insertion of a component or device in the optical network.

Channel uniformity is defined as the difference between maximum and minimum measured insertion losses on a specific wavelength. Lower is better.

Because communication over the fiber is bidirectional, the ORL must be measured in both directions using an ORL test meter or a compatible optical loss test set (OLTS) at each end of the link. When using an OLTS it is important to remember that the test set requires a sequence of two measurements, which are taken in two steps [2]:

1) Two OLTSs are first referenced to each other using their individual light sources.

2) Each OLTS then sends a calibrated power value from its light source over the section under test to the other OLTS, which measures the received power and calculates the loss.

The splitter causes an inherent loss because the input power is divided between several outputs. Splitter loss depends on the split ratio and is approximately 3 dB for a 1x2 splitter. The loss increases by 3 dB each time the number of outputs is doubled. A 1x16 splitter has a splitter loss of at least 12 dB [1].

V. TEST RESULTS

Testing was performed using an OLTS. Both instruments were configured by selecting appropriate wavelengths (1310nm and 1550nm) and establishing other test parameters. Then, referencing on both instruments was performed. This was necessary to measure loss through the splitter only, and not through the test jumpers and accessories. One OLTS instrument was then connected with the input port of the splitter and the other instrument to the first output port of the splitter. This was repeated for all output ports Fig. 4.

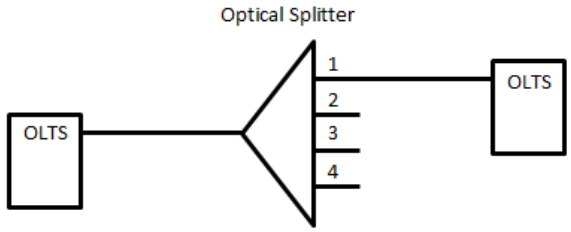


Figure 4 Measurement setup

First tested splitter was a PLC-R2 104 with a split ratio 1x4. Specification parameters for this splitter are: maximum 7.2 dB insertion loss, minimum 55 dB optical return loss and maximum 0.7 dB uniformity. Results are shown in Table I.

TABLE I. PLC SPLITTER1X4 INSPECTION REPORT

Item	IL(dB)		ORL (dB)	
Wavelength	1310	1550	1310	1550
Spec	MAX ≤ 7.2		MIN ≥ 55	
Port1	6.80	6.77	Pass	Pass
Port2	6.86	6.82	Pass	Pass
Port3	6.81	6.79	Pass	Pass
Port4	6.84	6.77	Pass	Pass
MAX.	6.86	6.82	N/A	
MIN.	6.80	6.77	N/A	
Average	6.83	6.79	N/A	

Typical insertion loss for a 1x4 splitter is between 6 and 8 dB. The highest measured loss for this splitter was 6.86 dB on the 1310nm wavelength and the lowest was 6.77 dB on the 1550nm wavelength. The average losses were 6.83 dB on 1310nm and 6.79 dB on the 1550nm wavelengths. Uniformity is 0.06 dB on the 1310nm and 0.05 dB on the 1550nm wavelengths.

Second tested splitter was a PLC-R2 116 with a split ratio 1x16. Specification parameters for this splitter are: maximum 13.5 dB insertion loss, minimum 55 dB optical return loss and maximum 1.4 dB uniformity. Results are show in Table II.

TABLE II. PLC SPLITTER1X16 INSPECTION REPORT

Item	IL(dB)		ORL (dB)	
Wavelength	1310	1550	1310	1550
Spec	MAX ≤ 13.5		MIN ≥ 55	
Port1	13.00	13.15	Pass	Pass
Port2	12.98	12.86	Pass	Pass
Port3	12.79	12.72	Pass	Pass
Port4	13.00	12.91	Pass	Pass
Port5	13.01	13.13	Pass	Pass
Port6	12.95	12.87	Pass	Pass
Port7	12.98	12.88	Pass	Pass
Port8	13.01	13.14	Pass	Pass
Port9	12.92	12.84	Pass	Pass
Port10	12.97	12.90	Pass	Pass
Port11	12.74	12.73	Pass	Pass
Port12	12.97	12.91	Pass	Pass
Port13	13.05	12.86	Pass	Pass
Port14	12.99	12.89	Pass	Pass
Port15	12.94	12.86	Pass	Pass
Port16	12.98	12.96	Pass	Pass
MAX.	13.05	13.15	N/A	
MIN.	12.74	12.72	N/A	
Average	12.96	12.91	N/A	

Typical insertion loss for a 1x16 splitter is between 12 and 14 dB. The highest measured loss for this splitter was 13.15 dB on the 1550nm wavelength, lowest was 12.72 dB on the 1310nm wavelength and the average losses were 12.96 dB on the 1310nm and 12.91 dB on the 1550nm wavelengths. Uniformity is 0.31 dB on the 1310nm and 0.43 dB on the 1550nm wavelengths. In comparison to the 1x4 splitter we can conclude that the 1x16 splitter amounts approximately twice more insertion loss is but we get four times as many output ports.

Taking into consideration that connectors usually create 0.2 dB of inherent power loss and splices usually create only 0.05 dB of inherent power loss we can conclude that splitters create the majority of optical power loss in PONs today.

VI. CONCLUSION

There are many advantages of using optical splitters in PONs. The loss of transmission is not sensitive to the wavelength of light, to meet the transmission needs of different wavelengths. Spectroscopic uniform signal can be uniformly allocated to the user. They are small and compact, and as such can be installed directly in an existing junction box. Only a single device shunt channel can achieve in excess of 32 channels. However there are some disadvantages. Not only are the power losses on splitters much greater than on other elements of the PON, but they split the maximum available bandwidth of a single fiber to multiple users so that each user receives a fraction of that bandwidth. Production process is complex, has a high technical threshold and also the signal needs to be coded, which results in more expensive end devices.

The tested splitters have passed the specification parameters and can be installed into segments of PONs that require a larger number of channels then there are available fibers. When choosing a particular split ratio it is necessary to provide enough output channels and also take into account reserve channels for possible future expansion, while staying within the optical loss budget the whole time. Another important factor to consider is the uniformity of splitter loss across every port of the splitter, and also across multiple splitters of the same type. This helps with planning and maintenance of the PON. As a result, the splitter with the most uniform loss across all of its ports and the lowest split ratio necessary is the best option to choose.

ACKNOWLEDGMENT

We would like to thank Professor Zvonimir Zelenika for providing us with very useful materials, and to Professor Alberto Teković for making us do this paper in the first place.

REFERENCES

[1] Andre Girard, "FTTx PON Technology and Testing", EXFO Electro-Optical Engineering Inc., Quebec City, Canada, 2005

[2] "FTTH PON Guide: Testing Passive Optical Networks, 5th Edition", EXFO Electro-Optical Engineering Inc., Quebec City, Canada, 2012

[3] Understanding Fiber-Optic Cable Signal Loss, Attenuation, and Dispersion 24.1.2014. found at: http://www.juniper.net/techpubs/en_US/release-independent/junos/topics/concept/fiber-optic-cable-signal-loss-attenuation-dispersion-understanding.html

[4] Attenuation of Optical Fibers 6.2.2014. found at: http://www.invocom.et.put.poznan.pl/~invocom/C/P1-9/swiatlowody_en/p1-1_2_1.htm

[5] Mechanisms of attenuation 6.2.2014. found at: http://en.wikipedia.org/wiki/Attenuation

[6] Fused Biconical Taper (FBT) Process & Packaging 8.4.2014 found at: http://www.gouldfo.com/fbtpackaging.aspx

[7] How does a PLC splitter work 16.4.2014 found at: http://www.exfiber.com/tutorial/How-does-a-PLC-Splitter-work-13.html

[8] "Polarization Dependent Loss Measurement of Passive Optical Components", Agilent Technologies, 2002

[9] Nilam Ruparelia, Altera, "Evolving passive optical networks (PONs) demand FPGA design flexibility", 2008

Comparison of different DC motor positioning control algorithms

N. Baćac*, V. Slukić*, M. Puškarić*, B. Štih*, E. Kamenar**, S. Zelenika**

* University of Rijeka, Faculty of Engineering, Rijeka, Croatia
** University of Rijeka, Faculty of Engineering and Centre for Micro and Nano Sciences and Technologies, Rijeka, Croatia

ekamenar@riteh.hr

Abstract – A comparison between different DC motor positioning control algorithms is performed in this work. Transient responses while employing a PID controller, a cascade controller and a state-space controller are considered. LabVIEW programming environment with a suitable acquisition card and a miniature DC motor with an integrated encoder are used for experimental assessment. Calculations and control system simulations are made using Matlab. The PID controller is implemented via the predefined PID block in LabVIEW. In turn, the state-space controller is modelled by using Matlab while the accuracy of the results is confirmed experimentally using LabVIEW. The cascade controller is developed as a series of two Proportional-Integral (PI) controllers, one representing the positioning and the other the velocity loop. The obtained results allow establishing that positioning control via the state-space controller has the fastest response and the lowest settling times.

I. INTRODUCTION

A widely used actuator in positioning systems is a Direct Current (DC) motor. It finds application in many of today's mechatronics systems such as robots, precision positioning machines or industrial applications. Positioning control of mechatronic devices is also used in situations when there is a need for an accurate response in a predictable and repeatable manner. In a pick-and-place machine for production of Printed Circuit Boards (PCB), for example, components must be placed precisely on the board before the soldering process. Another example is a robot manipulator that uses several motors for 2D or 3D positioning of the robotic arms. Although stepper motors are sometimes employed for these purposes, DC motors can also be a viable solution. However, when DC motors are used, a feedback sensor is needed in order to establish positioning control [1, 2].

A DC motor with an embedded quadrature rotational incremental encoder as a feedback sensor is employed in this work. Three different control algorithms are compared in the terms of simulations in Matlab and experimental results obtained by using the LabVIEW programming environment.

II. DC MOTOR POSITIONING CONTROL ALGORITHMS

A. PID control

One of the most common controllers in industrial applications and control systems is the Proportional-Integral-Derivative (PID) controller. The transfer function of an ideal parallel PID structure can be expressed as (refer to the list of symbols at the end of the paper) [1]:

$$G_R(s) = \frac{U(s)}{E(s)} = K_P\left(1 + \frac{1}{T_I s} + T_D s\right) \tag{1}$$

The PID parameters are defined as:

- K_P proportional gain
- T_I integral time
- T_D derivative time

$E(s)$ can be calculated as the difference between the reference position $Y_0(s)$ and the actual position $Y(s)$:

$$E(s) = Y_0(s) - Y(s) \tag{2}$$

In the time domain, the PID controller can be expressed as:

$$u(t) = K_P \cdot \left(e(t) + \frac{1}{T_I}\int_0^t e(\tau)d\tau + T_D\frac{de(t)}{dt}\right) \tag{3}$$

Using the finite time values:

$$t \to kT_s \tag{4}$$

a discretized form of PID controller can be obtained [1]:

$$u(kT_s) = K_P \cdot \left(e(kT_s) + \frac{T_s}{T_I}\sum_{i=0}^{k} e(iT_s) + \frac{T_D}{T_s}[e(kT_s) - e(kT_s - T_s)]\right) \tag{5}$$

The PID controller is implemented in this work via the predefined PID block in LabVIEW. The tuning of the PID controller parameters is conducted in two steps. The Ziegler-Nichols method is used first to achieve a rough estimate of the gains; in a second step, an experimental method of fine-tuning of PID parameters is performed.

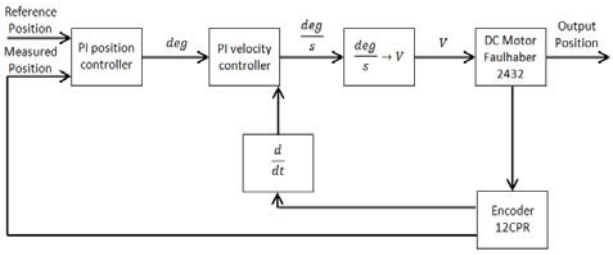


Figure 1. Cascade control scheme

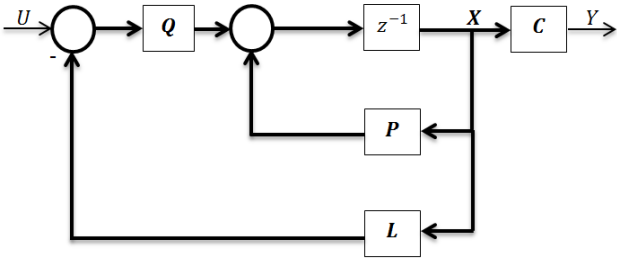


Figure 2. z-transformation of the state-space model

B. Cascade control

Cascade control is composed of two loops: the velocity and the position loop. In the simplest form, an analogue command (set point velocity) is compared with the signal from the feedback sensor (incremental rotational quadrature encoder) to generate voltage that induces rotational motion. The produced torque will speed up or slow down the actuator in order to reach the velocity set point. The most known velocity loop is the Proportional-Integral (PI) loop and it is constituted by two parameters: a proportional gain (K_P), which scales the velocity error, and an integral time constant (T_I), which defines the integration time. A velocity loop itself cannot ensure that the actuator stops in a certain position. Hence, one of the common configurations is to place a positioning loop in cascade (series) with the PI velocity controller (Fig. 1) [3].

Tuning of the regulators in cascade can often prove to be a challenging task since there are four parameters to tune [3]. The parameters are thus tuned in this work by using a custom developed Matlab model. The resulting parameters are then implemented in the LabVIEW environment where additional online fine tuning is performed in order to match better real system response.

C. State-space control

State-space control derives from the state-variable method of representing differential equations. When compared with transfer-function based control, state-space control is designed by working directly with the state-space-variable description of the system [2].

Advantages of state-space design are particularly relevant for Multiple-Input – Multiple-Output (MIMO) systems, although the system considered in this work is a Single-Input – Single-Output (SISO) system. A further advantage of this method is its robustness to dynamic perturbations in the system [2].

A process with one input $u(t)$ and one output $y(t)$ can hence be described in a state-space model as [4-6]:

$$\frac{d\mathbf{x}(t)}{dt} = \mathbf{A}\mathbf{x}(t) + \mathbf{B}u(t)$$
$$y(t) = \mathbf{C}\mathbf{x}(t) \tag{6}$$

where **A** is the system matrix which correlates the current state with its change, **B** is the control matrix and determines how the system input affects the state change and **C** is the output matrix which determines the relationship between the system state and its output. The considered process in a discrete form can thus be described as:

$$\mathbf{x}[k+1] = \mathbf{P}\boldsymbol{x}[k] + \mathbf{Q}u[k]$$
$$y[k] = \mathbf{C}\mathbf{x}[k] \tag{7}$$

The **P** and **Q** matrices can be defined by using the fundamental matrix $\boldsymbol{\Phi}(s)$ [4]:

$$\boldsymbol{\Phi}(s) = (s\mathbf{I} - \mathbf{A})^{-1}$$
$$\mathbf{P} = \mathcal{L}^{-1}\{\boldsymbol{\Phi}(s)\} \tag{8}$$
$$\mathbf{Q} = \mathcal{L}^{-1}\left\{\frac{\boldsymbol{\Phi}(s)}{s}\right\}\mathbf{B}$$

The process can also be described by applying the z-transformation of the equations of the state-space model [4]:

$$\mathbf{X}(z) = (z\mathbf{I} - \mathbf{P})^{-1}\mathbf{Q}U(z)$$
$$Y(z) = \mathbf{C}(z\mathbf{I} - \mathbf{P})^{-1}\mathbf{Q}U(z) \tag{9}$$
$$G(z) = \frac{Y(z)}{U(z)} = \mathbf{C}(z\mathbf{I} - \mathbf{P})^{-1}\mathbf{Q}$$

Now the state-space controller equation can be defined:

$$U(z) = -\mathbf{L}\mathbf{X}(z) \tag{10}$$

The basic principle of the synthesis of the state-space controller is to achieve the desired closed-loop dynamics with proper values of the **L** vector. These values can be obtained by using the following equations:

$$z\mathbf{X}(z) = (\mathbf{P} - \mathbf{Q}\mathbf{L})\mathbf{X}(z)$$
$$|z\mathbf{I} - \mathbf{P} + \mathbf{Q}\mathbf{L}| = 0 \tag{11}$$

where the roots of equation (11) are correlated with the poles of the desired closed-loop dynamics. The state-space controller gain vector **L** can then be calculated by using Ackermann's formula [4]. The block diagram of the resulting state-space model is shown in Fig. 2.

D. DC motor model in state-space

The behaviour of a DC motor can be modelled as [5]:

$$\frac{d\varphi}{dt} = \omega \tag{12}$$

$$J\frac{d\omega}{dt} + b \cdot \omega = \frac{K_m}{R_a}(u - K_e \cdot \omega)$$

For simplicity, the armature time constant is neglected.

In turn, in state-space a DC motor can be described as:

$$s\begin{bmatrix}\varphi\\ \omega\end{bmatrix} = \begin{bmatrix}0 & 1\\ 0 & -\frac{1}{T_m}\end{bmatrix} \cdot \begin{bmatrix}\varphi\\ \omega\end{bmatrix} + \begin{bmatrix}0\\ \frac{K_m}{J \cdot R_a}\end{bmatrix} \cdot u$$

$$Tm = \frac{J \cdot R_a}{b \cdot R_a + K_m \cdot K_e} \tag{13}$$

$$b = K_m \frac{I_{an}}{\omega}$$

The DC motor transfer function is:

$$G(s) = \frac{\Phi(s)}{\Omega(s)} = \frac{\frac{K_m}{J \cdot R_a}}{s^2 + s \cdot \frac{1}{T_m}} \tag{14}$$

while the respective discretized model in state-space is:

$$z\begin{bmatrix}\varphi\\ \omega\end{bmatrix} = \begin{bmatrix}1 & \alpha \cdot T_m\\ 0 & \beta\end{bmatrix} \cdot \begin{bmatrix}\varphi\\ \omega\end{bmatrix} + \frac{K \cdot T_m}{J \cdot R_a}\begin{bmatrix}\gamma \cdot T_m\\ \alpha\end{bmatrix} \cdot u$$

$$\beta = e^{-\frac{T_s}{T_m}}$$
$$\alpha = 1 - \beta \tag{15}$$
$$\gamma = \frac{T_s}{T_m} - \alpha$$

The requirements for aperiodic closed-loop dynamics imply then:

$$\xi = \frac{1}{\sqrt{2}}$$
$$\omega_n = \frac{1}{T_m\sqrt{2}} \tag{16}$$

while the recommended sample time is [5]:

$$T_s = \frac{0.2}{\omega_n} \tag{17}$$

Vector **L** can hence be approximately calculated as [5]:

$$\mathbf{L} = \frac{J \cdot R_a}{K_m \cdot T_m}\left[\frac{0.4983}{T_m} \quad 0.0672\right] \tag{18}$$

By inserting the characteristic values for the used DC motor (see Table I), vector **L** becomes:

$$\mathbf{L} = [1.1173 \quad 0.0011] \tag{19}$$

To estimate finally the variables used in the control algorithm, the Matlab software is used.

TABLE I. FAULHABER 2342 DC MOTOR CHARACTERISTICS

Moment of inertia	J	$5.7 \cdot 10^{-7}$ kg · m^2
Torque constant	K_m	$13{,}4 \cdot 10^{-3}\ \frac{\text{Nm}}{\text{A}}$
Back-EMF constant	K_e	$1{,}4 \cdot 10^{-3}\ \frac{\text{V}}{\text{rpm}}$
Armature resistance	R_a	1,9 Ω
Nominal voltage	U_n	12 V
Nominal armature current	I_{an}	75 mA
Friction coefficient	b	$1\ \frac{\text{mNm}}{\text{rpm}}$

III. EXPERIMENTAL SET-UP

The assessment of the characteristics of the described positioning control algorithms is performed on an experimental set-up, whose block-scheme is shown in Fig. 3. The system is composed of a DC motor coupled with a gearbox and an encoder, the operational amplifier and the control system based on the National Instruments (NI) hardware and the LabVIEW software.

A. Actuator and encoder

The main component of the experimental set-up is the Faulhaber 2342 DC motor with an embedded planetary gearbox and an incremental rotational encoder [7]. The main characteristics of the actuator are given in Table I. As shown in Fig. 4, the shaft of the actuator is connected to the gearbox unit (1) with a reduction ratio 1:64. On the opposite side, an encoder (2) is installed.

The used encoder is an incremental encoder with 2 channels (A and B) with a 90 degree phase delay and 12 cycles per revolution (cpr). Two encoder channels enable not only position and speed but also the direction of movement to be determined. A very important parameter for encoder measurements is the encoding type, which directly affects encoder resolution. In this application, rising and falling edges on both channels are counted (Fig. 5). This type of encoding is usually referred as X4 encoding [8-10].

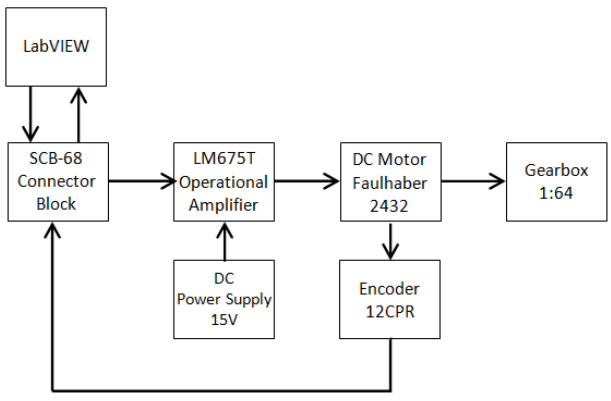


Figure 3. Experimental set-up scheme

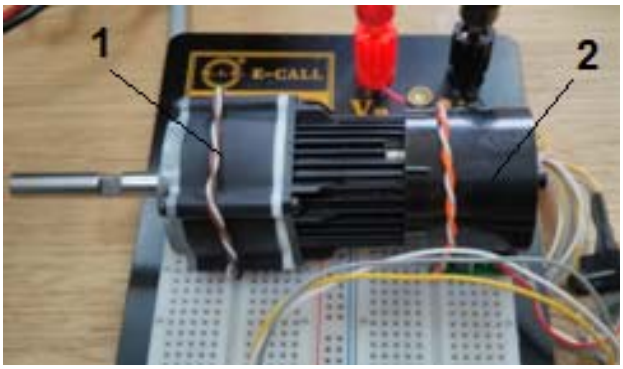


Figure 4. DC motor with gearbox and encoder

B. *National Instruments Hardware*

The electronics of the experimental set-up is based on the National Instruments PXI-1050 chassis, including a PXI-8196 embedded controller and a PXI-6221 Data Acquisition Card (DAQ) which is connected to a SCB-68 connector block [11]. The SCB-68 has 68 terminals with digital and analogue inputs and outputs. Counter inputs are used for connecting the encoder while the analogue output is used for driving the DC motor.

IV. LABVIEW CONTROL IMPLEMENTATION

The LabVIEW development environment is used for implementing the control algorithms and storing measurement data. This configuration meets the requirements for embedded system design such as real-time computing, wide choice of peripheral equipment and extensive support for accessing instrumentation hardware. Moreover, its library contains predefined mathematical blocks that process acquired data to enable appropriate outputs. In the considered case, the DAQmx *Start Task block* is used to trigger the used channels. The acquired data is then passed to the DAQmx *Read block* that, due to continuous position readings, is located within the time loop. Similarly, voltage output is driven using the DAQmx *Write block*. On the other hand, measurements are stored in a CSV file. This file is created by using a *Write-to-spread sheet file block* which enables to capture the data that can subsequently be used for further analysis in Matlab.

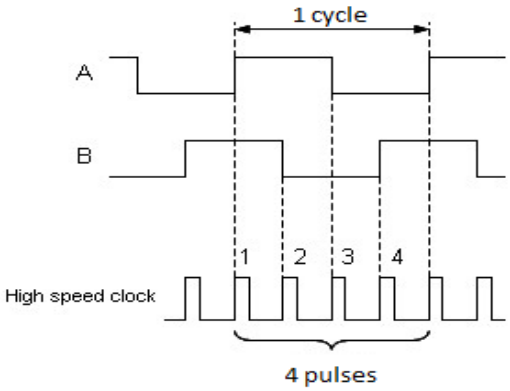


Figure 5. Reading and encoding signals from the encoder

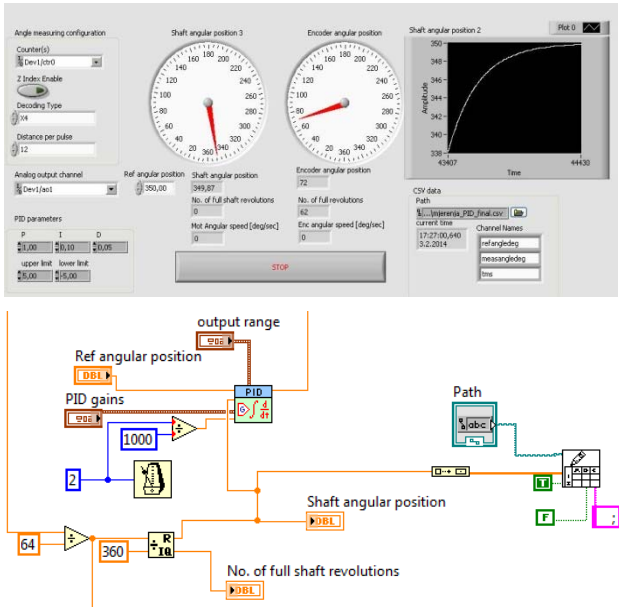


Figure 6. PID regulation VI: front panel (up), block diagram (down)

A. *PID control*

PID control is implemented in LabVIEW by using the predefined PID block from the *LabVIEW Control, Design and Simulation Module*. The respective front panel as well as the main elements of the block diagram are shown in Fig. 6.

Input and output channels have to be selected first.

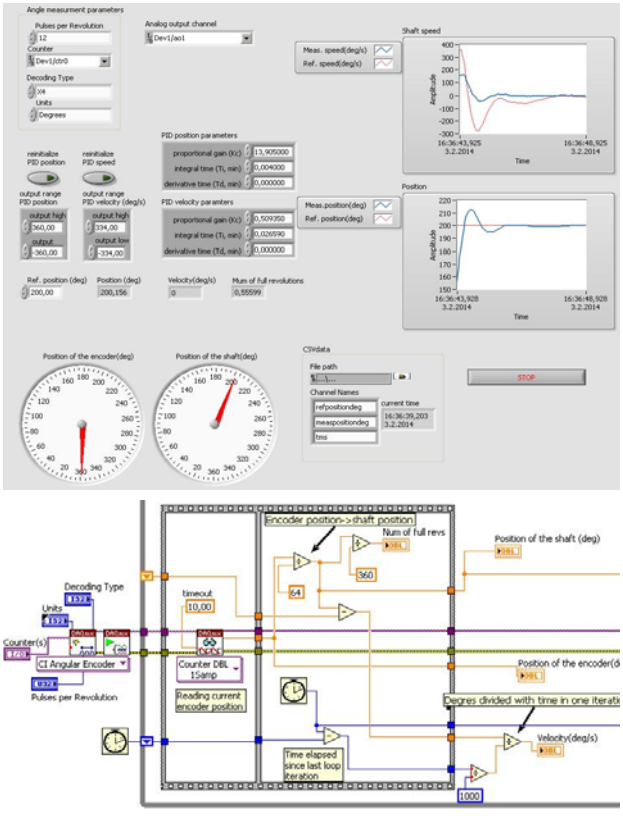


Figure 7. Cascade control VI: front panel (up), block diagram (down)

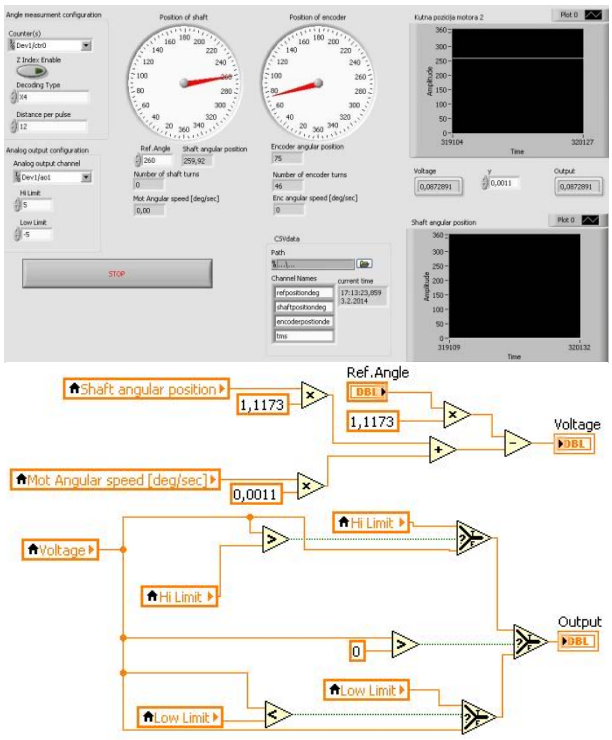


Figure 8. State-space control VI: front panel (up), block diagram (down)

Using the DAQmx counter as input, information about position is obtained from the channels of the encoder. This is compared to the defined reference position so as to determine the analogue input to the actuator.

B. Cascade control

The front panel and the block diagram of the cascade control scheme are shown in Fig. 7. The front panel allows the following input parameters to be defined: number of pulses per revolution, input and output channels, decoding type, position measurement units, parameters for the PI regulators as well as the file path. For a given reference position, the dynamic response of the system can be visualised via a waveform chart and/or a circular gauge. The number of full motor shaft revolutions is also shown. The inferred velocity is hence calculated as shown in the block diagram in Fig. 7. Experiments allowed establishing that a disadvantage of the used method is that, for sample times shorter than 50 ms, the velocity signal results in errors because the computational routines are slower than motor's response.

C. State-space control

The state-space regulator is implemented in LabVIEW using equations (12-19) – Fig. 8. The measured position and velocity are scaled by the gain vector **L** using equation (19). These values are then compared to the desired reference values. Since there are no additional time constants (no additional poles) added to the system, the state-space control algorithm has faster response times than the PID and the cascade controller. On the other hand, since state-space control has only proportional gain, a residual steady-state error is present. To minimize this error, the reference position is multiplied with a compensation gain.

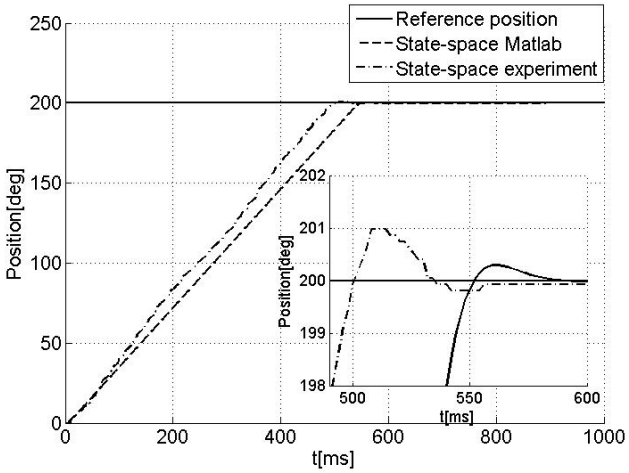


Figure 9. State-space transient response

V. EXPERIMENTAL RESULTS AND DISCUSSION

To test the performances of the described DC motor position control systems, a set of experiments using PID, cascade and state-space control typologies is performed. All tests are made with an input step command from 0° to 200°. DC motor input voltage is limited to ± 5V. The parameters of the PID and the cascade controllers are given in Tables II and III both for the simulations and the experiments. It is interesting to notice here that the simulations did not allow the correct integral time of the positioning loop of the cascade control to be determined. This value will thus be obtained in future investigations.

The comparison of simulated and experimentally obtained transient responses in the case of the state-space controller is shown in Fig. 9 and Table IV. From the magnified detailed view of the system dynamic response during settling, it can be seen that an aperiodic transient response occurs with a negligible difference in rise times between the simulated and actual results. In the case of the PID and the cascade controllers, due to simplified modelling assumption, some discrepancies of the simulated and experimental dynamic results occur. Future investigations will be done to obtain better matching of the simulated and experimental results.

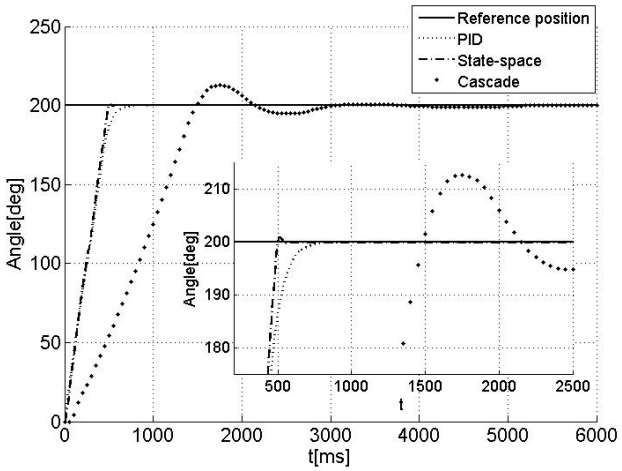


Figure 10 Combined experimental results

TABLE II PARAMETERS FOR THE PID CONTROL METHOD

	MATLAB	Experiment
Proportional gain K_p	0.576	0.6
Integral time T_I[ms]	0.313	0.3
Derivate time T_D [ms]	0.021	0.02

TABLE III PARAMETERS FOR THE CASCADE CONTROL METHOD

		MATLAB	Experiment
Positioning loop	Proportional gain K_p	13.56	13.905
	Integral time T_I [ms]	?	4
Velocity loop	Proportional gain K_p	0.508	0.508
	Integral time T_I [ms]	5	5

TABLE IV DYNAMIC RESULTS FOR THE STATE-SPACE CONTROL METHOD

	MATLAB	Experiment
Rise time τ_r [ms]	430	390
Percent overshot σ [%]	0.35	0.5
Settling time τ_s [ms]	580	550
Steady-state error [%]	0	0.04

TABLE V COMPARISON OF DYNAMICS FOR DIFFERENT CONTROL METHODS (EXPERIMENTS)

	PID	Cascade	S.S.
Rise time τ_r [ms]	425	1120	390
Percent overshot σ [%]	0	6.35	0.5
Settling time τ_s [ms]	800	3500	550
Steady-state error [%]	0	0.1	0.04

The comparison of transient responses obtained experimentally for all three used controllers is given in Fig. 10. Rise time, percent overshoot, settling time and steady-state error for each control algorithm are in turn given in Table V. It can be concluded that the fastest response is obtained by using the state-space control algorithm. The PID control algorithm results, however, in the smallest steady-state error.

VI. CONCLUSIONS AND OUTLOOK

An overview of different DC motor control approaches is given in this work. A conventional PID controller is employed first. The respective parameters are obtained using the Zieger-Nichols method and fine tuning. A cascade controller is developed next. Finally, the state-space controller with positioning and velocity loop is employed. A Matlab model of the used actuator is established in order to simulate different control approaches. Controllers are then implemented in the LabVIEW environment and experiments are conducted.

By comparing experimental results (Table V), it is concluded that positioning control via the state-space controller has the fastest response and the lowest settling times. Cascade control can be efficiently used, although the tuning of its parameters can often be cumbersome and computationally more intensive due to the presence of two PI blocks and the needed velocity calculation. This all limits the execution time which directly affects system's dynamic response. Improvements of cascade control could be achieved by using real-time hardware (e.g. the NI FPGA module) or if direct measurement of velocity would be possible. PID control results in negligible steady-state errors and acceptable rise and settling times.

LIST OF SYMBOLS

A	system matrix
B	control matrix
b	friction coefficient [mNm/rpm]
C	output matrix
$E(s)$, $e(t)$	difference between the reference and the process value
$G_R(s)$	PID transfer function
I_{an}	nominal armature current [mA]
J	motor inertia [kg·m^2]
k	finite time values factor (k=1, 2, 3, …)
K_e	back EMF constant [V/rpm]
K_m	torque constant [Nm/A]
K_P	proportional gain
L	state-space controller gain vector
P	system matrix in discrete domain
Q	control matrix in discrete domain
R_a	armature resistance [Ω]
s	Laplace variable
t	time variable
T_I, T_D	integral and derivative time constants
T_m	mechanical time constant
T_S	sample period
$U(s)$, $u(t)$	output from the controller
U_n	nominal voltage [V]
Y(s), y(t)	process value
Y_0(s)	reference position
z	discrete (z) domain variable
Δt	time change [ms]
$\Delta\varphi$	motor shaft angle change [°]
ξ	attenuation coefficient
σ	percent overshot [%]
τ_r	rise time [ms]
τ_s	settling time [ms]
φ	motor shaft position [°]
ω	velocity of the motor shaft [rpm]
Φ	fundamental matrix

REFERENCES

[1] E. Kamenar and S. Zelenika, "Micropositioning mechatronics system based on FPGA architecture," Proc. 36th International Convention on Information and Communication Technology, Electronics and Microelectronics MIPRO, pp. 138-143, 2013.

[2] G. F. Franklin, J. D. Powel and A. Emami-Naeini, "Feedback control of dynamic systems – 2nd ed.," Addison-Wesley, 1991.

[3] URL: www.machsupport.com

[4] D. Matika, "Sustavi digitalnog upravljanja (Digital control systems)," University of Rijeka, 2005.

[5] R. Cupec, "Diskretni sustavi upravljanja," University of Osijek, 2008. URL: http://www.etfos.unios.hr/upload/OBAVIJESTI/obavijesti_diplomski/diskretni_sustavi_upravljanja_19-11-2009.pdf

[6] K. J. Astrom and B. Wittenmark, "Computer controlled systems: theory and design – 3rd ed.," Prentice Hall, 1996.

[7] URL: www.faulhaber.com/uploadpk/EN_2342_CR_DFF.pdf

[8] URL: www.gurley.com/Encoders/Understanding_Quadrature.pdf

[9] G. S. Gordon, "Square Waves And Pulses: A Clarification," Measurements & Control, 1988.

[10] URL: www.dynapar.com/uploadedFiles/Downloads/Encoder_Glossary_of_Terminology.pdf

[11] URL: www.ni.com

Hrvatski Telekom kao predvodnik tehnoloških trendova

Kad se kaže: sve što može biti digitalizirano – bit će digitalizirano, možda to zvuči kao floskula ambicioznog profesora informatike za vrijeme štrajka, no s obzirom da je to rekao Nicholas Negroponte, jedan od najvećih vizionara današnjice, nad ovakvim se tvrdnjama valja zamisliti. Ipak, posrijedi je izjava koja datira iz devedesetih, i uistinu, premda proces još nije dovršen – sve što se može digitalizirati, uspješno se digitalizira.

Negroponteovu izjavu parafrazirao je predsjednik Uprave Deutsche Telekoma Timotheus Höttges na ovogodišnjem Mobile World Congressu u Barceloni rekavši: "Sve što može biti umreženo – bit će umreženo." Uistinu, umrežavanje ljudi i uređaja jedan je od glavnih tehnoloških trendova, koji se može uočiti svuda oko nas. Automobili su dovoljan primjer. Danas prosječni automobil sadrži 60 do 100 senzora, a u bliskoj budućnosti bit će ih više od 200. Razvijamo e-government, e-školstvo, e-zdravstvo, no u pozadini moramo imati i mreže koje to mogu podržati. Stoga u HT-u svakodnevno ulažemo u razvoj najnaprenije mreže kako bismo korisnicima u svakom trenutku omogućili pristup svemu što žele, te stvorili preduvjete za količine podatkovnog prometa koje nam tek dolaze. Jer, budućnost se mora graditi sada.



Među prvima u Europi i svijetu s IP migracijom

Hrvatski Telekom modernizira svoju fiksnu mrežu i to na multimedijsku platformu nove generacije temeljene na internet protokolu, te je u planu do kraja 2015. godine na novu tehnologiju prebaciti sve korisnike. Pored korištenja postojećih usluga, nova platforma korisnicima omogućuje konvergenciju visoko kvalitetnih multimedijskih usluga i inovativne načine korištenja bezgraničnih mogućnosti koje im pruža internet. Vrhunske multimedijske usluge i uobičajene telekom usluge bit će korisnicima dostupne praktički odmah i to bez ulaska tehničkog osoblja u dom, a korisnik će moći jednim klikom izabrati željenu brzinu interneta.

HT nudi najveću garantiranu brzinu u Hrvatskoj u sklopu svojih Ultra MAX paketa, i to 100 Mbps u downloadu i 20 Mbps u uploadu. Riječ je o paketima baziranim na FttH tehnologiji



Usluge na super brzom internetu

U fiksnom segmentu HT korisnicima nudi najveću garantiranu brzinu u Hrvatskoj u sklopu svojih Ultra MAX paketa, i to 100 Mbps u downloadu i 20 Mbps u uploadu baziranih na FttH tehnologiji. Maksimalna brzina od 100 Mbps downloada i 20 Mbps uploada upravo je ona za koju Hrvatski Telekom jamči da će je korisnik i dobiti, budući da FttH tehnologija omogućuje garantiranu brzinu do korisnika, za razliku od drugih ponuda na tržištu gdje se ponuđena maksimalna brzina dijeli među korisnicima na optičko-kabelskoj mreži. Budući da su zahtjevi korisnika za brzim internetom i umrežavanjem sve veći, u HT-u radimo i na mreži budućnosti TeraStream: novom mrežnom konceptu temeljenom na protokolu IPv6, koji smo predstavili prvi u svijetu i koji omogućuje korisniku brzinu od 1 Gbps.

Najveća pokrivenost 4G mrežom u Hrvatskoj

Dostupnost brzog interneta strateška je odrednica Hrvatskog Telekoma i u mobilnom segmentu. Svakodnevno povećavamo pokrivenost i brzim 3G i 4G mrežama i nudimo najveću dostupnost obje mreže u Hrvatskoj. Hrvatski Telekom je prvi u Hrvatskoj i među prvima u regiji i Europi u rad pustio 4G mrežu temeljenu na naprednoj LTE tehnologiji. Osim u urbanim i korisnicima u ruralnim područjima nudimo brzine prijenosa podataka do 75 Mbps u downloadu te do 25 Mbps u uploadu. Ponosni smo da naša ulaganja u razvoj i inovacije prepoznaje i šira javnost pa je tako Hrvatski Telekom prošle godine proglašen liderom na području inovativnosti u kategoriji velikih tvrtki prema rezultatima nacionalnog istraživanja o stanju inovativnosti hrvatskih tvrtki. Budućnost digitalnog društva kojeg stvaramo želimo zajedno s korisnicima živjeti već sada.

sponsors

golden sponsors

KONČAR

ERICSSON
Ericsson Nikola Tesla

T Hrvatski Telekom

silver sponsor

INFODOM

bronze sponsors

HEP

OiV

vip

STORM GRUPA

sponsors

Microsoft

in2 GRUPA

KING ICT
INFORMATION & COMMUNICATION TECHNOLOGIES

m.TEH mjerne tehnologije d.o.o.

hp

SELMET

ALTPRO d.o.o.
ISTRAŽIVANJE - RAZVOJ - PROIZVODNJA

nomen d.o.o.
MES and ERP Systems since 1994

micro-link

PROCADD d.o.o.
www.ib-procadd.si

MIPRO
Kružna 8/II, HR-51000 Rijeka, Croatia

phone: +385 51 423 984
mipro@mipro.hr • www.mipro.hr

ISBN 978-953-233-078-6